

LREC 2026

# 12th Workshop on the Representation and Processing of Sign Languages: Language in Motion (sign-lang@LREC 2026)

## Workshop Proceedings

**Editors**
**Eleni Efthimiou, Stavroula-Evita Fotinea,**
**Thomas Hanke, Julie A. Hochgesang,**
**Johanna Mesch, and Marc Schulder**

16 May 2026

**Proceedings of the LREC 2026 12th Workshop on the Representation and Processing of Sign Languages: Language in Motion (sign-lang@LREC 2026)**

© ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-82-1

# Preface

This collection of papers stems from the *12th Workshop on the Representation and Processing of Sign Languages* which takes place as a satellite workshop to the *Language Resource and Evaluation Conference (LREC 2026)* in Palma, Mallorca (Spain).

While there has been occasional attention to sign languages at the LREC main conference, the focus there is on spoken languages in their written and spoken forms. This series of workshops, however, offers a forum for researchers focussing on sign languages, especially on ressources such as corpora and lexica and on language technologies for sign languages.

This year's hot topic "Language in Motion" addresses the relevance of motion in sign language research, both regarding the literal role of movement to the visual-gestural modality, as well as the figurative roles of evolution in the language and research communities. Sign languages themselves constantly evolve and research needs to account for such language change as well as pushing the boundaries of what can be expected from the field, both technologically and ethically.

The contributions composing this volume are presented in alphabetical order by the last name of the first author. For the reader's convenience, an author index is provided as well.

Once again, we would like to thank all members of the program committee who helped us tremendously by reviewing the submissions to the workshop within a very short timeframe!

Finally, we would like to point the reader to the *sign-lang@LREC Anthology* at

https://www.sign-lang.uni-hamburg.de/lrec/

The anthology contains all publications of the workshop series as well as sign language papers from the LREC main conference and its other workshops. It offers author and topic indices across all papers, stable URLs for all workshop papers and their supplementary materials, as well as bibliographical (BibTeX) data for all entries. Happy browsing!

The Editors

# Organizers

### Organizing Committee

Eleni Efthimiou, Institute for Language and Speech Processing, Athens, Greece
Stavroula-Evita Fotinea, Institute for Language and Speech Processing, Athens, Greece
Thomas Hanke, University of Hamburg, Hamburg, Germany
Julie A. Hochgesang, Gallaudet University, Washington, USA
Johanna Mesch, Stockholm University & KTH Royal Institute of Technology, Stockholm, Sweden
Marc Schulder, University of Hamburg, Hamburg, Germany

### Program Committee

Sam Bigeard, Inria Centre, Nancy, France
Carl Börstell, University of Bergen, Bergen, Norway
Richard Bowden, University of Surrey, Guildford, United Kingdom
Annelies Braffort, CNRS/LISN, Orsay, France
Kearsy Cormier, University College London, London, United Kingdom
Maartje De Meulder, HU University of Applied Sciences Utrecht, Utrecht, The Netherlands
Athanasia-Lida Dimou, Institute for Language and Speech Processing, Athens, Greece
Sarah Ebling, University of Zurich, Zurich, Switzerland
Eleni Efthimiou, Institute for Language and Speech Processing, Athens, Greece
Michael Filhol, CNRS/LISN, Orsay, France
Stavroula-Evita Fotinea, Institute for Language and Speech Processing, Athens, Greece
Kathleen Currie Hall, University of British Columbia, Vancouver, Canada
Thomas Hanke, University of Hamburg, Hamburg, Germany
Julie A. Hochgesang, Gallaudet University, Washington, USA
Amy Isard, University of Hamburg, Hamburg, Germany
Tommi Jantunen, University of Jyväskylä, Jyväskylä, Finland
Vadim Kimmelman, University of Bergen, Bergen, Norway
Maria Kopf, University of Hamburg, Hamburg, Germany
Anna Kuder, University College London, London, United Kingdom
Gabriele Langer, University of Hamburg, Hamburg, Germany
Kevin Lee, EQ4ALL, Seoul, South Korea
Lisa Loy, University of Hamburg, Hamburg, Germany
John McDonald, DePaul University, Chicago, USA
Özge Mercanoğlu Sincan, University of Surrey, Guildford, United Kingdom
Johanna Mesch, Stockholm University & KTH Royal Institute of Technology, Stockholm, Sweden
Hope E. Morgan, University of Hamburg, Hamburg, Germany
Irene Murtagh, Technical University Dublin, Dublin, Ireland
Carol Neidle, Boston University, Boston, USA
Justin Power, University of Texas, Austin, USA
Floris Roelofsen, University of Amsterdam, Amsterdam, The Netherlands
Paweł Rutkowski, University of Warsaw, Warsaw, Poland
Marc Schulder, University of Hamburg, Hamburg, Germany
Rose Stamp, University of Birmingham, Birmingham, United Kingdam
Tsubasa Uchida, NHK Science & Technology Research Laboratories, Tokyo, Japan
Anna Vacalopoulou, Institute for Language and Speech Processing, Athens, Greece
Sabrina Wähl, University of Hamburg, Hamburg, Germany
Joanna Wójcicka, University of Warsaw, Warsaw, Poland
Rosalee Wolfe, Institute for Language and Speech Processing, Athens, Greece

# Table of Contents

*Capturing Methodology for Generating Synthetic and 3D Training Data in Catalan Sign Language (LSC): The Case of Verbal Agreement*
Gemma Barberà, Inés Broto Clemente, Xavier Vinaixa Roselló and Roger Cassany Viladomat . . . . . 1

*Leveraging Unannotated Sign Language Data via a Robust Data Augmentation Method for Contrastive Representation Learning*
Ariel Basso Madjoukeng, Pierre Poitier, Belise Edith Kenmogne, Adelaide Couplet, Margaux Leleu and Frenay Benoit . . . . . 10

*The SMILE Continuous DSGS Corpus: A Resource for Longitudinal Exploration of Continuous Swiss German Sign Language*
Alessia Battisti, Katja Tissi, Sandra Sidler-Miserez and Sarah Ebling . . . . . 17

*The In-Car Sign Language Corpus (ICSL): A Multi-Modal Resource for Constrained-Space Sign Language Recognition*
Raviteja Boddu, Guilherme Vieira Leite, Joed Lopes da Silva, Ângelo Benetti, Isabela Barbieri, Natália de Melo Afonso, Thyago Santos, Hélio Pedrini, Felipe Venâncio Barbosa, José Mario De Martino, Munir Georges and Alessandro Zimmer . . . . . 31

*Seeing Who Is Signing and With Which Hand*
Carl Börstell . . . . . 42

*SignGPT and the Visual Language Toolkit*
Matt Brown, Oline Ranum, Edward Fish, Heidi Proctor, Bencie Woll, Richard Bowden and Kearsy Cormier . . . . . 51

*Processing Kinematics of Nonmanual Markers in R*
Jan Bulla and Vadim Kimmelman . . . . . 61

*A Small Model for Big Articulators: Sign Language Detection With a Tiny Machine Learning Model*
Frederick Chan, Gina-Anne Levow and Qi Cheng . . . . . 71

*"A Sacred Bird Called the Phoenix". Auditing the most-used Parallel Corpus for German Sign Language Recognition and Translation*
Vera Czehmann, Shakib Yazdani, Yasser Hamidullah, Fabrizio Nunnari and Eleftherios Avramidis . . . . . 80

*Diffusion-Based 3D Sign Language Motion Anonymization: A Feasibility Study on Balancing Identity Confusion and Semantic Preservation*
Zixuan Dai and Shinji Sako . . . . . 93

*GeoQuery-LSFB: A French Belgian Sign Language Corpus with Procedural Semantic Annotations*
Liesbet De Vos, Laurence Meurant, Paul Van Eecke and Katrien Beuls . . . . . 100

*The De-Sign Platform: An Online Psychometric Tool for Dementia Screening of Deaf Older Adults in two Sign Languages, GSL and ÖGS*
Athanasia-Lida Dimou, Theodore Goulas, Marianna Tsatali, Tarsita Ntova, Doris Hoffmann-Lamplmair, Stavroula-Evita Fotinea, Eleni Efthimiou, Birgit Teichmann, Magda Tsolaki, Joanna Atkinson and Bencie Woll . . . . . 113

*Feature Analysis of MoCap Data for Optimised Sign Language Processing*
Yves A. Duppen, Mirella De Sisto, Ifigeneia Mavridou, Phillip Brown, Lisa Lepp and Dimitar Shterionov . . . . . 120

*Leveraging Text-side Augmentation For Sign Language Translation*
Diandra Fabre, Julie Lascar, Julie Halbout and Markarit Vartampetian . . . . . 129

*The Construction of the CORALSE Corpus, Now and Beyond: A Tool for Documenting Spanish Sign Language*
Ana Fernández Soneira, María C. Bao-Fente, Rayco H. González-Montesino and Inmaculada C. Báez-Montero . . . . . 140

*The Community and Ethics Shaping the Norwegian Sign Language Corpus*
Lindsay Ferrara . . . . . 148

*Generations in the DGS Corpus: Evolving Outreach Activities and Cross-Generational Stories on Social Media in a Long-Term Corpus Project*
Anike Fiedler, Marc Schulder, Julian Bleicken and Annika Herrmann . . . . . 155

*Formalising Sign Language Depiction, Characterising Categories and Measuring Iconicity with AZee*
Michael Filhol and Emmanuella Martinod . . . . . 164

*Extracting Signs from Weakly Aligned Sign Language Corpora: A Study on LSF and LSM*
Lorena de la Garza, Julie Halbout, Julie Lascar, Niels Martinez, Arturo Curiel, Michèle Gouiffès and Annelies Braffort . . . . . 174

*An Annotation Formalism for a French–LSF Bilingual Corpus Supporting Sign Language Generation*
Sylvie Gibet, Clément Reverdy and Pierre-François Marteau . . . . . 184

*A Pose-Based Pipeline for Annotation of Headshakes in Sign Language Corpora*
Gustaf Gren and Nikolaus Riemer Kankkonen . . . . . 193

*Learning to Spot Signs from Named Entities. A study on French Sign Language.*
Julie Halbout, Annelies Braffort, Michèle Gouiffès, Diandra Fabre and Julie Lascar . . . . 203

*The Iterative Development and Evaluation Framework for Kazakh-Russian Signing Avatars Targeted to Native Deaf Signers*
Alfarabi Imashev and Tohid Alizadeh . . . . . 212

*Movement Coherence in High Visual Load Environments: Implications for Attention in Mixed-Hearing Classes*
Mert Inan, Saki Imai, Anna Marshall, Tessa Karel and Malihe Alikhani . . . . . 226

*A Comparative Analysis of Traditional and Contemporary Visual Features for Computational Annotation of Irish Sign Language*
Sarmad Khan, Simon McLoughlin and Irene Murtagh . . . . . 239

*HeSLEx: A novel online questionnaire for heritage sign language research*
Evgeniia Khristoforova and Roman Poryadin . . . . . 248

*Comparison of Low Bitrate Quantizers for Encoding Swedish Sign Language*
Anna Klezovich, Johanna Mesch, Gustav Eje Henter and Jonas Beskow . . . . . 256

*Exploring Aspects of Spontaneous Signing in the DGS Corpus*
Maria Kopf, Reiner Konrad, Gabriele Langer, Marc Schulder and Lutz König . . . . . 262

*Two-Handed Signs and Handedness: Phonological Implications for Sign Language Structure*
Lisa Lepp, Mirella De Sisto and Dimitar Shterionov . . . . . 275

*HNS2CF: A Mapping Tool from HamNoSys to SL CatForm*
Lisa Loy and Hope E. Morgan . . . . . 287

*Emotion Recognition in German Sign Language with Facial Action Units*
Cristina Luna Jimenez, Lennart Eing, Sergio Esteban Romero, Tanja Schneeberger, Patrick Gebhard, Fabrizio Nunnari and Elisabeth Andre . . . . . 297

*DGS-BIGEKO: A Dataset for Hypothetical Emergency Scenarios in German Sign Language*
Cristina Luna Jimenez, Lennart Eing, Daksitha Senel Withanage Don, Marco González, Fabrizio Nunnari, Pamela Perniss, Patrick Gebhard and Elisabeth Andre . . . 306

*Perceptual Validation of 3D Pose, Guided Sign Language Synthesis*
Ezekiel Maina, Lilian Wanzare and James Obuhuma . . . . . 315

*The Displacement-Velocity Dissociation in Sign Language Learning: Kinematic Signatures of Event Structure in Novice ÖGS Signers*
Evguenia A. Malaia, Julia Krebs, Eric Harbour, Julia Martetschläger, Hermann Schwameder, Dietmar Roehm and Ronnie B. Wilbur . . . . . 324

*CEFR-Based Assessment in Sign Languages: The Case of LSE and Perspectives for LIS*
Maria Grazia Marrocu . . . . . 333

*Improving phonological distance measures for signs: the CatFormCompare tool*
Hope E. Morgan, Amy Isard and Anh Dang . . . . . 341

*Assisting Corpus Annotation: Automatic BIO-Tagging of Clause-Like Units in Polish Sign Language. A Pilot Study on Corpus Data*
Piotr Mostowski, Anna Kuder and Joanna Wójcicka . . . . . 351

*Introducing VISTA-SL: A Multilingual e-Learning Platform for Deaf and Hearing Learners of Sign Languages*
Irene Murtagh, Marc Schulder, Annika Herrmann, Liona Paulus, Julian Bleicken, Kostas Blekos, Athanasios Konstantakopoulos, Klimis Antzakas, Dimitrios Kosmopoulos, Eva Valls, Ricardo Marques, Josep Blat, Konstantinos Karampidis and Ben Elsendoorn . . . . . 361

*Evaluation of Pose Estimation Systems for Sign Language Translation*
Catherine O'Brien, Gerard Sant, Mathias Müller and Sarah Ebling . . . . . 371

*Designing a Data Model for a Diachronic Sign Language Database: A Case Study of Nineteenth-Century Bohemian Sources*
Lenka Okrouhlíková . . . . . 387

*A Video-Based Reverse Dictionary for Sign Language Using Gesture Similarity*
Batyrbek Orazumbekov, Daniyal Bayanov, Aruzhan Kaltay and Anara Sandygulova . . . 398

*Norwegian Sign Language: Overview of Resources and Experiments with Automatic Sign-Writing Transcription*
Elisabeth Othamar and Yves Scherrer . . . . . 408

*Long-Term Sign Language Data Crowdsourcing Through Collaborative Lexicons*
Pierre Poitier, Jérôme Fink, Ariel Basso Madjoukeng, Adelaide Couplet, Margaux Leleu and Benoît Frénay . . . . . 419

*Effect of Data Augmentation with Multi-View Perspectives of Signers on the DGS-Fabeln-1 Dataset*
Fabian Renner, Daksitha Withanage Don, Elisabeth Andre and Cristina Luna Jimenez 429

*Lost in Expression: Diagnosing Systemic Challenges with Non-Manual Generalization in Sign Language Understanding Tasks*
Dmitriy Sazonov, Sevgi Z. Gurbuz and Evie A. Malaia . . . . . 438

*The SignBeach Dataset of Dutch Sign Language (NGT) signs*
Annika Schiefner, Gomèr Otterspeer, Beyza Sümer and Floris Roelofsen . . . . . 450

*Comparing Computer Vision Instruments for Eye Blink Analysis*
Margaux Susman, Carla Miquel Blasco and Jan Bulla . . . . . 459

*Grounding Sign Language Representation Learning in Phonology*
Toon Vandendriessche, Mathieu De Coster and Joni Dambre . . . . . 468

*Towards Integrating Pose Estimation with Neuroimaging for the Analysis of Signed Language Video Stimuli*
Sébastien Vandenitte, Doris Hernández, Jarkko Keränen, Tommi Jantunen and Anna Puupponen . . . . . 477

*KWIC view on Constructed Action (CA) and its Collocates in German Sign Language (DGS) – Possibilities and Limitations*
Sabrina Wähl . . . . . 484

*Beyond BLEU: Linguistic Invisibility and Interactional Repair Sequence in End-to-End Sign Language Translation*
Zirui Wang and Mayumi Bono . . . . . 491

*Continuous Sign Language Recognition using Multimodal Input and Handshape-aware Boundary Detection*
Mingyu Zhao, Zhanfu Yang, Yang Zhou, Zhaoyang Xia, Can Jin, Xiaoxiao He, Shuhang Lin, Carol Neidle and Dimitri Metaxas . . . . . 501

# Capturing Methodology for Generating Synthetic and 3D Training Data in Catalan Sign Language (LSC): The Case of Verbal Agreement

**Gemma Barberà**[1], **Inés Broto Clemente**[1], **Xavier Vinaixa Roselló**[2], **Roger Cassany Viladomat**[1]
[1]Pompeu Fabra University, Barcelona, Spain
[2]Mortensen, Barcelona, Spain
gemma.barbera@upf.edu, inesbrotoi@upf.edu, xavi@sorensen.ai, roger.cassany@upf.edu

**Abstract**

This paper proposes a hybrid methodology to generate high-quality synthetic data. Unlike other approaches based purely on generative Artificial Intelligence, which may suffer from hallucinations or inconsistent movements, this project uses 3D biomechanics and kinematics algorithms that enforce the anatomical constraints of the human body to ensure physically plausible movements. The aim of this research is to demonstrate that it is possible to synthetically expand the dataset. In particular, this paper focuses on verb agreement, a grammatical domain which is known for its morphological and articulatory complexity. By concentrating on the possible configurations of the movements in signing space when expressing different person agreeing verbal forms, we aim to capture real movements to extract physical parameters and apply them as logical rules —similar to those of a video game engine— to automatically synthesize thousands of new conjugations from infinitives with complete anatomical precision. Beyond spatial conjugation, the methodology further augments data through procedural variation of prosody and body morphology.

**Keywords:** inverse kinematics, low-resource translation, sign language recognition, sign language translation, spatial morphosyntax

## 1. Introduction

The development of Sign Language Recognition (SLR) systems and signing avatars requires a large amount of data to cover the morphological richness of the language. However, the massive and at the same time detailed recording of all grammatical components is unfeasible (Bragg et al., 2019). This paper proposes a hybrid methodology to generate high-quality synthetic data to train models on Sign Language Recognition (SLR). To avoid reported hallucinations or inconsistent movements generated by Artificial Intelligence (AI) augmentation pipelines (Walsh et al., 2025), this project uses 3D biomechanics and kinematics algorithms. The aim of this research is to demonstrate that it is possible to synthetically expand a dataset of signed infinitives focusing on verb agreement, a grammatical domain known for its morphological and articulatory complexity consistent across different verbs and conjugations. By studying the procedural movements that are repeated in the signing space when expressing different verbal forms of person agreement, we aim to extract physical parameters and apply them as logical rules —similar to those of a video game engine— to automatically synthesize thousands of new conjugations from infinitives with complete anatomical precision. This paper presents the capturing methodology used for creating synthetic data in the realm of verbal agreement forms.

End-to-end computer vision systems require big data. In order to function properly, thousands of hours of annotated video are needed. However, this is inefficient for minority languages, such as signed languages, were a huge contrast in the amount of data when comparing English vs. American Sign Language (ASL) or Spanish Sign Language (LSE) is found (see Figure 1) and also when comparing the approximate number of hours of video data among ASL, LSE and Catalan Sign Language (see Figure 2.)

Catalan Sign Language (*llengua de signes catalana*, LSC) is a low-resource environment (Cassany Viladomat et al., In press). Besides the LSC Corpus (IEC, 2025), which includes about 60 hours of annotated data of different regional and social variants, there do not exist yet massive video corpora in LSC to train traditional models. Previous research has shown that injecting linguistic features into translation performance supplements low-resource strategies like transfer learning by also producing more coherent output (McGill, 2026). Moreover, it has been also argued that adding linguistic aspects such as pointing and spatial inflection to sign language animations leads to a significant improvement in user comprehension (Huenerfauth and Lu, 2012).

This work proposes a methodology to synthetically increase annotated dataset by leveraging the linguistic rules that govern sign–gloss transcriptions, particularly those related to verbal agreement expressed spatially. The proposed approach is designed for a resource-poor sign language, specifically LSC, where the scarcity of annotated data constitutes a significant bottleneck for the develop-



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 1–9
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

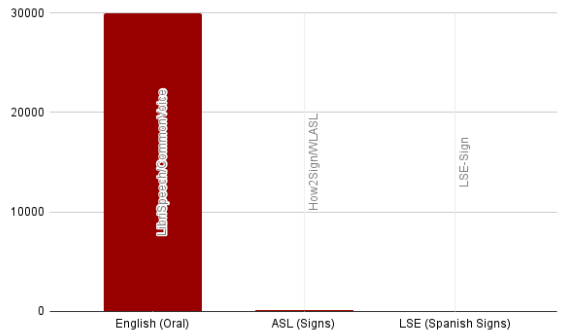



Figure 1: Sign languages as low-resource languages compared to spoken language

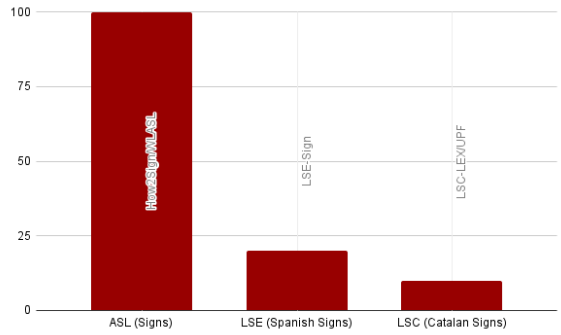



Figure 2: Data scarcity in sign languages

ment of robust SLR systems.

## 2. Background

**Verb agreement.** Linguistic research has shown that most sign languages known to date have ways to mark reference to arguments on verbal forms, usually by linking the start and end point of the movement path of the verb, the orientation of the hands, or both, to spatial referential locations (also known as R-loci) associated to those arguments (Quer, 2021). The most common categorization of verb classes stems from Padden (1988) which established three morphological classes, namely (i) plain verbs, which do not encode agreement; (ii) agreeing verbs, which overtly agree with subject and object arguments; and (iii) spatial verbs, which agree with locative arguments. This paper focuses on the manual behavior of the second group because it allows to investigate the path movement and hand orientation when marking the internal arguments. Within the category of agreeing verbs, two classes arise. On the one hand, regular agreeing verbs show a path movement that starts in the subject argument and ends in the object. On the other hand, backwards agreeing verbs show an opposite movement: it starts in the location of the object and moves or is directed to the subject. Some examples of regular agreeing verbs in LSC are TELL, TAKE-CARE, and SUPPORT, among many others. Some examples of backward verbs are UNDERSTAND, INVITE, and CHOOSE, to cite just a few. The methodology proposed in this paper concentrates on agreeing verbs and it includes the two classes as shown in the following section.

**Signing space.** The spatial arrangement of person locations resembles in most sign languages known to date, and LSC is not an exception. First person is localized by pointing or articulating the verb towards to the chest of the signer. While second person is localized in front of the signer, the locus for third person is established in the lateral areas, both in the ipsilateral and the contralateral, indistinctively. Signing space, as a three-dimensional space, is morphologically and productively used to show the possible person combinations. These combinations are expressed in a structured way, divided into the three spatial planes established by Brentari (1998). They are defined as follows: (i) The horizontal plane stands perpendicularly to the body of the signer and it is commonly considered as the default plane (Klima and Bellugi, 1979). (ii) The frontal plane extends vertically to the body of the signer. (iii) The midsagittal plane extends vertically and perpendicularly to the body of the signer. First and second person are contrasted along the midsagittal plane, therefore agreement relations between first and second person are expressed within this plane. The horizontal plane is the placeholder for the expression of agreement relations between first and third person, and also between second and third person. Last but not least, the frontal plane establishes the agreement relations between first/second person and third person, when third person conveys an indefinite or an impersonal subject, as shown for LSC (Barberà, 2015). This is graphically shown in Figure 3, which exhibits the configuration of the three grammatical person locations established within the three spatial planes. The complexity of grammatical notions and use of signing space offers an interesting domain to capture real movements to extract physical parameters and apply them as logical rules, with the main aim of automatically synthesizing thousands of new conjugations from infinitives with complete anatomical precision, as shown in section 3.1.

**Gloss-based SLT.** Sign language glossing is a system of written notation used to transcribe and analyze signs, representing their morphosyntactic structure rather than their direct translation into another language. Glosses do not serve as interpretations, but as transcriptions, and are commonly used to facilitate scientific study by showing the or-

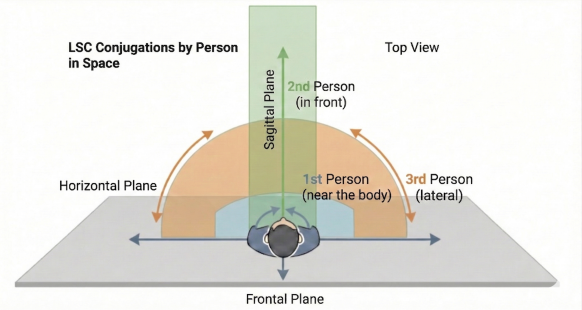



Figure 3: Configuration of grammatical person within the spatial planes

der of signs and non-manual behaviors.
Gloss-free Sign Language Translation (SLT) methods aim to directly translate video sequences into spoken-language text. In contrast, gloss-based SLT approaches rely on glosses as an intermediate textual representation between the input signed sequence and the target text. Glosses provide a well-structured and constrained ground-truth representation, which facilitates model training by reducing output variability and linguistic ambiguity. However, gloss-based datasets are challenging to construct, as gloss and translation annotation requires the expertise of trained sign language linguists. This process is both time-consuming and costly, which limits the scalability of such resources.

This work proposes a methodology to augment a gloss-level annotated dataset by leveraging the linguistic rules that govern sign–gloss transcriptions, particularly those related to verbal agreement expressed spatially. The proposed approach is designed for a resource-poor sign language, specifically LSC, where the scarcity of annotated data constitutes a significant bottleneck for the development of robust SLR and SLT systems. The main goal is to prove that there is a way to augment the dataset by leveraging the linguistic rules related to verbal agreement expressed spatially.

**Isolated and Continuous SLR.** SLR techniques can be classified into Isolated Sign Language Recognition (ISLR) (Sarhan and Frintrop, 2023) and Continuous Sign language Recognition (CSLR) (Alyami and Luqman, 2025). While the first group works over discrete objects (clips) in order to recognize single signs, the second includes sequential information between these isolated events aiming to interpret consecutive signs taking into account each context and producing its corresponding glosses. Rastgoo et al. (2021) argues that although continuous data contains information relevant to the SLR task, there are still unresolved challenges that lead most systems to rely on isolated data. Firstly, computational complexity hinders the use of sequential data. Furthermore, the fact that continuous data can be discretized (losing, of course, some information) means that systems predominantly opt for the former, even though the real world always presents a continuous scenario. This paper proposes the creation of a dataset that preserves sequential information through avatar control technology commonly used in video games and virtual reality. As a result, the dataset can be synthesized for ISLR when needed, while still providing the temporal information required to train CSLR models.

## 3. Methodology and corpus design

### 3.1. Calibration dataset

The calibration dataset includes a complex repertoire of grammatical constructions including different features. First, a list of 48 isolated verb types, including regular and backwards. Second, 12 grammatical person combinations and the infinitive form for each verb, summing up to 624 person combinations. Table 1 includes the 12 combinations of grammatical person and infinitive form for the particular verb TELL.

| Gloss | Translation |
|---|---|
| TELL | (infinitive) |
| 1TELL2 | 'I tell you.' |
| 2TELL1 | 'You tell me.' |
| 1TELL$3_a$ | 'I tell him.' |
| $3_a$TELL1 | 'He tells me.' |
| 1TELL$3_b$ | 'I tell her.' |
| $3_b$TELL1 | 'She tells me.' |
| 2TELL$3_a$ | 'You tell him.' |
| $3_a$TELL2 | 'He tells you.' |
| 2TELL$3_b$ | 'You tell her.' |
| $3_b$TELL2 | 'She tells you.' |
| $3_a$TELL$3_b$ | 'He tells her.' |
| $3_b$TELL$3_a$ | 'She tells him.' |

Table 1: Person-marked forms of TELL

Furthermore, for the verbs SAY, TELL and INFORM six other combinations were added by using high locus for the subject argument (indicated with the subscript $_{hi}$ in the glosses) conveying indefiniteness or an impersonal subject, as previously shown in Barberà (2015) (see Table 2). An example of the various agreement forms of the verb GIVE-ADVICE is illustrated in Figure 4.

Third, complete sentences including internal objects of the verb, with a total of 280. The numbers and features of this complex repertoire are summarized in table 3.

The final amount of items were recorded by three deaf native signers, all three from the area of Barcelona, with different profiles and belonging to the three age groups established in the LSC Corpus (Barberà et al., 2015). Signer TG is a 60-year-old

| Gloss | Translation |
|---|---|
| $3_{ahi}$TELL1 | 'They/Someone tells me.' |
| $3_{bhi}$TELL1 | 'They/Someone tells me.' |
| $3_{ahi}$TELL2 | 'They/Someone tells you.' |
| $3_{bhi}$TELL2 | 'They/Someone tells you.' |
| $3_{ahi}$TELL$3_b$ | 'They/Someone tells her.' |
| $3_{bhi}$TELL$3_a$ | 'They/Someone tells him.' |

Table 2: High/indefinite third-person subject forms of TELL

| Type | Isolated | Combinations | Sentences |
|---|---|---|---|
| Regular | 34 | 442 | 195 |
| Backwards | 14 | 182 | 85 |
| **TOTAL** | **48** | **624** | **280** |

Table 3: Features and numbers of the recorded dataset

male signer that comes from a family of five signing deaf generations, with more than 35 years of experience as a deaf teacher and more than 25 years as a researcher. Signer BF is a a 31-year-old male signer that comes from a family of three signing deaf generations, with no prior contact with LSC teaching or research but with an extensive experience as a deaf actor and poet. Signer CW is a 26-year-old female signer that comes from a family of two signing deaf generations, with a diploma of LSC teacher and currently studying a BA on journalism. The total amount of complex grammatical constructions sums up to 2853, as shown in table 4.

### 3.2. Dataset Augmentation

One of the main limitations of SLR systems lies in the complexity of understanding sign language in its context and the interdependencies that exist in the signing of continuous signs in the space

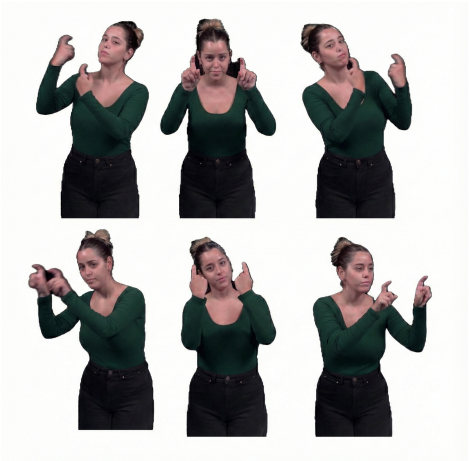


Figure 4: Various agreement forms of the verb GIVE-ADVICE

| | Gender | Age group | Recordings |
|---|---|---|---|
| TG | M | 51-80 | 949 |
| BF | M | 31-50 | 952 |
| CW | F | 18-30 | 952 |
| TOTAL | | | **2853** |

Table 4: Signers' profile and amount of data

(Alyami and Luqman, 2025). The present proposal addresses this problem through two complementary strategies. The first strategy is to introduce a data augmentation methodology designed to reinforce the nuances of spatial context encoded in gloss-level transcriptions, operating along three procedural axes: spatial conjugation, prosodic variation, and body morphology variation (detailed in section 4.3). The second strategy is to reduce the burden on the computer vision component by adopting gloss-level translation as an intermediate objective.

## 4. Procedural processing and synthesis

This section outlines the overall system architecture adopted for the procedural processing and augmented dataset curation. The system is designed as a pipeline comprising skeleton extraction, biomechanical parameterization, and procedural data generation, where each stage is detailed in the following subsections.

### 4.1. Skeleton Extraction

The recorded clips are processed to extract 3D joint coordinates (in `.pose` format) using the MediaPipe framework (Kim et al., 2023; Google LLC, 2023) through the `video_to_pose` utility from Moryossef et al. (2021). To ensure signer privacy, the anonymization module proposed in Moryossef (2024) has been applied, which removes identifying information from sign language `.pose` data. Authors claim it guarantees anonymity from an information-theoretic perspective, as the only retained information corresponds to the linguistic content conveyed by the selected signs.

It should be noted that the 3D joint coordinates estimated by MediaPipe from monocular video are subject to depth ambiguity and reduced accuracy for distal landmarks (particularly fingers). However, the procedural pipeline mitigates these limitations at two levels: (1) shoulder-width normalization reduces inter-signer scale variation, and (2) the IK solver recomputes arm geometry analytically from the target locus, meaning that the spatial conjugation is determined by the locus definitions rather than by the noisy 3D estimates. The original MediaPipe coordinates only influence the natural arm

| Locus | Person | Position $(x, y, z)$ | Plane |
|---|---|---|---|
| $\ell_1$ | 1st | $(0,\ 30,\ 20)$ | — |
| $\ell_2$ | 2nd | $(0,\ 30,\ 200)$ | Sagittal |
| $\ell_3^a$ | 3rd a | $(220,\ 30,\ 100)$ | Horizontal |
| $\ell_3^b$ | 3rd b | $(-220,\ 30,\ 100)$ | Horizontal |
| $\ell_3^{a,\mathrm{hi}}$ | 3rd indef. a | $(170,\ 230,\ 50)$ | Frontal |
| $\ell_3^{b,\mathrm{hi}}$ | 3rd indef. b | $(-170,\ 230,\ 50)$ | Frontal |

Table 5: Loci definitions, associated grammatical persons, 3D spatial positions, and reference anatomical planes used for spatial conjugation.

posture (via the elbow hint) and bone lengths, which are averaged over multiple frames to reduce noise.

### 4.2. Biomechanical Parameterization

The `.pose` files extracted from the calibration dataset that correspond to a grammatical-person combination of isolated verbs are loaded into a 3D system that respects the biomechanical constraints of the human body. This framework allows to identify and extract the mathematical rules that explain the relationship between the signer's spatial movement and the verbal agreement. How the wrist moves when changing from "I" to "you", the articulatory limits when reaching different spatial loci, or natural human acceleration and deceleration curves are extracted and serve as the basis for the procedural functions that will generate new conjugated samples.

The biomechanical parameterization defines six spatial loci in 3D space, each associated with a grammatical person in LSC. Let the signer's body center be the origin. Each locus $\ell_k$ is defined as a position vector relative to this origin (Table 5). These coordinates (in mm after shoulder-width normalization) encode the spatial arrangement of grammatical person within the three spatial planes established by (Brentari, 1998). Given a conjugation from person $i$ to person $j$, the IK offset at time $t$ is defined as:

$$o(t) = \mathbf{o}_{src} + (\mathbf{o}_{tgt} - \mathbf{o}_{src})\ S(t) \tag{1}$$

where $\mathbf{o}_{src} = \ell_i - \ell_1$, $\mathbf{o}_{tgt} = \ell_j - \ell_1$ (offsets relative to the neutral 1st-person locus), and $S(t)$ is the Hermite smoothstep function:

$$S(t) = 3t^2 - 2t^3, \quad t \in [0, 1] \tag{2}$$

which provides zero-velocity endpoints ($S'(0) = S'(1) = 0$), ensuring smooth onset and offset of the spatial transition. The parameter $t$ represents normalized time within the sign's production interval. For backward agreeing verbs, the temporal direction is reversed: the movement begins at $l_j$ (object) and ends at $l_i$ (subject).

| Body region | Joints | Weight |
|---|---|---|
| Torso | Hips | 0.15 |
| Head | Landmarks 0–10 | 0.25 |
| Shoulders | L/R shoulder | 0.30 |
| Wrists | L/R wrist (IK target) | 1.00 |

Table 6: Anatomically motivated weighting scheme for IK offset propagation across body regions

### 4.3. Data Generation

For the infinitive-form verbs, the system operates as a procedural animation engine along three complementary axes of variation, each producing new `.pose` files that are directly usable for training SLR models.

**Spatial conjugation.** Given an infinitive verb in `.pose` format, a procedural function repositions the movement trajectory in signing space according to the target person combination. An inverse kinematics algorithm (Aristidou et al., 2018) adjusts the hand positions to the target spatial loci while preserving natural shoulder and torso behavior. The system enforces biomechanical constraints to ensure that all generated poses remain within human physiological limits, much like the physics engines used in video games and virtual reality. Intermediate frames are computed through interpolation based on biologically informed motion curves.

**Inverse Kinematics Solver.** The system employs a 2-bone analytical IK solver based on the law of cosines to reposition the arm chain (shoulder $\rightarrow$ elbow $\rightarrow$ wrist) while preserving segment lengths. Given the shoulder position $p_s$, desired wrist position $p_w^*$, upper arm length $a = \|p_e - p_s\|$, and lower arm length $b = \|p_w - p_e\|$, the shoulder angle $\theta_s$ is computed as:

$$\cos(\theta_s) = \frac{a^2 + c^2 - b^2}{2ac} \tag{3}$$

where $c = \min\left(\|p_w^* - p_s\|,\ 0.99(a + b)\right)$ is the target distance, clamped to $99\%$ of the total arm length to prevent hyperextension. The elbow position is then given by:

$$p_e' = p_s + a\cos(\theta_s)\hat{d} + a\sin(\theta_s)\hat{n} \tag{4}$$

where $\hat{d}$ is the unit direction from shoulder to target wrist and $\hat{n}$ is the bend direction perpendicular to $\hat{d}$, derived from the original elbow position to maintain postural continuity.

The IK offset $o(t)$, which helps correct motion beyond the basic IK solution is not applied uniformly across the body. Instead, it is distributed with anatomically motivated weights depicted in Table 6. Each joint's displaced position is:

$$p_j' = p_j + w_j \cdot o(t) \tag{5}$$

This graduated propagation ensures that the torso responds subtly to the spatial shift while the hands reach the target loci, mimicking the natural kinematic chain behavior observed in signers (Aristidou et al., 2018).

The IK solution is blended with the original pose proportionally to the offset magnitude:

$$\alpha = \min\left(1,\ \frac{\|o(t)\|}{\tau}\right) \tag{6}$$

where $\tau = 30\,\text{mm}$ is a saturation threshold. For offsets below $\tau$, the IK influence ramps linearly, preventing abrupt transitions for small conjugations (e.g., 1st $\leftrightarrow$ 2nd person on the sagittal plane).

**Temporal Interpolation** Intermediate frames between the source and target spatial configurations are computed through the smoothstep easing function in Eq 2. This profile is consistent with the bell-shaped velocity curves observed in human reaching movements (Flash and Hogan, 1985). Additionally, an exponential moving average (EMA) filter with configurable smoothing factor $\lambda \in [0, 0.7]$ is applied to joint positions:

$$\tilde{p}_t = (1-\lambda)\, p_t + \lambda\, \tilde{p}_{t-1}. \tag{7}$$

This suppresses high-frequency jitter from the MediaPipe estimation while preserving the overall trajectory dynamics.

**Prosodic Variation.** Beyond spatial retargeting, the system procedurally modifies the temporal dynamics of each sign within linguistically attested ranges, so that multiple prosodic variants of the same conjugated form are generated, enriching the temporal diversity of the dataset. Three parameters are varied within linguistically attested ranges (Wilbur, 2009)

*Tempo scaling* $\sigma \in [0.7,\ 1.4]$. The sign duration is scaled by $T' = \frac{T}{\sigma}$ resampling via linear interpolation at the original FPS.

*Hold extension* $h \in \{0, 1, \ldots, 6\}$ frames. Additional frames are inserted at the movement apex (the frame of maximum wrist velocity), simulating the phonological hold phase.

*Ease exponent* $\gamma \in [1.0,\ 2.0]$. A generalized smoothstep remap is applied to the temporal axis:

$$f(t) = \frac{t^\gamma}{t^\gamma + (1-t)^\gamma}. \tag{8}$$

**Body morphology variation.** Finally, the system procedurally varies the body proportions of the signer, such as arm length, shoulder width, and overall stature, to simulate different physical profiles (Li et al., 2020). This morphological augmentation ensures that the resulting dataset captures inter-signer anatomical diversity, preventing models from overfitting to a single body type. The different physical profiles are simulated by parametrically scaling body proportions. Each segment is scaled relative to its proximal joint following $p'_{\text{distal}} = p_{\text{proximal}} + s \cdot (p_{\text{distal}} - p_{\text{proximal}})$ ranging the scaling factor withing the parameters shown in Table 7.

| Parameter | Range |
|---|---|
| Arm length $s_{\text{arm}}$ | $[0.85,\ 1.15]$ |
| Shoulder width $s_{\text{sh}}$ | $[0.90,\ 1.10]$ |
| Torso height $s_{\text{torso}}$ | $[0.92,\ 1.08]$ |

Table 7: Parameter ranges for body morphology variation.

Note that scaling propagates hierarchically through the kinematic chain (shoulder $\rightarrow$ elbow $\rightarrow$ wrist), ensuring that modified bone lengths remain anatomically consistent. Hand landmarks are translated by the resulting wrist displacement to maintain hand–arm coherence.

Together, applying the steps described in Sections 4.1, 4.2 and 4.3) enables a multiplicative augmentation strategy: each infinitive verb can be conjugated into all person combinations, varied prosodically, and rendered across multiple body morphologies, producing thousands of new `.pose` samples in a scalable, systematic, and reproducible manner. Table 8 summarizes the potential size of the augmented dataset depending on the number of infinitive verbs and the number of prosodic-morphological variants applied per conjugated form.

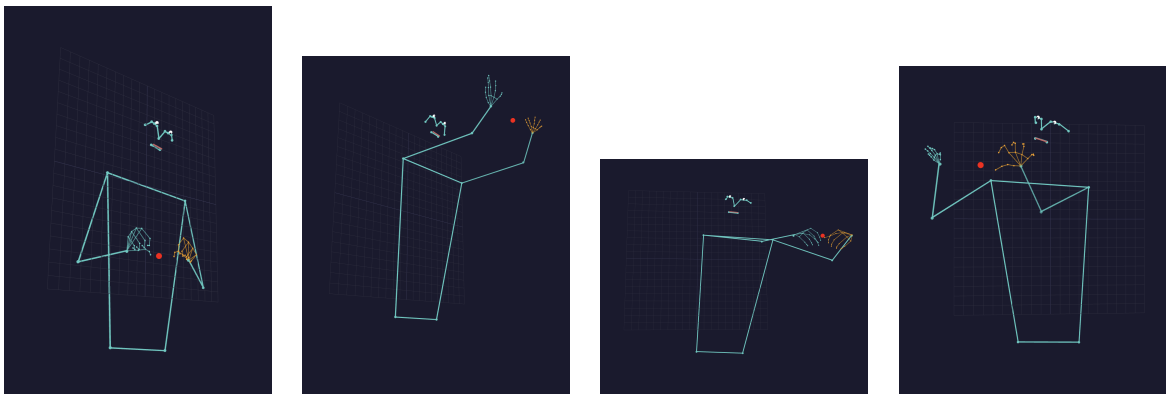
Figure 5 shows four procedurally generated spatial conjugations from the same infinitive verb. Each subfigure shows the skeleton (which later on will be applied to the avatar) with hands repositioned to different spatial loci. Each loci corresponds to distinct person agreement combinations while preserves the original movement dynamics and biomechanical constraints.

| $V$ | **Base ($\times$13)** | $n$**=5** | $n$**=10** | $n$**=25** | $n$**=50** |
|---|---|---|---|---|---|
| 1 | 13 | 65 | 130 | 325 | 650 |
| 10 | 130 | 650 | 1,300 | 3,250 | 6,500 |
| 25 | 325 | 1,625 | 3,250 | 8,125 | 16,250 |
| 48 | 624 | 3,120 | 6,240 | 15,600 | 31,200 |

Table 8: Augmented dataset size as a function of $V$ (number of infinitive verbs) and $n$ (number of prosodic $\times$ morphological variants). The *Base* column represents the conjugated forms without prosodic or morphological variation. $V$=48 corresponds to our full calibration dataset.

## 5. Evaluation

While the proposed system is based on deterministic algorithms that guarantee reproducibility and



(a) (b) (c) (d)

Figure 5: Examples of procedurally generated skeletons using different spatial conjugations from the same infinitive verb (TELL). Each subfigure shows the skeleton with hands repositioned to distinct spatial loci via IK: (a) $1\text{TELL}2$ — "I tell you" (sagittal plane, $\ell_1 \rightarrow \ell_2$); (b) $1\text{TELL}3^a$ — "I tell him" (horizontal plane, $\ell_1 \rightarrow \ell_{3a}$); (c) $2\text{TELL}3^b$ — "You tell her" (horizontal plane, $\ell_2 \rightarrow \ell_{3b}$); (d) $3^a\text{TELL}1$ — "He tells me" (horizontal plane, $\ell_{3a} \rightarrow \ell_1$). This skeleton is subsequently applied to an avatar to produce the resulting synthetic data.

geometric consistency, the naturalness and linguistic validity of the generated animations must be empirically validated. The evaluation of the system takes two perspectives. First, by assessing the anatomical feasibility and fluidity of the synthesized avatars compared to renderings from the original videos. Second, by addressing movement and linguistic legibility, incorporating feedback from LSC deaf signers, which is essential in the deaf-signing community because automated metrics are not sufficient to validate semantics (Bragg et al., 2019).

### 5.1. Anatomical Feasibility

The procedural system guarantees biomechanical plausibility by design: the generation functions operate within the anatomical constraints of the human body, ensuring that no generated configuration exceeds physiological limits. Nevertheless, this guarantee must be verified empirically. For each generated sample, spatial trajectories are validated to ensure they remain within the articulatory workspace and that the resulting movements are consistent with the original captured data. This evaluation prevents the introduction of anatomically impossible configurations that could reduce realism and pollute language quality.

### 5.2. Fluency

Motion fluency can be evaluated by analyzing the trajectory continuity in the joint space. Since the system relies on interpolation using biologically informed motion curves, smoothness may be assessed by examining first- and second-order derivatives (velocity and acceleration) of joint trajectories. Bézier curves and spline-based interpolation methods are widely used in computer animation to ensure smooth motion transitions (Izdebski et al., 2020). With this approach, it is verified that no discontinuities occur in position, velocity, or acceleration. No abrupt kinematic jumps are introduced during retargeting or conjugation, and temporal transitions follow smooth parametric curves consistent with biological motion profiles.

### 5.3. Legibility

Even if an animation is anatomically feasible and kinematically smooth, it must preserve its linguistic meaning. Therefore, legibility must be evaluated through user testing with signers or fluent users of the target sign language (Bragg et al., 2019), for this particular case deaf LSC signers. Participants should be shown synthesized conjugated forms and asked to identify the verb and its agreement direction, judge naturalness and acceptability and report whether spatial modification altered the intended meaning. This evaluation ensures that the mathematical manipulation of spatial parameters does not compromise semantic integrity or grammatical correctness and complements the objective assessments from the previous assessments with user-centered evaluation, which is a standard practice in sign language avatar research.

Together, these three evaluation dimensions, namely: biomechanical validity, kinematic smoothness, and linguistic intelligibility provide a multilayered validation framework that addresses both physical realism and communicative effectiveness.

## 6. Discussion and conclusions

This paper has presented a novel hybrid methodology for generating high-quality synthetic training data for Catalan Sign Language (LSC). By focusing on the complex morphological domain of verbal agreement, a significant bottleneck has been addressed in the development of SLR systems: the scarcity of large-scale, annotated video corpora for low-resource languages. The contributions of the paper may be grouped in three domains.

Procedural augmentation: Unlike purely generative AI approaches often struggle with "hallucinations" or anatomical inconsistency, our system utilizes 3D biomechanics and kinematics algorithms. This ensures that all synthesized movements remain physically plausible and strictly adhere to the anatomical constraints of the human body. Moreover, the proposed methodology allows for the generation of the 12 verbal conjugation combinations for each infinitive, multiplied by $n$, $n$ being different possible prosodies and morphologies which can be added in each new context, therefore showing a great potential of data augmentation.

Multidimensional data generation: By leveraging three procedural axes: spatial conjugation, prosodic variation, and body morphology, single infinitive verbs are systematically transformed into thousands of unique, usable `.pose` files. This approach captures essential inter-signer diversity and linguistic nuances, such as person agreement and temporal dynamics.

Preservation of linguistic structure: The methodology reinforces the spatial context of LSC by mapping grammatical person to specific spatial planes (horizontal, frontal, and midsagittal). By using gloss-level translation as an intermediate objective, we reduce the complexity of the visual recognition task while maintaining the semantic integrity of the data.

The proposed validation framework—spanning anatomical feasibility, kinematic fluency, and linguistic legibility—ensures that the generated data is not only realistic but also communicatively effective within the deaf-signing community. Preliminary results suggest that procedural synthesis can effectively bridge the data gap for low-resource sign languages like LSC. Future research will focus on integrating these synthetic datasets into end-to-end SLR systems to measure the actual improvement in translation performance. Additionally, we aim to expand the procedural engine to cover other complex grammatical markers (including manual, non-manual and spatial components), further enhancing the richness of synthetic sign language resources.

## 7. Acknowledgements

This research has partially been funded by the European Union Next Generation EU and supported by the Government of Catalonia Grant Agreement No. SDC007/25/000093. Barberà acknowledges the project PID2020-119041GB-I00 funded by MICIU/AEI/10.13039/501100011033. We warmly thank the deaf LSC participants who participated in the video recordings for the calibration dataset without whom this research would not have been possible.

## 8. Bibliographical References

Sarah Alyami and Hamzah Luqman. 2025. A comparative study of rgb-based continuous sign language recognition techniques. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4923–4932.

A. Aristidou, J. Lasenby, Y. Chrysanthou, and A. Shamir. 2018. Inverse kinematics techniques in computer graphics: A survey. *Computer Graphics Forum*, 37(6):35–58.

Gemma Barberà. 2015. *The meaning of space in sign language. Reference, specificity and structure in Catalan Sign Language discourse*. Mouton de Gruyter & Ishara Press, Berlin/Boston.

Gemma Barberà, Josep Quer, and Santiago Frigola. 2015. Primers passos cap a la documentació de discurs signat. El projecte pilot de constitució del corpus de la llengua de signes catalana [First steps towards the documentation of signed discourse. The pilot project for the creation of the Catalan Sign Language corpus]. *Treballs de Sociolingüística Catalana*, 25:287–302.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. The MIT Press, Boston.

Roger Cassany Viladomat, Xavier Vinaixa Roselló, and Marcel Mauri de los Ríos. In press. Identidad sonora personalizada mediante IA para

personas sordas signantes. In *CILCS Libro de Actas 2025*. Tirant Lo Blanch.

T Flash and N Hogan. 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *The Journal of Neuroscience*, 5(7):1688–1703.

Google LLC. 2023. Mediapipe.

Matt Huenerfauth and Pengfei Lu. 2012. Effect of spatial reference and verb inflection on the usability of sign language animations. *Universal Access in the Information Society*, 11:169–184.

IEC. 2025. *Corpus de referència de la llengua de signes catalana (LSC)*. (CORPUS LSC).

Łukasz Izdebski, Ryszard Kopiecki, and Dariusz Sawicki. 2020. Bézier curve as a generalization of the easing function in computer animation. In *Advances in Computer Graphics*, pages 382–393, Cham. Springer International Publishing.

Jong-Wook Kim, Jin-Young Choi, Eun-Ju Ha, and Jae-Ho Choi. 2023. Human pose estimation using mediapipe pose and optimization method based on a humanoid model. *Applied Sciences*, 13(4).

Edward Klima and Ursula Bellugi. 1979. *The Signs of Language*. Harvard University Press, Cambridge/London.

Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. 2020. Transferring cross-domain knowledge for video sign language recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6204–6213.

Euan McGill. 2026. *The Data Problem behind Sign Language Translation*. Ph.D. thesis, Pompeu Fabra University, Barcelona.

Amit Moryossef. 2024. pose-anonymization: Remove identifying information from sign language poses. https://github.com/sign-language-processing/pose-anonymization.

Amit Moryossef, Mathias Müller, and Rebecka Fahrni. 2021. pose-format: Library for viewing, augmenting, and handling .pose files. https://github.com/sign-language-processing/pose.

Carol A Padden. 1988. *Interaction of morphology and syntax in American Sign Language*. Garland Publishing, New York/ London.

Josep Quer. 2021. Verb agreement – theoretical perspectives. In Josep Quer, Annika Hermann, and Roland Pfau, editors, *Routledge Handbook of Theoretical and Experimental Sign Language Research*, pages 95–121. Routledge, London/New York.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.

Noha Sarhan and Simone Frintrop. 2023. Unraveling a decade: A comprehensive survey on isolated sign language recognition. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3202–3211.

Harry Walsh, Maksym Ivashechkin, and Richard Bowden. 2025. Using sign language production as data augmentation to enhance sign language translation. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Ronnie B Wilbur. 2009. Effects of varying rate of signing on asl manual signs and nonmanual markers. *Language and speech*, 52(2-3):245–285.

# Leveraging Unannotated Sign Language Data via a Robust Data Augmentation Method for Contrastive Representation Learning

**Ariel Basso Madjoukeng, Pierre Poitier, Edith Belise Kenmogne**
**Adélaïde Couplet, Margaux Leleu, Benoît Frénay**
NADI/HuMaLearn at University of Namur
Rue Grangagnage 21, Namur, Belgium
arielbassodev@gmail.com

**Abstract**

Contrastive learning is a deep learning paradigm that allows learning of useful representations without annotations. In many fields, including sign language recognition (SLR), contrastive approaches have proven to be very effective for developing pretrained models. To learn representations, they generate augmented variants of an instance through augmentation techniques and then maximize their similarities. The quality of the learned representations is strongly correlated with the augmentations used during training. In several fields, specialized augmentations have been developed and adopted. However, in SLR, we observed two trends: contrastive-based SLR approaches often rely on augmentations that are not realistic for the application (e.g., vertical flip, excessive rotations); specialized augmentation methods lack of robustness. Hence, when used as a starting point for contrastive algorithms, the learned representations are often irrelevant and sometimes sensitive. These issues considerably affect the accuracy of SLR models on downstream tasks. In response, this paper proposes a robust augmentation method specially designed for contrastive approaches applied to SLR. The results show an improvement in accuracy during linear evaluation and semi-supervised learning with only 30% of annotations.

**Keywords:** Sign language recognition, Contrastive representation learning, Data augmentation

## 1. Introduction

Worldwide, there exist more than 300 sign languages, each with its own properties. In this field, the annotation process is a laborious and time-consuming task that requires linguistic expertise. Studies report around 100 hours of work to annotate one hour of video (Renz et al., 2021). Among the existing sign languages, many are underrepresented. Even when they are represented, they are often partially annotated (Fink et al., 2021; Albanie et al., 2021). For this reason, sign language suffers from a scarcity of annotated data. Faced with this limitation, recent studies increasingly leverage unannotated data to build pretrained models (Wong et al., 2025; Jiang et al., 2024; Gueuwou et al., 2025). These models are designed to learn meaningful representations from unannotated data, then be fine-tuned with fewer annotations and achieve good results. To build pretrained models, several paradigms exist; contrastive representation learning is one of the most well-known (He et al., 2020). It is simple, flexible, and requires fewer computational resources compared to language-model-based approaches (Chen et al., 2020).

Contrastive approaches learn by maximizing the similarity between augmented views of an instance (positive pairs) while minimizing their similarity with other instances (negative pairs). They have the particularity of learning representations that are invariant to the augmentations used during the pretraining. To be useful, augmentations used during training should reflect the target application (Mansfield et al., 2023; Madjoukeng et al., 2025b). Hence, to enable the learning of relevant representations, specialized augmentation methods have been proposed in several fields (time series forecasting, plant anomaly detection, etc.). These specialized augmentations generate positive pairs by degrading the parts of the sample that are not relevant for the application (e.g., background in the case of images (Madjoukeng et al., 2025b,a)).

In SLR, Madjoukeng et al. (2026) observed that due to coarticulation (Poitier et al., 2024) and repositioning movements, not all parts of sign videos are relevant for their identification. Based on this observation, they generate positive pairs by degrading the non-relevant parts of the videos while keeping the relevant parts unchanged. This strategy enables contrastive approaches to learn representations that are invariant to these non-informative segments, thereby improving their overall quality. This method has led to a significant success. However, in real-world SLR scenarios, there exist variabilities between signers (hand shapes, positions, speed, etc.) and due to factors such as occlusions or others, noise may affect the parts that are relevant parts of signs. For a robust and accurate representation learning, augmentation methods should not focus solely on non-relevant parts but also consider variations in the relevant ones. Existing augmentation, by focusing solely on relevant parts completely ignores these factors. This results in representations that lack robustness, fail



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 10–16
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

to account for real-world scenarios, and poor accuracy in downstream tasks. To the best of our knowledge, these issues have not been addressed in SLR. In response, this paper presents a robust augmentation method designed to satisfy two objectives: (i) generating positive pairs while degrading non-relevant parts, and (ii) simulating subtle and realistic variations on the relevant parts. When applied to a contrastive approach, this method allows learning useful and robust representations from unannotated data. By allowing better leveraging of unannotated data, the proposed method is highly valuable for low-resource sign languages. The results show an improvement in accuracy during the linear evaluation and fine-tuning across several sign language datasets.

The rest of this paper is organized as follows: Section 2 presents contrastive approaches; Section 3 presents SLR; Section 4 presents existing data augmentations for SLR; Section 5 presents the proposed augmentation and Section 6 presents the results and discussion.

## 2. Contrastive Representation Learning

In contrastive learning, several methods exist; SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) are two popular approaches. They learn through four common steps: positive pairs generation, encoding, projection and loss computation. For each instance, augmentation methods are applied to generate positive pairs. They are then passed through an encoder (e.g., ResNet, Transformer) followed by a projection head that computes numerical representations. On these representations, a contrastive loss is computed to maximize the similarity between positive pairs while minimizing their similarity with other instances. The main difference between these approaches is that MoCo stores instances from previous iterations in a queue, increasing the quality of the learned representations. Furthermore, several contrastive methods that rely only on positive pairs have emerged. Among them, BYOL (Grill et al., 2020), SimSiam (Chen and He, 2021), and SL-FPN (Madjoukeng et al., 2026) are three well-known approaches. BYOL uses two encoders: the online and the target encoder. During training, the online encoder is trained to predict the output of the target encoder. SimSiam uses a single encoder for the positive pairs and, following a Siamese architecture (Koch et al., 2015), aligns the representations of the different views. SL-FPN leverages positive pairs and original instances, aligning their representations. These approaches have been used to build pretrained SLR models and have achieved great success. The next section presents previous works on SLR.

## 3. Sign Language Recognition

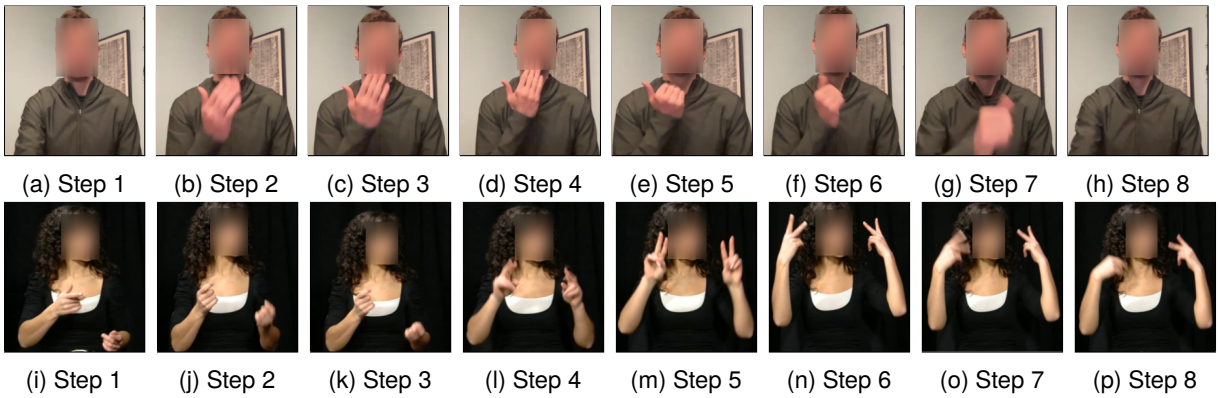
For several sign languages, depending on the amount of data available, SLR models have been proposed. For the French Belgian Sign Language (LSFB), Fink et al. (2021) conducted pioneering work, introducing a dataset consisting of more than 4,567 distinct signs. Using this dataset, they leveraged a Vision Transformer (Dosovitskiy et al., 2020) and achieved 54.4% accuracy on a subset of 700 different signs. For the Argentinian sign language (LSA), Masood et al. (2018) proposed a dataset consisting of multiple instances of 64 different signs. Building upon this dataset, Alyami et al. (2024) benchmarked several architectures and reported 98.25% accuracy with a Transformer-based model. For the Greek sign language, Adaloglou et al. (2021) presented a comprehensive dataset consisting of 310 different signs. Using this dataset, they combined an I3D (Carreira and Zisserman, 2017) with a BiLSTM (Huang et al., 2015) and achieved 89.74% accuracy. Papadimitriou et al. (2024) proposed a multimodal approach with both appearance and skeleton information and achieved 96.21% accuracy.

For the American sign language (ASL), Desai et al. (2023) presented the ASL-Citizen dataset, which consists of 2,731 signs. On this dataset, they trained an ST-GCN and achieved 59.52% accuracy. Furthermore, Jiang et al. (2024) proposed SignCLIP, an architecture that leverages two modalities (text and video), with one encoder for each modality and aligns their representations using a contrastive loss. On this dataset, they achieved 46% accuracy during fine-tuning. The main challenge with this approach is the fact that it requires a vast amount of data from diverse sources to be trained.

In many cases, sign language datasets are not fully annotated (Albanie et al., 2021; Fink et al., 2021). Often, practitioners prefer a model that can be trained on unannotated data from their specific sign language and fine-tuned on data of the same type. Contrastive approaches have been shown to be very effective for this task (Chen et al., 2020; He et al., 2020; Grill et al., 2020). However, the relevance of the representations learned by these methods is directly linked to the augmentations used during pretraining. For SLR, several augmentations are commonly used. The next section presents them.

## 4. SLR Data Augmentation Methods

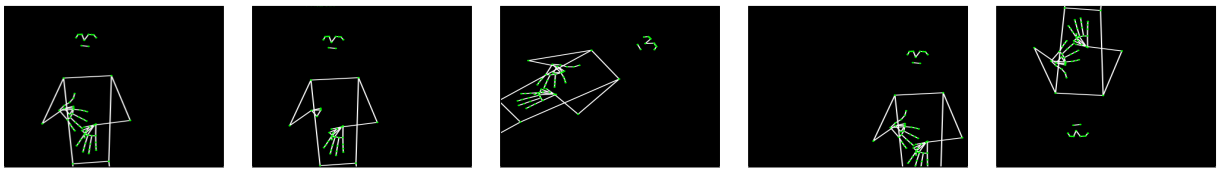
In SLR, data augmentation methods are generally divided into two categories: spatial and temporal augmentations. Spatial augmentations are those that modify the spatial distribution of frames. They include Gaussian noise, translation, flipping, rota-



(a) Step 1 (b) Step 2 (c) Step 3 (d) Step 4 (e) Step 5 (f) Step 6 (g) Step 7 (h) Step 8

(i) Step 1 (j) Step 2 (k) Step 3 (l) Step 4 (m) Step 5 (n) Step 6 (o) Step 7 (p) Step 8

Figure 1: Repositioning and coarticulation mouvement on the ASL (Desai et al., 2023) (first row), and LSFB (Fink et al., 2021) (second row). For privacy reasons, the faces of the signers have been blurred.



(a) Original sign (b) Masking (c) Rotation (d) Translation (e) Vertical flip

Figure 2: A frame and its augmented variants extracted from a video of the LSFB dataset using MediaPipe (Lugaresi et al., 2019).

tion (Lingg et al., 2022) and many others. Figure 2 shows an example of spatial augmentations applied to a skeleton. This skeleton corresponds to one of the frames extracted from a video in the LSFB dataset and processed using MediaPipe (Lugaresi et al., 2019).

Temporal augmentations alter the temporal dynamics of video sequences, including frame permutation, random frame dropping, frame speed-up, and frame slow-down. These augmentations are generic and, when used in contrastive learning for SLR to generate positive pairs, do not allow the learning of targeted representations. Faced with this issue, Madjoukeng et al. (2026) proposed a data augmentation method specifically designed for contrastive learning applied to SLR. They observe that in sign language videos, due to certain movements (repositioning, coarticulation (Poitier et al., 2024), etc.), not all parts are relevant for classification. Based on this observation, they propose a data augmentation method specially designed for contrastive approaches in SLR. Their method consists of generating positive pairs by altering the order of the frames in the non-relevant parts while keeping the relevant frames unchanged. This augmentation has proved to be effective with contrastive algorithms in SLR. However, the fact that any transformation is applied on the relevant parts of the videos represent a major limitation. Indeed, contrastive algorithms are therefore not constrained to be robust to small perturbations or to variations in the signer's position on these crucial segments. Due to the highly visual and dynamic nature of the movements in the SLR context, this limitation is particularly problematic. To address this limitation, the next section introduces a dedicated robust augmentation method.

## 5. Proposed Augmentation Method

To learn meaningful representations, contrastive approaches require data augmentations capable of generating positive pairs while taking into account the particularities of the application domain. In SLR, a video contains both relevant and non-relevant segments. To learn meaningful representations, a contrastive approach should be able to ignore non-relevant segments while being efficient (robust to subtle perturbations, variations, etc.) on the relevant ones. To enable contrastive approaches to achieve these objectives, this section introduces a novel data augmentation specifically designed for contrastive approaches in SLR. It first explains why not all parts of a video are relevant for a classification model. Then, it describes an existing method for identifying the relevant and non-relevant segments of a video. Finally, presents the proposed augmentation.

To illustrate the fact that in sign language, not all parts of a video are relevant for sign recognition, let us consider Figure 1. The first row is a video from the ASL- Citizen and the second from

the LSFB dataset. The different videos were split into 8 frames (steps 1 to 8). From this figure, we note that the signers begin the motion in the second frame (step 2) and continues signing until the fifth frame (for the ASL video) and the sixth frame (for the LSFB video). After these frames, the signers tend to reposition their hands. Such movements are not the essence of the sign and a reliable model should not focus on these frames to classify the signs.

At this step, a natural question arises: how to determine the relevant and non-relevant parts of the signs. For this task, Madjoukeng et al. (2026) proposed an empirical strategy based on the variation of the accuracy according to the degradation of the frames. Their approach consists of degrading progressively (through permutations) the frames from the first to the last and from the last to the first. After each permutation, they computed the accuracy. They assume that, for a sign to make sense, the order of the frames is important (due to the continuous nature of the movement). Therefore, if the relevant parts for identifying a sign are altered, the accuracy during the evaluation will be affected. Hence, they determine from which frame the movement starts and ends to be informative for a model.

In their approach, they generated positive pairs by focusing only on the non-relevant segments without helping the contrastive model to learn robust representations on the relevant parts. For contrastive approaches in SLR, an effective augmentation should satisfy at least these two properties: (i) generating positive pairs by degrading non-relevant parts, and (ii) introducing subtle and realistic perturbations on the relevant ones. Degrading non-relevant parts allows contrastive approaches to abstract away non-relevant movements, while subtle perturbations to relevant parts enable contrastive approaches to become robust to factors such as positional variations and joint-level errors that may occur in the data.

The proposed method is specifically designed to meet these criteria. Algorithm 1 summarizes the proposed approach. For a sign $x$, the relevant $(x^r)$ and non-relevant $(x^{nr})$ parts are determined using a function called `relevant&non_relevant_parts` following Madjoukeng et al. (2026). The non-relevant parts are split into two parts: the first non-relevant part $(x^f)$ and last non-relevant part $(x^l)$. They are typically corresponding to the coarticulation and repositioning movements. After this decomposition, temporal permutations $(\pi_1, \pi_2)$ and $(\pi'_1, \pi'_2)$ are applied to the non-relevant parts $(x^f, x^l)$ while a composition of translation and noise ($t_1 \circ g_1$ and $t_2 \circ g_2$ ) is applied on the relevant parts $(x^r)$. Since the meaning of a sign is linked to the order of the frames in the video, temporal permutation induces the learning of representations that are invariant to the order of the frames in these parts. This encourages the model to ignore these parts. The composition of translation and noise simulates subtle perturbations and enhances the robustness of the models. Indeed, noise simulates inaccuracies arising from pose estimation, sensor noise, while translation models small global shifts in hand and body position that commonly occur across different signers and recording setups. By applying these perturbations exclusively to relevant frames, the model is encouraged to become robust to natural variability in informative segments. Note that this method can be used by any contrastive approach. The next section presents an evaluation on several contrastive approaches.

**Algorithm 1:** Proposed Augmentation

**Data:** $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$: set of signs
**Result:** $\mathcal{X}'$: augmented sequences

**for** $x \in \mathcal{X}$ **do**
- $x^r, x^{nr} \leftarrow$ `relevant&non_relevant_parts`$(x)$
- $x^f, x^l \leftarrow x^{nr}$
- $A_1 \leftarrow t_1 \circ g_1$
- $A_2 \leftarrow t_2 \circ g_2$
- $x_1 \leftarrow \pi_1(x^f) \parallel A_1(x^r) \parallel \pi_2(x^l)$
- $x_2 \leftarrow \pi'_1(x^f) \parallel A_2(x^r) \parallel \pi'_2(x^l)$
- Add $(x_1, x_2)$ to $\mathcal{X}'$

**return** $\mathcal{X}'$

## 6. Experiments

This section evaluates the proposed augmentation. It presents the experiments, datasets, results and discussion.

### 6.1. Conducted Experiments

To evaluate the effectiveness of the proposed approaches, two types of evaluations were used: a linear evaluation protocol and semi-supervised learning. Linear evaluation consists of pretraining a backbone on unannotated sign language data, freezing the backbone, and training a simple classifier on top of it. Semi-supervised learning consists of training a pretrained backbone with only 30% of annotations. This evaluation is important to simulate data scarcity and partially annotated sign language datasets.

### 6.2. Datasets

For this study, four datasets with different sizes were used. First, the American sign language dataset (Desai et al., 2023) which consists of 83,399 videos covering 2,731 signs. Second, the Greek sign language dataset (Adaloglou et al., 2021) containing 40,785 videos corresponding to 310 unique signs. Third, the Argentinian sign language dataset (Masood et al., 2018) comprising 3,200 videos representing 64 different signs. Finally, the LSFB dataset (Fink et al., 2021) containing more than 47,551 videos spanning 4,567 distinct signs. These datasets were split according to their original papers (e.g., 70/30 for the LSFB, 80/20 for the LSA, etc.). During the experiments, we used all the classes of the LSA and GSL datasets, and a subset of 500 different classes from the LSFB and ASL datasets. Each model was trained five consecutive times and the results were reported with 95% confidence.

### 6.3. Training setup

For the experiments, Python 3.10 and PyTorch 2.5 were used. The Lightly library was used to implement the different contrastive approaches[1]. For training, a ViT Transformer (Dosovitskiy et al., 2020) was used as the backbone. Indeed, it has been widely used as a backbone in various studies (Fink et al., 2023; Madjoukeng et al., 2026). As in several studies, the videos were first transformed into skeletons using MediaPipe (Lugaresi et al., 2019). This reduces the computational cost and the influence of elements such as the background. The different contrastive approaches were trained for 200 epochs, using the parameters specified in their original papers (Chen et al., 2020; He et al., 2020; Grill et al., 2020). The batch size was set to 512. During the fine-tuning stage, the models were trained for 1000 epochs using an SGD optimizer.

### 6.4. Results and Discussion

The first evaluation consisted of comparing the proposed method with existing specialized SLR augmentation. For this evaluation, we used five contrastive methods and the different datasets. On each dataset, we evaluated the proposed augmentation against the one proposed by Madjoukeng et al. (2026). Table 1 presents the results obtained. The best results are highlighted in bold. We observe that, in linear evaluation, the proposed approach outperforms existing method in most cases. This shows that, to learn relevant representations, contrastive approaches require augmentation methods that are likely to reflect the phenomena of the specific application domain. In semi-supervised learning, we observe that with our approach, contrastive methods perform better than using existing augmentation. These results are promising and highlight the potential impact of contrastive approaches in addressing the problem of scarcity of annotated data in sign language.

[1] https://docs.lightly.ai/self-supervised-learning/index.html

## 7. Conclusion

This paper presented a robust data augmentation method designed to enable contrastive approaches to learn relevant representations. The proposed method consists of generating positive pairs by degrading the non-relevant parts of the videos, while applying subtle and realistic perturbations on the relevant ones. The proposed approach has been evaluated with various contrastive approaches on different datasets. The results showed an improvement in terms of accuracy during linear evaluation and semi-supervised learning on various datasets. These results demonstrate the potential of contrastive approaches to leverage unannotated data in sign language. Given the scarcity of annotated data in this field, this advancement represents a major contribution. As further work, we plan to extend the proposed method to continuous sign language recognition.

## 8. Acknowledgments

The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n°1910247.

Table 1: Comparison against (Madjoukeng et al., 2026) on linear evaluation, fine-tuning with 30% of annotations. The best results are highlighted in bold.

| Method | Dataset | SimCLR | MoCo | SimSiam | BYOL | SL-FPN |
|---|---|---|---|---|---|---|
| ***Linear Evaluation Results*** | | | | | | |
| Madjoukeng et al. (2026) | LSFB | $14.16\% \pm 0.24$ | $13.68\% \pm 0.48$ | $15.26\% \pm 0.67$ | $14.72\% \pm 0.65$ | $23.73\% \pm 0.53$ |
| | ASL | $14.13\% \pm 0.42$ | $14.69\% \pm 0.35$ | $15.91\% \pm 0.56$ | $16.43\% \pm 0.96$ | $20.46\% \pm 1.21$ |
| | GSL | $34.19\% \pm 0.85$ | $36.15\% \pm 0.69$ | $32.01\% \pm 0.54$ | $34.09\% \pm 0.93$ | $47.76\% \pm 0.79$ |
| | LSA | $34.02\% \pm 1.24$ | $35.69\% \pm 1.06$ | $30.06\% \pm 2.14$ | $37.47\% \pm 1.51$ | $41.74\% \pm 1.08$ |
| Ours | LSFB | **16.89**$\% \pm 0.33$ | **16.13%** $\pm 0.81$ | **16.77%** $\pm 0.39$ | **17.71%** $\pm 0.10$ | $23.90\% \pm 0.81$ |
| | ASL | **16.28%** $\pm 0.39$ | **17.09%** $\pm 0.30$ | $16.17\% \pm 0.56$ | **18.39%** $\pm 0.20$ | **22.13%** $\pm 0.63$ |
| | GSL | **36.23%** $\pm 0.51$ | **37.14%** $\pm 0.17$ | **35.57%** $\pm 0.59$ | **36.17%** $\pm 0.69$ | $48.57\% \pm 0.42$ |
| | LSA | **37.08%** $\pm 1.13$ | $36.21\% \pm 0.89$ | $32.22\% \pm 0.19$ | $38.58\% \pm 1.11$ | **43.98%** $\pm 0.88$ |
| ***Partial Fine-tuning*** | | | | | | |
| Madjoukeng et al. (2026) | LSFB | $42.69\% \pm 2.50$ | $42.23\% \pm 2.14$ | $43.69\% \pm 2.69$ | $41.40\% \pm 2.04$ | $49.93\% \pm 2.98$ |
| | ASL | $47.43\% \pm 0.77$ | $47.49\% \pm 0.54$ | $47.23\% \pm 0.63$ | $47.02\% \pm 0.51$ | $49.28\% \pm 0.79$ |
| | GSL | $78.82\% \pm 2.96$ | $77.42\% \pm 2.87$ | $77.02\% \pm 2.95$ | $78.04\% \pm 2.65$ | $83.86\% \pm 2.01$ |
| | LSA | $87.69\% \pm 1.48$ | $88.04\% \pm 1.69$ | $87.96\% \pm 0.06$ | $88.64\% \pm 1.36$ | $92.76\% \pm 1.63$ |
| Ours | LSFB | **44.57**$\% \pm 1.16$ | **44.09**$\% \pm 0.98$ | **48.03**$\% \pm 1.07$ | $44.14\% \pm 1.14$ | **51.97**$\% \pm 0.51$ |
| | ASL | $47.21\% \pm 0.19$ | **48.05**$\% \pm 0.34$ | **48.13**$\% \pm 0.31$ | **48.14**$\% \pm 0.99$ | **52.56**$\% \pm 0.98$ |
| | GSL | $79.33\% \pm 1.14$ | $78.41\% \pm 0.59$ | **79.07**$\% \pm 1.19$ | **80.05**$\% \pm 1.02$ | $83.07\% \pm 0.60$ |
| | LSA | **89.01**$\% \pm 0.17$ | $88.96\% \pm 0.56$ | **88.57**$\% \pm 0.36$ | **89.17**$\% \pm 0.51$ | $92.88\% \pm 0.11$ |

## 9. Bibliographical References


Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2021. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE transactions on multimedia*, 24:1750–1762.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*.

Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. 2024. Isolated arabic sign language recognition using a transformer-based model and landmark keypoints. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–19.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. Proceedings of Machine Learning Research.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2023. Asl citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36:76893–76907.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. 2021. Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*.

Jerome Fink, Pierre Poitier, Maxime André, Loup Meurice, Benoît Frénay, Anthony Cleve, Bruno Dumas, and Laurence Meurant. 2023. Sign language-to-text dictionary with lightweight transformer models. In *IJCAI*, pages 5968–5976.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025. Signmusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. Signclip: Connecting text and sign language by contrastive learning. *arXiv preprint arXiv:2407.01264*.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pages 1–30. Lille.

Nico Lingg, Miguel Sarabia, Luca Zappella, and Barry-John Theobald. 2022. Contrastive self-supervised learning for skeleton representations. *arXiv preprint arXiv:2211.05304*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Ariel Basso Madjoukeng, Jerome Fink, Pierre Poitier, Edith Belise Kenmogne, and Benoît Frénay. 2025a. Benchmarking data augmentation for contrastive learning in static sign language recognition. In *ESANN 2024: 32nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. i6doc. com.

Ariel Basso Madjoukeng, Jérôme Fink, Pierre Poitier, Edith Belise Kenmogne, and Benoit Frenay. 2026. SSL-SLR: Self-supervised representation learning for sign language recognition. *Transactions on Machine Learning Research*.

Ariel Basso Madjoukeng, Edith Bélise Kenmogne, Pierre Poitier, Benoît Frénay, and Jérôme Fink. 2025b. Local-global data augmentation for contrastive learning in static sign language recognition. In *International Symposium on Intelligent Data Analysis*, pages 54–66. Springer.

Philip Andrew Mansfield, Arash Afkanpour, Warren Richard Morningstar, and Karan Singhal. 2023. Random field augmentations for self-supervised representation learning. *CoRR*.

Sarfaraz Masood, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad. 2018. Real-time sign language gesture (word) recognition from video sequences using cnn and rnn. In *Intelligent Engineering Informatics: Proceedings of the 6th International Conference on FICTA*, pages 623–632. Springer.

Katerina Papadimitriou, Galini Sapountzaki, Kyriaki Vasilaki, Eleni Efthimiou, Stavroula-Evita Fotinea, and Gerasimos Potamianos. 2024. A large corpus for the recognition of greek sign language gestures. *Computer Vision and Image Understanding*, 249:104212.

Pierre Poitier, Jérôme Fink, and Benoît Frénay. 2024. Towards better transition modeling in recurrent neural networks: The case of sign language tokenization. *Neurocomputing*, 567:127018.

Katrin Renz, Nicolaj C Stache, Samuel Albanie, and Gül Varol. 2021. Sign language segmentation with temporal convolutional networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139. IEEE.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2025. Signrep: Enhancing self-supervised sign representations. *arXiv preprint arXiv:2503.08529*.

# The SMILE Continuous DSGS Corpus: A Resource for Longitudinal Exploration of Continuous Swiss German Sign Language

**Alessia Battisti*, Katja Tissi†, Sandra Sidler-Miserez†, Sarah Ebling***
*University of Zurich
Andreasstrasse 15, 8050 Zurich
{battis, ebling}@cl.uzh.ch

†University of Teacher Education in Special Needs
Schaffhauserstrasse 239, 8050 Zurich
katja.tissi@hfh.ch | sandysidler@gmail.com

**Abstract**

This paper presents the *SMILE Continuous DSGS Corpus*, a longitudinal dataset that allows for investigating how hearing adults acquire Swiss German Sign Language as a second language. It includes recordings of sign language learners and native signer controls collected at four points over a period of 18 months and annotated for manual and non-manual components, errors, and sentence-level acceptability. The resource provides high-quality, synchronized video suitable for both linguistic and automatic sign language processing research, for example, supporting studies of interlanguage development and training of automatic sign language recognition models. We present here an exploratory analysis of the learner subcorpus using Bayesian mixed-effects modeling. The corpus and accompanying annotations are available for research purposes under a Creative Commons license (CC BY-NC-SA 4.0).

**Keywords:** Swiss German Sign Language dataset, automatic sign language assessment, sign language learner analysis, Bayesian mixed-effects modeling, multimodal annotation, sign language processing

## 1. Introduction

Sign languages are full-fledged natural languages with phonological, morphological, and syntactic structures expressed through manual and non-manual articulations. While research on sign language linguistics has advanced substantially over the past two decades, the development of standardized, annotated resources still lags behind that of spoken languages[1]. The emergence of large-scale corpora such as the British Sign Language (BSL) Corpus (Schembri, 2008), the German Sign Language (*Deutsche Gebärdensprache*, DGS) Corpus (Konrad et al., 2020), and the Corpus of the Sign language of the Netherlands (*Nederlandse Gebarentaal*, NGT) (Crasborn et al., 2006-2017) has changed linguistic and computational research by providing empirical foundations for studying grammatical variation and use within deaf[2] communities. These resources have also supported the creation of digital tools for lexicography as well as catalyzed progress in automatic sign language processing (SLP), fostering advances in recognition and generation.

Despite this progress, comparable resources remain scarce for sign languages such as Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS), the first sign language of deaf and hard-of-hearing persons in German-speaking Switzerland. DSGS is used by approximately 7,000 native or early signers and about 13,000 hearing individuals who engage with the language for various purposes (Boyes Braem, 2024). DSGS has gained institutional visibility in education, media, and accessibility contexts, yet still lacks a comprehensive corpus suitable for both linguistic research and computational modeling.

To address this need, the SMILE Continuous DSGS Corpus was created within a four-year interdisciplinary project aimed at advancing sign language learning and assessment technologies. The resource combines DSGS productions of both hearing adult (L2) and native (L1) signers, recorded across multiple tasks and data collection phases. It provides continuous signing data accompanied by fine-grained manual and non-manual annotations, error labeling, and sentence-level acceptability judgments. This combination of linguistic richness and technical precision makes the corpus a dual-purpose resource: it enables in-depth analysis of sign language structure and learning while simultaneously serving as high-quality training/fine-tuning data for SLP tasks.

The recordings and annotations were designed to capture the variability of both L1 and L2 productions, offering realistic input for training models that can generalize across signer profiles, fluency levels, and communicative styles.

[1] In this paper, "spoken language" denotes any non-signed language, expressed either orally or in writing.

[2] Following more recent conventions (e.g., Kusters et al. (2017); Napier and Leeson (2016)), we use the term *deaf* inclusively to refer to all members of the sign language community, irrespective of audiological status.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 17–30
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

From a linguistic perspective, the resource supports the investigation of interlanguage (Selinker, 1972). The longitudinal design allows for tracing changes in articulatory precision, grammatical accuracy, and fluency over time, while parallel native data provide a benchmark for assessing progress and defining target competence.

This paper introduces the SMILE Continuous DSGS Corpus, which is made available for research purposes under a Creative Commons license[3]. It also presents a first exploratory analysis of the learner data, focusing on the way in which error patterns and fluency evolve over time.

The remainder of this paper is structured as follows: Section 2 reviews previous work on sign language corpora and learner-oriented resources. Section 3 describes the methodology and steps involved in creating the SMILE Continuous DSGS Corpus. Section 4 presents a first interlanguage study conducted on the collected data. Finally, Section 5 concludes with a discussion of limitations and directions for future work.

## 2. Related Work

Despite progress in the development of sign language corpora, sign languages remain severely under-resourced, particularly in the context of SLP (De Sisto et al., 2024; Holmes et al., 2023; Joshi et al., 2020; Sayers et al., 2021). Existing datasets have been created for a variety of purposes, including linguistic research and technological development, and therefore differ widely in their design, annotation granularity, and availability (Ebling et al., 2018). Compared to spoken language corpora, most sign language datasets are considerably smaller, often rely on manual transcription and annotation, and lack standardized practices, which hinders large-scale data sharing and machine learning applications (Tanzer and Zhang, 2024; Schulder et al., 2023).

Broadly speaking, sign language resources can be categorized into three types: (i) single-sign datasets primarily used for SLP, (ii) continuous signing datasets primarily used for SLP, and (iii) linguistic research corpora.

The first category, single-sign datasets primarily used for SLP, includes resources such as Pop-Sign ASL (Starner et al., 2023), BOSPHORUS-Sign22k (Özdemir et al., 2020), and AUTSL (Sincan and Keles, 2020). These are essential for lexical modeling and recognition but focus on isolated signs, neglecting grammatical phenomena expressed across larger temporal units and often making use of non-manual articulators. The second category, continuous datasets mainly created for SLP, includes interpreted or translated content such as PHOENIX-14T (Camgöz et al., 2018) or BOBSL (Albanie et al., 2021). While they provide valuable parallel data, these datasets are often limited to narrow domains (e.g., weather forecasts) and reflect interpreter language rather than spontaneous signing. Moreover, many lack detailed linguistic annotation and metadata, limiting their usefulness for fine-grained linguistic analysis. From a technological perspective, advances in deep learning and computer vision have driven the creation of large-scale continuous SLP datasets such as MS-ASL (Joze and Koller, 2019), WLASL (Li et al., 2020), and YouTube-SL-25 (Tanzer and Zhang, 2024). These resources have enabled significant progress in automatic recognition and translation (Bragg et al., 2019), but they generally focus on lexical tasks, lack detailed linguistic annotation, and seldom include learner or non-native data.

In contrast, corpora developed primarily for linguistic research, such as the DGS Corpus (Konrad et al., 2020), Corpus NGT (Crasborn and Zwitserlood, 2008), and the BSL Corpus (Schembri et al., 2014), provide richly annotated, representative samples of native signers. These resources are invaluable for describing sign language structure and variation, yet they are often not designed for computational reuse. Licensing restrictions and format heterogeneity reduce their applicability for SLP research.

These combined issues have led to a divide between datasets that are optimized for large-scale use but linguistically shallow and those that are linguistically rich but not immediately usable for computational research.

Sign language learner data is even more scarce. Neither the comprehensive overviews by Kopf et al. (2021, 2022) nor existing learner corpus registries include any corpus devoted to sign language learners. Nonetheless, several studies have collected data for linguistic analyses (Chen Pichler, 2011; Ferrara and Nilsson, 2017; Gulamani et al., 2020; Schönström and Leeson, 2015; Kurz et al., 2023). Among these, the Swedish Sign Language L2 Corpus (STSC-L2) compiled by Mesch and Schönström (2018) stands out as the first longitudinal resource documenting adult learners of Swedish Sign Language (*Svenskt teckenspråk*, STS) over four phases of instruction. It has supported research on the development of mouthing and depicting signs (Mesch and Schönström, 2020; Schönström and Mesch, 2022). Similarly, the Corpus of German Sign Language as L2 (Oviedo et al., 2018) includes large-scale assessment data from adult learners, but the recordings are primarily exam-based and only partially longitudinal.

For DSGS, the dataset introduced by (Ebling et al., 2018) provides a foundational contribution

[3] https://www.swissubase.ch/en/catalogue/studies/21056

containing both L1 and L2 data. However, its recordings are limited to isolated signs rather than continuous sequences, and thus do not capture the interplay between manual and non-manual articulations or the temporal evolution of learner fluency.

## 3. Corpus Creation

The SMILE Continuous DSGS Corpus was developed within a four-year interdisciplinary project aimed at advancing research and technology in sign language learning and assessment. Its design was guided by two complementary goals: (1) to create a longitudinal, linguistically annotated resource of signing by both L2 and L1 signers of DSGS; (2) to provide a machine-readable benchmark suitable for training and evaluating SLP models.

It consists of productions in a semi-controlled setting over a period of 18 months, providing the first longitudinal record of DSGS learning. The corpus was collected between March 2022 and November 2023 at the University of Zurich.

### 3.1. Participants

The corpus includes productions from two groups of participants.

The L1 subcorpus contains recordings from ten adult native or early signers[4], all of whom use DSGS as their primary means of communication. This subset serves as a control subcorpus.

The L2 subcorpus includes hearing adults acquiring DSGS as a second language in a second modality (M2L2; cf. Chen Pichler, 2019). Participants ranged in age from 22 to 55 years and had all completed at least one formal DSGS course by the time of the first recording. Among them, 14 participants were enrolled in the DSGS interpreter training program at the University of Teacher Education in Special Needs (Interkantonale Hochschule für Heilpädagogik, HfH), while eleven learners were recruited from public DSGS courses.

All participants provided informed consent and completed an online metadata questionnaire available in both German and DSGS. Deaf participants received financial compensation, while learners either obtained ECTS credits within their training program, received monetary compensation, or obtained individualized performance feedback after each data collection phase.

### 3.2. Elicitation Tasks

To examine developmental trajectories, L2 participants were recorded across four data collection phases, spaced approximately six months apart, resulting in four comparable datasets (Phase 1-Phase 4). Each phase included the same set of core elicitation tasks, allowing for longitudinal comparisons across identical stimuli, while a small number of new items were added at each phase to gradually increase linguistic complexity.

At each data collection, seven tasks were administered. They were designed to elicit a wide range of linguistic features, specifically with respect to the use of non-manual components, within controlled and semi-spontaneous contexts.

The tasks were presented in DSGS along with written German instructions for the L2 group, using on-screen prompts.

Table 1 summarizes the seven tasks presented to the participants. Controlled items, such as from a *Sentence Repetition Task* (SRT) (Haug et al., 2020), featured identical items across phases, or a mix of repeated and new items, as in the *Sentence Transformation Task* (STT). Narrative tasks (e.g., video retelling) were alternated to maintain participant engagement and minimize memorization effects. In addition, L2 participants were invited to an interview (Haug et al., 2019), which was conducted individually with a deaf expert in DSGS teaching and research.

### 3.3. Data Collection Setup

All recordings were conducted in a controlled studio environment to ensure consistent visual conditions suitable for both linguistic analysis and computer vision processing.

Each recording session was filmed using five synchronized cameras of two types (cf. Figure 1). Three FLIR Blackfly S machine vision cameras (front, front-left, front-right) provided high-resolution frontal views of the signers' faces and upper bodies for detailed analysis of manual and non-manual components. These cameras recorded at 60 fps and with 1024p resolution. Two auxiliary webcams, positioned laterally and overhead, captured the depth and trajectory of hand and arm movements for complementary linguistic analysis. These recordings were captured at 30 fps.

Uniform controlled lighting was used to minimize shadow interference, particularly in the face and hand regions. A fixed recording distance and signer position were maintained across all sessions, marked on the floor and backdrop to guarantee consistent framing across participants and data collection phases.

All cameras were hardware-synchronized via a unified trigger cable to eliminate temporal lag and monitored through a single recording interface, an adaptation of the capture system used in (Ebling et al., 2018). Each session was logged using standardized metadata templates documenting partici-

[4]Early signers are individuals who acquired DSGS during early childhood and use it as their dominant language.

| Task | Elicitation |
|---|---|
| Route description | Spatial descriptions, pointing signs, fingerspelling |
| Sentence Transformation Task | Grammatical structures, non-manual components |
| Picture retelling | Description and comparison |
| Clip retelling | Description, long sentences |
| Video retelling | Long sentences, non-manual components |
| Sentence Repetition Task | Language comprehension, processing, and production skills |
| Proficiency interview | Language comprehension, processing, and production skills |

Table 1: Summary of the tasks performed in each recording session by the participants.

pant identifier, task type, and session date. Timestamped metadata were automatically generated to ensure reliable mapping between the raw video recordings and the corresponding task metadata files. Each recording session lasted approximately 45–60 minutes. To avoid potential order effects, the presentation order of individual stimuli within each task was randomized for every participant and session.

All recordings were saved in MP4 format (H.264 codec).

### 3.4. Annotation and Validation

A detailed description of the annotation conventions, data validation procedures, and reliability measures is available in Battisti et al. (2024). Building on this framework, two tasks produced by both L1 signers and L2 learners who participated in all four data collection phases were annotated accordingly. The annotations were carried out in iLex (Hanke and Storz, 2008) using a multilayered structure. Each file includes annotation tiers for:

- manual components: glosses as well as handshape, movement, location, orientation;
- non-manual component: facial expression, eyebrow position, head movement, eye gaze, mouthing;
- errors: deviations from canonical forms for manual and non-manual components;
- acceptability: judgments of comprehensibility and naturalness at the level of individual errors and entire sentences.

All annotations were produced by deaf expert annotators and project members with linguistic expertise. Annotation quality was ensured through double annotation and internal review cycles, followed by inter-annotator agreement (IAA) testing on a representative subset. IAA reached $\kappa > 0.7$ across the main tiers (manual and non-manual components, error types, and acceptability), indicating moderate to substantial reliability. All validated annotations are included in the released corpus, aligned with the corresponding metadata and video recordings.

### 3.5. Corpus Profile

Table 2 provides a quantitative overview of the SMILE Continuous DSGS Corpus, detailing the number of participants, sessions, hours, and annotated items across phases. The corpus comprises 153 video sessions collected from 35 participants, spanning over four data collection phases. Each phase includes all elicitation tasks, resulting in approximately 67 hours and 15 minutes of recorded material in total, of which 13 hours and 59 minutes correspond to effective signing time after removing pauses and preparation segments.

Within the L2 subset, the annotated portion currently contains 1,723 sentences, 11,883 gloss tokens, and 27,656 non-manual component instances. This yields an average of 6.9 glosses and 16 non-manual components per sentence. The overall annotation includes 4,074 error tokens, divided into 1,968 manual and 2,106 non-manual errors. Approximately two-thirds of the learner sentences (1,145 items) contain at least one annotated error.

The L1 control subset consists of around nine hours of recordings produced by ten L1 participants performing the same tasks as those assigned to the L2 group.

## 4. Interlanguage Study

This section presents a first empirical analysis of the L2 learner subset of the SMILE Continuous DSGS Corpus, focusing on error patterns, fluency development, and non-manual articulation across four data collection phases. The aim is to examine how adult hearing learners of DSGS develop competence over time and to assess how the corpus can support quantitative modeling of interlanguage.

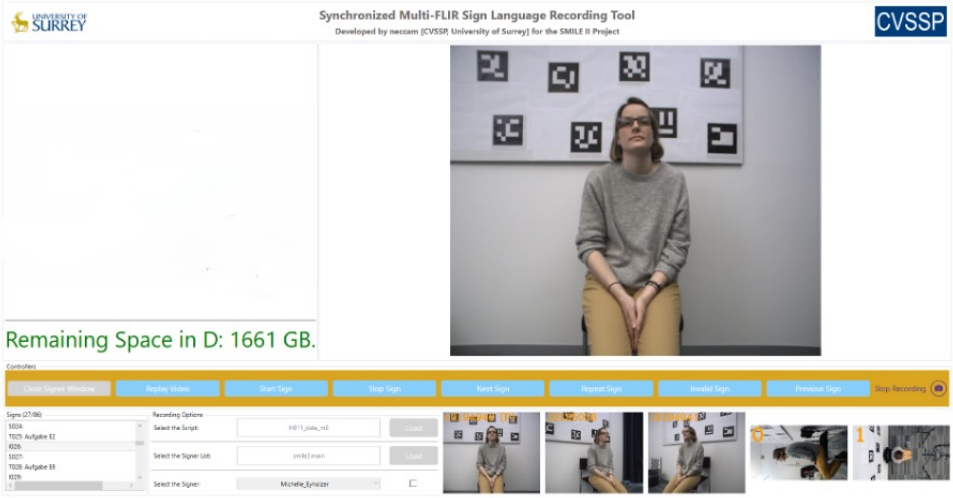



Figure 1: Sample of the recording interface with the input shown to the participant on the left and the multi-view perspectives at the bottom.

| | (M1)L1 | M2L2 (P1) | M2L2 (P2) | M2L2 (P3) | M2L2 (P4) |
|---|---|---|---|---|---|
| Signers | 10 | 22 | 18 | 14 | 14 |
| Items | 1,089 | 757 | 720 | 602 | 631 |
| Total hours | 09:43:19 | 14:53:58 | 15:30:43 | 13:35:35 | 13:31:33 |
| Eff. signing | 02:56:46 | 03:06:18 | 02:53:33 | 02:22:57 | 02:40:15 |
| Sentences | 460 | 250 | 313 | 336 | 364 |
| Tokens | 3,191 | 1,414 | 1,820 | 2,377 | 3,081 |
| Non-manuals | 6,514 | 3,089 | 4,270 | 5,813 | 7,970 |
| Errors (MC) | - | 395 | 402 | 476 | 695 |
| Errors (NMC) | - | 351 | 465 | 575 | 715 |

Table 2: Corpus profile: participants, hours, items, and annotation coverage across phases P1|2|3|4.

More specifically, the goal is to model how accuracy and fluency in DSGS production change over time and to identify which articulatory components, manual or non-manual, show the most progress or persistence of errors. In other words, how do the frequency and distribution of errors in manual and non-manual components change over an 18-month learning period?

## 4.1. Data

As described in Section 3.4, two tasks were fully annotated for the present study: the STT and the SRT (see Section 3.2). The analysis presented in this section focuses on the L2 annotated subset. The subset comprises 1,263 learner sentences with token-aligned error tiers, yielding 4,074 error tokens in total: 1,968 affecting manual components and 2,106 affecting non-manual components (cf. Table 2 in Section 3.5).

For each learner and phase, the annotations record the number of glosses, non-manual components, and error instances per sentence. Since all learners completed the same tasks at each phase, the dataset has a hierarchical structure with repeated measures: multiple observations (sentences) are nested within learners, and each learner is observed across four phases.

## 4.2. Methods

**Descriptive Analysis.** We first examined the frequencies and distributions of (i) non-manual component occurrences, (ii) component-level error types and their co-occurrences across manual and non-manual domains, and (iii) sentence-level error categories. These distributions offer an overview of interlanguage profiles; however, as relative frequencies do not account for sentence length, learner-specific baselines, or task-related variation, they should be interpreted with caution (Gries and Wulff, 2021; Von Stutterheim et al., 2021).

**Fluency Investigation.** To approximate fluency, we computed signing speed as average time per gloss: sentence duration in seconds divided by gloss count (Kanto and Haapanen, 2019; Haug et al., 2024). We modeled speed differences across phases using linear mixed-effects models (`lmer`,

Bates et al., 2015), followed by Tukey-adjusted pairwise contrasts with `emmeans` (Lenth, 2024).

**Error-count Modeling.** To examine phase effects on error production while accounting for repeated measures, we fitted Bayesian generalized mixed-effects models with `brms` (Bürkner, 2017). Manual-component and non-manual-component errors were modeled separately due to their weak correlation ($\rho = 0.24$) and to simplify the models for interpretation. Phase was entered as a fixed effect. Random intercepts were specified for learners and tasks, with by-learner random slopes for phase to capture individual learning trajectories (Baayen et al., 2008; Gries, 2021; Meteyard and Davies, 2020).

Distributional families were selected via leave-one-out cross-validation (Vehtari et al., 2017): a negative binomial for manual component errors to address overdispersion (McElreath, 2020; Winter and Bürkner, 2021) and a hurdle Poisson for non-manual-component errors to accommodate the 21% zero-inflation. Offsets standardized counts: $\log(\text{glosses})$ for the manual component model (*mc model*); $\log(\text{NMC-tokens})$ for the non-manual component model (*nmc model*). We used weakly informative priors encoding an expectation of declining error rates across phases while allowing substantial variability across learners and tasks; sensitivity checks indicated robust inferences.

### 4.3. Results: Descriptive Analysis

**Use of Non-manual Components.** As shown in Figure 2, the frequency and variety of non-manual components increased steadily over time. Later phases display higher median counts and reduced variability, indicating that learners gradually incorporated a broader range of articulators into their signing, particularly mouth gestures and eye gaze. In contrast, eyelid and nose movements remained infrequent across all phases.

**Error Distributions.** Visualizations of the error distributions at the manual, non-manual, and sentence levels are provided in Appendix A.

For manual components (Figure 5), the *movement* parameter consistently show the highest frequency (33.1%), followed by its combination with *location*. Overall, movement-related errors, alone or in combination, account for 72.1% of all manual-component errors. Manual error profiles remain stable across phases.

For non-manual components, errors involving the *eyebrows* are the most frequent (35.2% individually; 65.5% when co-occurring with other components), with *eyebrow* + *head* combinations forming a common pairing (16.1%) (Figure 6). Across phases, *mouthing* emerges as the second most frequent error category in Phase 3, overtaking head-movement errors.

At the sentence level (Figure 7), *fluency*-related issues are most prevalent (approximately 24%), followed by *facial expression* (12%) and *sentence construction* (10%). Fluency errors decreased across phases, from 43% in Phase 1 to 26% in Phase 4. In this context, *fluency errors* corresponded to annotator judgments indicating disruptions in the temporal progression of signing, such as prolonged pauses, visible hesitation, repetitions, or restarts, that hindered the smooth delivery or continuity of the utterance (Haug et al., 2024). Co-occurrence patterns between phonological and syntactic errors highlight the interaction of linguistic domains in the learners' developing interlanguage. For example, inaccurate handshape or movement realization (phonological error) often co-occurred with incorrect use of non-manual markers required for syntactic contrast.

### 4.4. Results: Modeling

**Signing Speed.** As shown in Figure 3, average signing speed, measured as time per gloss, improved by decreasing from Phase 1 (median ≈1.46 s) to Phase 4 (median ≈1.12 s), with significant contrasts for Phase 4-Phase 1 ($\beta = 0.15$, $p = 0.003$) and Phase 2-Phase 1 ($\beta = 0.04$, $p = 0.02$); Phase 3-Phase 2 and Phase 4-Phase 3 are not credibly different. This trajectory indicates that learners rapidly gained fluency during early acquisition stages before plateauing once a functional pace was reached.

**Manual Component Errors.** The *mc model* examined changes in manual component errors across phases while controlling for sentence length and random variation among learners and tasks. As shown in Figure 4a, manual errors decreased steadily from Phase 1 to Phase 3 ($\beta_{P3} = -0.33$, $\text{CrI}_{95\%}$=[-0.52,-0.15]), followed by a stabilization in Phase 4 ($\beta_{P4} = -0.24$, $\text{CrI}_{95\%}$=[-0.56,-0.06]). Credible intervals (CrI) are broad, reflecting strong individual and, to a lesser extent, task variability, but the posterior mass for Phase 4 is predominantly below zero (94%), indicating fewer errors than at baseline.

An investigation of the random effects revealed moderate inter-learner variability, particularly in the first and last phases, suggesting that some participants either progressed faster or maintained higher baseline accuracy. Task-related variability was comparatively small. Posterior parameter estimates are reported in Appendix B.

**Non-manual Component Errors.** The *nmc model* showed a similar non-linear trajectory but

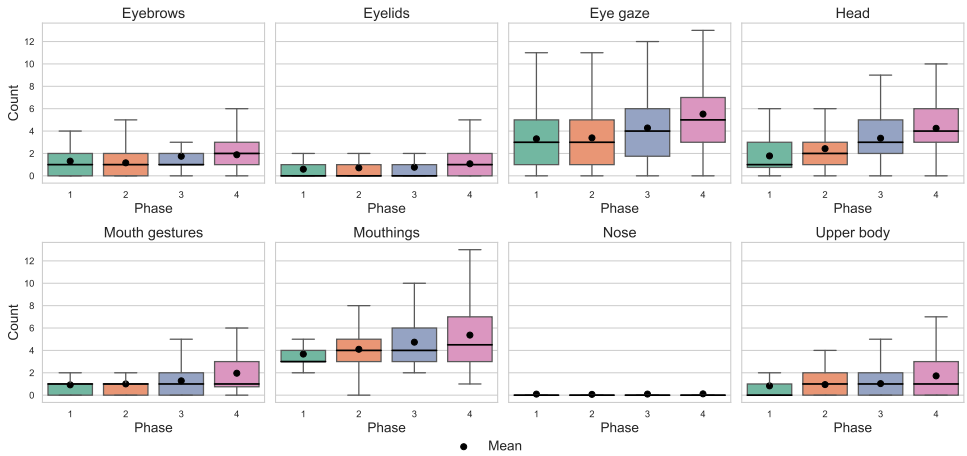



Figure 2: Distribution of individual non-manual components across data collection phases. Mouth gestures and eye gaze increase over time, while eyelid and nose movements remain rare.

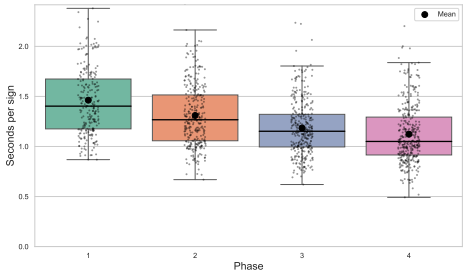
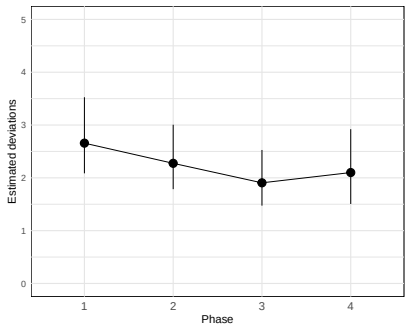
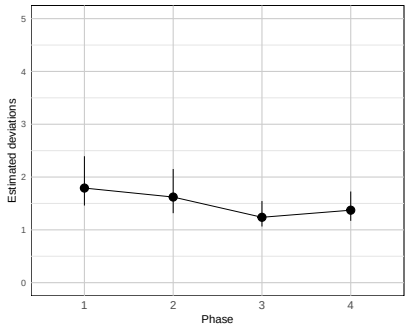



Figure 3: Average signing time per gloss across four phases. Median duration per gloss decreases over time, indicating increasing fluency.




(a) Manual components




(b) Non-manual components

Figure 4: Conditional effects of phase on expected error counts for manual (a) and non-manual (b) components. Both show decreasing trends that stabilize in later phases.

with a steeper mid-phase decline (Figure 4b). Non-manual errors decreased sharply between Phase 1 and Phase 3 ($\beta_{P3} = -0.70$, $\text{CrI}_{95\%}$=[-1.03,-0.37]) and stabilized thereafter ($\beta_{P4} = -0.46$, $\text{CrI}_{95\%}$=[-0.79,-0.15]).

Unlike manual components, non-manual variability remained high across all phases, reflecting persistent differences in how learners used facial expressions and head movements to mark grammatical and discourse functions. This variability likely stems from individual differences in expressiveness, as well as from pedagogical factors; non-manual features often receive less systematic feedback and are introduced later or less explicitly in instruction compared to manual parameters. Learners accounted for most of the random effect variance, suggesting that non-manual coordination follows highly individualized developmental paths. Phase alone explained about 9% of the variance, rising to 15% when learner and task effects were included.

## 4.5. Discussion

This study examined how L2 learners of DSGS developed manual and non-manual articulations over four data collection phases using a combination of descriptive and Bayesian mixed-effects analyses. The main goal was to identify whether error counts declined and whether fluency and non-manual com-

ponent use improved over time.

The results confirm that interlanguage development in sign language is neither linear nor uniform. In early phases, learners relied on manual articulation, producing few non-manual components and many movement-related errors. This pattern reflects an initial focus on lexical retrieval and basic articulatory control rather than multi-channel integration. The predominance of movement and location errors is consistent with earlier studies of DSGS (Ebling et al., 2021) and aligns with cross-linguistic evidence from ASL and BSL learner corpora (Rosen, 2004; Ortega and Morgan, 2015). These errors can be interpreted as early-stage motor simplifications rather than conceptual misunderstandings (Marshall et al., 2021).

Between Phases 2 and 3 of our study, a marked restructuring occurred: manual accuracy improved, fluency errors declined, and the frequency of non-manual components increased. This phase may represent a consolidation of articulatory routines. The observed acceleration in signing speed supports this interpretation, showing that learners begin to produce sentences with greater temporal cohesion.

Non-manual components showed the steepest improvement but also the greatest variability. Eyebrow and head movement errors remained frequent across all phases, while mouthing errors peaked in Phase 3. This pattern suggests that learners temporarily rely on spoken-language strategies, such as overusing mouth gestures, before converging toward target-like grammatical facial expressions (Mesch and Schönström, 2020; Kimmelman et al., 2020). Persistent variability in non-manual components likely stemmed from perceptual constraints, i.e., signers could not directly monitor their own facial configurations (Marshall et al., 2021), and from pedagogical practices that prioritized manual precision in early training.

The models showed that the largest reduction in both manual and non-manual errors occurred between Phases 1 and 3, followed by a plateau. The wide credible intervals and learner-level variance highlighted heterogeneous trajectories rather than a uniform trend. This variability points to individual differences in motor learning, feedback, and exposure, aligning with usage-based theories of second-language acquisition (Chang and Zhang, 2021; Lambelet and Berthele, 2015).

Overall, the findings suggest that using and coordinating non-manual components is a late-acquired skill in the DSGS learning process. As learners gain control of articulatory timing, their interlanguage becomes more stable. These insights can thus support a pedagogical focus on integrating facial grammar and prosody earlier in instruction and motivate data-driven feedback tools.

## 5. Limitations and Future Work

Several limitations constrain the generalizability of this analysis. The annotated subset covers only two tasks and a limited sample of learners, which restricts statistical power and the range of linguistic phenomena examined. Manual annotation, particularly for non-manual components, can result to be inherently subjective due to overlapping articulations and fuzzy temporal boundaries. Although Bayesian modeling mitigates small-sample uncertainty, the wide credible intervals reflect substantial learner-level variance. Future analyses should incorporate additional predictors, such as learners' prior exposure, contact with deaf signers, and task complexity, to better explain this variability.

Future work will extend annotation to spontaneous and narrative tasks, such as video retelling, include automatic tracking of facial landmarks to complement manual labels, and refine the annotation scheme based on validation results.

## 6. Conclusion

This paper presented the *SMILE Continuous DSGS Corpus*, a longitudinal resource that includes continuous productions by adult learners of DSGS, as well as productions by native and early signers. Combining controlled recordings, detailed annotation, and rich metadata, the corpus provides a foundation for linguistic and computational research on sign language learning and processing. A first analysis of learner interlanguage using Bayesian mixed-effects modeling revealed a non-linear developmental pattern: a marked improvement in manual and non-manual accuracy between the early and mid-learning phases, followed by stabilization and strong inter-learner variability.

Methodologically, the study illustrates how probabilistic modeling can yield interpretable developmental insights even from a small-scale longitudinal dataset.

The SMILE Continuous DSGS Corpus thus offers both a benchmark for modeling sign language learning and a resource for developing data-driven feedback in sign language education.

## 7. Data and Code Availability

The *SMILE Continuous DSGS Corpus* is available for research purposes through SwissUBase[5] under the Creative Commons license CC BY-NC-SA 4.0: SMILE Continuous DSGS Corpus L1 https://doi.org/10.48656/ch11-5k92; SMILE Continuous DSGS Corpus L2 https://doi.org/

[5] https://www.swissubase.ch/en/catalogue/studies/21056/21315

10.48656/rk0q-3g09. All analysis scripts will be made available upon request to the corresponding author.

## 8. Acknowledgments

This work was funded through the Swiss National Science Foundation (SNSF) Sinergia project "Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment II" (SMILE II) (grant agreement no. CRSII5_193686).

## 9. Bibliographical References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BBC-Oxford British Sign Language Dataset. ArXiv:2111.03635 version: 1.

Rolf Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Alessia Battisti, Katja Tissi, Sandra Sidler-Miserez, and Sarah Ebling. 2024. Advancing annotation for continuous data in Swiss German sign language. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 1–12, Torino, Italia. ELRA and ICCL.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Penny Boyes Braem. 2024. Gebärdensprache. In Elvira Glaser, Johannes Kabatek, and Barbara Sonnenhauser, editors, *Sprachenräume der Schweiz*, volume 1 of *Sprachen*, pages 185–209. Narr Francke Attempto, Tübingen.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31, New York, NY, USA. Association for Computing Machinery.

Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Proceedings of CVPR 2018*.

Necati Cihan Camgöz, Ahmet Alp Kındıroğlu, Serpil Karabüklü, Meltem Kelepir, Ayşe Sumru Özsoy, and Lale Akarun. 2016. BosphorusSign: A Turkish Sign Language Recognition Corpus in Health and Finance Domains. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1383–1388, Portorož, Slovenia. European Language Resources Association.

Pengyun Chang and Lawrence Jun Zhang. 2021. A CDST Perspective on Variability in Foreign Language Learners' Listening Development. *Frontiers in Psychology*, 12.

Deborah Chen Pichler. 2011. Sources of Handshape Error in First-Time Signers of ASL. In Gaurav Mathur and Donna Jo Napoli, editors, *Deaf around the world: the impact of language*. Oxford University Press, Oxford; New York.

Deborah Chen Pichler. 2019. Effect of minimal sign language instruction on hearing learners' constituent order. In *Presented at The 13th International Conference on Theoretical Issues in Sign Language Research (TISLR)*, University of Hamburg, Germany.

Onno Crasborn and Inge Zwitserlood. 2008. The Corpus NGT: an online corpus for professionals and laymen. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 44–49, Marrakech, Morocco. European Language Resources Association (ELRA).

Nivja H. de Jong. 2018. Fluency in Second Language Testing: Insights From Different Disciplines. *Language Assessment Quarterly*, 15(3):237–254.

Nivja H. de Jong. 2023. Fluency in Speaking as a Dynamic Construct. *Language Teaching Research Quarterly*, 37:179–187.

Mirella De Sisto, Vincent Vandeghinste, Caro Brosens, Myriam Vermeerbergen, and Dimitar Shterionov. 2024. Xsl-horeco and gost-parc-sign: Two new signed language - written language parallel corpora. In Krister Lindén, Thalassia Kontino, and Jyrki Niemi, editors, *Selected papers from the CLARIN Annual Conference 2023*. Linköping Electronic Conference Proceedings, Leuven, Belgium.

Sarah Ebling, Necati Cihan Camgöz, Penny Boyes Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, Marzieh Razavi, and Mathew Magimai-Doss. 2018. SMILE Swiss German Sign Language Dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Sarah Ebling, Katja Tissi, Sandra Sidler-Miserez, Cheryl Schlumpf, and Penny Boyes Braem. 2021. Single-parameter and parameter combination errors in L2 productions of Swiss German Sign Language. *Sign Language & Linguistics*, 24(2):143–181.

Lindsay Ferrara and Anna-Lena Nilsson. 2017. Describing spatial layouts as an L2M2 signed language learner. *Sign Language & Linguistics*, 20(1):1–26.

Stefan Th. Gries and Stefanie Wulff. 2021. Examining individual variation in learner production data: A few programmatic pointers for corpus-based analyses using the example of adverbial clause ordering. *Applied Psycholinguistics*, 42(2):279–299.

Stefan Thomas Gries. 2021. *Statistics for linguistics with R: a practical introduction*, 3rd revised edition edition. Mouton Textbook. de Gruyter Mouton, Berlin.

Sannah Gulamani, Chloë Marshall, and Gary Morgan. 2020. The challenges of viewpoint-taking when learning a sign language: Data from the 'frog story' in British Sign Language. *Second Language Research*, 38(1):55–87.

Thomas Hanke and Jakob Storz. 2008. iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, pages 64–67, Marrakesh, Marocco. European Language Resources Association.

Tobias Haug, Aaron Olaf Batty, Martin Venetz, Christa Notter, Simone Girard-Groeber, Ute Knoch, and Mireille Audeoud. 2020. Validity evidence for a sentence repetition test of Swiss German Sign Language. *Language Testing*, 37(3):412–434.

Tobias Haug, Nivja H. de Jong, Franz Holzknecht, Katja Tissi, Sandra Sidler-Miserez, Alessia Battisti, Regula Perrollaz, Sarah Ebling, Sabine Reinhard, and Sarah Caminada. 2024. Development and validation of a fluency rating scale for Swiss German Sign Language. *Frontiers in Education*, 9.

Tobias Haug, Heidi Stocker, and Daniela Nussbaumer. 2019. Die Entwicklung von Instrumenten zur Überprüfung von kognitiven Fähigkeiten, gebärdensprachlicher Kompetenz und Dolmetschleistung von Gebärdensprachdolmetscherinnen. *Das Zeichen*, 111:130–143.

Ruth Holmes, Ellen Rushe, Mathieu De Coster, Maxim Bonnaerens, Shin'Ichi Satoh, Akihiro Sugimoto, and Anthony Ventresque. 2023. From Scarcity to Understanding: Transfer Learning for the Extremely Low Resource Irish Sign Language. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2000–2009, Paris, France.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Hamid Reza Vaezi Joze and Oscar Koller. 2019. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. ArXiv:1812.01053.

Laura Kanto and Ulla-Maija Haapanen. 2019. Fluency in Sign Language. In *Fluency in L2 learning and use*, pages 96–110. Multilingual Matters.

Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study. *PLOS ONE*, 15(6):e0233731. Publisher: Public Library of Science.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2021. Overview of Datasets for the Sign Languages of Europe | ZFDM Repository. Technical report, University of Hamburg, Germany.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. The Sign Language Dataset Compendium:

Creating an overview of digital linguistic resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association.

Kim B. Kurz, Geo Kartheiser, and Peter C. Hauser. 2023. Second language learning of depiction in a different modality: The case of sign language acquisition. *Frontiers in Communication*, 7.

Annelies Maria Jozef Kusters, Dai O'Brien, and De Meulder. 2017. Innovations in deaf studies: Critically mapping the field. In Annelies Kusters, Maartje De Meulder, and Dai O'Brien", editors, *Innovations in Deaf Studies*, pages 1–53. Oxford University Press, United Kingdom.

Amelia Lambelet and Raphael Berthele. 2015. *Age and Foreign Language Learning in School*. Palgrave Macmillan UK, London.

Russell V. Lenth. 2024. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.10.3.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1448–1458.

Chloe Marshall, Aurora Bel, Sannah Gulamani, and Gary Morgan. 2021. How are signed languages learned as second languages? *Language and Linguistics Compass*, 15(1).

Richard McElreath. 2020. *Statistical rethinking: a Bayesian course with examples in R and Stan*, 2 edition. CRC texts in statistical science. Taylor and Francis, CRC Press, Boca Raton.

Johanna Mesch and Krister Schönström. 2018. From Design and Collection to Annotation of a Learner Corpus of Sign Language. In *Proceedings of the 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 121–126, Miyazaki, Japan. European Language Resources Association.

Johanna Mesch and Krister Schönström. 2020. Use and acquisition of mouth actions in L2 sign language learners: A corpus-based approach. *Sign Language & Linguistics*, 24(1):36–62.

Lotte Meteyard and Robert A. I. Davies. 2020. Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112:104092.

Jemina Napier and Lorraine Leeson. 2016. *Sign Language in Action*, 1 edition. Number 2946-2258 in Research and Practice in Applied Linguistics (RPAL). Palgrave Macmillan London.

Gerardo Ortega and Gary Morgan. 2015. Phonological Development in Hearing Learners of a Sign Language: The Influence of Phonological Parameters, Sign Complexity, and Iconicity: Phonological Development in Sign L2 Learners. *Language Learning*, 65(3):660–688.

Alejandro Oviedo, Thomas Kaul, Leonid Klinner, and Reiner Griebel. 2018. The Cologne Corpus of German Sign Language as L2 (C/CSL2): Current development stand. In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 163–166, Miyazaki, Japan. European Language Resources Association.

Russel S. Rosen. 2004. Beginning l2 production errors in asl lexical phonology. *Sign Language & Linguistics*, 7.

Dave Sayers, Rui Sousa-Silva, Sviatlana Höhn, and et al. 2021. The Dawn of the Human-Machine Era - A forecast of new and emerging language technologies. Report for EU COST Action CA19102 'Language In The Human-Machine Era'.

Adam Schembri. 2008. British Sign Language Corpus project: Open access archives and the observer's paradox. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 165–169, Marrakech, Morocco. European Language Resources Association (ELRA).

Marc Schulder, Sam Bigeard, Thomas Hanke, and Maria Kopf. 2023. The Sign Language Interchange Format: Harmonising Sign Language Datasets For Computational Processing. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, Rhodes Island, Greece. IEEE.

Krister Schönström and Ingela Holmström. 2022. L2M1 and L2M2 Acquisition of Sign Lexicon: The Impact of Multimodality on the Sign Second Language Acquisition. *Frontiers in Psychology*, 13.

Krister Schönström and Lorraine Leeson. 2015. Signed language L2 learners' corpora and sign language assessment. Presentation at Interkantonale Hochschule für Heilpädagogik. Zurich.

Krister Schönström and Johanna Mesch. 2022. Second language acquisition of depicting signs: A corpus-based account. *Language, Interaction and Acquisition*, 13(2):199–230.

Larry Selinker. 1972. INTERLANGUAGE. *International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–232.

Özge Mercanoğlu Sincan and Hacer Yalim Keles. 2020. AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods. *IEEE Access*, 8:181340–181355.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam S. Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, Daksh Sehgal, Saad Hassan, Bill Neubauer, Sofia Anandi Vempala, Alec Tan, Jocelyn Heath, Unnathi Kumar, Priyanka Mosur, Tavenner Hall, Rajandeep Singh, Christopher Cui, Glenn Cameron, Sohier Dane, and Garrett Tanzer. 2023. PopSign ASL v1.0: An Isolated American Sign Language Dataset Collected via Smartphones. In *Neural Information Processing Systems*.

Garrett Tanzer and Biao Zhang. 2024. YouTube-SL-25: A Large-Scale, Open-Domain Multilingual Sign Language Parallel Corpus. ArXiv:2407.11144 version: 1.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Christiane Von Stutterheim, Monique Lambert, and Johannes Gerwien. 2021. Limitations on the role of frequency in L2 acquisition. *Language and Cognition*, 13(2):291–321.

Bodo Winter and Paul-Christian Bürkner. 2021. Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass*, 15(11):e12439.

Oğulcan Özdemir, Ahmet Alp Kındıroğlu, Necati Cihan Camgöz, and Lale Akarun. 2020. BosphorusSign22k Sign Language Recognition Dataset. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 181–188, Marseille, France. European Language Resources Association.

## 10. Language Resource References

Onno Crasborn, Inge Zwitserlood, Johan Ros, and Annemieke van Kampen. 2006-2017. *Corpus NGT*. The Language Archive.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. *MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release*. Universität Hamburg.

Adam Schembri, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. 2014. *British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2014*. University College London.

# A. Appendix A: Error Distributions

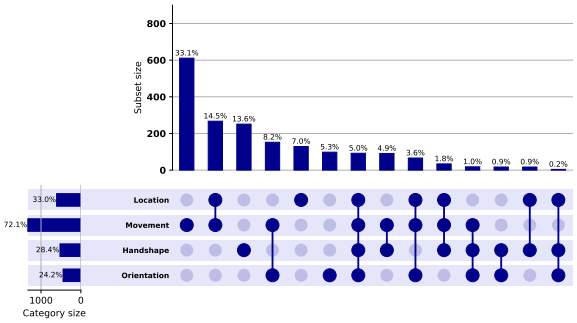



Figure 5: UpSet plots summarizing co-occurring error categories in manual components.

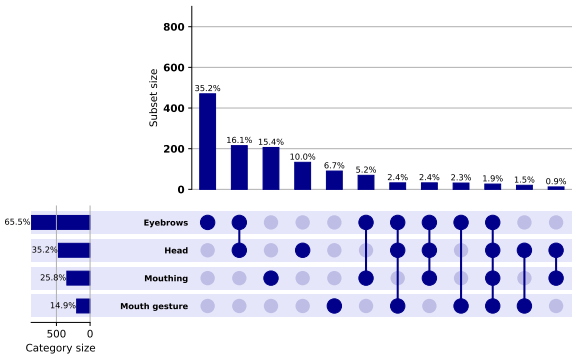



Figure 6: UpSet plots summarizing co-occurring error categories in non-manual components.

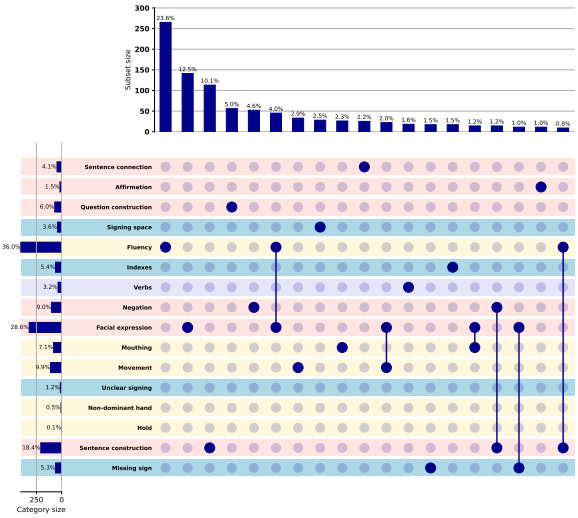



Figure 7: UpSet plots summarizing co-occurring error categories at sentence level.

## B. Appendix B: Posterior Estimates

| | **mc model** | |
|---|---|---|
| *Parameter* | *Mean± SE* | *95% CrI* |
| *Fixed effects* | | |
| $\alpha_{P1}$ | -1.02 ± 0.13 | [-1.27, -0.75] |
| $\beta_{P2}$ | -0.16 ± 0.10 | [-0.34, 0.03] |
| $\beta_{P3}$ | -0.33 ± 0.09 | [-0.52, -0.15] |
| $\beta_{P4}$ | -0.24 ± 0.15 | [-0.56, -0.06] |
| *Random effects* | | |
| $\sigma_{id:P1}$ | 0.36 ± 0.09 | [0.22, 0.57] |
| $\sigma_{id:P2}$ | 0.16 ± 0.10 | [0.01, 0.38] |
| $\sigma_{id:P3}$ | 0.11 ± 0.08 | [0.00, 0.31] |
| $\sigma_{id:P4}$ | 0.50 ± 0.13 | [0.30, 0.80] |
| $\sigma_{ex:P1}$ | 0.36 ± 0.06 | [0.26, 0.49] |
| $\sigma_{ex:P2}$ | 0.07 ± 0.06 | [0.00, 0.22] |
| $\sigma_{ex:P3}$ | 0.12 ± 0.09 | [0.01, 0.33] |
| $\sigma_{ex:P4}$ | 0.09 ± 0.07 | [0.00, 0.24] |
| | **nmc model** | |
| *Parameter* | *Mean± SE* | *95% CrI* |
| *Fixed effects* | | |
| $\alpha_{P1}$ | -2.01 ± 0.19 | [-2.32, -1.59] |
| $\beta_{P2}$ | -0.16 ± 0.18 | [-0.53, 0.19] |
| $\beta_{P3}$ | -0.70 ± 0.17 | [-1.03, -0.37] |
| $\beta_{P4}$ | -0.46 ± 0.16 | [-0.79, -0.15] |
| *Random effects* | | |
| $\sigma_{id:P1}$ | 0.50 ± 0.14 | [0.29, 0.83] |
| $\sigma_{id:P2}$ | 0.57 ± 0.15 | [0.32, 0.91] |
| $\sigma_{id:P3}$ | 0.38 ± 0.17 | [0.05, 0.74] |
| $\sigma_{id:P4}$ | 0.45 ± 0.13 | [0.23, 0.75] |
| $\sigma_{ex:P1}$ | 0.15 ± 0.07 | [0.01, 0.29] |
| $\sigma_{ex:P2}$ | 0.12 ± 0.09 | [0.00, 0.35] |
| $\sigma_{ex:P3}$ | 0.31 ± 0.14 | [0.04, 0.60] |
| $\sigma_{ex:P4}$ | 0.14 ± 0.09 | [0.01, 0.35] |

Table 3: Posterior mean, standard error, and 95% credible interval (CrI) for the $phase$ predictor and the random effects in the manual (*mc*) and non-manual (*nmc*) error models.

# The In-Car Sign Language Corpus (ICSL): A Multi-Modal Resource for Constrained-Space Sign Language Recognition

**Raviteja Boddu*, Guilherme Vieira Leite†, Joed Lopes da Silva*, Ângelo Benetti†, Isabela Barbieri†, Natália de Melo Afonso†, Thyago Santos‡, Helio Pedrini†, Felipe Venâncio Barbosa‡, José Mario De Martino†, Munir Georges*, Alessandro Zimmer***

*Technische Hochschule Ingolstadt (THI), Bayern, Germany
†Universidade Estadual de Campinas (UNICAMP), São Paulo, Brazil
‡Universidade de São Paulo (USP), São Paulo, Brazil
*{raviteja.boddu, joed.lopesdasilva, munir.georges, alessandro.zimmer}@thi.de
†martino@unicamp.br, †guilherme.leite@ic.unicamp.br

**Abstract**

This paper addresses the challenges of using sign language within shared mobility services, such as taxis, carpools, or ride-sharing platforms. The use of sign language recognition (SLR) in real-world, confined environments, specifically vehicle interiors remains largely unexplored. To motivate research in this area, we present the In-Car Sign Language (ICSL) dataset for Brazilian Sign Language (Libras), with the long-term goal of improving public transport accessibility for the Deaf and Hard-of-Hearing community. The dataset consists of: (1) high-precision laboratory motion capture (MoCap) data to establish an idealized linguistic baseline and (2) real-world multi-modal in-car recordings captured using a 2D camera and 3D Time-of-Flight sensors. The dataset provides a basis for comparative analyses between synthesized signing avatar animations and recorded real signing interpreter videos, which enable future research into robust "in-the-wild" SLR models and domain adaptation. We describe in detail the use cases, the setup, the data collection protocol, and the metadata structure of the corpus. In total, we recorded a multimodal dataset exceeding 1.5 million frames, comprising the synchronized multimodal streams described above featuring Libras users across various in-car scenarios. The corpus is provided with gloss annotation of lexical signs and non-lexical sign language elements specially designed to support the training and evaluation of deep neural networks for constrained space recognition. In-vehicle signing offers a technically significant example of a constrained, occluded, and non-frontal environment. While recognizing the diverse communication strategies already employed by the Deaf community, identifying automotive-specific limitations provides a useful stepping stone for research into enhancing in-car accessibility and passenger quality of life.

**Keywords:** Brazilian Sign Language (Libras), Shared Mobility Service, Motion Capture (MoCap), In-Car Communication, Constrained Signing Space, Signing Avatars, Multimodal Sensors

## 1. Introduction

The advancement in natural language processing and computer vision has brought us closer to seamless human-machine interaction. While the Deaf and Hard-of-Hearing (DHH) community has developed effective strategies for "on-the-go" communication, such as the use of mobile devices or visual cues, the interior of vehicles within shared mobility services represents a technically significant and under-researched environment for Sign Language Recognition (SLR).

Specifically, the car cabin serves as a critical case study for constrained, occluded, and non-frontal signing environments, which are often overlooked in traditional laboratory-based datasets. While Brazilian Sign Language (Libras) has an established linguistic foundation, there is currently no specialized resource documenting how signing is produced and processed in the highly constrained physical and visual environments of a car cabin (dos Santos et al., 2025; Lee et al., 2025).

The relevant studies and corpora that exist for Libras are predominantly recorded in laboratory-controlled settings with optimal lighting and neutral backgrounds. In such environments, signers have a full, unobstructed signing space (dos Santos et al., 2025). However, the interior of a vehicle introduces severe constraints, such as seatbelts cutting across the torso, occlusions, dashboard obstruction, and constrained space that limit the range of torso, elbow, and arm movements, and dynamic environmental lighting creates substantial visual noise. These factors represent fundamental challenges for current SLR systems, which often fail to generalize to such "in-the-wild" conditions.

Within the framework of Project UNITY, a collaborative research initiative between the UNICAMP (Brazil) and the THI (Germany), we seek to establish a specialized multimodal dataset that captures real-world constraints while providing a high-fidelity baseline. We aim to enhance the robustness of SLR models by utilizing signing avatars as a controlled reference to evaluate the environmental chal-



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 31–41
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

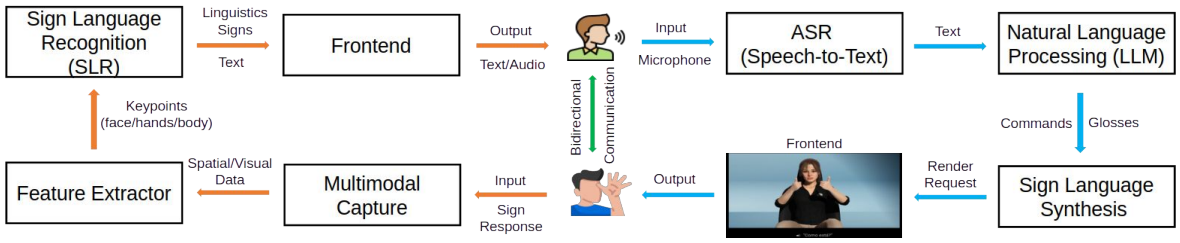



Figure 1: Proposed Project UNITY architecture for a conceptual bidirectional loop: (right) Driver-to-Passenger Speech-to-Sign Synthesis and (left) Passenger-to-Driver Sign Language Recognition.

lenges present in real-world in-car video. Led by professional Libras signers and Deaf researchers, who are also co-authors of this study, the data acquisition process seeks to document the authentic challenges of the constrained transport domain by documenting passenger-driver communication within the vehicle cabin.

The long-term vision of this project is to investigate the feasibility of a real-time, bidirectional translation framework, as seen in Figure 1, which serves as the target architecture for our ongoing work. Unlike traditional one-way approaches, our target architecture facilitates a conversational loop between driver speech and passenger signs (Libras). Building such a system requires resources that capture both idealized linguistic baselines and the severe physical restrictions of the vehicle cabin.

Rather than presenting a final recognition solution, this paper describes an initial step toward the In-Car Sign Language (ICSL) corpus, which we hope can serve as a helpful reference for future data collection efforts in the automotive domain. It details the acquisition protocol, use-case design, and setup for in-car recordings; the technical setup of the laboratory MoCap baseline; and the architecture of the signing avatar pipeline. By providing this comparative resource, we hope to contribute to future studies on accessible communication technologies for the DHH community in transportation systems.

## 2. Related Work

For the Brazilian Sign Language (Libras) community, there is a big gap between lab research and real-world use, especially in ride-hailing and ride-sharing services. The space, movement, and social setting within a vehicle influence how signs are produced. This section provides an overview of existing language resources, the algorithms used, and the real-world challenges involved in building a Libras dataset and tools designed specifically for use in vehicles.

### 2.1. The Landscape of Brazilian Sign Language (Libras) Resources

The development of language resources for Libras has changed over time. It started mainly with written documentation and later moved to video-based, multimodal data. Libras is recognized as a legal means of communication and expression in Brazil, with its own grammar and linguistic system for transmitting ideas and facts (Presidência da República, 2002).

Early work on documenting Libras mostly focused on creating dictionaries and bilingual corpora to support education and translation (de Quadros and de Sousa, 2021). A comparative summary of prominent Libras datasets and their recording environments is provided in Table 1.

#### 2.1.1. Textual and Bilingual Corpora

One of the important recent resources for Libras is VLibras-DB (Lima et al., 2025). It is a bilingual text dataset with about 127,000 translation pairs between Brazilian Portuguese and Libras. This dataset was mainly created to help neural machine translation systems convert written text into glosses. Broadly defined, glossing is the use of written language to represent primary data, whether signed or spoken. Within SLR, it serves as a symbolic bridge, utilizing labels from a spoken language to represent the manual and non-manual components of a sign.

To build VLibras-DB, a group of 10 professional interpreters worked together to define annotation rules for representing signs using glosses that could leverage neural network training. The gloss convention established includes directional verbs, intensifiers, and negation. The dataset primarily covers manual lexical signs but does not address challenges posed by non-manual markers or timing-related information. Non-manual markers (NMMs) are linguistically meaningful articulations produced by the face, head, torso, and gaze used as part of a sign language's phonology, morphology, syntax, and prosody (Herrmann, 2014). As a result, the

VLibras-DB is not fully suitable for sign recognition tasks in real-world settings.

#### 2.1.2. Visual and Isolated Sign Recognition Benchmarks

For isolated sign language recognition (ISLR), researchers typically use datasets such as MINDS-Libras (Rezende et al., 2021) and LIBRAS-UFOP (Cerna et al., 2021). These datasets have videos of single signs, mostly recorded in a controlled setup, with a plain background and the signer facing the camera. Recently, researchers have improved recognition performance by utilizing skeleton-based image representations paired with 2D CNN models, achieving accuracies up to 0.93 on MINDS-Libras. This technique involves transforming skeletal sequences comprising body, hand, and facial landmarks into a single spatio-temporal image, allowing a 2D CNN to recognize the sign from this unified input. This image then goes into a classifier to recognize the sign (Alves et al., 2024). However, such datasets do not work well for car environment. Most ISLR benchmarks assume the signer is facing the camera directly, which is very different from the angles you get inside a vehicle (Ranum et al., 2024). While vehicle interiors present unique spatial challenges, they share commonalities with other atypical or constrained-space signing environments. For instance, the PopSign dataset (Starner et al., 2023) explores the complexities of one-handed signing with the narrow field of view of a smartphone camera. Much like PopSign, our ICSL dataset addresses the problem of reduced signing volume, where the signer must adapt the scale of their gestures due to spatial boundaries. Our work extends this by focusing on the specific two-handed constraints, such as seatbelt occlusions and non-frontal camera perspectives, unique to the automotive cabin that are not present in handheld mobile datasets.

#### 2.1.3. Regional and Documentary Inventories

The LIBRAS-UFSC corpus and the Libras National Inventory are focused on preserving Brazil's Deaf heritage and the language's diversity. They ensure the corpus includes people of different ages, genders, and regions, using careful methods to build it. The LIBRAS-UFSC corpus provides phonetic and phonological details, inspired by the Dutch Signbank model, and classifies signs by hand shape, movement, and location. These resources are relevant for linguistic research, but they are usually recorded in controlled settings, so they do not capture challenges such as visual noise, occlusions, or restricted movement that you would see in a car (de Quadros and de Sousa, 2021) (Lobo-Neto and Pedrini, 2024).

### 2.2. Sign Language Recognition Architectures and Methodologies

The technological evolution of SLR has transitioned from traditional manual feature extraction to end-to-end deep learning models. This shift is critical for handling the high variability of sign language, which is based on precise hand movements, facial expressions, and body language.

#### 2.2.1. Feature Extraction and Temporal Modeling

Feature extraction focuses on capturing spatial information about hand shape, position, and orientation, as well as facial landmarks. Temporal modeling addresses the multichannel sign language characteristics (hands, facial expression, gaze, torso shift, head movement) presented in continuous stream of video frames (Camgoz et al., 2020; Rastgoo, 2021; Alyami et al., 2024; Taher and Zeebaree, 2025). Common architectures include:

- **CNN-BiLSTM:** Convolutional Neural Networks (CNNs) extract spatial features from individual frames, while Bidirectional Long Short-Term Memory (BiLSTM) networks learn the temporal dependencies over time. This combination is particularly effective for dynamic gestures, where the order of movements is crucial to the meaning (Lu et al., 2023).
- **Graph Convolutional Networks (GCNs):** These consider body landmarks as nodes in a graph, allowing the model to learn the structural interconnections between different joints and body parts (de Amorim et al., 2019).
- **Transformers:** Utilizing spatial and temporal attention mechanisms, Transformers can capture long-range dependencies across frames and correlations between different visual channels (Ma et al., 2024).

#### 2.2.2. Landmark Detection and Efficiency

In real-time automotive applications, computational efficiency is as vital as accuracy. MediaPipe and OpenPose are the two most frequently cited tools for skeletal feature extraction due to their extensive joint coverage. OpenPose is often prone to temporal jitter and high latency, particularly in occluded environments (Alves et al., 2024). In contrast, MediaPipe offers a lightweight alternative that can achieve 5x faster recognition by utilizing an optimized subset of body landmarks (dos Santos et al., 2025). Research has shown that a well-chosen landmark subset (e.g., focusing on the upper body and hands) can maintain state-of-the-art accuracy while significantly reducing inference time. While efficient, MediaPipe's performance can degrade

Table 1: Overview of existing and proposed Brazilian Sign Language (Libras) datasets.

| Dataset Name | Modality | Primary Task | Language | Scope | Annotations |
|---|---|---|---|---|---|
| VLibrasBD | Text/Gloss | Neural Machine Translation | Libras/BP | 127k Pairs | Gloss, Syntax |
| MINDS-Libras | RGB Video | Isolated Sign Recognition | Libras | 1,200+ Samples | Gloss |
| LIBRAS-UFOP | RGB Video | Isolated Sign Recognition | Libras | 2,000+ Samples | Gloss |
| LIBRAS-UFSC | Video/Bank | Linguistic Documentation | Libras | National Inventory | Phonetics |
| LSWH100 | Synthetic RGB | Handshape Classification | Libras | 144k Images | Keypoints |
| **Ours (ICSL)** | **RGB/Depth/IR /Point Cloud** | **Constrained Sign Recognition** | **Libras** | **1.5M+ Frames, 127 Phrases*** | **Gloss** |

*Current subset of a total 1,344 defined phrases.

significantly when parts of the body are blocked or in poorly lit environments, which often happens inside a car (Amalfitano et al., 2023). Consequently, there is a need for models that can handle missing hand or body points or combine information from different sensors to deal with visual noise.

## 2.3. The Automotive Domain: Physical and Social Constraints

The interior of a vehicle presents a highly restricted physical and visual environment for communication. Unlike laboratory settings, where signers have full, unobstructed signing space, the car cabin introduces several "in-the-wild" constraints that serve as a starting point for specialized SLR research.

### 2.3.1. Constrained Signing Space and Occlusions

The physical setup inside a car changes how people sign. Seatbelts cross the body and can obscure important parts of a sign and hinder torso movements. The dashboard, headrests, and tight space also limit the movement of the arms and elbows, reducing overall signing space.

### 2.3.2. Environmental Noise and Dynamic Lighting

Inside a vehicle, visual conditions can change significantly. Lighting can go from bright sunlight to near darkness very quickly. Shadows, reflections from windows, and uneven light can add visual noise, making it harder to clearly see hands and movements. Because of this, sign language recognition systems that work well in lab lighting often struggle to handle these changing real-world conditions (Amalfitano et al., 2023).

## 2.4. Hardware and Multimodal Sensing

Because single RGB cameras do not work reliably inside a car, researchers are using multiple sensor types to make sign language recognition systems more stable and accurate.

### 2.4.1. 2D vs. 3D Sensing

While 2D cameras are affordable and widely available in modern vehicles, they lack depth information. Time-of-Flight (TOF)/3D cameras provide 3D skeletal data that can be used to track hand motion trajectories in three dimensions. Depth maps enable the system to segment hands from the background more effectively, especially when the hands are similar in color to the vehicle interior. Depth sensing is critical for sign language because the movement of the hand toward or away from the camera often distinguishes between different signs (Zhang, 2022) (Ranum et al., 2024).

### 2.4.2. Synthetic Data and Motion Capture (MoCap)

Synthetic data generation has emerged as a key strategy for overcoming the difficulty and high cost of collecting large-scale natural datasets.

Previous studies have successfully utilized motion capture systems to acquire precise three-dimensional point cloud and skeletal data from deaf signers (Kipp et al., 2011; De Martino et al., 2016; Will et al., 2018; Jedlička et al., 2020; Andersen et al., 2025; De Martino et al., 2025). This high-fidelity spatiotemporal data, capturing both manual and non-manual signals, serves as a critical training corpus for neural networks, such as CNNs and RNNs, enabling them to learn the mapping from continuous sign language kinematics to written language translation.

Shterionov et al. (2024) discuss about the types of synthetic work, two of the main trends in synthesis are: (i) the more traditional 3D animation-based which resolves in generating a 3D animated character, commonly referred to as an avatar; and (ii) a video of a virtual human that can be synthesized with generative AI methods based on real human video/image data

## 3. Motivation for the Set-up

The synthesis of the existing literature reveals the primary motivations for developing our specialized Libras resource. This work distinguishes itself by starting from a stable, fundamental baseline, which is currently missing in the field of automotive sign language recognition.

- **Establishing a Multimodal In-Car Baseline:** There is currently no resource documenting Libras production within the physical and visual constraints of a car cabin. Our setup provides the very first multimodal baseline (RGB + 3D TOF) recorded inside a vehicle. By starting with a simplified, stable environment, we establish the necessary scientific foundation before introducing more complex parameters in future iterations, and it provides the first comparative benchmark between idealized laboratory MoCap and these constrained "in-the-wild" conditions.
- **Comparative Analysis of Constrained Space:** This resource is motivated by the need to quantify the linguistic and physical differences between idealized laboratory signing and signing within the confined space of a car cabin. This dataset allows for a direct comparison between high-precision MoCap and real-world cabin constraints.
- **Facilitating Initial Domain Adaptation:** By providing a synchronized real-world baseline and a corresponding synthetic dataset, we enable researchers to begin exploring domain adaptation for cars using expressive 3D avatars.

### 3.1. Potential Research Directions

To address these motivations, the ICSL dataset provides the necessary multimodal data to investigate several key research questions in future studies, such as the following:

- **Camera Placement Optimization:** Determining the optimal camera positions within a vehicle cabin to maximize the visibility of facial expressions and head movements in Libras, despite the spatial constraints.
- **Physical Baseline Comparison:** Analyzing how sign production (hand configuration and path movement) differs when comparing a high-precision laboratory MoCap baseline to the constrained physical space of a vehicle.
- **Multimodal Stability Evaluation:** Assess to what extent multimodal sensor fusion improves the stability of landmark detection in a confined interior compared to traditional monocular systems.
- **Perspective Gap Assessment:** Evaluating whether signing avatars, generated from idealized MoCap, can effectively simulate the non-frontal viewing angles required for in-car Human-Machine Interfaces (HMI).
- **Linguistic Adaptation Analysis:** Investigating how physical restrictions, such as the car seat and seatbelt, impact the duration and spatial volume of signs compared to laboratory-recorded data.
- **Model Robustness Testing:** Quantifying the performance drop when models trained on stable in-car data are tested against other driving-related micro-gestures or environmental noise.

By establishing this baseline, we provide the scientific community with a foundation for building inclusive, real-time assistive technologies that bridge the communication gap in shared mobility services.

## 4. Corpus Design

The design of the In-Car Sign Language (ICSL) corpus is driven by the need for accessible passenger-driver communication within shared mobility services. Unlike general-purpose sign language datasets, this resource focuses exclusively on the constrained signing space of vehicle interiors to analyze how physical constraints impact LIBRAS language.

### 4.1. Scenario Selection and Use Cases

In collaboration with our Deaf researchers and professional Libras interpreters, we defined 1,344 essential communication scenarios for shared mobility services. To date, we have recorded 127 essential cases that reflect the immediate needs of DHH passengers. These are categorized into five functional groups:

- **Navigation and Routing**
  - "Qual o destino?" (Where to?)
  - "Melhor evitar o centro." (It is best to avoid downtown.)
  - "Já passamos o ponto." (We are past my drop-off.)

– etc.

- **Safety and Regulations**
  - “Reduzindo por segurança.” (Reducing for safety reasons.)
  - “Pode ir devagar?” (Can you slow down?)
  - etc.
- **Passenger Comfort and Interaction**
  - “Preciso descer urgente.” (I need to stop urgently.)
  - “Posso ligar o ar-condicionado?” (May I turn on the air conditioning?)
  - etc.
- **Assistance**
  - “Pode ajudar com a cadeira de rodas?” (Can you help with the wheelchair?)
  - “Tem troco para R$ 25,00?” (Do you have change for R$ 25,00?)
  - “Tenho uma mala pesada.” (My luggage is heavy.)
  - etc.
- **Alphabet Spelling**
  - A, B, C, etc.

#### 4.1.1. Subject Recruitment and Diversity

The corpus was recorded by one hearing and two deaf signers, all proficient in Libras, which were instructed to dress as usual for the capture session. They also served as annotators to our data. To ensure that our data captures a wide range of physical constraints, recordings were conducted across three different vehicle models with varying cabin geometries: a Jeep Compass, a Nissan March, and a Renault Megane. All sessions were performed under clear daylight conditions to minimize extreme external lighting noise while preserving the natural dynamic shadows of the vehicle interior as seen in Figure 2.



(a) (b) (c)

Figure 2: Illustration of the three signers and challenges within vehicle cabin: (a) FLIR camera, (b) ToF camera, and (c) recording tablet point of views.

## 5. Data Acquisition and Technical Setup

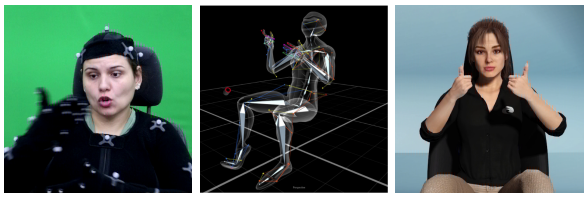
Our data collection followed a two-stage approach. We first established a laboratory baseline using high-precision motion capture to define the “gold standard” for each sign. This was followed by an acquisition phase inside a vehicle cabin, using a multimodal sensor setup to capture the natural constraints for in-car communication.

### 5.1. Laboratory MoCap Baseline

Our ICSL MoCap corpus was built to simulate in-vehicle body positioning and its effects on Libras sign production. To do so, we employed several sensors, including RGB cameras and a motion capture suit (MoCap), to record joint movements with detail and precision during sign language communication. The setup made use of several professionals, including:

- Two Libras signers who were fluent in Libras and performed the signs.
- One system operator to up-keep the quality of the MoCap data being recorded.
- Two hardware operators to ensure frame synchronization and camera operation, and lastly.
- One content facilitator to assist with translation and language production evaluation.

Regarding our frame composition for this task, the laboratory has a green-screen background and a mock-up car seat. The Libras’ signers are recorded while seated, wearing a special MoCap suit that covers most of their bodies. The suit is made from black cloth and allows the attachment of reflective beads (or spherical markers) with velcro, which are detectable by the infrared cameras as seen in Figure 3a. In total, there are 73 beads distributed across the main signer’s joints.



(a) (b) (c)

Figure 3: MoCap setup illustrating the (a) capturing suit, (b) the markerset and skeleton from the Vicon tracking software, and (c) the rendered avatar (Clara).

We also distributed fourteen Vicon (Vicon, 2026) cameras radially around the signer. These cameras are equipped with near-IR sensor that detect

the spherical markers attached to the MoCap suit. Later on the markers will be digitalized to move the character skeleton in the capture software as illustrated in Figure 3b, which in turn will animate the avatar as shown in Figure 3c. Moreover, two Canon EOS T6i Rebel (Canon, 2026) are also employed for RGB capture, alongside the FLIR 2D (Teledyne FLIR, 2026), and the Vicon Vue cameras.

## 5.2. In-Car Data Acquisition

The in-car component of the ICSL corpus was designed to capture the linguistic and environmental challenges of a transport system setting. We employed a multi-sensor approach to provide both standard video and high-fidelity 3D spatial data.

### 5.2.1. Hardware and Sensor Setup

We utilized multi-modal sensors as seen in Figure 4 and the setup as seen in Figure 5b was deployed across three distinct test vehicles: a Jeep Compass(illustrated in Figure 5a), a Nissan March, and a Renault Megane, providing variability in cabin dimensions, window size, lightning, color, texture, seat shape, and interior geometry.




(a) (b) (c)

Figure 4: The multi-modal sensor suite including (a) the FLIR 2D camera, (b) Orbbec ToF sensor, and (c) the Lenovo Tablet.

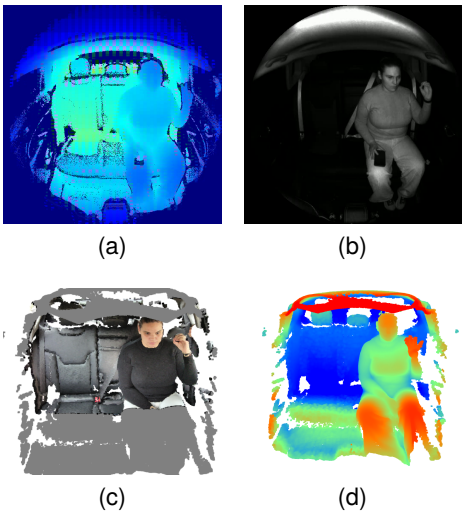
- **2D Imaging:** A FLIR Blackfly S (Teledyne FLIR, 2026) was utilized to capture the primary 2D visual stream at a resolution of $1440 \times 1080$, operating at 60 fps as seen in Figure 4a. To accommodate the confined vehicle interior, we utilized fisheye lenses to maximize the field of view as seen in Figure 2a.
- **3D Time-of-Flight (ToF):** An Orbbec Femto Bolt (Orbbec, 2026) was utilized to capture spatial information as seen in Figure 4b. The sensor was configured to capture synchronized RGB, Depth, Infrared (IR), and Point Cloud streams at a resolution of $1024 \times 1024$ pixels, operating at 30 fps, allowing for precise tracking of hand movements in 3D space as seen in Figures 2b and 6a-d.
- **Recording Tablet:** A Lenovo Tab P12 (Lenovo, 2026) was utilized as an additional perspective, as seen in Figure 4c.



(a)

(b)

Figure 5: (a) test vehicle, and (b) full technical installation inside the Jeep Compass cabin, illustrating sensor placement and the passenger signing space.



(a) (b)

(c) (d)

Figure 6: Multimodal perspectives of Libras communication within a vehicle cabin, illustrating the two sensors: (a) ToF depth image shown in color, where blue indicates closer objects and green indicates farther objects. Dark blue areas mean no depth was measured, (b) ToF infrared (IR) image shown in grayscale, where white represents strong IR reflection and black represents weak or no reflection, (c) ToF RGB point cloud, where each 3D point contains spatial coordinates $(X, Y, Z)$ and its corresponding color (R, G, B), and (d) ToF Depth point cloud, where each 3D point contains spatial coordinates $(X, Y, Z)$ derived only from depth measurements, without color information.

Positioned behind the front-seat headrest at a resolution of up to $3840 \times 2160$ pixels, it recorded the signer from an elevated, centered viewpoint, providing a third data stream to supplement the mounted sensors as seen in Figure 2c.

- **Prompting Interface (Mobile Phones):** In addition to the main sensors, mobile phones were positioned within the signer's line of sight specifically to display the selected use cases. These served as a remote prompting system to guide the signer through the recording sequence.

### 5.3. Acquisition Protocol and Workflow

The recording session followed a strictly controlled protocol managed via a custom web-based control panel, as seen in Figure 7. To ensure the signer remained focused and in the correct posture, we utilized a remote signaling system:

- **Instruction Delivery:** Use cases were published on the mobile phone screens from the central Control Panel.
- **Start/Stop Signals:** The operator sent a "Start" signal (green) to the phone to indicate which sign to perform. Once the gesture was completed, a "Stop" signal (red) was sent, and the system automatically switched to the "Next Case" for the signer.




Figure 7: Web-based remote control interface for Start and Stop signaling.

## 6. Corpus Analysis and Limitations

Our dataset represents a first step toward a comprehensive in-vehicle Libras resource. To ensure the scientific validity of our comparative analysis, we chose to establish a solid baseline under stable, controlled conditions before introducing more dynamic environmental variables. This approach allows us to demonstrate the causal relationship between the physical layout of the car and signing production without the interference of visual noise. By focusing on these idealized conditions, we provide a "ground truth" reference that enables the evaluation of technical benchmarks in a confined space. This is critical because sign language representations must reflect 3D awareness to handle the orientations and spatial trajectories inherent in real-world cabins.

The current iteration of the dataset is built upon a theoretical vocabulary of 1,344 Portuguese context-specific phrases for ride-sharing applications. To date, a subset of 127 phrases has been successfully recorded in video, with plans to complete the remaining recordings in the coming months. The data collection and annotation methodology relies on a cohort of three participants (two deaf, one hearing) who perform both the signing and the ongoing data annotation.

A significant current limitation is the lack of dynamic lighting conditions, such as sun glare, tunnel transitions, or nighttime driving noise. While these factors are vital for "in-the-wild" robustness, in this initial phase, we theorize that models pre-trained on normal illumination data serve as a superior foundation for subsequent fine-tuning in low-light scenarios. As an early-stage resource within Project UNITY, the dataset currently features a limited number of signers, two deaf signer and one hearing. We acknowledge this as a potential source of signer bias that could limit the generalization of initial AI models. While we capture stationary occlusions (seatbelts, headrests), the dataset does not yet account for dynamic occlusions caused by vehicle movement or secondary passenger interactions.

By acknowledging these constraints, we frame this work not as a final solution but as the essential foundation required to guide our future, more complex data-collection efforts in the automotive sign language domain.

## 7. Conclusion

This paper introduces the In-Car Sign Language (ICSL) dataset, a novel multimodal resource designed to investigate the technical challenges of sign language recognition across shared mobility services. By focusing specifically on the constrained signing environment of vehicle interiors, we have established the first specialized dataset that systematically documents how Brazilian Sign Language (Libras) is produced within the physical and visual limitations of car cabins. The corpus comprises over 1.5 million synchronized frames across multiple modalities, providing researchers with a comprehensive foundation for investigating sign language recognition in real mobile settings.

Our contribution is split into two parts. First, a comparative benchmark between laboratory and in-car signing that addresses potential research direc-

tions in Section 3 focused on physical, multimodal, and linguistic adaptations. Second, we provide 127 communication scenarios recorded across three vehicle models, reflecting real-world accessibility needs within diverse cabin geometries.

The limitations acknowledged in Section 6 including controlled lighting conditions, and signer diversity, this work establishes a necessary foundation for robust "in-the-wild" SLR systems. As a component of Project UNITY, the ICSL corpus provides an empirical baseline for assistive technologies aiming to reduce communication barriers between hearing drivers and DHH passengers. Available upon request, this resource supports the development of inclusive transportation systems where language barriers no longer hinder DHH access and dignity.

## 8. Future Work

Moving forward, our next step is to introduce dynamic environmental parameters. We plan to record subsequent datasets under varied lighting conditions. Additionally, future iterations will move from a stationary vehicle to active driving environments to capture dynamic occlusions and noise caused by vehicle vibration, which are known bottlenecks for real-time landmark tracking. Although our dataset focuses on sentence-based communication scenarios, we aim to expand the ICSL corpus to include continuous Libras conversations between passengers, more complex phrases, and a broader range of subjects.

## 9. Acknowledgement

This project is partially financed by the São Paulo Research Foundation (FAPESP), grants #2024/23068-4 and #2024/00914-7, the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), grant #88887.091672/2014-01, National Council for Scientific and Technological Development (CNPq), grant #458691/2013-5, and FINEP, grant #2778/20. The authors would also like to thank the Verkehrsverbund Großraum Ingolstadt (VGI) for their support through the VGI newMIND project.

## 10. Ethical Considerations

The Libras signers in this project are not just participants, they are paid team members and co-authors of this study. Deaf and hard-of-hearing researchers were involved in every stage of the project, from designing the data collection process to writing the paper. This helps us make sure that the linguistic details of Libras are treated with care and that the technology we develop truly reflects the needs of the community. All team members gave their informed consent for the use of their images and biometric data in the corpus. Since the data was created by the research team specifically for this resource, the project follows the institutional ethical guidelines for collaborative research. To request access to the corpus or for further technical inquiries regarding the acquisition software, annotation tools, or the Project UNITY framework, please contact José Mario De Martino at martino@unicamp.br.

## 11. Bibliographical References

Carlos Eduardo GR Alves, Francisco De A Boldt, and Thiago M Paixão. 2024. Enhancing Brazilian Sign Language recognition through skeleton image representation. In *37th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 1–6. IEEE.

Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. 2024. Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects. *Information Processing & Management*, 61(5):103774.

Domenico Amalfitano, Vincenzo D'Angelo, Antonio Rinaldi, Cristiano Russo, and Cristian Tommasino. 2023. Enhancing gesture recognition for sign language interpretation in challenging environment conditions: A deep learning approach. In *15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, pages 395–402.

Jari Ivar Andersen, Gomer Otterspeer, Robert Belleman, and Floris Roelofsen. 2025. Designing a marker based motion capture setup for sign language research. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, pages 1–9.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language Transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Canon. 2026. EOS Rebel T6i product manual. https://www.cla.canon.com/en/p/eos-rebel-t6i. Digital camera technical documentation. Accessed: 2026-02-13.

Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. 2019. Spatial-temporal graph convolutional networks for sign language recognition. In *International Conference on Artificial Neural Networks*, pages 646–657. Springer.

José Mario De Martino, Paula D Paro Costa, A Benetti, Luciana Aguera Rosa, Kate Mamhy Oliveira Kumada, and IR Silva. 2016. Building a Brazilian Portuguese–Brazilian Sign

Language parallel corpus using motion capture data. In *Proceedings of the 12th International Conference on the Computational Processing of the Portuguese Language, Tomar*, pages 56–63.

José Mario De Martino, Ivani Rodrigues Silva, Janice Gonçalves Temoteo Marques, Antonielle Cantarelli Martins, Enzo Telles Poeta, Dener Stassun Christinele, and João Pedro Araújo Ferreira Campos. 2025. Neural machine translation from text to sign language. *Universal Access in the Information Society*, 24(1):37–50.

Daniele LV dos Santos, Thiago B Pereira, Carlos Eduardo GR Alves, Richard JMG Tello, Francisco de A Boldt, and Thiago M Paixão. 2025. Proper body landmark subset enables more accurate and 5X faster recognition of isolated signs in LIBRAS. *arXiv preprint arXiv:2510.24887*.

Annika Herrmann. 2014. *Modal and Focus Particles in Sign Languages: A Cross-Linguistic Study*, volume 2. De Gruyter Mouton, Berlin, Boston.

Pavel Jedlička, Zdeněk Krňoul, Jakub Kanis, and Miloš Železnỳ. 2020. Sign language motion capture dataset for data-driven synthesis. In *9th Workshop on the Representation and Processing of Sign Languages*, pages 101–106.

Michael Kipp, Alexis Heloir, and Quan Nguyen. 2011. Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer.

Marie Lee, Ziming Li, Wendy Dannels, Tae Oh, and Roshan L Peiris. 2025. Exploring one handed signing during driving for interacting with in-vehicle systems for Deaf and hard of hearing drivers. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

Lenovo. 2026. Tab P12: 12.7-inch Storm Grey tablet (8GB/128GB) product manual. https://www.mediamarkt.de/de/product/_lenovo-tab-p12pencil-8gb-128gb-127-zoll-tablet-128-gb-127-zoll-storm-grey-133566330.html. Hardware product specifications. Accessed: 2026-02-12.

Chenghong Lu, Misaki Kozakai, and Lei Jing. 2023. Sign language recognition with multimodal sensors and deep learning methods. *Electronics*, 12(23):4827.

Xiaohan Ma, Rize Jin, and Tae-Sun Chung. 2024. Multi-channel spatio-temporal transformer for sign language production. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11699–11712.

Orbbec. 2026. Orbbec: Femto Bolt Time-of-Flight depth camera specifications. https://www.orbbec.com/products/tof-camera/femto-bolt/. Product technical datasheet. Accessed: 2026-04-07.

Presidência da República. 2002. Lei número 10.436. Diário Oficial da União, Imprensa Nacional, n. 79, Seção 1.

Reza Rastgoo. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.

Dimitar Shterionov, Lorraine Leeson, and Andy Way. 2024. *The pipeline of sign language machine translation*, pages 1–25. Springer.

Hanan A. Taher and Subhi R. M. Zeebaree. 2025. A critical study of recent deep learning-based continuous sign language recognition. *The Review of Socionetwork Strategies*, 19(1):131–161.

Teledyne FLIR. 2026. Firefly S technical specifications. https://www.teledynevisionsolutions.com/products/firefly-s/?segment=iis&vertical=machine+vision. Machine vision camera documentation. Accessed: 2025-02-12.

Vicon. 2026. Vicon Valkyrie advanced motion capture (mocap) camera specifications. https://www.vicon.com/hardware/cameras/valkyrie/. Hardware technical datasheet. Accessed: 2026-02-13.

Ackley Dias Will, José Mario De Martino, and Juliano Renato S Bezerra. 2018. An optimized marker layout for 3D facial motion capture. In *Proceedings of the Smart Tools and Apps for Graphics (STAG)*, pages 107–113.

Wanyu Zhang. 2022. Sign language recognition based on depth image processing. *Highlights in Science, Engineering and Technology*, 23:25–33.

## 12. Language Resource References

Luis R. Cerna, E. E. Cardenas, D. G. Miranda, David Menotti, and Guillermo Camara-Chavez. 2021. *A Multimodal Libras-UFOP Brazilian Sign Language Dataset of Minimal Pairs using a Microsoft Kinect Sensor*. Expert Systems with Applications.

Ronice Müller de Quadros and Alexandre Melo de Sousa. 2021. *Corpus da Língua Brasileira de Sinais: Inventário de Libras do Acre*. Revista de Estudos da Linguagem.

M. A. Lima, D. Cruz, D. R. Silva, D. D. Albuquerque, D. F. Lacerda, R. Costa, G. L. de Souza Filho, and T. M. de Araújo. 2025. *VLibrasBD: A Brazilian Portuguese-Brazilian Sign Language (Libras) Bilingual Text Dataset*. Data in Brief.

Vicente Coelho Lobo-Neto and Helio Pedrini. 2024. *LSWH100: A Handshape Dataset for Brazilian Sign Language (Libras) using SignWriting*. Data in Brief.

Oline Ranum, Gomer Otterspeer, Jari I. Andersen, Robert G. Belleman, and Floris Roelofsen. 2024. *3D-LEX v1.0: 3D Lexicons for American Sign Language and Sign Language of the Netherlands*. arXiv preprint arXiv:2409.01901.

Tamires Martins Rezende, Sílvia Grasiella Moreira Almeida, and Frederico Gadelha Guimarães. 2021. *Development and Validation of a Brazilian Sign Language Database for Human Gesture Recognition*. Neural Computing and Applications.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, Daksh Sehgal, Saad Hassan, Bill Neubauer, Sofia Vempala, Alec Tan, Jocelyn Heath, Unnathi Kumar, Priyanka Mosur, Tavenner Hall, Rajandeep Singh, Christopher Cui, Glenn Cameron, Sohier Dane, and Garrett Tanzer. 2023. *PopSign ASL v1.0: An Isolated American Sign Language Dataset Collected via Smartphones*. Curran Associates, Inc.

# Seeing Who Is Signing and With Which Hand

**Carl Börstell**
Department of Linguistic, Literary and Aesthetic Studies, University of Bergen
Bergen, Norway
carl.borstell@uib.no

**Abstract**

This study is a computer vision analysis of 4.5 hours of video data from 40 signers in the Swedish Sign Language Corpus, aiming to evaluate the reliability of classifying 1) who the main signer is at any given time during dyadic conversation, and 2) the dominant hand (i.e., handedness) of each signer. First, the distance moved by the hands of each signer is used to compare the manual activity between a) the two signers to determine whose hands are more active, and b) the hands of each signer to determine which hand is more likely to be dominant. Second, the height of the hands is used to compare their prominence in signing space between a) the two signers to determine whose hands are more prominent, and b) the hands of each signer to determine which hand is more likely to be dominant. The results show that while both distance and height approaches can reliably classify – individually or combined – the main signer in any segment of a conversation, the height approach is better at determining the overall handedness (right- or left-dominant) of signers. For the handedness classification, the optimal method turns out to be a two-step approach, first classifying the main signer per segment, then using only signer-relevant segments to classify handedness.

**Keywords:** sign language, handedness, corpus, conversation, computer vision

## 1. Introduction

Signed language is produced through simultaneous movements of manual and non-manual articulators, of which the two hands are generally seen as the main articulators of the lexical items (i.e., the signs). While the two hands are symmetrical articulators, most signers display some handedness-preference, generally signing with either right or left hand as their more dominant hand (Papadatou-Pastou and Sáfár, 2016). The dominant hand is the one used for most one-handed signs, and is the more active hand in unbalanced two-handed signs, for which the other (non-dominant) hand acts as a place of articulation (Sandler, 1993; van der Hulst, 1996; Crasborn, 2011). However, in natural discourse, the dominant hand can switch, a concept known as *dominance reversal*. Dominance reversal can happen for multiple reasons, ranging from articulatory ease to discourse and information structure (Crasborn and Sáfár, 2016). For instance, a hand may retain a fragment of a previous sign or establish a referential anchor (known as *buoys*) to structure discourse (Liddell, 2003; Liddell et al., 2007; Nilsson, 2007; Kimmelman et al., 2016), which can also lead to dominance reversal (Crasborn and Sáfár, 2016). While a signer's hand dominance is often assumed to be the same as their general handedness (right- or left-handed), this is not always the case, and some signers may be referred to as *mixed-handed*, exhibiting a high degree of ambidexterity in their signing (see Papadatou-Pastou and Sáfár, 2016).

Just as one of a signer's two hands may be more dominant at a given point during signing, one signer may be more dominant during some stretch of signed conversation. Turn-taking practices have been studied in several sign languages to date, and concern the switching of the "main signer" in a conversation – i.e., a transfer of who holds the floor – which may involve more or less overlap between participants' signing (Baker, 1977; Van Herreweghe, 2002; Groeber and Pochon-Berger, 2014; de Vos et al., 2015; Manrique and Enfield, 2015; Ferrara, 2022). Conversational feedback (backchanneling) may lead to overlap in signing, whether by manual or non-manual means, without necessarily involving a switch of main signer (Mesch, 2016; Bauer et al., 2024; Börstell, 2024; Lutzenberger et al., 2024). Using a computer vision analysis of Swedish Sign Language (STS) conversational video data, Börstell (2024) demonstrated that manual backchanneling – more specifically *continuers*, signals that encourage the conversational partner to proceed – are articulated lower in signing space and with less movement relative to other signs. Börstell (2024) argued that these properties may decrease visual prominence, signaling that the signing is not an attempt to claim the floor.

Börstell's (2024) study on continuer backchannels involved computer vision to analyze some phonetic properties of backchannels. The use of computer vision for processing sign language data directly from videos has proven useful for analyzing phonetic and spatial properties of both manual (Östling et al., 2018; Börstell and Lepic, 2020) and non-manual articulation (Kimmelman et al., 2020; Bauer et al., 2024; Kuznetsova and Kimmelman, 2024; Sargano et al., 2024). However, while signer handedness has been estimated on the basis of distance or velocity of hand movements as a pre-



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 42–50
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

processing step for further analysis (e.g., Östling et al., 2018; Fragkiadakis et al., 2020; Fragkiadakis and van der Putten, 2021; Akamine et al., 2026), it is rarely evaluated in its own right. An exception to this is Börstell (2023), who used computer vision estimates of manual movement distances to predict signers' dominant hand from individual dictionary sign videos, compared to their known handedness in the lexical database metadata as ground truth. Börstell (2023) reports around 80% accuracy overall, although there were large discrepancies between one-handed signs, which the method predicted with high accuracy, and balanced two-handed signs, in which both hands move in a similar/symmetrical fashion, hence were much more difficult to classify in terms of handedness.

In this study, I evaluate whether classification of main signer and hand dominance can be done accurately from conversational video data using computer vision. The research questions are:

- Using a computer vision analysis of Swedish Sign Language (STS) dyadic, conversational corpus video data, is it possible to determine
  1. ...who the main signer is at a given time?
  2. ...which hand is a signer's dominant one?

## 2. Methodology

### 2.1. Data

For the goals of this study, 40 dyadic, conversational texts were sampled from the Swedish Sign Language (STS) Corpus (see Öqvist et al., 2020). The dataset used here consists of 80 MPEG video files (Mesch et al., 2012), each showing the (slightly angled) frontal view of each signer in the dyad, as well as the accompanying EAF annotation files (Mesch et al., 2014), all retrieved from *The Language Archive*.[1] The texts range from about 30 seconds to 8 minutes in duration (about 3.5 minutes on average), comprising over 2 hours and 15 minutes of conversation, or about 10% of the STS Corpus. The texts were selected with two main goals in mind: first, they should all involve both signers engaging in signing; second, they should cover as many signers as possible and involve different text types. From these criteria, two specific tasks were selected: a) the *presentation* texts in which each signer provides a short self-introduction to their addressee; b) the *elicited narratives* in which the signers retell short comic strips. Thus, each text should have at least one switch of main signer, and the texts include both free and narrative-type conversation. There was comparative data available for 40 out of the 42 signers in the STS Corpus (across 20 pairs of signers) for these texts, and these were all selected to maximize signer diversity. Table 1 shows the sampled signers and texts, using the labels found in the corpus itself, with each row representing a single pair of signers and their associated texts.

| **Signer** | | **Text type** | |
|---|---|---|---|
| S1 | S2 | Presentation | Elicited narrative |
| S003 | S004 | SSLC01_020 | SSLC02_028 |
| S005 | S006 | SSLC01_040 | SSLC02_049 |
| S007 | S008 | SSLC01_060 | SSLC02_070 |
| S009 | S010 | SSLC01_080 | SSLC02_089 |
| S011 | S012 | SSLC01_100 | SSLC02_117 |
| S013 | S014 | SSLC01_120 | SSLC02_132 |
| S015 | S016 | SSLC01_140 | SSLC02_149 |
| S017 | S018 | SSLC01_160 | SSLC02_169 |
| S019 | S020 | SSLC01_180 | SSLC02_190 |
| S021 | S022 | SSLC01_200 | SSLC02_214 |
| S023 | S024 | SSLC01_220 | SSLC02_229 |
| S025 | S026 | SSLC01_240 | SSLC02_250 |
| S027 | S028 | SSLC01_260 | SSLC02_275 |
| S029 | S030 | SSLC01_280 | SSLC02_295 |
| S031 | S032 | SSLC01_300 | SSLC02_310 |
| S033 | S034 | SSLC01_320 | SSLC02_333 |
| S035 | S036 | SSLC01_340 | SSLC02_352 |
| S037 | S038 | SSLC01_360 | SSLC02_372 |
| S039 | S040 | SSLC01_380 | SSLC02_394 |
| S041 | S042 | SSLC01_400 | SSLC02_411 |

Table 1: The sampled texts by signer pair and text type with the STS Corpus signer and file labels.

### 2.2. Analysis

In total, 80 video files (40 signers $\times$ two texts), about 4.5 hours in total duration, were analyzed using the MediaPipe pose estimation model (Lugaresi et al., 2019). This model outputs coordinates of 33 landmarks of the body, shown in Figure 1. The coordinates used for this study are the so-called *world landmarks*, which reconstruct the landmark coordinates in three-dimensional space, using the middle of the person (around the hips) as the origin to which all other landmarks are positioned (see Figure 1). For the purposes of this study, the movement of the hands in space was the relevant measure, and thus the landmarks for the wrist of each hand (Figure 1) are used for measuring manual activity. The reason for using the wrist rather than finger landmarks is that the wrists tend to be more visible and thereby have higher reliability for pose estimation of the hands compared to the fingers.[2]

[1] https://archive.mpi.nl/tla/

[2] In the current dataset, the visibility of the wrists ($M = .95; SD = .11$) is indeed higher than that of the index fingers ($M = .91; SD = .13$).

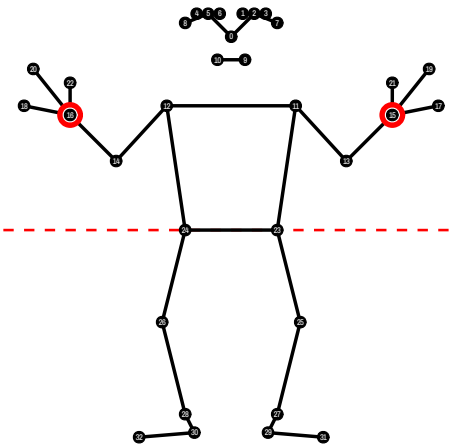


Figure 1: The 33 landmarks in the MediaPipe pose estimation model. The red circles mark the wrist landmarks, used for calculating both distance and height. The dashed red line indicates the vertical origin of the pose estimation model coordinates.

The two objectives of this study were to evaluate whether a computer vision analysis of sign language conversations can identify 1) who the main signer is at a given time in a conversation (i.e., the current *main signer*), and 2) which hand each signer uses as their overall dominant one (i.e., the *handedness* of each signer). The two approaches to attempt such classifications use *distance moved* by the hands and their *height* in signing space.

For the first approach, the distance moved by the hands was calculated as the Euclidean distance between each wrist position across sequential frames.[3] Thus, when the coordinates of one wrist in a frame are measured as $(p_1, p_2, p_3)$ and in following frame as $(q_1, q_2, q_3)$, the distance moved between those two frames can be calculated as:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$

Distance was calculated across frames per text, signer and wrist and then summarized across the two hands per frame and signer, since the signer's handedness is unknown for this task and we do not know which movements are actual signing (rather than grooming/adjusting), nor which signs are one- vs. two-handed (cf. Börstell, 2023).

For the second approach, the height of the hands was calculated as the maximum and mean height, across the two wrists of each signer.[4]

[3]Approximately .01% of MediaPipe pose estimation frames had a (mostly single-frame) gap compared to video frames. Due to the low frame-drop rate, gaps were disregarded and frames were treated as consecutive.

[4]The *y* coordinates were inverted in the height calculations, as MediaPipe outputs *y* coordinates that are vertically inverted as per standard image processing.

In both approaches, texts were cut into 5-second segments, resulting in many individual data points per text. The 5-second duration was chosen somewhat arbitrarily, but estimated as a unit that should span multiple signs but preferably not multiple turns (see Börstell et al., 2024; Börstell, 2024). In the classification of main signer, distance is taken as the *total* across frames per segment and signer, while height is taken as the *maximum* and *mean* per segment and signer. Each segment is then assigned a main signer for each method: distance, maximum height and mean height, respectively. Each method allowed for ties (i.e., no *main* signer), although no such cases were ultimately observed. In the classification of handedness, distance and height are kept separate for each hand, and the two hands are compared within each signer and segment. For each segment, signers are assigned a dominant hand for each method – distance, maximum height and mean height – allowing for ties, although no such cases were observed. Signers are ultimately classified based on the majority of left-/right-predicted segments, defaulting to right-handed in cases of equal prediction counts.

For the classification of main signer, the ground truth was based on the number of sign glosses produced by each signer for each text segment, labeling the signer with more signs as the main signer of that segment. Segments in which both signers had the same number of sign glosses (<1% of all segments) were removed from further analysis. For the classification of signer handedness, the ground truth was given as the handedness label found in the metadata files of the STS Corpus, which use the labels *right-*, *left-* and *mixed-handed*. However, three signers' handedness were corrected before applying the labels as ground truth. First, signers S003 and S004 have erroneously had their handedness swapped, which is obvious from observing the video data as well as seeing that the annotation file gloss tiers have the reverse hand dominance compared to the metadata. That is, the annotation tiers are correctly assigned, but the metadata files have incorrect handedness labels for this pair of signers. Second, signer S037 is the only one labeled as *mixed-handed*, but has been assigned a right-handed status in the annotation files, in which the dominant hand tier is the right hand. According to Mesch and Wallin (2015, 109), dominant hand tiers in the STS Corpus files were determined based on the perceived dominant hand by annotators, which in undecided cases defaulted to the right hand.

MediaPipe was run in Python v3.10.5. Data analysis was done with R v4.5.3 (R Core Team, 2026) and `tidyverse` v2.0.0 (Wickham et al., 2019), `tidysigns` v.0.2.2 (Börstell, 2025), `scales` v1.4.0 (Wickham et al., 2025), `gt` v1.3.0 (Iannone et al., 2026) and `here` v1.0.2 (Müller, 2025).

## 3. Results

### 3.1. Classifying Main Signer

The first objective of this study was to classify the main signer for each 5-second segment, using the annotated sign glosses as the ground truth for evaluation. Table 2 shows the accuracy of the approaches: distance vs. (max/mean) height.

| | Classification | |
|---|---|---|
| **Method** | correct | incorrect |
| distance | 1,573 (95.3%) | 78 (4.7%) |
| height (max) | 1,567 (94.9%) | 84 (5.1%) |
| height (mean) | 1,535 (93.0%) | 116 (7.0%) |
| *combined* | 1,600 (96.9%) | 51 (3.1%) |

Table 2: Accuracy of main signer classification for each method by number of segments correctly vs. incorrectly labeled (percentages in brackets).

Based on these results, both approaches obtain a high degree of accuracy, with the distance method performing the best (95.3% correct classifications), slightly ahead of height, for which maximum height (94.9%) performs better than mean height (93%). The influence of text type appears insignificant, with the accuracy for the distance approach scoring equally well for both types, and the height approach scoring similarly across text types (max: presentation 95.1% and narrative 94.7%; mean: presentation 94.1% and narrative 91.6%).

Looking at some individual texts to gauge potential sources of error, it is clear that many incorrect classifications occur at turn-taking boundaries, such as in the example from text SSLC01_020, visualized in Figure 2. Determining who the main signer is around natural turn-taking points in a conversation will obviously be difficult, as there may be overlapping signing and slightly different classifications by different methods (e.g., a signer ending their turn may still be signing more signs, but with lowered manual articulators). However, other issues also exist, such as the text SSLC01_100, in which a signer is classified as main by the mean height method while having almost no manual activity, but having both arms crossed on the chest, giving a high average height despite not actively articulating with the hands. Another case can be seen in the text SSLC01_180, in which a signer is being incorrectly labeled as main by the distance method, due to several instances of self-grooming (e.g., face scratching) or occasional manual backchannels, thus creating sudden bursts of greater distances moved by the hands from a rest position on the lap.

If the methods fail in different contexts and for different reasons, a combined approach could be more accurate. Indeed, combining the three methods together and classifying main signer based on the majority label does increase the accuracy to 96.9%, higher than any individual method (see the bottom row of Table 2).

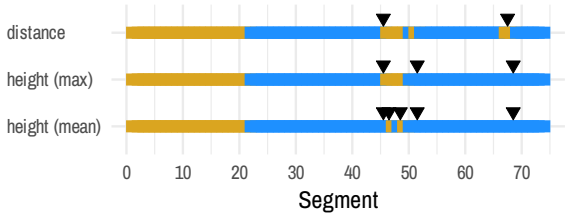



Figure 2: Classification of main signer by segment in one text (SSLC01_020). Segment colors show main signer predictions by method. Black triangles indicate segments with incorrect classification compared to gloss dominance as ground truth.

Since the 5-second segment duration was somewhat arbitrary, different segment lengths were compared in order to explore the effects of this parameter setting on the classification.[5] Figure 3 shows the main signer prediction accuracy across methods and the five different segment durations evaluated: 1, 2.5, 5, 7.5 and 10 seconds. As seen in Figure 3, the distance approach appears to increase in accuracy when using slightly longer segment durations, whereas the height approach may rather benefit from shorter durations. However, it turns out that the combined majority-rule approach across methods actually performs optimally with the initial 5-second segment setting, perhaps as the best trade-off across methods. Since it appears that the initial 5-second segment length performs well, it will be used throughout the analyses that follow.

### 3.2. Classifying Signer Handedness

The second objective of this study was to classify each signer by their overall handedness, whether they are mainly right- or left-dominant signers. Here, the ground truth is the classification in the STS Corpus metadata files, labeling each signer as either *right-handed* (n = 35), *left-handed* (n = 4) or *mixed-handed* (n = 1). Upon visual inspection of the videos and annotation files, it is clear that two signers in the same pair have had their classifications swapped (S003 and S004) and that the mixed-handed signer (S037) has been labeled right-handed in the STS Corpus annotation structure (see Section 2.2). Thus, the handedness labels used as ground truth include the corrected classification of S003 and S004, and classifying S037 as right-handed.

Since signs are annotated on separate tiers for each hand in the STS Corpus, we can use the

[5] I thank an anonymous reviewer for this suggestion.




Figure 3: Accuracy of main signer classification by method and segment length. Shapes and colors denote the method, with the combined approach using majority rule across the three methods.

distribution of sign gloss annotations across tiers as a frequency measure of handedness, labeling whichever hand has more sign gloss annotations as the dominant one for each signer. Using this measure across all sampled texts, the classification is 100% accurate, correctly predicting all 40 signers: 36 right-handed and 4 left-handed.

As for the computer vision approaches, signer handedness was classified based on the majority of 5-second segments in which one hand was deemed the more active in terms of a) distance moved, b) maximum height, and c) mean height. For each method, if ≥50% of segments were labeled *right*, the signer was classified as right-handed, otherwise left-handed. All three methods correctly classify the four left-handed signers as such, but incorrectly label a number of right-handed signers as left-handed: distance (15/36 incorrect); maximum height (8/36 incorrect); mean height (6/36 incorrect) – see classifications in Figure 4 and F-scores in Table 3. Thus, the height approach is overall better than the distance approach, and using mean height performs slightly better than using maximum height. Unlike the main signer classification, a combined approach classifying by majority rule across the three methods does not outperform the best individual approach (see the bottom row in Table 3).

Figure 5 shows the proportional right hand preference for each signer as calculated across segments for each method. As Figure 5 illustrates, most signers are correctly classified, but seven right-handed signers are incorrectly classified as left-handed when basing the classification on majority rule. It is noteworthy that signer S037, the only

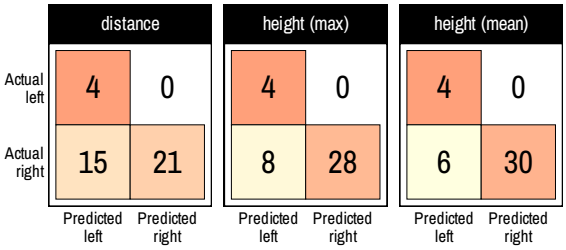



Figure 4: Predicted vs. actual handedness of signers (n = 40) by method.

| | F-score | |
|---|---|---|
| **Method** | balanced | weighted |
| distance | 0.54 | 0.70 |
| height (max) | 0.69 | 0.84 |
| height (mean) | 0.74 | 0.88 |
| *combined* | 0.71 | 0.86 |

Table 3: Performance of handedness classification by method as balanced and weighted F-scores.

*mixed-handed* signer according to the metadata, is classified as a left-handed signer, despite being annotated as right-handed in the STS Corpus.

Looking at some of the misclassified signers' sign gloss annotation frequencies, it becomes obvious that some (e.g., S021) are among the signers with the fewest annotations in the dataset sampled for this study. Since these computer vision methods are using the entire video files as data, it is likely that much of the recording will be of the signers as addressees, with their hands mostly at rest, apart from the occasional grooming or backchanneling (Mesch, 2016; Börstell, 2024). Thus, a two-step approach was attempted, in which the best-performing method for determining main signer (i.e., distance moved; see Section 3.1) is used to subset the analyzed segments to only those in which a signer is classified as the main signer – that is, segments in which they are likely actively signing – and these segments are subsequently used to classify the signers' handedness using the best-performing method for determining handedness (i.e., mean height). The classification in this two-step approach turns out to be 100% accurate, with all 40 signers being correctly classified according to their known handedness. The classification for each signer in this two-step approach is shown in Figure 6, for which we can also see that signer S037 – the one originally listed as *mixed-handed* in the metadata – is the one closest to the midpoint between right and left hand preference, whereas all other misclassified right-handed signers (see Figure 5) now show a clear right hand preference.

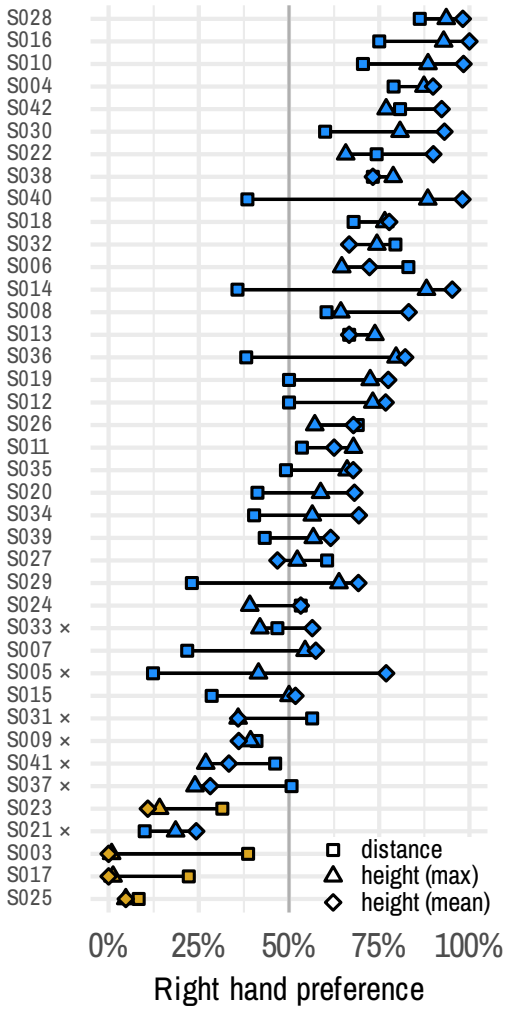



Figure 5: Handedness preference across signers according to the three methods: distance, height (max) and height (mean). Colors show true handedness (blue: right-handed; yellow: left-handed); shapes represent the method. Xs (×) mark signers classified incorrectly by majority rule.

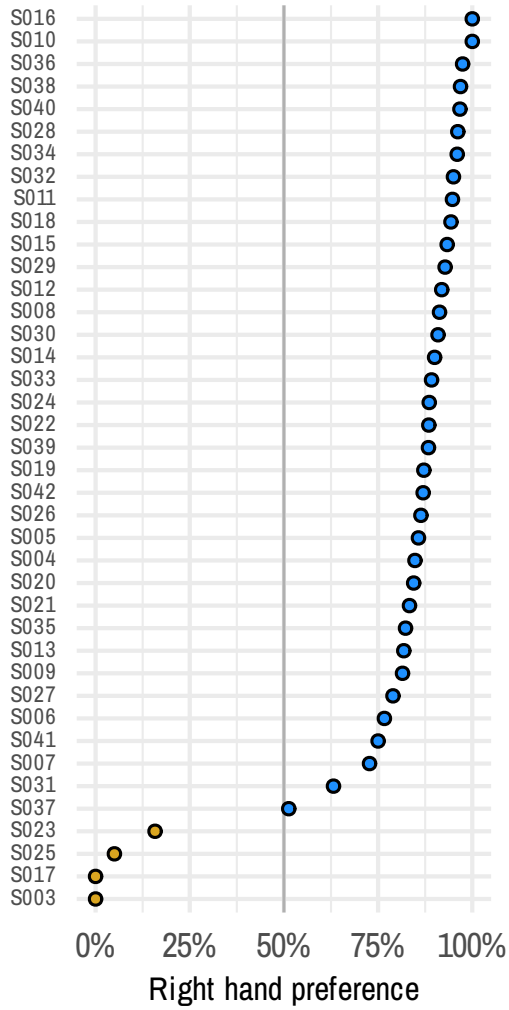



Figure 6: Handedness preference across signers according to the combined two-step approach in which active signing is determined by distance, and handedness is determined by mean height for active signing segments only. Colors show true handedness (blue: right-handed; yellow: left-handed).

## 4. Conclusions

This study set out to evaluate the applicability of computer vision in determining 1) the main signer at any given point in dyadic, conversational signing, and 2) the dominant hand (or, preferential handedness) of each signer in the dataset. While the sampled dataset is small, it constitutes quite a substantial portion of the entire STS Corpus (≈ 10%) and is different among computer vision analyses by looking at whole conversations rather than extracted clips or individual sign videos – e.g., Börstell (2023), who also addressed hand dominance, but in individual dictionary videos. Despite limited data, both tasks turned out to be quite reliably solved, whether through individual or combined approaches.

For the main signer classification, both distance moved by the hands and their height in space as measures could reliably identify the correct signer with high accuracy (≈ 95%). These approaches could thus be used as a first-pass analysis of not-yet-annotated sign language conversations, to get a sense of conversational turn-taking patterns. They could also be applied to already annotated sign language datasets and corpora, as a way to spot interactions or mismatches between manual annotations and phonetic/visual prominence, and analyze the connections between kinematics and grammatical and conversational structure.

For the signer handedness classification, it turned out that the distance and height approaches both struggled to accurately identify all signers ac-

cording to their known handedness. Since the videos contain a lot of segments of manual inactivity, there is a lot of noise in the data, which may skew the metrics simply based on irrelevant features, such as the choice of manual rest position (e.g., on the lap vs. arms crossed). However, with the high accuracy of the main signer classification already established, a two-step approach could be devised, in which only active signing segments are identified by looking at the distance moved by the hands, followed by an analysis of mean hand height for these segments only. For this dataset, the accuracy of this two-step approach was 100%, correctly classifying all 40 signers according to their known hand dominance. Employing a handedness classification analysis could be used during the initial stages of building a dataset or corpus, automatically classifying handedness for signer metadata purposes without first having to commit to manual annotation work. Furthermore, even if sign gloss annotations have already been added to the data, or metadata about signers' self-reported handedness has been collected, an automated classification of hand dominance could contribute to a more gradient view of preferential handedness among signers, categorizing signers as more or less right-/left-dominant. Similarly, a hand dominance analysis could assist in locating relevant segments in conversations involving dominance reversal or interesting simultaneous use of the two hands.

## 5. Data Availability

Data and code for this study can be found at: https://osf.io/3da27

## 6. Acknowledgments

Funded by the European Union (ERC, NONMANUAL, project number 101039378). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## 7. Bibliographical References

Sho Akamine, Mark Dingemanse, and Aslı Özyürek. 2026. Validating dynamic time warping as a measure of gesture form similarity.

Charlotte Baker. 1977. Regulators and turn-taking in American Sign Language discourse. In Lynn A. Friedman, editor, *On the other hand: New perspectives on American Sign Language*, pages 215–236. Academic Press, New York, NY.

Anastasia Bauer, Anna Kuder, Marc Schulder, and Job Schepens. 2024. Phonetic differences between affirmative and feedback head nods in German Sign Language (DGS): A pose estimation study. *PLOS ONE*, 19(5):e0304040.

Carl Börstell. 2023. Extracting sign language articulation from videos with MediaPipe. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, NEALT Proceedings Series, No. 52, pages 169–178, Tórshavn, Faroe Islands. University of Tartu Library.

Carl Börstell. 2024. Finding continuers in Swedish Sign Language. *Linguistics Vanguard*, 10(1):537–548.

Carl Börstell. 2025. *tidysigns: Sign language corpus data in a tidy format*. R package version 0.2.2.

Carl Börstell and Ryan Lepic. 2020. Spatial metaphors in antonym pairs across sign languages. *Sign Language & Linguistics*, 23(1–2):112–141.

Carl Börstell, Adam Schembri, and Onno Crasborn. 2024. Sign duration and signing rate in British Sign Language, Dutch Sign Language and Swedish Sign Language. *Glossa Psycholinguistics*, 3(1).

Onno Crasborn. 2011. The other hand in sign language phonology. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice, editors, *The Blackwell companion to phonology*, pages 223–240. Wiley, Malden, MA & Oxford.

Onno Crasborn and Anna Sáfár. 2016. An annotation scheme to investigate the form and function of hand dominance in the Corpus NGT. In Roland Pfau, Markus Steinbach, and Annika Herrmann, editors, *A matter of complexity*, pages 231–251. De Gruyter Mouton & Ishara Press, Boston, MA/Berlin & Preston.

Connie de Vos, Francisco Torreira, and Stephen C. Levinson. 2015. Turn-timing in signed conversations: coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology*, 6.

Lindsay Ferrara. 2022. Indexing turn-beginnings in Norwegian Sign Language conversation. *Gesture*, 21(1):1–27.

Manolis Fragkiadakis, Victoria Nyst, and Peter van der Putten. 2020. Signing as input for a dictionary query: Matching signs based on joint

positions of the dominant hand. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 69–74, Marseille, France. European Language Resources Association (ELRA).

Manolis Fragkiadakis and Peter van der Putten. 2021. Sign and search: Sign search functionality for sign language lexica. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 23–32, Virtual. Association for Machine Translation in the Americas.

Simone Groeber and Evelyne Pochon-Berger. 2014. Turns and turn-taking in sign language interaction: A study of turn-final holds. *Journal of Pragmatics*, 65:121–136.

Richard Iannone, Joe Cheng, Barret Schloerke, Shannon Haughton, Ellis Hughes, Alexandra Lauer, Romain François, JooYoung Seo, Ken Brevoort, and Olivier Roy. 2026. *gt: Easily create presentation-ready display tables*. R package version 1.3.0.

Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study. *PLOS ONE*, 15(6):e0233731.

Vadim Kimmelman, Anna Sáfár, and Onno Crasborn. 2016. Towards a classification of weak hand holds. *Open Linguistics*, 2(1).

Anna Kuznetsova and Vadim Kimmelman. 2024. Testing MediaPipe Holistic for linguistic analysis of nonmanual markers in sign languages. ArXiv:2403.10367v2.

Scott K. Liddell. 2003. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press, Cambridge.

Scott K. Liddell, Marit Vogt-Svendsen, and Brita Bergman. 2007. A crosslinguistic comparison of buoys. Evidence from American, Norwegian, and Swedish Sign Language. In Myriam Vermeerbergen, Lorraine Leeson, and Onno Crasborn, editors, *Simultaneity in signed languages: Form and function*, volume 281 of *Current issues in linguistic theory*, pages 187–215. John Benjamins, Amsterdam/Philadelphia, PA.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.

Hannah Lutzenberger, Lierin De Wael, Rehana Omardeen, and Mark Dingemanse. 2024. Interactional infrastructure across modalities: A comparison of repair initiators and continuers in British Sign Language and British English. *Sign Language Studies*, 24(3):548–581.

Elizabeth Manrique and N.J. Enfield. 2015. Suspending the next turn as a form of repair initiation: evidence from Argentine Sign Language. *Frontiers in Psychology*, 6:1326.

Johanna Mesch. 2016. Manual backchannel responses in signers' conversations in Swedish Sign Language. *Language & Communication*, 50:22–41.

Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, 20(1):103–121.

Kirill Müller. 2025. *here: A simpler way to find your files*. R package version 1.0.2.

Anna-Lena Nilsson. 2007. The non-dominant hand in a Swedish Sign Language discourse. In Myriam Vermeerbergen, Lorraine Leeson, and Onno Crasborn, editors, *Simultanteity in signed languages: Form and function*, volume 281 of *Current issues in linguistic theory*, pages 163–185. John Benjamins, Amsterdam/Philadelphia, PA.

Zrajm Öqvist, Nikolaus Riemer Kankkonen, and Johanna Mesch. 2020. STS-korpus: A sign language web corpus tool for teaching and public use. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 177–180, Marseille, France. European Language Resources Association (ELRA).

Robert Östling, Carl Börstell, and Servane Courtaux. 2018. Visual iconicity across sign languages: Large-scale automated video analysis of iconic articulators and locations. *Frontiers in Psychology*, 9:725.

Marietta Papadatou-Pastou and Anna Sáfár. 2016. Handedness prevalence in the deaf: Meta-analyses. *Neuroscience & Biobehavioral Reviews*, 60:98–114.

R Core Team. 2026. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Wendy Sandler. 1993. Hand in hand: The roles of the nondominant hand in sign language phonology. *The Linguistic Review*, 10(4):337–390.

Allah Bux Sargano, Sébastien Vandenitte, Tommi Jantunen, and Vadim Kimmelman. 2024. Evaluation of head pose estimation algorithms for sign language analysis. In *2024 International Conference on IT and Industrial Technologies (ICIT)*, pages 1–6, Chiniot, Pakistan. IEEE.

Harry van der Hulst. 1996. On the other hand. *Lingua*, 98:121–143.

Mieke Van Herreweghe. 2002. Turn-taking mechanisms and active participation in meetings with Deaf and hearing participants in Flanders. In Ceil Lucas, editor, *Turn-taking, fingerspelling and contact in signed languages*, volume 8 of *Sociolinguistics in deaf communities series*, pages 73–103. Gallaudet University Press.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686.

Hadley Wickham, Thomas Lin Pedersen, and Dana Seidel. 2025. *scales: Scale functions for visualization*. R package version 1.4.0.

## 8. Language Resource References

Johanna Mesch, Maya Rohdell, and Lars Wallin. 2014. *Annotated files for the Swedish Sign Language Corpus. Version 2.* Sign Language Section, Department of Linguistics, Stockholm University.

Johanna Mesch, Lars Wallin, Anna-Lena Nilsson, and Brita Bergman. 2012. *Dataset. Swedish Sign Language Corpus project 2009–2011 (version 1)*. Sign Language Section, Department of Linguistics, Stockholm University.

# SignGPT and the Visual Language Toolkit

**Matt Brown[1], Oline Ranum[2], Edward Fish[2], Heidi Proctor[1], Bencie Woll[1], Richard Bowden[2], Kearsy Cormier[1]**

[1] Deafness, Cognition, & Language Centre (DCAL), University College London, 49 Gordon Square, London WC1H 0PD
[2] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, Surrey GU2 7XH
{m.brown.13, heidi.proctor.14, b.woll, k.cormier}@ucl.ac.uk
{o.ranum, edward.fish, r.bowden}@surrey.ac.uk

**Abstract**

SignGPT's Visual Language Toolkit (VLTK) aims to remove fundamental barriers to large scale sign language modelling by developing data-driven, linguistically grounded methods for continuous sign language recognition. We first identify fundamental issues around the ecological validity of potential data sources (e.g. broadcast media with interpreted signing or captions, scraping of social media). We contrast these with the currently highly resource-intensive development of curated sign language corpora based on linguistic principles. The VLTK addresses this scarcity of high quality sign language data by providing semi-automated glossing and other recognition tools, driving large scale corpus expansion without sacrificing linguistic principles. Unlike prior systems that rely on sparse glossing, the project integrates dense temporal annotation, non-manual and non-lexical feature tracking, and transformer-based architectures to capture the multimodal and spatial structure of signing. By aligning machine vision innovation with linguistic insights and community-embedded evaluation, SignGPT establishes a foundation for robust and extensible sign language models.

**Keywords:** sign language linguistics, large language models, glossing, transformers, corpus linguistics

## 1. Overview

SignGPT (https://sites.google.com/view/signgpt) is a 5-year grant from the Engineering and Physical Sciences Research Council, part of UK Research and Innovation. The project brings together specialists in machine vision, generative AI, and sign language linguistics from Surrey University, Oxford University, and University College London. It aims to develop the first large-scale, end-to-end AI system capable of bidirectional translation between British Sign Language (BSL) and spoken/written English. To that end, we treat languages in the visual-spatial modality as deserving of their own large language models (LLMs) rather than as derivatives of a spoken language. Unlike prior systems that have relied on limited vocabularies and rule-based translation, the project integrates computer vision, sign linguistics, and generative AI to address the full complexity of continuous signing, including non-manual features and spatial grammar. The long term ambition is to create a conversational sign language system analogous to ChatGPT, allowing deaf users to interact naturally in sign and receive fluent, photorealistic signed responses.

Training LLMs and other foundational models requires a large amount of data. Sign languages do not have a written form, so historically the creation of suitable datasets has been hampered by the intensive labour required for accurate and exhaustive linguistic annotation. For instance, only about 15% of the video data in the BSL Corpus (Schembri et al., 2014) has been annotated at all to date. To address this issue, SignGPT is developing semi-automatic annotation tools that reduce dependence on costly manual glossing while still preserving linguistic validity. This Visual Language Toolkit (VLTK) will be released openly, benefiting the wider research community. The project itself will use these tools to curate and annotate a massive sign language dataset, expanding existing annotated BSL corpora by an order of magnitude.

On the recognition side, the project moves beyond isolated sign recognition to dense temporal gloss annotation, i.e. comprehensive time-aligned labelling which includes non-manual and non-lexical material, rather than only representing manual forms from the core lexicon. We combine skeletal uplift, facial landmark tracking, and transformer-based models that are robust enough to handle co-articulation and signer variation. On the production side, the project avoids the traditional use of animated avatars, favouring photorealistic digital humans generated using diffusion models, volumetric capture, and differentiable rendering. Another key innovation is the use of self-supervised learning to build predictive language models directly from sign representations. Rather than relying solely on glossing as a proxy for language, the project is exploring how symbolic representations at the lexical, sub-lexical, and gestural levels can support large-scale modelling of sign structure and meaning.

## 2. The Data Problem

As noted above, a fundamental linguistic challenge addressed by the project is that sign



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 51–60
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

languages do not have a standardised written form: they cannot be reduced to sequences of discrete “words”. As a result, standard Natural Language Processing (NLP) assumptions based on text analysis around linear order, tokenisation, and stable form-meaning mappings do not necessarily hold. Natural continuous signing integrates both lexical and non-lexical elements: just ~60% of spontaneous BSL consists of lexical signs (Fenlon et al., 2014). The remainder consists largely of productive constructions that exploit 3D space, iconicity, and context. These include manual forms such as pointing, depicting signs, and enactment, as well as non-manual features such as brow and eye movements, mouthings, mouth gestures, head movements, and body postures. These elements interact simultaneously rather than sequentially, with meaning distributed across multiple articulators, and are thus not easily symbolised or glossed. The project addresses this by explicitly modelling non-lexical structures, rather than treating them as noise or ignoring them. Linguistic analysis of indexing, depiction, and enactment (following e.g. Ferrara & Hodge, 2018) feeds directly into the translation models, ensuring that these crucial features of spatial grammar and reference tracking unique to sign languages are incorporated.

## 2.1 Current State of the Sign Language Processing (SLP) Field

Interest in the computational modelling of sign languages has accelerated over the last decade, with SLP research increasingly represented at top-tier machine learning conferences (e.g. Camgoz et al., 2018; Fish & Bowden, 2025; Jang et al., 2025; Li et al., 2025; Zuo et al., 2025). Despite the breadth of modelling challenges, research has largely concentrated on the three primary tasks - recognition, production, and translation - typically operating on video or coarse pose data together with gloss annotations or free-text translations (Bragg et al., 2019; Desai et al., 2024; Fox et al., 2025).

Sign-to-text translation is commonly framed as Sign Language Translation (SLT), encompassing both Sign Language Recognition (SLR, meaning sign-to-gloss) and translation to spoken language (either gloss-to-text or sign-to-text). Historically, these components have often been treated as separate modelling problems (Y. Chen, Wei, et al., 2023; Y. Chen, Zuo, et al., 2023; Yin et al., 2021; B. Zhang et al., 2023; H. Zhou et al., 2021, 2022), organised as a pipeline from recognition to translation and frequently without explicit attention to distinctions between lexical and non-lexical structures or broader linguistic organisation (De Coster et al., 2024). More recent work has introduced gloss-free approaches that model sign-to-text mapping end-to-end rather than via gloss intermediates (Asasi et al., 2025; Z. Chen et al., 2025; Kim et al., 2025; Sincan et al., 2025; Wong et al., 2024; B. Zhou et al., 2023). However, both gloss-based and gloss-free paradigms still tend to prioritise mapping coarse manual features to spoken language text, often without modelling the full compositional and multimodal structure of signed communication (De Coster et al., 2024). Text-to-sign research similarly tends to prioritise surface-level appearance and reconstruction, with limited evaluation of linguistic composition (e.g. inclusion of productive constructions) or systematic incorporation of non-manual features (as noted by Bragg et al., 2019). This misunderstanding is reflected in the fact that text-to-sign research is often framed as Sign Language Production or Sign Language Generation, when actually it is translation to the same as extent as sign-to-text, just in the opposite direction.

Given the low-resource nature of sign language data compared to spoken languages, advancing the field requires methodologies and evaluation metrics that explicitly assess the extent to which models capture more complex visual linguistic functions and non-lexical language components. The following sections outline key challenges and research directions for broadening SLP beyond current task formulations towards modelling full, naturalistic, spontaneous, and conversational signing.

## 2.2 Influence and the Use of Available Sign Language Data Repositories

The development of the SLP field has been strongly shaped by the availability and characteristics of existing datasets, which for continuous signing largely consist of interpreted broadcasts (e.g. Albanie et al., 2021; Camgoz et al., 2018; Li et al., 2025), content sourced from social media (e.g. Shi et al., 2022; Tanzer & Zhang, 2024; Uthus et al., 2023) or curated linguistic corpora (e.g. Crasborn & Zwitserlood, 2008; Schembri et al., 2014; Konrad et al., 2020). Because curated corpora are difficult to scale and linguistic annotation is costly, interpreted and online sources are frequently used for training and benchmarking models (De Coster et al., 2024). For most of these benchmarks, annotations are typically limited to weakly aligned captions or gloss-level labels.

Reliance on such data can bias modelling priorities. Interpreted signing produced under live broadcast constraints differs from everyday interaction. Simultaneous interpreting and sight translation are cognitively demanding even for experienced practitioners regardless of modality, and the broadcast medium itself may limit the use of repair strategies typical in natural conversation. Interpreted material is therefore quite likely to contain errors (e.g. omissions, substitutions), flattening of propositional structure, explicitations (overt target text statements or connectives that were merely implied in the source), and the over-use of formulaic sequences (Ding, 2017; Gumul,

2021; Huang et al., 2023; Y. Li & Halverson, 2022; Tang & Li, 2017; Wang, 2025). There may also be a detectable difference between hearing and deaf interpreters with regard to sign language target text design: for example, Stone (2009) suggests that the former are more clearly influenced by source text structure, while the latter are more likely to shape their texts as audience-focused renditions. For non-interpreted spontaneous sign language data, the use of material sourced from the internet raises ethical concerns around consent and copyright, and it usually lacks reliable information about signer proficiency or context (De Coster et al., 2024; Fox et al., 2025). Finally, broadcast captions are not necessarily verbatim representations of speech: principled decisions are routinely made around omissions and simplifications due to limitations of the medium such as available screen estate and reading speeds (Romero-Fresco, 2016, 2020; Schilperoord et al., 2005). As a result, current datasets provide only limited coverage of natural, spontaneous signing, potentially shaping models toward features that diverge from those needed for real-world applications. Within this broader context, the SignGPT project contributes through the VLTK to support scalable annotation of naturalistic signing data and linguistically grounded corpora.

## 2.3 Challenges in the Use of Gloss Annotations

From a modelling perspective, the reliance on glosses as an intermediate representation is both useful and limiting. Polysemy and variation present particular linguistic challenges. A single sign form can have multiple meanings, depending on context. The BSL lexicon also exhibits sociolinguistic variation (Stamp et al., 2014). The project mitigates this by distinguishing contextual glosses from ID glosses, supporting lemmatisation that groups together semantically and phonologically related variants while preserving contextual interpretation (Fenlon et al., 2015). Transformer-based models are then trained to resolve meaning using discourse-level context and non-manual information rather than simply relying on one-to-one sign/word mappings. Translation quality will be evaluated not only via automatic metrics but also through expert human judgement by qualified deaf translators, reflecting linguistic adequacy and naturalness rather than surface similarity.

## 2.4 Constituent Order, Simultaneity, and Linear Modelling Paradigms

The reliance on gloss annotations also highlights a broader issue: the misalignment between simultaneity and organisation in sign languages and the largely linear, sequential modelling paradigms inherited from spoken-language NLP (De Coster et al., 2024). While all natural languages exhibit ordering preferences, grammatical relations in sign languages are generally less dependent on strict linear sequencing and are comparatively flexible (Baker et al., 2016). Consequently, signed grammatical structures may map onto multiple spoken-language sequences, and vice versa, a variability not well captured by traditional n-gram evaluation metrics such as BLEU (Papineni et al., 2002).

Linearised intermediate representations, including gloss strings or caption-aligned tokens, provide practical supervision but introduce an information bottleneck that can obscure spatial grammar, productive constructions, and flexible constituent ordering. Progress in SLP will likely require modelling approaches that move beyond strictly sequential assumptions toward representations capable of capturing multimodality, abstract use of space, cross-articulator temporal alignment, and inductive biases informing productive language use.

## 2.5 Inclusion of Non-lexical Material as Input Features

A related challenge concerns the representation of non-lexical material. This includes manual articulations that will not be found in any dictionary, such as depicting constructions (object, handling, and size-and-shape specifiers); constructed action (also known as role shift), which may or may not co-occur with lexical signs; prosodic features (rhythmic and spatial modulations analogous to stress); and gestures which may not be emblematic or even particularly conventional. All of these linguistic phenomena encode information that is only weakly or coarsely captured in gloss-based annotations, if at all. Non-lexical material also includes non-manual features (signals or gestures conveyed by articulators other than the hands) which can contribute to lexical, grammatical, prosodic, and discourse-level meanings (see Pfau & Quer, 2010 for an overview). Despite the fundamental role of non-manuals, many pose-based SLP approaches have neglected them and focused primarily on manual pose (De Coster et al., 2024). However, although still relatively limited, work incorporating non-manual features into model inputs has demonstrated improved performance (Dey et al., 2022; Miranda et al., 2022; Zheng et al., 2021).

From a modelling perspective, this necessitates multimodal representations that extend beyond handshape, pose trajectories, or low-information RGB features to include non-manual articulators, spatial relationships, and dependency patterns. Many grammatical functions, including directionality, clause type marking and adverbial modification, are conveyed through spatial and articulatory modulation rather than discrete lexical items, while constructions such as depicting signs encode meaning through the interaction of handshape, movement, and space (Baker et al., 2016). Without these dimensions, models risk privileging lexical recognition over broader linguistic competence.

### 2.6 Towards Real-world Generalisation

To improve real-world applicability, SLP must move towards greater use of fluent deaf conversational data, deaf-presented content, and ethically sourced naturalistic signing, supported by scalable annotation tools such as the VLTK. More broadly, progress depends not only on dataset scale but on modelling and evaluation approaches that reflect the multimodal, spatial, and partially non-lexical structure of sign language. Within this context, the SignGPT project develops a conversational agent incorporating grounded translation pipelines, both lexicalised and productive language structures, and attention to linguistic diversity. Addressing these factors is essential for systems that generalise beyond constrained benchmarks toward real-world interaction, motivating the requirements outlined in the following section.

## 3. Visual Language Toolkit (VLTK)

The accessibility of tools for machine translation and processing of sign language remains limited, with users facing multiple technical barriers when attempting to deploy existing methods for downstream applications. These barriers include poorly documented code, large and fragile dependency chains, dataset-specific implementations, and computational requirements that place many approaches out of reach for researchers without access to high-performance infrastructure. Compounding this, the majority of research in this area is driven primarily by publication incentives, and consequently few studies have released their methodologies in forms that are readily usable by the broader communities who stand to benefit, including linguists, deaf communities, educators, and researchers in adjacent fields.

The VLTK is designed to address these limitations, providing a modular suite of software tools for the annotation, analysis, modelling, and generation of visual languages, with an initial focus on sign languages. Developed within the SignGPT programme and associated research funded by Google.org, the VLTK will provide the technical infrastructure needed to bridge continuous sign language video, linguistic representations, and large-scale machine learning models. This effort is inspired in part by toolkits that have proven highly successful in adjacent fields, such as NLTK for natural language processing (Bird, 2006) and SpeechBrain for speech recognition (Ravanelli et al., 2024), which demonstrate the value of unifying complex ML pipelines into accessible, well-documented frameworks.

At its core, the VLTK will integrate tools for automatic and semi-automatic annotation of sign language video, encompassing dense temporal gloss annotation, sign spotting, phrase and sentence segmentation, signer and video anonymisation, skeletal estimation and correction, and sign-text alignment. The toolkit will additionally provide a dedicated Python library and API for accessing datasets, extracting features from state-of-the-art representation models, and computing a unified set of validated metrics for evaluating translation and production results. A central design principle is interoperability with established linguistic annotation platforms such as ELAN, enabling linguists and sign language researchers to work with this data at scale without necessarily adopting additional software. Early development has focused on integrating state-of-the-art methods for sign segmentation and sign spotting, described in the following subsections; other tools still under development are looking at methods for tracking non-lexical grammatical phenomena such as indexing and other forms of deictic reference (Ranum et al., 2026).

### 3.1 Segmentation Tool

Segmentation is concerned with the identification of temporal boundaries within continuous signing. This is a challenging problem both for humans and automated systems since no single cue can reliably predict boundary locations. For example, boundaries can be signalled by multiple combinations of manual and non-manual markers. Cues relevant to segmentation at the lexical level might include phonological and prosodic transitions (Brentari, 1998); at the intonational phrase level they might include eye blinks, pauses, changes in head position and eye gaze direction, and/or shifts in body posture (Ormel & Crasborn, 2012). Perception studies have shown that while signers are able to segment discourse more consistently than non-signers, non-signers do demonstrate some sensitivity to visually salient cues such as pauses and hand drops, suggesting that at least some boundary markers are accessible independently of linguistic knowledge (Brentari et al., 2011; Fenlon et al., 2007). Early work on automatic boundary detection in signed languages has explored video-based feature extraction of facial and manual cues (Jantunen et al., 2010; Piater et al., 2010), though robust automatic segmentation remains an open challenge given the multi-articulatory nature of sign and the absence of any single dominant predictor (Ormel & Crasborn, 2012). Despite these difficulties, more recent work has demonstrated that machine learning systems can predict human-annotated boundaries on certain datasets with up to 85% accuracy using skeletal pose and hand data alone (He et al., 2025).

Within the VLTK, we incorporate the state-of-the-art segmentation approach presented by He et al. (2025), which annotates frames with Begin, In, and Out (BIO) labels and employs a self-attention mechanism to learn, at each timestep, how to optimally combine body and hand features for

boundary prediction. Figure 1 illustrates representative outputs on a sample from the MEINE DGS dataset (Konrad et al., 2020), showing predictions and ground-truth, including successful segmentations alongside cases of over- and under-segmentation: the method performs well considering the ambiguous nature of signing boundaries. The VLTK simplifies this pipeline so that users can upload video directly and receive ELAN-compatible annotation files with proposed boundaries for review. A browser-based visualisation tool is also provided for interactive inspection of segmentation outputs.

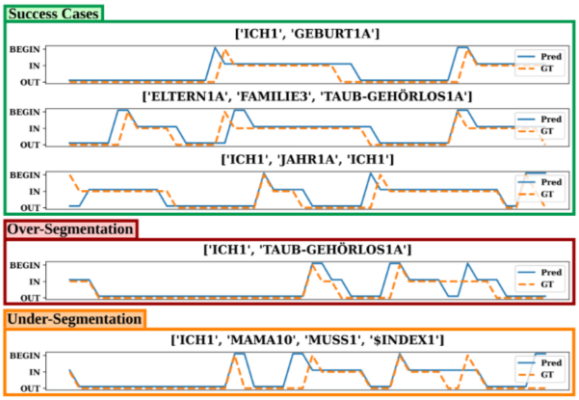



Figure 1: Examples of segmentation accuracy on a sample from the MEINE DGS dataset

## 3.2 Sign Spotting Tool

Sign spotting is the task of identifying specific lexical signs or other annotations within continuous signing by comparing sequences against a library of isolated examples drawn from lexical databases or dictionaries. By grounding spotted instances in verified reference examples, this process can be used to expand the coverage and diversity of sign dictionaries or to increase the annotation density of existing corpora.

The VLTK provides an interface and API for sign spotting built upon SignRep (Wong et al., 2025) , a state-of-the-art sign language representation model. SignRep is a masked autoencoder trained to reconstruct skeletal pose estimations from video. Through this reconstruction objective, the model learns general-purpose representations of sign segments within a compressed latent space that can be compared using cosine similarity or other distance metrics. Crucially, because the network operates on reconstructed skeletal features rather than raw RGB input, the resulting embeddings are substantially less sensitive to signer appearance, mitigating a well-known source of bias in sign recognition systems.

The spotting tool is exposed via an API that enables comparison of these learned embeddings against multiple reference dictionaries. Results can be exported in ELAN format, facilitating evaluation of inter-annotator agreement or analysis of production variation across signers. Importantly, the toolkit is designed so that the segmentation pipeline described in Section 3.1 can be composed directly with the spotting module: users can first extract temporal segments from continuous video, then perform spotting over those segments using pre-extracted dictionary embeddings provided as part of the toolkit. Figure 2 illustrates this combined pipeline on an example sequence. First, continuous signs are segmented using the method described in section 3.1. Then we embed these segments as representations using the method described above and compare with the most likely sign candidates in a dictionary of isolated signs.

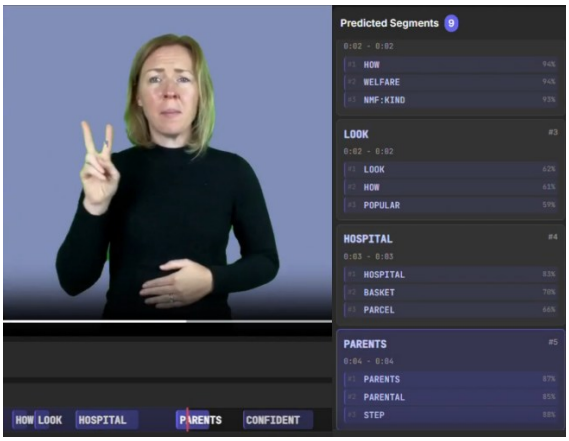



Figure 2: The sign segmenter and sign spotter interface

## 3.3 Other Tools

Additional tools within the VLTK leverage pre-trained models for 3D body pose estimation (Loper et al., 2015) and hand reconstruction (Pavlakos et al., 2024) to generate high-quality mesh representations of signers from video. These reconstructions serve multiple purposes within the toolkit, from feature extraction and linguistic analysis to appearance transfer and signer anonymisation.

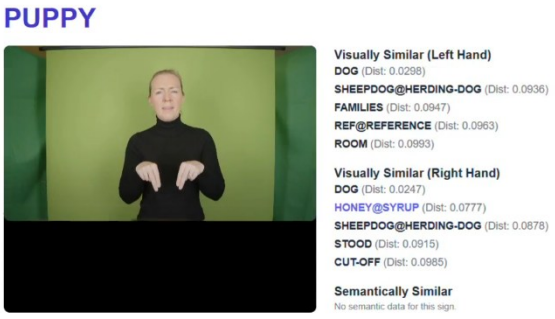



Figure 3: Example of a heuristic method for comparing hand shapes extracted using the HaMER framework (Pavlakos et al., 2024)

The extracted 3D hand pose keypoints can be used independently for linguistic analysis. In Figure 3, we demonstrate an experiment in which isolated hand keypoints, without spatial or temporal context, are compared against a library

of isolated sign examples to identify hand shape similarities across signs. The results reveal expected correspondences, such as between *puppy* and *dog*, but also surface less obvious relationships: for instance, *honey* appears among the nearest neighbours because, in BSL, the dominant hand of this sign employs the same handshape. This capability has direct applications in lexicographic analysis and dictionary construction.

Figure 4 illustrates the mesh reconstruction pipeline: keypoint and pose estimations are extracted for various articulators (body skeleton, hand keypoints, head pose) using multiple methods according to their relative limitations and strengths. Building on these reconstructions, Figure 5 demonstrates how SMPL body meshes (Loper et al., 2015) can be extracted from signing video and remapped to novel identities using Gaussian splatting (D. Zhang et al., 2025). First, the signer's appearance is separated from the underlying pose, normalised, and then this can be remapped to another identity. Signs from isolated dictionaries can then be stitched together to form smooth sequences of signs using techniques such as flow-matching (Lipman et al., 2023). This remapping has direct applications in signer anonymisation as well as in driving human avatars for sign language production.

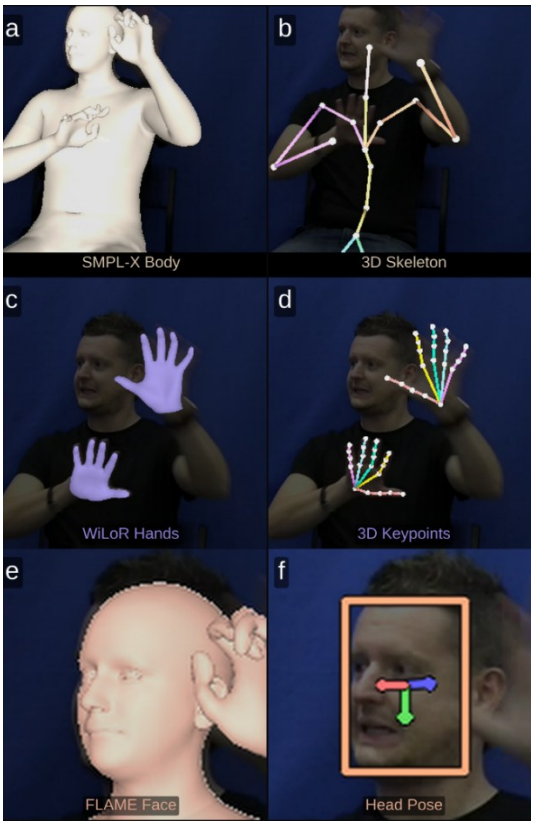



Figure 4: Example of extraction pipeline via SMPL-X (Pavlakos et al., 2019) with 3D keypoints for hands via WiLoR (Potamias et al., 2024) and head pose via FLAME (T. Li et al., 2017)

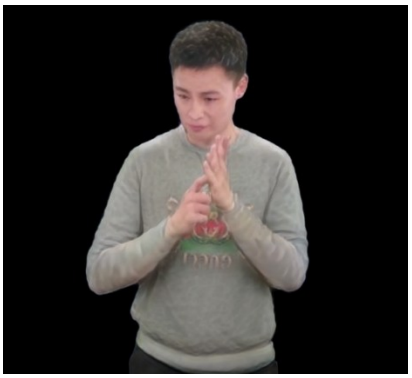


Figure 5: Example of “re-skinning” multiple different signers into a continuous sequence with a new appearance using SMPL and Guava

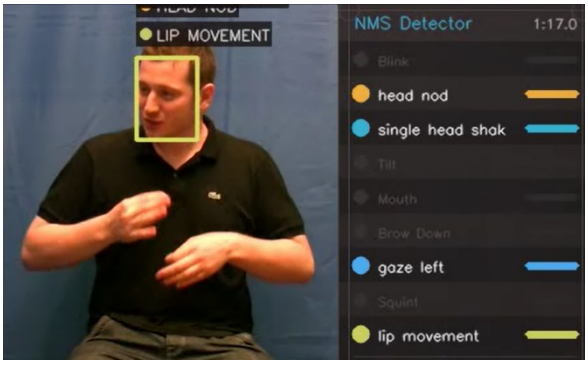



Figure 6: Example of predicted non-manual articulations based on keypoints extracted from the method described in Liu et al. (2025)

Additional tools include heuristic based algorithms that use 3D face estimation features extracted using the method “Teaser” (Liu et al., 2025) to predict when a signer is blinking, tilting their head, nodding, shaking their head, along with methods for gaze detection and eye-brow movement. As shown in Figure 6, these can be visualised over the videos or returned as ELAN files to the user.

As research in 3D reconstruction and visual representation continues to advance, the VLTK will incorporate and standardise emerging methods within its modular architecture, following the development framework outlined by Ravanelli et al. (Ravanelli et al., 2024).

### 3.4 Uses for the VLTK

Beyond annotation, the VLTK aims to support the full sign language processing pipeline. It will incorporate technologies for continuous translation between signed and spoken languages, providing interfaces between visual representations, symbolic linguistic units, and large language models. Although developed primarily for sign languages, the toolkit is explicitly designed to generalise to other forms of visual and multimodal communication, including co-speech gesture in interaction.

The VLTK is intended both as a research platform and as a community-facing resource.

Components of the toolkit are to be released as open-source software to support reuse by the wider research community, while selected capabilities will be exposed through web-based demonstrations and applications. These may include dictionary search tools, sign language learning and assessment tools, signer personalisation, and conversational interfaces that connect sign language input and output to large language models. We hope that through this dual role, the VLTK will underpin advances in sign linguistics, computer vision, and machine learning, while also enabling practical, user-driven applications co-created with deaf communities (Desai et al., 2024; Fox et al., 2025).

## 4. Summary

We have presented SignGPT's Visual Language Toolkit project as a transformative step towards genuinely accessible sign language AI, directly addressing the linguistic and technological barriers that have historically limited progress in this field. Rather than treating sign languages as a simple sequence of glosses or as representations of spoken language, the work recognises their multimodal, spatial, and simultaneous nature with its approach to handling non-lexical signs, non-manual features, and other forms of context. By developing robust semi-automatic glossing tools that support large scale data expansion, as well as taking a community-embedded design and evaluation approach to transformer-based modelling, we aim to overcome the bottlenecks of data scarcity and linguistic over-simplification. Through this integration of linguistic insight and machine learning innovation, SignGPT establishes a foundation for robust and extensible sign language models capable of supporting real-world communication needs while respecting the linguistic integrity of BSL and other sign languages.

## 5. Acknowledgements

This work is supported by the UK Engineering and Physical Sciences Research Council [EP/Z535370/1], and by funding from Google.org via the AI for Global Goals scheme. We also acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

## 6. Bibliographical References

Albanie, S., Varol, G., Momeni, L., Bull, H., Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R., McParland, A., & Zisserman, A. (2021). *BBC-Oxford British Sign Language Dataset*. https://arxiv.org/pdf/2111.03635

Asasi, S., Lakhal, M. I., Sincan, O. M., & Bowden, R. (2025). *Beyond Gloss: A Hand-Centric Framework for Gloss-Free Sign Language Translation*. http://arxiv.org/abs/2507.23575

Baker, A., van den Bogaerde, B., Pfau, R., & Schermer, G. M. (Eds.). (2016). *The linguistics of sign languages: an introduction*. John Benjamins Publishing Company.

Bird, S. (2006). NLTK: The Natural Language Toolkit. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 69–72. https://doi.org/10.3115/1225403.1225421

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., & Morris, M. R. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. *ASSETS 2019 - 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 16–31. https://doi.org/10.1145/3308561.3353774

Brentari, D. (1998). *A prosodic model of sign language phonology*. MIT Press.

Brentari, D., González, C., Seidl, A., & Wilbur, R. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech*, *54*(1), 49–72. https://doi.org/10.1177/0023830910388011

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural Sign Language Translation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7784–7793. https://doi.org/10.1109/CVPR.2018.00812

Chen, Y., Wei, F., Sun, X., Wu, Z., & Lin, S. (2023). A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*, 5110–5120. http://arxiv.org/abs/2203.04287

Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., & Mak, B. (2023). Two-Stream Network for Sign Language Recognition and Translation. *Advances in Neural Information Processing Systems*, *35*. http://arxiv.org/abs/2211.01367

Chen, Z., Zhou, B., Huang, Y., Wan, J., Hu, Y., Shi, H., Liang, Y., Lei, Z., & Zhang, D. (2025). C2RL: Content and Context Representation Learning for Gloss-Free Sign Language Translation and Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *35*(9), 8533–8544. https://doi.org/10.1109/TCSVT.2025.3553052

De Coster, M., Shterionov, D., Van Herreweghe, M., & Dambre, J. (2024). Machine translation from signed to spoken languages: state of the art and challenges. *Universal Access in the Information Society*, *23*(3), 1305–1331. https://doi.org/10.1007/s10209-023-00992-1

Desai, A., De Meulder, M., Hochgesang, J. A., Kocab, A., & Lu, A. X. (2024). Systemic Biases

in Sign Language AI Research: A Deaf-Led Call to Reevaluate Research Agendas. *11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, Sign-Lang@LREC-COLING 2024*, 54–65. https://arxiv.org/pdf/2403.02563

Dey, S., Pal, A., Chaabani, C., & Koller, O. (2022). Clean Text and Full-Body Transformer: Microsoft's Submission to the WMT22 Shared Task on Sign Language Translation. *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 969–976. https://doi.org/10.18653/v1/2022.wmt-1.93

Ding, Y. (2017). Using Propositional Analysis to Assess Interpreting Quality. *International Journal of Interpreter Education*, *9*(1). https://open.clemson.edu/ijie/vol9/iss1/4

Fenlon, J., Cormier, K., & Schembri, A. (2015). Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography*, *28*(2), 169–206. https://doi.org/10.1093/ijl/ecv008

Fenlon, J., Denmark, T., Campbell, R., & Woll, B. (2007). Seeing sentence boundaries. *Sign Language & Linguistics*, *10*(2), 177–200. https://doi.org/https://doi.org/10.1075/sll.10.2.06fen

Fenlon, J., Schembri, A., Rentelis, R., Vinson, D., & Cormier, K. (2014). Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua*, *143*, 187–202. https://doi.org/10.1016/j.lingua.2014.02.003

Ferrara, L., & Hodge, G. (2018). Language as description, indication, and depiction. *Frontiers in Psychology*, *9*(MAY), 329842. https://doi.org/10.3389/fpsyg.2018.00716

Fish, E., & Bowden, R. (2025, October 28). Geo-Sign: Hyperbolic Contrastive Regularisation for Geometrically Aware Sign Language Translation. *NeurIPS 2025*. http://arxiv.org/abs/2506.00129

Fox, N., Woll, B., & Cormier, K. (2025). Best practices for sign language technology research. *Universal Access in the Information Society*, *24*(1), 69–77. https://doi.org/10.1007/s10209-023-01039-1

Gumul, E. (2021). Explicitation and cognitive load in simultaneous interpreting. *Interpreting. International Journal of Research and Practice in Interpreting*, *23*(1), 45–75. https://doi.org/10.1075/intp.00051.gum

He, L. J., Walsh, H., Sincan, O. M., & Bowden, R. (2025, August 20). Hands-On: Segmenting Individual Signs from Continuous Sequences. *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition, FG 2025*. https://doi.org/10.1109/FG61629.2025.11099255

Huang, D. F., Li, F., & Guo, H. (2023). Chunking in simultaneous interpreting: the impact of task complexity and translation directionality on lexical bundles. *Frontiers in Psychology*, *14*, 1252238. https://doi.org/10.3389/fpsyg.2023.1252238

Jang, Y., Raajesh, H., Momeni, L., Varol, G., & Zisserman, A. (2025). *Lost in Translation, Found in Context: Sign Language Translation with Contextual Cues*. 8742–8752. https://doi.org/10.1109/cvpr52734.2025.00817

Jantunen, T., Koskela, M., Laaksonen, J., & Rainò, P. (2010). Towards the automated visualization and analysis of signed language motion - Method and linguistic issues. *Proceedings of the International Conference on Speech Prosody*. https://doi.org/10.21437/speechprosody.2010-103

Kim, J., Jeon, H., Bae, J., & Kim, H. Y. (2025). Leveraging the Power of MLLMs for Gloss-Free Sign Language Translation. *ICCV 2025*, 21048–21058. https://arxiv.org/abs/2411.16789v1

Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, *36*(6), 17. https://doi.org/10.1145/3130800.3130813

Li, Y., & Halverson, S. L. (2022). Lexical bundles in formulaic interpreting: A corpus-based descriptive exploration. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, *19*(1), 33–56. https://doi.org/10.1075/TIS.19037.LI

Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., & Li, H. (2025). Uni-Sign: Toward Unified Sign Language Understanding At Scale. *13th International Conference on Learning Representations, ICLR 2025*, 1354–1373. http://arxiv.org/abs/2501.15187

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. (2023). Flow Matching for Generative Modeling. *11th International Conference on Learning Representations, ICLR 2023*. http://arxiv.org/abs/2210.02747

Liu, Y., Zhu, L., Lin, L., Zhu, Y., Zhang, A., & Li, Y. (2025). TEASER: Token Enhanced Spatial Modeling for Expressions Reconstruction. *13th International Conference on Learning Representations, ICLR 2025*, 69084–69103. http://arxiv.org/abs/2502.10982

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, *34*(6). https://doi.org/10.1145/2816795.2818013

Miranda, P. B. C., Casadei, V., Silva, E., Silva, J., Alves, M., Severo, M., & Freitas, J. P. (2022). TSPNet-HF: A Hand/Face TSPNet Method for Sign Language Translation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13788 LNAI*, 305–316. https://doi.org/10.1007/978-3-031-22419-5_26

Ormel, E., & Crasborn, O. (2012). Prosodic Correlates of Sentences in Signed Languages. *Sign Language Studies*, *12*(2), 279–341. http://www.jstor.org/stable/26190836

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311–318. https://doi.org/10.3115/1073083.1073135

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., & Black, M. J. (2019). Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 10967–10977. https://doi.org/10.1109/CVPR.2019.01123

Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., & Malik, J. (2024). Reconstructing Hands in 3D with Transformers. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 9826–9836. https://doi.org/10.1109/CVPR52733.2024.00938

Pfau, R., & Quer, J. (2010). Nonmanuals: Their grammatical and prosodic roles. In D. Brentari (Ed.), *Sign Languages* (pp. 381–402). Cambridge University Press. https://doi.org/10.1017/CBO9780511712203.018

Piater, J., Hoyoux, T., & Du, W. (2010). Video Analysis for Continuous Sign Language Recognition. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz, & A. Schembri (Eds.), *Language Resources And Evaluation* (pp. 1–4). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/10052.html

Potamias, R. A., Zhang, J., Deng, J., & Zafeiriou, S. (2024). *WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild*. 12242–12254. https://doi.org/10.1109/cvpr52734.2025.01143

Ranum, O., Hadfield, S., & Bowden, R. (2026). *What's the Point? Spatial Grammar & Index Resolution for Sign Language Processing [Manuscript in preparation]*. Centre for Vision, Speech and Signal Processing, University of Surrey.

Ravanelli, M., Parcollet, T., Moumen, A., de Langen, S., Subakan, C., Plantinga, P., Wang, Y., Mousavi, P., Della Libera, L., Ploujnikov, A., Paissan, F., Borra, D., Zaiem, S., Zhao, Z., Zhang, S., Karakasidis, G., Yeh, S.-L., Champion, P., Rouhe, A., Braun, R., Mai, F., Zuluaga-Gomez, J., Mahed Mousavi, S., Nautsch, A., Nguyen, H., Liu, X., Sagar, S., Duret, J., Mdhaffar, S., Laperriere, G., Rouvier, M., De Mori, R., Esteve, Y. (2024). Open-Source Conversational AI with SpeechBrain 1.0. *Journal of Machine Learning Research*, *25*. http://arxiv.org/abs/2407.00463

Romero-Fresco, P. (2016). Accessing communication: The quality of live subtitles in the UK. *Language & Communication*, *49*(4), 56–69. https://doi.org/10.1016/j.langcom.2016.06.001

Romero-Fresco, P. (2020). Negotiating quality assessment in media accessibility: the case of live subtitling. *Universal Access in the Information Society 2020 20:4*, *20*(4), 741–751. https://doi.org/10.1007/s10209-020-00735-6

Schilperoord, J., de Groot, V., & van Son, N. (2005). Nonverbatim captioning in Dutch television programs: A text linguistic approach. *Journal of Deaf Studies and Deaf Education*, *10*(4), 402–416. https://doi.org/10.1093/deafed/eni038

Shi, B., Brentari, D., Shakhnarovich, G., & Livescu, K. (2022). Open-Domain Sign Language Translation Learned from Online Video. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 6365–6379. http://arxiv.org/abs/2205.12870

Sincan, O. M., Low, J. H., Asasi, S., & Bowden, R. (2025). Gloss-free Sign Language Translation: An unbiased evaluation of progress in the field. *Computer Vision and Image Understanding*, *261*, 104498. https://doi.org/10.1016/j.cviu.2025.104498

Stamp, R., Schembri, A., Fenlon, J., Rentelis, R., Woll, B., & Cormier, K. (2014). Lexical Variation and Change in British Sign Language. *PLOS ONE*, *9*(4), e94053. https://doi.org/10.1371/JOURNAL.PONE.0094053

Stone, C. (2009). *Toward a Deaf Translation Norm*. Gallaudet University Press. https://doi.org/10.2307/j.ctv2rcng24

Tang, F., & Li, D. (2017). A corpus-based investigation of explicitation patterns between professional and student interpreters in Chinese-English consecutive interpreting. *Interpreter and Translator Trainer*, *11*(4), 373–395. https://doi.org/10.1080/1750399X.2017.1379647

Tanzer, G., & Zhang, B. (2024). *YouTube-SL-25: A Large-Scale, Open-Domain Multilingual Sign Language Parallel Corpus*. http://arxiv.org/abs/2407.11144

Uthus, D., Tanzer, G., & Georg, M. (2023). YouTube-ASL: A Large-Scale, Open-Domain American Sign Language-English Parallel Corpus. *Advances in Neural Information Processing Systems*, *36*.

Wang, J. (2025). Strategic additions in simultaneous interpreting from a signed language into a spoken language. *Translation and Interpreting Studies*, *20*(1), 24–49. https://doi.org/10.1075/tis.23053.wan

Wong, R., Camgoz, N. C., & Bowden, R. (2024, May 7). Sign2GPT: Leveraging Large

Language Models for Gloss-Free Sign Language Translation. *12th International Conference on Learning Representations, ICLR 2024*. http://arxiv.org/abs/2405.04164

Wong, R., Camgoz, N. C., & Bowden, R. (2025). *SignRep: Enhancing Self-Supervised Sign Representations*. http://arxiv.org/abs/2503.08529

Yin, A., Zhao, Z., Liu, J., Jin, W., Zhang, M., Zeng, X., & He, X. (2021). SimulSLT: End-to-End Simultaneous Sign Language Translation. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, *10*(21), 4118–4127. https://doi.org/10.1145/3474085.3475544

Zhang, B., Müller, M., & Sennrich, R. (2023, May 2). SLTUNET: A Simple Unified Model for Sign Language Translation. *11th International Conference on Learning Representations, ICLR 2023*. http://arxiv.org/abs/2305.01778

Zhang, D., Liu, Y., Lin, L., Zhu, Y., Li, Y., Qin, M., Li, Y., & Wang, H. (2025, August 1). GUAVA: Generalizable Upper Body 3D Gaussian Avatar. *ICCV 2025*. http://arxiv.org/abs/2505.03351

Zheng, J., Chen, Y., Wu, C., Shi, X., & Kamal, S. M. (2021). Enhancing Neural Sign Language Translation by highlighting the facial expression information. *Neurocomputing*, *464*, 462–472. https://doi.org/10.1016/j.neucom.2021.08.079

Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., & Zhang, D. (2023). Gloss-free Sign Language Translation: Improving from Visual-Language Pretraining. *Proceedings of the IEEE International Conference on Computer Vision*, 20814–20824. https://doi.org/10.1109/ICCV51070.2023.01908

Zhou, H., Zhou, W., Qi, W., Pu, J., & Li, H. (2021). Improving Sign Language Translation with Monolingual Data by Sign Back-Translation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1316–1325. https://doi.org/10.1109/CVPR46437.2021.00137

Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2022). Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation. *IEEE Transactions on Multimedia*, *24*, 768–779. https://doi.org/10.1109/TMM.2021.3059098

Zuo, R., Potamias, R. A., Ververas, E., Deng, J., & Zafeiriou, S. (2025). *Signs as Tokens: An Autoregressive Multilingual Sign Language Generator*. http://arxiv.org/abs/2411.17799

## 7. Language Resource References

Crasborn, O., & Zwitserlood, I. (2008). The Corpus NGT: An online corpus for professionals and laymen. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, & E. Thoutenhoofd (Eds.), Construction and exploitation of sign language corpora: 3rd Workshop on the Representation and Processing of Sign Languages (pp. 44–49). ELDA, Paris.

Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., Böse, O., Jahn, E., & Schulder, M. (2020). MEINE DGS — annotiert: Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release. Universität Hamburg. https://doi.org/10.25592/dgs.corpus-3.0

Schembri, A., Fenlon, J., Rentelis, R., & Cormier, K. (2014). British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2014 (Second Edition). University College London. http://www.bslcorpusproject.org

# Processing Kinematics of Nonmanual Markers in R

**Jan Bulla, Vadim Kimmelman**
University of Bergen
Bergen, Norway
{jan.bulla, vadim.kimmelman}@uib.no

**Abstract**

Nonmanual markers, such as head and eyebrow movements, eye blinks, and mouth shapes, are an important part of natural languages, both spoken and signed. Recent developments in computer vision have made it possible to extract facial and body landmark positions, as well as head-rotation measures, from 2D video recordings, which can be further processed to analyse the kinematics of nonmanual articulators. In this paper, we present an R-based workflow for processing raw outputs of computer vision toolkits with the goal of producing reliable and interpretable kinematic measurements of nonmanual articulators.

**Keywords:** nonmanual markers, kinematics, peak detection, smoothing

## 1. Introduction

Natural languages (both signed and spoken) employ nonmanual articulators (the head, the body, and facial features) to produce linguistically meaningful and relevant information. For sign languages, nonmanual markers have been shown to play a crucial role at all levels of language, including phonological, morphological, lexical, syntactic, semantic, and prosodic functions (Pfau and Quer, 2010). For spoken languages, co-speech gestures also include an important nonmanual component, such as headshakes, nods, tilts, eye blinks, and eyebrow movements, which are used to express meaning and regulate conversations (see e.g. Wagner et al., 2014; Nota et al., 2021).

Until a few years ago, measuring nonmanual markers in signed and spoken languages involved using a motion-capture system (Hadar et al., 1983, 1985; Wagner et al., 2014; Puupponen et al., 2015). Motion-capture systems require expensive equipment, and output large amounts of data which require extensive post-processing and statistical analysis (Puupponen et al., 2015), which, in practice, limits the length of recordings that can be made and analysed using this approach. Furthermore, most existing datasets of natural sign languages consist of 2D video recordings.

However, recently computer vision tools such as OpenPose (Cao et al., 2018), OpenFace (Baltrusaitis et al., 2018), and MediaPipe (Lugaresi et al., 2019) have made it possible to extract facial and body landmarks and head-rotation estimates from 2D video, enabling analysis of the kinematics of head movements (and other manual and nonmanual movements). Several papers describing methods for using these computer vision tools for measuring kinematics have been published, both for sign languages (Börstell, 2023) and co-speech gestures (Trujillo et al., 2019). The EnvisionBox project (https://envisionbox.org) is an especially useful resource, producing tutorials on leveraging various computer vision tools for linguistic kinematic analysis, see specifically Trujillo and Pouw (2021); Pouw (2024).

While some studies have also been published using these tools specifically for analysing nonmanual markers (Kimmelman et al., 2020; Bauer et al., 2024), several methodological challenges remain. Specifically, we identify:

**Dealing with noisy measurements**. Any measurement instrument produces noise, and this is especially true for computer vision tools that are not developed specifically for kinematic analysis of sign languages or gestures. Consequently, it is necessary to separate the noise from the relevant motion in the outputs of these models. For instance, in Figure 1 we can compare the raw outputs of OpenFace yaw measurements for a single headshake in Brazilian Sign Language (Libras) with an inferred smooth motion of the head.

**Peak detection**. Many nonmanual movements are periodic (such as headshakes), or have relevant peaks and plateaus that researchers need to locate and measure (such as head nods). For instance, in Figure 1 the direct interpretation of local minima and maxima as peaks produces a large number of peaks that do not reflect real head turns.

**Calculating measurements**. Numerous kinematic measures can be calculated for any nonmanual marker (e.g. range of motion, velocity, frequency, etc.). Furthermore, these measures can be summarized in different ways for individual motion events. Figure 1 shows that the measures derived from the data crucially depend on the smoothing and peak detection steps. For example, range of motion in the raw data equals 8.7 degrees, and in the smoothed data it is 6.2 degrees. In addition, the number of peaks – and thus the derived peak frequency – differs markedly.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 61–70
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

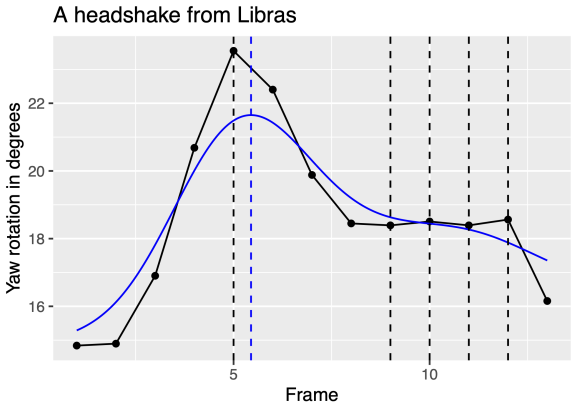



Figure 1: Visualization of yaw (in degrees) for a single headshake from Libras, see https://osf.io/krejq for the video. Black dots and line: raw measurement; blue line: inferred movement after smoothing using kernel regression; black dashed lines: raw local maxima and minima; blue dashed line: inferred peak after smoothing

In this paper, we present a solution addressing these challenges, implemented as a small set of R scripts. These are openly available at https://doi.org/10.17605/OSF.IO/QFG98, together with detailed documentation, a demo script, and the data used for our analyses. The demo script introduces the functions presented in Section 2 and may serve as a template for adapting the workflow to other datasets. The scripts require a table with raw outputs of measurement instruments as input. First, they apply smoothing to address the noisy measurement problem; next, they apply a peak-detection algorithm that identifies and classifies peaks into various categories; finally, they produce a large number of kinematic measures for each motion event in the original dataset. We have tested this approach using outputs of OpenFace (Baltrusaitis et al., 2018) and MediaPipe (Lugaresi et al., 2019), but it is compatible with any instrument outputting per-frame measurements of an articulator.

In Section 2 we present our scripts and briefly explain the underlying statistical theory of the developed functions. In Section 3 we present a case study applying this approach to the study of kinematics of negative headshakes in Libras. Section 4 discusses limitations of the approach and concludes the paper.

## 2. Presentation of the Script

This section introduces the three R (R Core Team, 2025) functions `smooth.curves.est`, `smooth.curves.classif`, and `smooth.curves.res.prep` that together allow users to analyse head-movement time-series data. The functions are applied sequentially as follows.

**Step 1**: The function `smooth.curves.est` estimates a smoothing curve for a given sequence of observations and identifies the resulting peaks. It provides a graphical display of the fitted curve, the peak locations, and (optionally) the first and second derivatives of the smoothing curve.

**Step 2**: The function `smooth.curves.classif` classifies the inferred peaks from Step 1 according to user-defined criteria. It can also display the classification graphically (optional). In addition, `smooth.curves.classif` generates various quantitative measures based on the inferred peaks as well as basic statistical measures of the underlying data.

**Step 3**: The function `smooth.curves.res.prep` consolidates the output from the function `smooth.curves.classif` for one or several sequences of observations into a format that is more suitable for downstream analysis and visualisation.

Note that the computationally most demanding components occur in Step 1. In the following, we provide a description of each function, including a brief summary of the statistical foundations required for Step 1.

### 2.1. Smoothing

The first step carries out the estimation of smoothing curves for a set of input values provided by the user. In principle, these input values can consist of a comparatively general set of pairs of $x$- and $y$-values. In the application considered, we assume that the $x$-values correspond to time index measured in frames, and the $y$-values measure an angle of movement or a distance between articulators (e.g. the eyebrows and the eye line). We implemented two different approaches for estimating the smoothing curve: kernel regression and smoothing splines.

Kernel regression belongs to the class of non-parametric statistical methods; the underlying theory for continuous data dates back to Nadaraya (1965) and Watson (1964). It estimates the relationship between variables by calculating a weighted average of nearby data points, using a so-called kernel function to determine the weights. This allows kernel regression to model complex, non-linear patterns without assuming a specific functional form, thus constituting a more flexible alternative to linear regression for smoothing data and making predictions. Roughly speaking, kernel regression is similar to a moving average but uses data-driven weighting by placing more emphasis on points closer to the prediction location. The weighting process is

controlled by the crucial bandwidth parameter (usually denoted by $h$). What follows is a brief technical formulation of the kernel regression estimator. Let $m(x) = \mathbb{E}[Y \mid X = x]$ denote the (unknown) regression function. A standard formulation is the Nadaraya-Watson (local constant) estimator, which at a target point $x$ takes the form

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\sum_{i=1}^{n} K_h(x - X_i)} = \sum_{i=1}^{n} w_i(x; h) Y_i,$$

$$\text{with } w_i(x; h) = \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)}$$

$$\text{and } K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right).$$

Here, the kernel $K$ is a weight function and $w_i(x; h)$ are the normalised kernel weights. The latter are non-negative, sum to one, and allocate larger weight to observations $X_i$ closer to $x$ (as determined by the kernel $K$ and bandwidth $h > 0$). Then, $m(x)$ corresponds to the kernel estimator, i.e., the estimated conditional mean of $Y$ at $x$. The bandwidth parameter $h > 0$ controls how rapidly weights decay with distance from $x$. This parameter determines the degree of smoothing and controls what is known as the bias-variance trade-off. More precisely, the bandwidth $h$ determines how far the kernel's influence extends, thus balancing local data fitting (small $h$ leading to low bias and high variance) with overall smoothness (large $h$ leading to high bias and low variance). The aim of bandwidth selection is to find an optimal balance of bias and variance for the regression curve.

For carrying out kernel regression in R, we rely on the function `npreg` from the package `np` (Hayfield and Racine, 2008). Bandwidth selection is implemented in the function `npregbw`, which allows choosing between a local linear and a local constant (i.e. Nadaraya-Watson) estimator. Moreover, `npregbw` carries out cross-validation to assess the optimal bandwidth via two criteria: Kullback-Leibler or least-squares. We refer the reader to Li and Racine (2004) and Racine and Li (2004) and the references therein for further details on the theoretical background of both the available estimators and bandwidth selection.

For our data, we observed that both cross-validation criteria led to relatively strong smoothing, which turned out to be slightly weaker for the Kullback-Leibler criterion. Therefore, we selected this criterion for our bandwidth selection. Regarding the choice of estimator, we could not determine a clear preference between the local linear or the local constant version: depending on the data analysed, sometimes one or the other showed favourable results. Nevertheless, we observed that occasionally one, sometimes both estimators resulted in more or less severe over-smoothing (i.e., $h$ tends towards infinity). Hence, the function `smooth.curves.est` internally carries out bandwidth selection with both estimators and selects the smaller of the two bandwidths. This constitutes a pragmatic approach, because we did not observe under-smoothing in any of our test datasets. Finally, it is noteworthy that the function `npregbw` possesses several additional arguments that enable fine-tuning of the bandwidth selection process. For our applications, we did not find any relevant effect of the kernel type and therefore relied on the default Gaussian kernel. Moreover, we set the number of sets of random initial points for starting the process of finding extrema of the cross-validation function to 25 for numerical stability – although a lower value might suffice if computational power is scarce.

Smoothing splines belong to the class of non-parametric regression methods and provide a flexible approach to estimating smooth functional relationships between variables. Their theoretical foundations date back to early work on spline functions and penalised least squares, notably by Reinsch (1967) and Wahba (1990). Smoothing splines estimate the regression function by fitting a smooth curve that balances fidelity to the observed data with a penalty on excessive curvature. This balance allows smoothing splines to capture complex, non-linear patterns without requiring the specification of a particular functional form, making them a powerful alternative to parametric regression techniques for smoothing and prediction.

Conceptually, smoothing splines differ from local averaging methods such as kernel regression in that they are global smoothers. Rather than computing predictions based on local neighbourhoods, smoothing splines determine a single smooth function over the entire domain of the predictor. The degree of smoothness is governed by a smoothing parameter, which controls the trade-off between goodness of fit and smoothness of the estimated curve. More formally, the smoothing spline $\hat{f}$ is defined as the function that minimises

$$\hat{f} = \arg\min_{f} \underbrace{\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2}_{\text{data-fit term}} + \lambda \underbrace{\int \left(f^{(3)}(t)\right)^2 dt}_{\text{roughness penalty}},$$

i.e. $\hat{f}$ denotes the estimated regression function (the optimiser of this criterion). The data-fit term measures how closely the fitted function matches the observed responses at the sampled covariate values, while the roughness penalty quantifies the rate of change of curvature (wiggliness) of $f$ through the integrated squared third derivative. The scalar $\lambda \geq 0$ is the smoothing (regularization) parameter that weights the roughness penalty relative to the data-fit term: larger $\lambda$ yields smoother fits; smaller $\lambda$ prioritizes fidelity to the data.

In R, we implement smoothing splines using the function `ss` from the package `npreg` (Helwig, 2020, 2024). In our case, this function fits a quintic smoothing spline via penalised least squares, where roughness is quantified by $\int (f^{(3)}(t))^2 dt$ and the smoothing parameter $\lambda$ weights this penalty relative to the data-fit term. By default, $\lambda$ is chosen by generalised cross-validation (GCV) to balance fit and smoothness in a data-driven manner. Unlike standard leave-one-out cross-validation, GCV is an analytical approximation that exploits the linearity of smoothing splines: instead of repeatedly refitting the model, it adjusts the residual sum of squares using the smoother's effective degrees of freedom to estimate predictive performance efficiently. An alternative approach is to supply a fixed $\lambda$ or to target an effective degrees of freedom value for more direct control over smoothness. In our applications, after comparing several selection methods for $\lambda$, we relied on the default GCV selection because we found that it yielded stable and interpretable fits without requiring manual tuning. During our selection process, we also identified the Bayesian Information Criterion (BIC) as second-best performing method. Overall, we observed occasional under-smoothing regardless of the selection method.

We also evaluated various alternative nonparametric approaches for estimating smooth functions in the context of our applications. These included local polynomial regression via base R's `loess` function, which implements LOESS (Cleveland and Devlin, 1988). We selected the span via a custom implementation of $k$-fold cross-validation (with $k = 5$ and $k = 10$, respectively, in testing), using polynomial degree two, and relied on locally weighted least squares. We ultimately did not pursue this approach further due to a tendency to under-smooth on our datasets. We additionally considered an alternative classical kernel smoothing via base R's `ksmooth` function, fixing bandwidths manually across a grid to assess sensitivity. However, no single bandwidth yielded consistently satisfactory results across our test datasets. Finally, we studied smoothing within the generalised additive model (GAM) framework: (i) with the `gam` function from the `gam` package (Hastie, 2025), following the original GAM formulation (Hastie and Tibshirani, 2017); and (ii) with the similarly named gam function from the `mgcv` package, which fits penalised regression splines to balance flexibility and smoothness (Wood, 2004, 2017). In our experiments, the `gam` package used its default smoothing spline terms, while `mgcv` was evaluated with both thin plate regression splines and P-splines. Neither of the GAM implementations was ultimately pursued due to a tendency to over-smooth for our applications.

After estimating the smoothing curve, we identify the locations of its peaks (i.e., local minima and maxima in the mathematical sense) by numerically finding the roots of the approximate first derivative. The classification of each peak is determined by the sign of the approximate second derivative at that point. We obtain both derivatives by numerical approximation, more precisely, using the classical backward finite-difference method, where the user can control the approximation accuracy. After inferring the extrema, the function `smooth.curves.est` optionally produces a plot of the input data overlaid with the estimated smoothing curve and the inferred peaks. In addition, the first- and second-derivative curves can be added to the plot by the user. Figure 2 shows the estimated smoothing curves for our single headshake from Libras, using kernel regression (top panel) and smoothing splines (bottom panel).

Finally, the function returns the estimation results of the kernel regression or smoothing spline, respectively, along with additional quantities relevant for further analysis. Note that various consistency checks are performed before estimating the smoothing curve: (i) the number of frames must match the number of movement angles; (ii) the number of observations must satisfy $n \geq 5$. In addition, the presence of missing observations is recorded – identified by non-equidistant (gapped) frame indices. Sequences failing checks (i)–(ii) are excluded from fitting. However, missingness is merely recorded and does not constitute a failure.

## 2.2. Peak Detection

The function `smooth.curves.classif`, implemented for Step 2, requires similar input to Step 1: (1) paired $x$- and $y$-values, corresponding to frames and movement angles, respectively; (2) the estimation-results list from Step 1. The estimation results also contain certain control parameters from Step 1. These have to be supplied again because some computations are repeated. While this is not computationally the most efficient approach, it significantly reduces the storage size of the results object produced by Step 1. Core functionality of Step 2 is the classification of the inferred peaks from Step 1 according to user-defined criteria. This classification is based on the principle that the label assigned to a specific peak depends on the magnitude of movement since the previous peak. Depending on this movement magnitude, the peak falls into the category 'relevant' for a sufficiently large enough movement, 'hold' for a sufficiently small movement, and 'not classified' for the remaining cases. The necessary amount of movement can be quantified either in absolute (e.g. degree) units or in relative terms, i.e., as a proportion of the total movement range of all observations in the sample. Example: suppose the peak under consid-

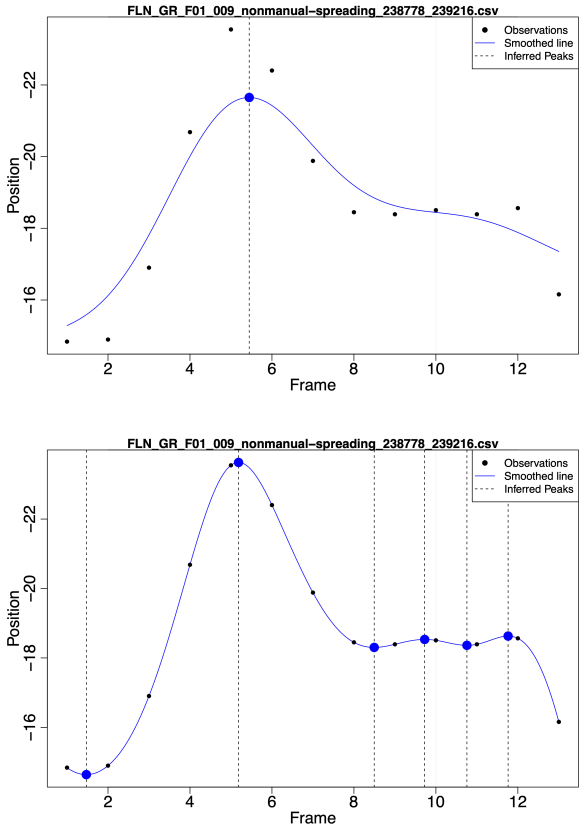



Figure 2: Smoothing example using kernel regression (top) and smoothing splines (bottom).

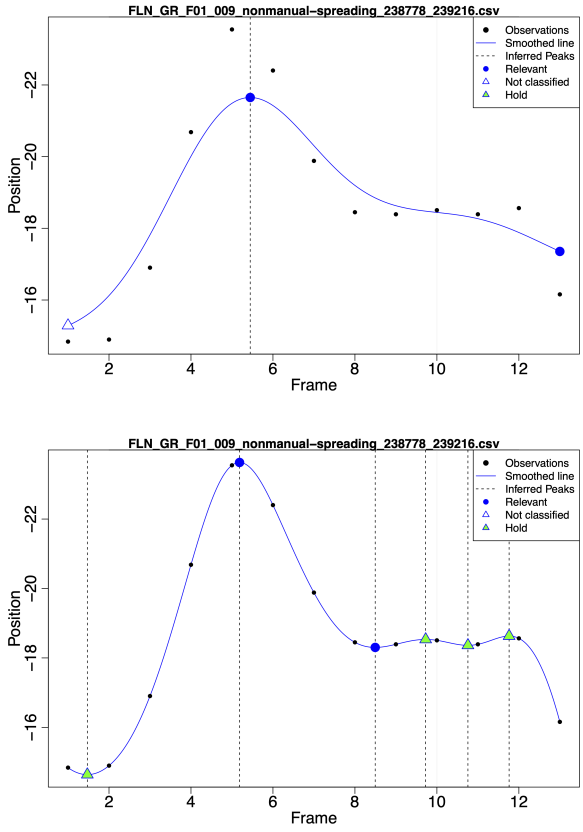



Figure 3: Peak detection example using kernel regression (top) and smoothing splines (bottom).

eration is a maximum. Then the previous peak is a minimum, and we consider the difference between these two peaks for classifying the maximum. If this difference equals 15 degrees and the threshold for a relevant peak was set at 10 degrees, the maximum is classified as a 'relevant' peak. However, if the difference equals only 2 degrees and the threshold for a hold equals 5 degrees, we classify the maximum as a 'hold'. All observed movements between 5 and 10 degrees result in a 'not classified' maximum.

The first inferred peak and the border points require special attention. For the first inferred peak, no previous peak is present by definition. Therefore, the first available observation at the left border serves as the reference point for assigning a label. Moreover, border points, i.e., the first and last observed values of the sequence under consideration, can be set to be peaks manually by the user via the argument `border.check`. This argument allows labelling a border point as a peak if no other inferred peak is available within a user-defined range (set via the argument `border.frames`) following the first or preceding the last observation, respectively. By default, a border peak at the very beginning of a sequence of observations falls into the 'not classified' category, because no previous movement data are available. This behaviour can, however, be overridden by the user (via the argument `border.classif.first`). When enabled, the first border peak is classified based on the movement magnitude up to the first subsequent peak. Consequently, the border peak at the first observation and that subsequent peak then receive identical labels, as both classifications are based on the same movement magnitude. On the other hand, the classification of a border peak at the end of an observation sequence follows the same logic as for all other peaks inferred from the smoothing curve. While border-point treatment may be of reduced importance for very long sequences of observations, it may be important when only a few observations are available.

Figure 3 illustrates the peak classification for the two previously estimated smoothing curves. In the top panel (kernel regression), only one peak was inferred via first-derivative roots, near Frame 6. In the absence of other roots near the boundaries, both the first and last observations are treated as border peaks. Following the logic described above, the peak at Frame 1 is labelled 'not classified' by default. If the override is enabled, this peak would be labelled 'relevant' because the movement from Frame 1 to the subsequent peak

near Frame 6 exceeds the relevance threshold. The second peak in the middle is classified as 'relevant' because of a sufficiently large magnitude of movement since the first peak. The same is true for the third peak at the right border. In the bottom panel (smoothing splines), no border peaks are inserted. Then, after a 'hold' at the beginning, two 'relevant' peaks follow. Afterwards, a period with three holds occur between approximately Frames 9 and 12.

Similar to Step 1, the function `smooth.curves.classif` displays the classification results as a figure. Moreover, it returns a list object with various quantities, which are mostly linked to the peaks in the sample but also include basic descriptive statistics of the underlying data (see following section). If no estimation results are available, only the descriptive measures are returned.

### 2.3. Measurements

We detail the output of the function `smooth.curves.classif` in the following. The returned quantities are heterogeneous and can be grouped into six categories. A brief description of each category follows.

**A.** Basic information: the series name and, when available, the frame rate.

**B.** Peak-based measures: quantities related to peaks inferred by the smoothing curve and border peaks (e.g., number and location of peaks, movement range between adjacent peaks).

**C.** Movement-based measures - standard: descriptive measures related to the observed movement. These are computed twice, from a) the originally observed positions and b) the corresponding positions inferred by the smoothing curve (e.g., average position, range, maximum frame-to-frame velocity).

**D.** Movement-based measures – robust: statistically robust counterparts to the measures presented under Item C, again computed from both the observations and the smoothing-curve positions (e.g., median position, 80%/90% interquantile range, 90% quantile of the absolute frame-to-frame velocity).

**E.** Other measures: measures mainly relying on either the observations or the inferred smoothing curve, that do not fit Items C-D (e.g., number of missing observations, predicted position values).

**F.** Control measures: quantities related to script execution and the estimation procedure.

We provide a detailed per-quantity description in the function documentation available in the repository at `https://doi.org/10.17605/OSF.IO/QFG98`.

The total number of quantities (of varying dimension) returned by the function `smooth.curves.classif` exceeds 50. When multiple sequences of observations are processed, it is useful to post-process the resulting output into a more user-friendly format. This third step is performed by the function `smooth.curves.res.prep`, which takes as input a list with output from Step 2 as list elements. The function returns a list with six elements. We describe these elements below.

1. `res.mat.obs`: a matrix collecting all single-valued descriptors (one row per input sequence; one column per quantity).
2. `res.mat.peaks`: a matrix with peak-related quantities for all series. It includes quantities derived from all peaks and, separately, from 'relevant'-labelled peaks.
3. `res.list.peaks`: a list of matrices, where each matrix contains peak-specific quantities resulting from one of the sequences of observations analysed.
4. `res.list.peaks.rel`: analogous to Item 3, this element consists of a list of matrices. However, it is based only on 'relevant' peaks.
5. `res.list.pred.obs`: a list of numeric vectors with predicted positions from the inferred smoothing curve at each frame of the original sequence.
6. `res.fail`: a logical vector indicating whether Step 1 (estimation) failed for a given sequence.

## 3. A case study

In the previous section, we provided the details of our scripts. In this section, we provide a short but realistic example of how they can be used to analyse kinematic properties of negative headshakes in Brazilian Sign Language (Libras). The data come from Corpus de Libras (Quadros et al., 2020), and were annotated for another study (Figueiredo, 2026a,b)

The dataset contains 323 individual negative headshakes produced by 10 signers of Libras. The headshakes were annotated according to a number of linguistic variables, including *spreading*: whether they co-occur with only one manual sign or spread over at least two manual signs. The research question was whether spreading and non-spreading headshakes are kinematically different. We processed the video files with OpenFace (Baltrusaitis et al., 2018) and used `pose_Ry` converted to degrees to analyse headshake kinematics, i.e. yaw measurements. More specifically, we applied the scripts developed in the current study with default settings for kernel regression and smoothing

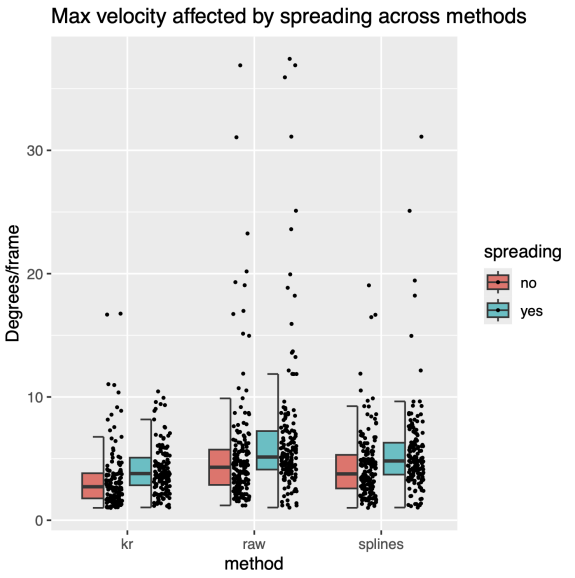



Figure 4: Maximal velocity by spreading status in raw data, spline-smoothed, and kernel-regression-smoothed data.

splines.

For the purposes of illustration, we focus on a small number of measurements. Figure 4 summarises how maximal velocity varies by spreading. We observe that spreading headshakes have a higher maximal velocity than non-spreading headshakes both in raw data, and in smoothed data (for splines and kernel regression). However, we also observe that the raw data exhibit the highest number of potential outliers, followed by splines, with kernel regression producing the fewest. This indicates that kernel regression produces, on average, smoother fits than spline smoothing, which is also what we observed for other datasets (not shown here). Thus, while the effect of spreading is observed with and without smoothing, it will be estimated more stably when smoothing is applied, because smoothing mitigates measurement noise and spurious extrema.

Figures 5 and 6 show measures that cannot be directly derived from the raw data, as discussed above, namely the number of peaks (reflecting the number of head turns) and the peak frequency (peaks per frame). We observe that both splines and kernel regression-based methods show that spreading headshakes exhibit a higher number of peaks but a lower frequency of peaks. This is not surprising, as spreading headshakes are typically longer and thus we expect the head to be able to turn more times. By contrast, for non-spreading headshakes – since they are relatively short – the frequency of peaks has to be higher in order for the head to be able to conduct perceivable turns. For these measures, we again observe that the spline-based method produces less smoothing, and thus a wider distribution of values.

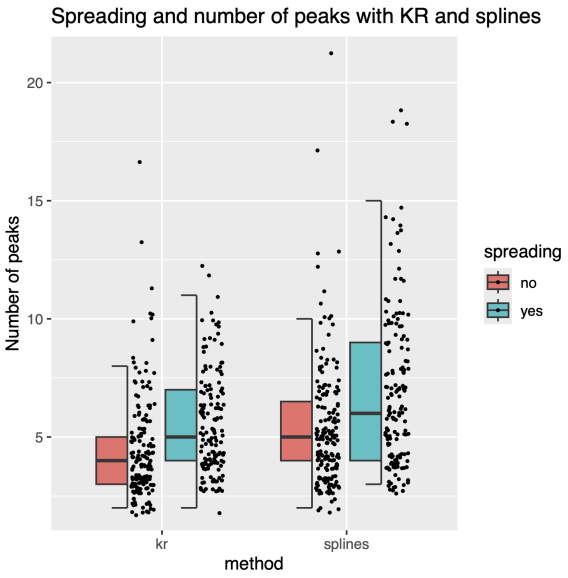



Figure 5: Number of peaks by spreading status in the data smoothed with splines, and with kernel regression, respectively.

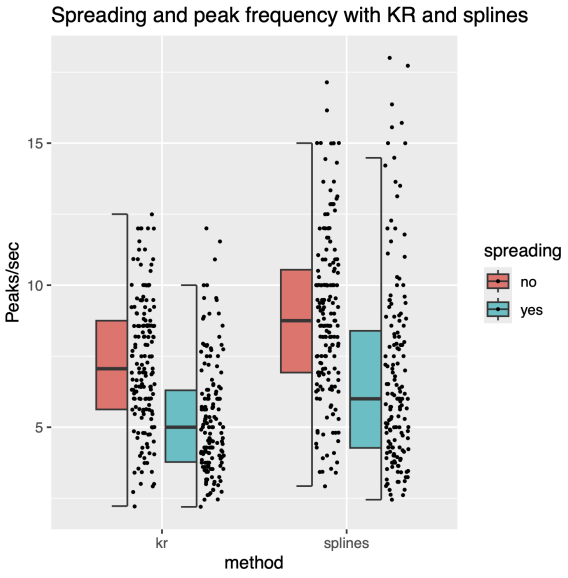



Figure 6: Frequency of peaks (peaks per second) by spreading status in the data smoothed with splines, and smoothed with kernel regression, respectively.

It is not possible to determine which smoothing approach performs better for this dataset, as no ground-truth measurements of head movements are available. Instead, visual inspection of individual movements with smoothing and peak detection should be used to ensure reasonable results, which

is why the scripts also output these visualizations. It seems, however, reasonable to choose spline smoothing if one believes that the outputs of the CV instrument used are stable and largely free of extremely faulty measurements, as this method preserves most of the variation present in the raw data. In contrast, if the CV instrument is relatively unreliable and noisy, kernel regression may produce superior results.

For other studies that fully employ the presented smoothing and peak-detection approach to negative headshakes, see Kimmelman et al. (2026). For an application to eye blinks, see Susman (2026).

## 4. Conclusions and Limitations

To sum up, in this paper we propose a workflow for processing outputs of computer vision tools applied to nonmanual markers in sign languages and co-speech gestures. The scripts we developed offer a solution to three crucial problems when processing such data, namely, smoothing, peak detection, and the calculation of measurements. The full pipeline is implemented in R (R Core Team, 2025) and is openly available and documented in the repository at https://doi.org/10.17605/OSF.IO/QFG98.

The proposed approach, however, has some clear limitations. First, while the goal is fully automated processing, it still requires expert input for calibration. For the smoothing part, we have tested the methods using several datasets of head movements, and decided on the best approaches using our experience in sign-language linguistics. However, these datasets are not representative of the breadth of signed and spoken languages, nor of the many different types of nonmanual movements. Nevertheless, it is possible – if needed – to modify the methods and to implement other methods (beyond the ones we tested) using the same general approach, which, in principle, should be applicable to other movement types of nonmanual articulators.

For the peak-detection part, expert input is even more crucial, as the peak classification into relevant peaks and holds bases entirely on expert-defined thresholds. In fact, in the published version of the script, we use a simple relative threshold, which does not perform perfectly even on all sequences belonging to the dataset that we tested. Hence, more complex approaches might be required for more robust peak classification in future.

We plan to explore automatic, data-driven peak classification in future work on this approach. However, a central challenge is the need for creating sufficiently large, well-annotated datasets in which the true movement patterns are known with high reliability. The construction of such datasets does not constitute a trivial task, especially for low amplitude movements.

Another limitation of this approach is that it was developed and tested on sequences of approximately 5-50 frames in length. It is not reliably applicable to shorter sequences due to mathematical and algorithmic constraints of some of the components; in particular, kernel regression is less forgiving than smoothing splines for very short sequences. For longer sequences, there is no inherent restriction, but we have not tested the robustness of the smoothing, and – especially for peak detection – additional calibration may be required.

## 5. Acknowledgments

Funded by the European Union (ERC, NONMANUAL, project number 101039378). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## 6. Author contributions

Vadim Kimmelman and Jan Bulla conceptualised the study. Jan Bulla developed the software, with input and testing from Vadim Kimmelman. Vadim Kimmelman conducted the case study. Both authors wrote the paper.

## 7. Data availability

The scripts described in this paper, together with a demo script and a test dataset, and the script used to produce the figures, are available at https://doi.org/10.17605/OSF.IO/QFG98.

## 8. Bibliographical References

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

Anastasia Bauer, Anna Kuder, Marc Schulder, and Job Schepens. 2024. Phonetic differences between affirmative and feedback head nods in German Sign Language (DGS): A pose estimation study. *PLOS ONE*, 19(5):e0304040.

Carl Börstell. 2023. Extracting Sign Language Articulation from Videos with MediaPipe. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 169–178, Tórshavn, Faroe Islands. University of Tartu Library.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

William S. Cleveland and Susan J. Devlin. 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Lorena Figueiredo. 2026a. A corpus-based study on negation in Brazilian Sign Language.

Lorena Figueiredo. 2026b. Negative Nonmanual Markers in Brazilian Sign Language (Libras). Zenodo.

U. Hadar, T. J. Steiner, and F. Clifford Rose. 1985. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228.

U. Hadar, T.J. Steiner, E.C. Grant, and F.Clifford Rose. 1983. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1-2):35–46.

T.J. Hastie and R.J. Tibshirani. 2017. *Generalized Additive Models*, 1 edition. Routledge.

Trevor Hastie. 2025. *gam: Generalized Additive Models*.

Tristen Hayfield and Jeffrey S. Racine. 2008. Nonparametric Econometrics: The **np** Package. *Journal of Statistical Software*, 27(5).

Nathaniel Helwig. 2020. Multiple and Generalized Nonparametric Regression. In *SAGE Research Methods Foundations*. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom.

Nathaniel E. Helwig. 2024. *npreg: Nonparametric Regression via Smoothing Splines*.

Vadim Kimmelman, Anastasia Bauer, Carl Börstell, Jan Bulla, Bux Allah, Laurence Crettenand, Lorena Figueiredo, Anna Kuder, Hannah Lutzenberger, Marloes Oomen, and Roland Pfau. 2026. Kinematics of negative headshake in seven sign languages.

Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study. *PLOS ONE*, 15(6).

Qi Li and Jeffrey S. Racine. 2004. Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14(2):485–512.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines.

É. A. Nadaraya. 1965. On Non-Parametric Estimates of Density Functions and Regression Curves. *Theory of Probability & Its Applications*, 10(1):186–190.

Naomi Nota, James P. Trujillo, and Judith Holler. 2021. Facial Signals and Social Actions in Multimodal Face-to-Face Interaction. *Brain Sciences*, 11(8):1017.

Roland Pfau and Josep Quer. 2010. Nonmanuals: their prosodic and grammatical roles. In Diane Brentari, editor, *Sign Languages*, pages 381–402. Cambridge University Press, Cambridge.

Wim Pouw. 2024. Wim Pouw's EnvisionBOX modules for social signal processing.

Anna Puupponen, Tuija Wainio, Birgitta Burger, and Tommi Jantunen. 2015. Head movements in Finnish Sign Language on the basis of Motion Capture data: A study of the form and function of nods, nodding, head thrusts, and head pulls. *Sign Language & Linguistics*, 18(1):41–89.

Ronice M. de Quadros, Deonísio Schmitt, Juliana T. Lohn, and Tarcísio de A. Leite. 2020. Corpus de Libras.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Jeff Racine and Qi Li. 2004. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130.

Christian H. Reinsch. 1967. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183.

Margaux Susman. 2026. Phonetic properties of blinks in French Sign Language.

James P. Trujillo and Wim Pouw. 2021. Kinematic Feature Extraction for Motion Tracking Analysis.

James P. Trujillo, Julija Vaitonyte, Irina Simanova, and Asli Özyürek. 2019. Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior Research Methods*, 51(2):769–777.

Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232.

Grace Wahba. 1990. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Geoffrey S. Watson. 1964. Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372.

Simon N Wood. 2004. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association*, 99(467):673–686.

Simon N. Wood. 2017. *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC.

# A Small Model for Big Articulators: Sign Language Detection With a Tiny Machine Learning Model

**Frederick Chan, Gina-Anne Levow, Qi Cheng**
University of Washington
Seattle, WA
fredchan@uw.edu, levow@uw.edu, qicheng2@uw.edu

**Abstract**

This paper introduces a small (1,013 parameter) machine learning model for sign language detection in videos of isolated American Sign Language (ASL) signs. Our model aims to alleviate the time-consuming nature of producing sign clips for psycholinguistic study stimuli, sign dictionaries, and sign databases. Given a video where the signer starts from a resting position, signs a sign, and returns to the resting position for an arbitrary number of repetitions, the model detects frames in which signing occurs that can be used to segment video into clips of individual signs. We train and evaluate our model on data with precise coding of signing onset and offset from ASL-LEX 2.0, so that our model's annotations are suitable for psycholinguistics research. The model works on both real signs and pseudosigns, two types of stimuli needed for certain psycholinguistic studies. Our model's small size compared to the state-of-the-art (100K parameters or more) enables quick, bulk processing even on resource-constrained hardware. It achieves this by computing Instantaneous Visual Change (IVC), a 1D measure of changes in brightness in the input video, extracting features from the IVC-over-time signal with a convolution, and classifying the video frames as signing or non-signing with three neural layers.

**Keywords:** sign language detection, sign language segmentation, machine learning, small models, convolutional neural networks (CNN)

## 1. Introduction

Sign languages are natural languages used by deaf and hard-of-hearing people that make use of both manual and non-manual movements, including posture and facial modulations (Sandler and Lillo-Martin, 2006). Like spoken languages, they have complex systems of phonology, grammar, and lexicon, but their use of the visual modality gives rise to many unique linguistic features that can prove challenging for machine processing (Sandler and Lillo-Martin, 2006; Toshpulatov et al., 2025). Development in sign language processing has lagged significantly behind spoken language processing, partly due to the complex nature of the phonology of sign languages, the difficulty of working with videos as opposed to audio, and a lack of high-quality, large-scale data sets. (Toshpulatov et al., 2025; Neidle et al., 2012).

In the past decade, the application of neural networks in other fields and increased availability of computing capacity have led to an explosion of neural network-based approaches to sign language tasks (Toshpulatov et al., 2025), which are able to handle the size and complexity of sign language video data. Since neural networks tend to benefit from larger parameter sizes, state-of-the-art sign language models tend to feature upwards of 100K parameters or more (Moryossef et al., 2020; Borg and Camilleri, 2019), which requires significant computational power and runtime.

In this paper, we demonstrate that small neural networks (around 1K parameters), can be used to detect signing in a kind of signing video commonly created for psycholinguistic studies and various sign language resources, like dictionaries.

A small model could also be added into existing sign language annotation tools, such as ELAN (Wittenburg et al., 2006), thus integrating easily into existing workflows.

### 1.1. Sign Language Detection and Segmentation

In an automatic sign language detection task, a video frame is given, with or without surrounding frames, to a model which labels it as signing or non-signing. By classifying all the frames of a video, these labels can be used to segment it into time spans in which signing occurs.

Larger, state-of-the-art sign language models perform well on complex, continuous signing videos, where signers produce phrases or sentences, often as they would in normal discourse (Stassi et al., 2025; Borg and Camilleri, 2019; Moryossef et al., 2020, 2023). In this paper, we focus on detecting sign language in a simpler type of video that is produced when creating psycholinguistics research stimuli (Emmorey et al., 2011; Gu et al., 2022; Caselli et al., 2021), sign language dictionaries, and sign databases (Sehyr et al., 2021). That is, videos where a signer in front of a static background begins in a resting position, signs a sign, and moves back to the resting position, an



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 71–79
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

arbitrary number of times. These videos are then edited down into clips, each containing one sign, for use in the final product or study.

Despite the simple nature of these videos, the process of editing the video into individual sign clips remains a major pain point. A psycholinguistic study may make use of dozens or hundreds of such clips, while a dictionary may need thousands. Producing a whole set takes many hours of work, and a means of clipping these videos quickly and without needing an army of undergraduates to do the grunt work would be ideal. The model we demonstrate in this paper aims to fill this need.

Since such videos are simpler compared to continuous signing, certain factors that are common in natural discourse but could confound a computer model, such as phonological assimilation between signs and simultaneous morphology, are not present to confuse the model. We show that a small, simple model is sufficient to achieve good performance on this simpler domain.

### 1.2. Definition of Signing Onset and Offset

Since part of the goals of our model is to be useful for creating stimuli for psycholinguistics, we need a precise definition for the onset and offset of a sign. Unlike spoken language, where the articulators are hidden in the vocal tract and the onset of the audio is the same as the onset of the word, motion begins before the sign onset, since the arms and hands must move into position before producing the sign (Emmorey et al., 2022).

In this paper, we use the definitions from in ASL-LEX 2.0 (Sehyr et al., 2021). For signs with body contact and two-handed signs, the onset is the first video frame where the fully-formed handshape has contacted the body, and the offset is the last frame where the hand has contacted the body. For signs without contact, the onset is the first frame where the fully-formed handshape has arrived at the target location near the body or in neutral space, and the offset is the last frame before the hand(s) have begun to transition to a resting position (Sehyr et al., 2021).

Using these definitions allows researchers to precisely control for the onset of a sign when measuring brain activity while a subject comprehends sign language stimuli. For example, Emmorey et al. (2022) looks for specific Event Related Potentials (ERPs), such as N400 priming effects, relative to the sign onset. Studies involving reaction time, such as Caselli et al. (2021) and Sehyr and Emmorey (2022), also need control for the signing onset to get accurate and comparable data between trials.

To serve these needs, we want as little error between our model's predicted sign onsets and the true sign onsets as possible. We therefore made use of two video corpora, ASL-LEX 2.0 and the stimuli videos from Caselli et al. (2021) (described further in Section 3.1), that have rigorously annotated sign onsets and offsets based on the definitions previously described.

### 1.3. Existing Models

Presently, the state-of-the-art in sign language detection and segmentation achieves high accuracy on the continuous, naturalistic videos. However, they are also quite large, which costs a considerable amount of computational power to run. Furthermore, the onsets and offsets in the training data of these models are not defined as precisely as the datasets we use here, and their suitability for the preparation of stimuli for psycholinguistics is unknown.

Borg and Camilleri (2019) created sign language segmentation models that extract features from raw video data, optical flow, motion history, and frame difference using the pretrained VGG-16 Convolutional Neural Network (Simonyan and Zisserman, 2015), which then classifies frames as signing and non-signing with a Recurrent Neural Network (RNN). The smallest of these models uses 3.3 million parameters, and classifies 20 frames of a 5 fps video in upwards of 3 seconds on CPU.

Moryossef et al. (2020) uses an RNN to classify frames using only the position of the joints of the signer as input, which are extracted from the raw video using the pose estimation model, OpenPose (Cao et al., 2019). This allows the model to ignore irrelevant motion in the video and focus on the major articulators involved in signing. While much smaller than Borg and Camilleri (2019), this model involves another neural network as a preprocessing step and still uses 102 thousand parameters.

Moryossef et al. (2023) proposes several models using joint positions extracted from MediaPipe (Lugaresi et al., 2019) combined with other techniques to improve segmentation performance. These include using optical flow features, 3D hand normalization, and location of nonmanual keypoints. The smallest of these models uses 454 thousand parameters.

He et al. (2025) uses two pretrained models that extract hand mesh and skeleton pose features from video, and trains two multi-layer perceptrons, a feature mixer, and a Transformer encoder within the same model to classify video frames from those features. The total parameter count cannot be inferred from the paper as only the feature sizes but not the layer sizes of the model are not reported, nor is the model's code published at the time of writing. However, given its complexity, it is reasonable to assume that the model is much larger than

1,013 parameters.

### 1.4. Instantaneous Visual Change (IVC)

The model introduced in this paper is able to achieve results with much smaller parameter sizes by using a simple 1-dimensional signal extracted from the input video called *Instantaneous Visual Change* (IVC). IVC is the sum of squared differences in each pixel across sequential frames of video:

$$IVC(t) = \sum_i ([\, x_i\,(\,t\,) - x_i\,(\,t-1\,)\,])^2$$

where $x$ is the grayscale value of pixel $i$ at time $t$.

IVC was introduced in Brookshire et al. (2017), which established a correlation between the quasiperiodic nature of the IVC-over-time signal in sign language videos with neural entrainment in the visual cortex of native signers who view them. The IVC signal is comparable to the spoken audio signal, with the spectral power densities of both being similar (Brookshire et al., 2017). The authors hypothesize that babies can "tune in" to the presence of signing in visual stimuli during language acquisition, because the quasiperiodicity is characteristic of a linguistic signal.

Given this, it stands to reason that a machine could also exploit this property of the IVC signal to detect the presence of signing in a video. In our model, we train a convolutional layer, which is popular for shape and object detection in computer vision, to learn shapes in the IVC signal that are characteristic of signing. Then, it classifies frames into signing or non-signing based on the presence or absence of these shapes.

## 2. Model Architecture

### 2.1. Pre-processing

First, each video is downsampled to 29.93 frames per second, the frame rate of the videos in ASL-LEX 2.0 and the Caselli et al. (2021) study.

Next, the IVC of the frames of the input video is computed using a version of Brookshire (2017)'s IVC module that we modified to work in Python 3. This module converts the color space to grayscale using OpenCV's (Bradski, 2000) $RGB \leftrightarrow GRAY$ color conversion mode (OpenCV team, 2025). Note that the first frame of the video has an IVC of 0; no video frame precedes it, so there is no visual change.

Then, the framewise IVC is normalized by dividing the IVC by the standard deviation of all the IVC in the video. This reduces the effect of variations in the physiology of the signer, distance from the camera, and video resolution. No denoising or resizing of the video frames themselves are performed.

A sliding window across all the frames of the video is then computed, creating windows that are 16 frames long (about 0.54 seconds). For each window, the model predicts the probability that the middle frame (frame 8) contains signing. This gives the model 7 frames of past IVC history, the IVC of the frame being classified, and the IVC of the 8 frames that follow.

### 2.2. The Model

The model starts with a feature extractor: a single 1D convolutional layer of 4 kernels, with a width of 4 frames, and a stride of 1. This feature extractor learns to detect shapes in the IVC signal that are characteristic of signing.

The convolution produced by the 4 kernels are then concatenated into a single feature vector. This feature vector is fed into a neural network of 3 densely connected neural layers:

- Layer 1: 16 neurons with ReLU activation
- Layer 2: 8 neurons with ReLU activation
- Regression: 1 neuron with sigmoid activation

The regression layer outputs the middle frame's probability of signing. During inference, frames with $\geq$50% probability are considered signing frames.

In total, the model consists of 1,013 parameters.

### 2.3. Hyperparameter Selection

The hyperparameters of this model were chosen using the GridSearch method in KerasTuner (O'Malley et al., 2019), over the settings below. All possible configurations within the search space were tested on a limited set of training data of 270 videos, and the hyperparameters of the best performing model were used to train a randomly initialized model on the full training set. The search space is as follows:

- Number of convolutional kernels: 1, 2, 4, 8, 16
- Convolutional kernel size: 4, 8, 16
- Number of ReLU dense layers: 0, 1, 2
- Size of ReLU dense layers: 4, 8, 16, 32

The results of the hyperparameter search are shown in Figure 1.

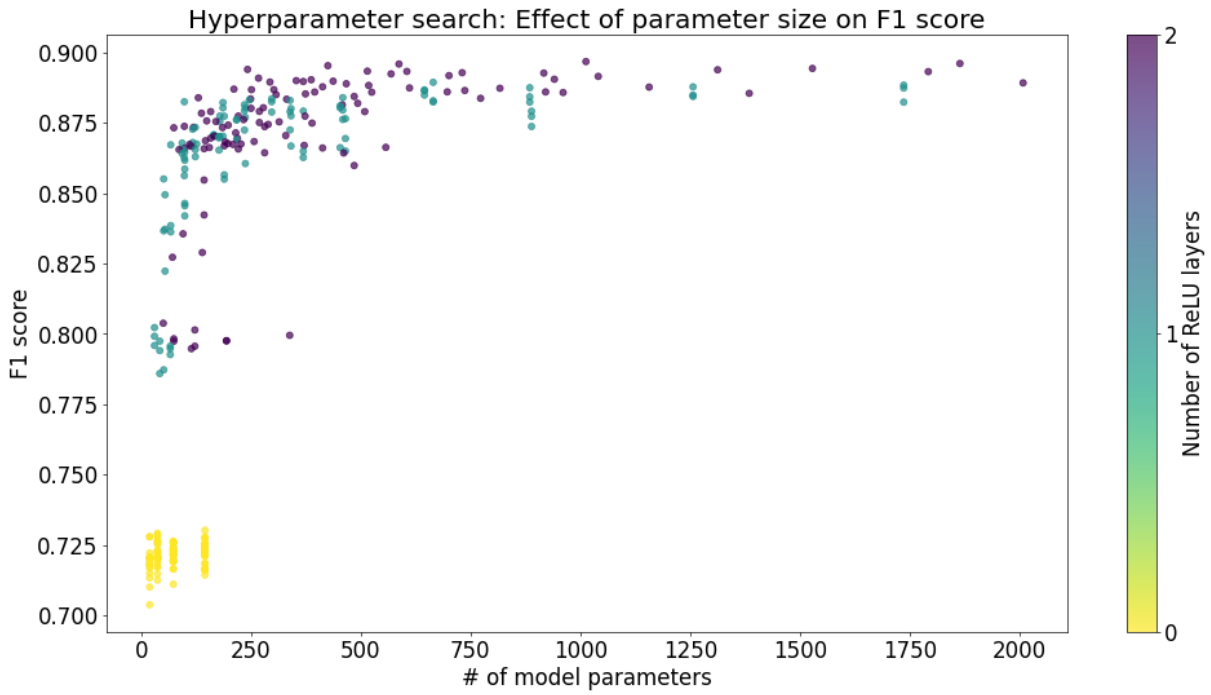



Figure 1: Results of the hyperparameter grid search. A number of different small-model hyperparameter configurations were tested on a limited set of training data (270 videos), before the best performing model was retrained on the full set. The hyperparameters in the search varied by number of convolutional kernels, kernel size, number of ReLU layers, and ReLU layer size. The best models use 2 ReLU layers of varying size. The cluster on the bottom left are models with no ReLU layers, only convolutions and a logistic regression.

## 3. Training

### 3.1. Training Data

The videos used to train the model were originally created for ASL-LEX 2.0 (Sehyr et al., 2021), a large-scale database of ASL signs that also contains the signing onset and offset times for each sign video. The ASL-LEX 2.0 set consists of 2,792 videos of unique single signs, featuring the same deaf native signer (female, middle-aged, White, born in the North-East USA, resides in California) filmed from the front and from the waist up, against a static blue background. The videos appear to have been filmed in different batches; while many videos have identical camera and lighting conditions, this varies between different batches, and the signer may have different outfits between them. Onsets and offsets for each video were annotated by two independent, trained coders, with 90.4% agreement for onsets and 99% agreement for offsets. The minimum non-normalized IVC in this dataset is 0, with a maximum of 4,246,152,807. From this dataset, we set aside 20% of the videos to use as a validation set.

For the test set, 569 stimulus videos that were created for the Caselli et al. (2021) psycholinguistics study were used. This set contains 276 unique real ASL signs and 293 ASL-based pseudosigns, which were created by changing one or two phonological parameters *(e.g. handshape, location, movement)* of a real sign so that it is not a sign in ASL. The signer in these videos is a different deaf, native signer from the one in the ASL-LEX 2.0 videos, and was filmed against a static white background in slightly different lighting conditions owing to the time of day. The minimum non-normalized IVC in this dataset is 0, with a maximum of 475,561,711.

In the videos of both datasets, the signer starts from a neutral, sitting position before beginning the sign. They produce the sign, then return to a neutral position. Since the neutral position and the transitions to and from signing do not carry lexical meaning, frames containing these are considered non-signing.

These two datasets were selected because they were among the only easily-available corpora of isolated sign videos with high-quality timing annotations necessary for training and evaluation.

Both datasets undergo the pre-processing described in Section 2.1, creating a large number of frame windows to classify. Then, windows are randomly dropped until there is an equal number of signing and non-signing windows in the split. Balancing the signing and non-signing examples

improves its ability to classify both classes, rather than considering one class the "default," with worse performance. This resulted in 111,968 frame windows to train on.

### 3.2. Hardware, Framework, and Schedule

We trained the model on a batch size of 32, using the Adam optimizer for 100 epochs, by which time the validation accuracy has stopped showing significant improvement. Each epoch took about 30 seconds each, for a total of 50 minutes.

The model was trained on an NVIDIA GeForce 3070Ti Laptop GPU. The Keras neural network library was used with the PyTorch backend.

## 4. Evaluation

The evaluation metrics in this section can also be seen in Table 1. As a performance baseline we could directly compare to, we also trained and evaluated a logistic regression with no convolution created with scikit-learn (Pedregosa et al., 2011), using the same training and test data preprocessed with the procedure in Section 2.1.

### 4.1. Frame-wise Performance

The frame-wise performance metrics were computed after balancing the number of signing and non-signing windows, as described in Section 3.1. This results in 28,494 windows to classify in the validation set and 11,480 windows in the test set, each with an equal number of signing and non-signing examples.

On the validation set, the model achieved an Accuracy/Precision/Recall/F1 of 87/84/91/88%. On the test set, it achieved 87/85/88/87%.

Among pseudosigns in the test set, it achieved 85/85/86/86%, compared with real signs where it achieved 88/85/92/88%.

Our model achieves better performance than the logistic regression on both the test (F1 of 88% > 71%, 17 percentage points better) and validation sets (F1 of 86% > 70%, 16 percentage points better).

### 4.2. Video-wise Performance

To ascertain how well the model at segmenting videos, instead of just classifying frames, we evaluated the model on whole videos using Intersection over Union and Segment Ratio. Unlike in the frame-wise performance section, no windows were dropped. Thus, the model classifies all the frames of each video.

#### 4.2.1. Intersection Over Union (IOU)

Intersection Over Union (IOU) is a measure of the similarity between two sets $A$ and $B$, using the formula:

$$IOU(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Here, we are comparing the the ground truth set of signing frames in a video against the model's predicted set. IOU is always in the interval [0, 1], and higher scores are better, indicating higher similarity.

In this paper, we compute the IOU between the prediction and the ground truth for each video in the test set, then take the mean IOU over all the videos.

On the test set, the mean IOU was 0.78 ($\sigma$=0.13). Among real sign videos in the test set, the mean IOU was 0.79 ($\sigma$=0.14), compared with pseudosign videos with a mean IOU of 0.77 ($\sigma$=0.13). Figure 2 shows the segmentation output for three sign videos in the test set at the best, median, and worst case IOU.

On the validation set, the mean IOU was 0.80 ($\sigma$=0.17).

#### 4.2.2. Segment Ratio

Segment Ratio (Seg) is a measure of how well the number of segments $S_{pred}$ predicted by a model matches the actual number of segments $S_{true}$ in the video. It is computed using the formula:

$$Seg(S_{pred}, S_{true}) = \frac{S_{pred}}{S_{true}}$$

Ideally, Seg is 1, with a higher number indicating over-segmentation, and a lower number indicating under-segmentation.

In this paper, we compute the Seg between the prediction and the ground truth for each video in the test set, then take the mean Seg over all the videos.

On the test set, the mean Seg was 1.43 ($\sigma$=0.76). Among real sign videos in the test set, the mean Seg was 1.45 ($\sigma$=0.78), compared with pseudosign videos with a mean Sig of 1.41 ($\sigma$=0.74).

On the validation set, the mean Seg was 1.52 ($\sigma$=0.89).

### 4.3. Onset Error

To see how well the model aligns the onset of the predicted segments with the actual onsets of the sign, we computed Onset Error (OE) on videos in the test and validation set with the formula:

$$OE(O_{pred}, O_{true}) = abs(O_{pred} - O_{true})$$

Where $O_{true}$ is the frame number of the true onset of the sign in the video, and $O_{pred}$ is the

| | Frame-level | | | | Video-level | |
|---|---|---|---|---|---|---|
| | **Accuracy** | **Precision** | **Recall** | **F1** | **IOU** | **Seg** |
| **Our model** | | | | | | |
| Validation set | 87% | 84% | 91% | 88% | .80 | 1.52 |
| Test set | 87% | 85% | 88% | 86% | .78 | 1.43 |
| Real signs | 88% | 85% | 92% | 88% | .79 | 1.45 |
| Pseudosigns | 85% | 85% | 86% | 86% | .77 | 1.41 |
| Concatenated | 86% | 80% | 88% | 84% | .72 | 2.09 |
| **IVC + Logistic Regression** | | | | | | |
| Validation set | 70% | 70% | 72% | 71% | .65 | 1.38 |
| Test set | 70% | 69% | 70% | 70% | .64 | 1.49 |
| Concatenated | 71% | 64% | 70% | 67% | .50 | 2.05 |
| **State-of-the-art (Continuous signing, not a direct comparison)** | | | | | | |
| Bull et al. (2020) ("body" model) | | | | | .69 | 2.96 |
| Moryossef et al. (2023) (E0 model) | | | | | .77 | 1.37 |
| He et al. (2025) | | | | 86% | .76 | 0.98 |

Table 1: Performance metrics for our small model, compared with a simple logistic regression. State-of-the-art models are also shown for **indirect** comparison (see Section 5 for caveats). Metrics for Bull et al. (2020) and Moryossef et al. (2023) are from Stassi et al. (2025), which evaluated both models using a common dataset and methodology. Metrics for the model in He et al. (2025) are from its own paper. Note that He et al. (2025) is a BIO tagger that classifies three labels, unlike our binary classifier, and that this may affect the F1 score comparison.

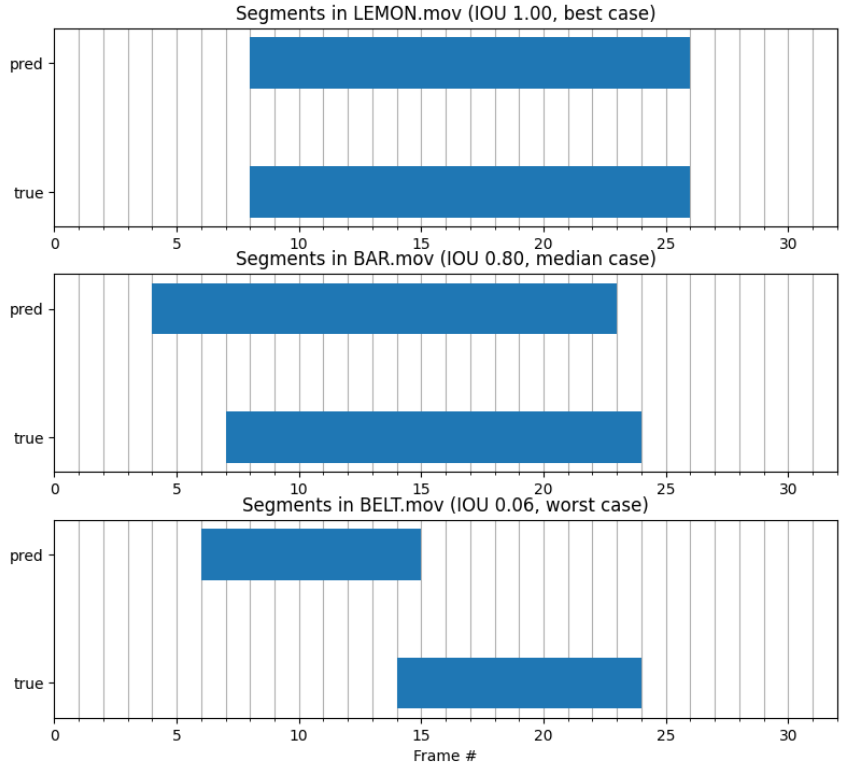



Figure 2: Predicted versus ground truth segments in three videos in the test set. Top: best case IOU (1.00) for a video containing the ASL sign LEMON, where the model predicts the segment perfectly. Middle: median case IOU (0.80) for BAR, with 4 frames of difference. Bottom: worst case IOU (0.06) for BELT, with little overlap.

frame number of the onset of the first predicted segment in the video.

On the test set, the mean OE was 2.82 frames ($\sigma$=2.97), or about 0.09s ($\sigma$=0.1s). The mean OE real signs in the test set, 2.86 frames ($\sigma$=3.34), and 2.77 frames ($\sigma$=2.58) for pseudosigns.

On the validation set, the mean OE was 2.34 frames ($\sigma$=3.76), or about 0.08s ($\sigma$=0.12s).

If we compute OE without the absolute value, we can see if the model tends to predict the onset too early or too late. On average, the model predicts onset 2.5 frames earlier than the ground truth in the test set, and 1.05 frames earlier in the validation set.

#### 4.3.1. Concatenated Long Videos

Although neither the training nor the test data include long-form videos for segmentation, we can simulate these videos by concatenating the IVC of randomly selected clips from the test set, normalizing the concatenated IVC as if it were one video, and running the model. 500 concatenated videos were tested this way, each consisting of a random number of clips, ranging from 10 to 100 videos.

We chose to concatenate the IVC instead of editing the video clips together and calculating the IVC afterwards, as the latter would create artifacts between clips that would not appear in a long video recorded in a single take. Namely, the transition between clips filmed in different conditions would create a frame with an unusually high IVC, since most of the pixels in the scene would be different. Similarly, we dropped the first frame of each video, starting from the second video in each concatenation, to remove 0-IVC frames that appear simply because it is the first frame.

On the set of 500 concatenated videos, the mean IOU was 0.72 ($\sigma$=0.02), and the mean Seg was 2.09 ($\sigma$=0.19).

### 4.4. Runtime

When running inference on the same hardware as training, pre-processing of the entire test set takes about 3 minutes and 40 seconds, and the model predicts the labels in about 6.6 seconds, for a total of 3 minutes and 46.6 seconds.

## 5. Conclusion

In this paper we demonstrate that, when the problem domain is limited to detecting signing in a stream of single signs, a tiny model of only 1,013 parameters is able to achieve an F1 score of 88% and a mean IOU of 0.78 on videos of single signs. On simulated videos of multiple isolated signs and an F1 score of 84% and a mean IOU of 0.72. Using logistic regression as a base of comparison, our model has significantly better F1 and IOU, showing that the model architecture yields meaningful improvements over a simple baseline.

We can compare our metrics to state-of-the-art sign language segmentation models. According to a paper that evaluated both models on the same task, Stassi et al. (2025) finds that the IOU of the best performing model from Bull et al. (2020) achieves an IOU of .69, and Moryossef et al. (2023) achieves an IOU of .77. Note that this is not a direct comparison; the state-of-the-art models were evaluated on continuous signing videos, which are more complex. The purpose of the comparison is to see whether our small model works as well on the simpler task as a complex model works on a complex task. Since our score appears within this range, it seems that our small model is able to do this successfully.

## 6. Limitations

While the datasets used for training and evaluating this model present good coverage for signs with different lexical and phonological properties (*e.g.* parts of speech, movement types, sign location) the fact that there are so few people represented in the data means that we do not know how well the model generalizes to signers of different ethnicity, ages, body shapes, etc,. The fact that the model achieves similarly high accuracy and F1 scores for the validation set as the test set, which do not have any signer overlap, is a good sign, but it is still far from rigorous.

If precise signing onset and offset annotations were available for a dataset like ASL Citizen (Desai et al., 2023), which features isolated sign videos crowdsourced from many Deaf and Hard of Hearing signers, such a dataset would be ideal for training and evaluating the model. Unfortunately, we have so far been unable to source such a similar dataset of good quality.

Since the sign language segmentation task can apply to other sign languages also, sign corpora from languages other than ASL should be considered for the future.

## 7. Future Work

The IVC signal alone is insufficient to capture the full complexity of the movements of sign language. Fundamentally, it is a measure of the amount of change in the visual field, analogous to volume in a spoken language's audio signal, so there is a limit to the amount of information that can be recovered from the original signing.

Although pose estimation introduces a lot of complexity and runtime to the process, making direct use of the position and motion of the articulators

involved in sign language (*e.g.* arms, hands, face, torso) could allow a model to pick up on more nuanced information. With MediaPipe (Lugaresi et al., 2019) now available even in a web browser, a small classifier that uses pose estimation features may be practical on midrange consumer hardware.

## 8. Funding

Research reported in this publication was supported by the National Institute for Deafness and Other Communication Disorders under Award Number R21DC022046, and the University of Washington Linguistics Fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the University of Washington.

## 9. Bibliographical References

Mark Borg and Kenneth P Camilleri. 2019. Sign language detection "in the wild" with recurrent neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1637–1641. IEEE.

Gary Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Geoffrey Brookshire. 2017. ivc: a small python module to compute the instantaneous visual change of videos. https://github.com/gbrookshire/ivc/tree/master. Accessed: 2026-03-22.

Geoffrey Brookshire, Jenny Lu, Howard C Nusbaum, Susan Goldin-Meadow, and Daniel Casasanto. 2017. Visual cortex entrains to sign language. *Proceedings of the National Academy of Sciences*, 114(24):6352–6357.

Hannah Bull, Michèle Gouiffès, and Annelies Braffort. 2020. Automatic segmentation of sign language into subtitle-units. In *European Conference on Computer Vision*, pages 186–198. Springer.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Naomi K Caselli, Karen Emmorey, and Ariel M Cohen-Goldberg. 2021. The signed mental lexicon: Effects of phonological neighborhood density, iconicity, and childhood language experience. *Journal of Memory and Language*, 121:104282.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2023. ASL citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36:76893–76907.

Karen Emmorey, Katherine J Midgley, and Phillip J Holcomb. 2022. Tracking the time course of sign recognition using erp repetition priming. *Psychophysiology*, 59(3):e13975.

Karen Emmorey, Jiang Xu, and Allen Braun. 2011. Neural responses to meaningless pseudosigns: evidence for sign-based phonetic processing in superior temporal cortex. *Brain and Language*, 117(1):34–38.

Shengyun Gu, Deborah Chen Pichler, L Viola Kozak, and Diane Lillo-Martin. 2022. Phonological development in american sign language-signing children: Insights from pseudosign repetition tasks. *Frontiers in psychology*, 13:921047.

Low Jian He, Harry Walsh, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Hands-on: Segmenting individual signs from continuous sequences. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, page 1–5. IEEE.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020. Real-time sign language detection using human pose estimation. In *European Conference on Computer Vision*, pages 237–248. Springer.

Carol Neidle, Ashwin Thangali, and Stan Sclaroff. 2012. Challenges in development of the American Sign Language lexicon video dataset

(ASLLVD) corpus. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 143–150, Istanbul, Turkey. European Language Resources Association (ELRA).

Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. 2019. Kerastuner.

OpenCV team, 2025. OpenCV 4.12.0 documentation: Color conversions. Accessed: 2026-03-22.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Wendy Sandler and Diane Carolyn Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 project: A database of lexical and phonological properties for 2,723 signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277.

Zed Sevcikova Sehyr and Karen Emmorey. 2022. The effects of multiple linguistic variables on picture naming in american sign language. *Behavior Research Methods*, 54(5):2502–2521.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Ariel E Stassi, J Matías Di Martino, and Gregory Randall. 2025. A brief review and analysis of two methods for automatic sign language segmentation. *Image Processing On Line*, 15:59–77.

Mukhiddin Toshpulatov, Wookey Lee, Jaesung Jun, and Suan Lee. 2025. Deep learning pathways for automatic sign language processing. *Pattern Recognition*, 164:111475.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*, pages 1556–1559.

## 10. Language Resource References

Naomi K Caselli, Karen Emmorey, and Ariel M Cohen-Goldberg. 2021. *The signed mental lexicon: Effects of phonological neighborhood density, iconicity, and childhood language experience*. Elsevier.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. *The ASL-LEX 2.0 Project: A database of lexical and phonological properties for 2,723 signs in American Sign Language*. Oxford University Press.

# "A Sacred Bird Called the Phoenix". Auditing the most-used Parallel Corpus for German Sign Language Recognition and Translation

**Vera Czehmann**[1,2], **Shakib Yazdani**[1], **Yasser Hamidullah**[1],
**Fabrizio Nunnari**[1], **Eleftherios Avramidis**[3,1]
[1]German Research Center for Artificial Intelligence (DFKI GmbH),
[2]Technische Universität Berlin, Berlin, Germany,
[3]alangu GmbH, Cologne, Germany
{vera.czehmann, shakib.yazdani, yasser.hamidullah,
fabrizio.nunnari, eleftherios.avramidis}@dfki.de, avramidis@alangu.de

**Abstract**

This paper presents an empirical audit of the widely used RWTH-PHOENIX-2014T corpus, examining its suitability as a benchmark for sign language recognition and translation. Through human annotation of the training set and extensive sign-to-text back translation of the test set, we provide detailed statistics that indicate substantial quality issues, including information loss and lexical errors. Automatic scores comparing human sign-to-text back translations to the original speech-transcribed references are remarkably low, suggesting strong translationese effects and substantial paraphrasing, revealing limitations of lexical metrics in adequately scoring translation quality. Replacing the original speech-transcribed references with human sign-to-text back translations while scoring existing sign language translation systems reveals the lack of robustness of system evaluation with lexical metrics against this test set. Our findings highlight risks associated with relying on this corpus for model evaluation and call for more rigorous, linguistically grounded evaluation practices in sign language technology research. The back-translated test set and error annotations are made publicly available.

**Keywords:** sign language, dataset, corpus analysis

## 1. Introduction

Sign language technology is in its early stages, and the research community has recently become increasingly active in its efforts to achieve significant progress. Automatic recognition (SLR) and translation of sign language (SLT) are two of the most promising, but also most challenging, application areas. Progress in these areas depends on state-of-the-art machine learning methods, which in turn require sound automatic evaluation protocols and high-quality test sets. Motivated by the belief that evaluation protocols should themselves be critically examined for robustness and validity, we investigate commonly used evaluation practices for research on sign language processing.

We focus on the most widely used test set in German sign language processing research, the RWTH-PHOENIX-2014T corpus (hereafter referred to as "PHOENIX"), which has been used as a benchmark for hundreds of research experiments. Despite its significance to the research community, several limitations of the corpus have already been noted in prior work. We revisit these concerns and provide a more detailed empirical analysis of the corpus, with a particular focus on the test set, its limitations, and the implications for evaluation practice. Beyond the corpus audit itself, we examine benchmark stability under original and sign-to-text back-translated reference conditions. Our goal is to raise awareness of potential weaknesses in current evaluation setups and to encourage more rigorous and linguistically grounded evaluation practices in sign language technology research.

In Section 2 we review relevant background, prior work and resources: we recap standard evaluation protocols, describe the RWTH–PHOENIX–2014/2014T corpus, summarise other published DGS datasets, and outline the main criticisms motivating our audit. The research questions and analysis design are presented in Section 3. The experimental setup, including sampling, human error annotation of the training set, test set back translation,



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 80–92
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

automatic scoring, and benchmark evaluation is described in Section 4. We then report quantitative findings in Section 5, and conclude with recommendations for evaluation practices and dataset use in Section 6.

## 2. Background

In this section we summarise common evaluation protocols in sign language translation, describe the RWTH-PHOENIX-2014/2014T corpus and related DGS resources, and review previously noted limitations that motivate our audit.

### 2.1. Evaluation Protocols

Current SLT development relies heavily on state-of-the-art machine learning methods and models trained on large amounts of parallel data (training sets), i.e. datasets that represent the same content in both source and target language. In machine learning, automatic evaluation protocols are used in order to provide quick feedback during development, and to guide decisions and settings for the creation of the models (Kohavi, 1995; Hastie et al., 2009). Such protocols typically consist of one or more automatic evaluation metrics, and a test set (Landy et al., 1978; Arlot and Celisse, 2010).

The test set is a parallel corpus that is not part of the training data, and therefore also referred to as a "blind" or "hold-out" set. The source side is provided to the trained translation model, and the resulting output is compared against the target side of the test corpus (reference translation), providing a numerical result indicative of the translation performance (Papineni et al., 2002; Rei et al., 2020). These scores allow comparison across models and can inform model selection, or, under comparative ablation analyses, indicate which settings or design decisions should be kept (Och, 2003; Demšar, 2006).

The use of such an evaluation protocol rests on the following assumptions: (a) that the automatic evaluation metric provides a quality judgment that approximates that of users (Callison-Burch et al., 2008; Mathur et al., 2020) and (b) that the test set has the necessary properties so that the findings of the evaluation generalise to the use cases relevant to the audience of the model or the performed analysis (Dwork et al., 2015). In this work, we focus on (b).

An important requirement of a test set is that it is itself of high quality. For machine translation, two properties are particularly important:

- the source content of the test set should be understandable in its own right, for example by a competent human translator, and
- the reference translations should be accurate translations of the source, since automatic metrics rely on them to provide meaningful statistics over model output (Freitag et al., 2020).

These aspects form the basis of our analysis.

### 2.2. The PHOENIX Corpus

RWTH-PHOENIX-Weather 2014/2014T (Stein et al., 2010; Forster et al., 2012, 2014) is a benchmark corpus for German Sign Language (DGS)-based continuous sign language recognition and translation. It is built from German public TV weather forecasts with simultaneous sign-language interpretation, and is widely used for SLR/SLT research and baselines. The corpus name derives from the name of the TV channel, *Phoenix*, which relays the news of the first channel of the public broadcaster with additional live sign language interpretation. The channel itself has been named after the mythical bird phoenix that rises from its own ashes,[1] symbolising a "media-political correction" and a fresh start for in-depth information.[2]

The signer videos are cropped to the interpreter inset ("interpreter box only") recorded at 25 fps with a frame size of 210×260 pixels, on a controlled grey background with signers wearing dark clothing to ease vision-based processing. The corpus provides sentence- and gloss-level annotations for continuous signing,

[1] Herodotus, Histories 2.73 "*There is also another sacred bird called the phoenix, which I did not myself see except in painting...*"; used for the title of the paper.

[2] History of the broadcaster, Phoenix website, retrieved on 15 Feb 2026.

as well as parallel spoken-language text. On the spoken-language side, the German text is a semi-automatic ASR transcript of the presenter's speech. An open-source speech recognition system was applied to the audio stream of the videos, and the recognition output was then manually corrected by native German speakers. Each SLT instance can therefore be represented as a triplet consisting of video, gloss sequence, and German text.

The latest release (2014T) contains roughly 0.95 million frames, >67k signs from a ∼1k-sign vocabulary, and >2.8k German word types, with standard train/dev/test splits, and the source comprises ∼386 weather editions recorded between 2009–2011. In addition to the corpus itself, the maintainers also distribute precomputed features, scoring tools and documented baselines for both multisigner SLR and SLT setups, which has contributed to making PHOENIX a de-facto standard for quantitative comparisons in sign-language processing.

Citation statistics[3] reveal that the corpus appears to have been used for more than 300 experiments since it was first published in 2010 (Stein et al., 2010), and it is still very actively used in recent publications. Since 2020, when experiments on machine learning over sign languages spiked, more than 30 experiments per year report their results on PHOENIX. Its use is stable over the years, as these numbers are maintained during 2025.

### 2.3. Other German Sign Language Datasets

To situate PHOENIX within the broader DGS resource landscape, we summarise other published datasets for German Sign Language. The flagship linguistic resource is the **Public DGS Corpus** (MY DGS–annotated; Hanke et al., 2020), released in several waves by the DGS-Korpus project with multi-portal access, rich manual annotations, and even pose tracks. Earlier DGS resources include **SIGNUM**, a signer-independent corpus with 25 native signers, 450 lexemes and 780 sentences (von Agris and Kraiss, 2010), and the **ATIS** Sign Language corpus that provides DGS alongside other languages within an air-travel domain (Bungeroth et al., 2008). Multilingual projects that include DGS comprise **Dicta-Sign**, a four-language corpus (BSL, DGS, GSL, LSF) with harmonised elicitation materials and multi-camera capture (Matthes et al., 2012). Recent additions include **DGS-Fabeln-1**, a parallel collection of German text and DGS fairy-tale interpretations recorded from multiple angles (Nunnari et al., 2024), and larger broadcast-based compilations such as the **TUB Sign Language Corpus Collection** (Avramidis et al., 2025). This overview shows that the prominence of PHOENIX in benchmark-driven SLT research should be understood against the background of other corpora with different properties and intended uses.

### 2.4. Prior Criticism of PHOENIX

Several limitations of PHOENIX have already been discussed in the literature, including by the creators of the corpus themselves. Stein et al. (2010) criticise the real-time interpreting situation, pointing to bias toward spoken German grammar, omissions in signed output, and limitations for high-quality linguistic analysis and translation evaluation. Forster et al. (2012) note the compactness of the weather domain, motion blur due to fast signing and the 25 fps frame rate, the amount of singleton glosses, and loose spoken–signed parallelism as issues for recognition and translation tasks. Forster et al. (2014) explicitly state that the simultaneous interpreting setup affects spatio-temporal organisation, differs from "real-life" deaf signer language, and that the corpus is not intended primarily for linguistic research. They also mention loose translations and out-of-vocabulary/signature sparsity.

Other work has raised related concerns. Bragg et al. (2019) survey SL datasets and highlight general limitations: small signer pools, constrained domains (such as weather/news), interpreted rather than naturally produced sign language, and mismatch with real-world language use, with PHOENIX as a central example. When reviewing sign language datasets, De Sisto et al. (2022) note that PHOENIX lacks annotation of non-manual features, whereas the fact that the

[3] Google Scholar and Semantic Scholar

annotation guidelines of the existing glosses are not publicly available may raise a transparency issue. Müller et al. (2023) discuss the live interpretation and translationese effects such as omission of information and influence from German grammatical structure, the tokenization and removal of punctuation from the spoken language text, and the glossing quality, stating that "from a scientific point of view achieving higher gloss translation quality on the PHOENIX dataset is near meaningless". Additionally, they note that PHOENIX is overused as a dataset, which is reminiscent of the overuse of MNIST (LeCun et al., 2010) in machine learning, and the WMT14 English-German test set (Bojar et al., 2014) in machine translation of text (Vaswani et al., 2017). A recent study of automatic SLT evaluation (Yazdani et al., 2026) uses PHOENIX to highlight the limitations of text-based metrics.

While not explicitly referring to PHOENIX, the *Sign Language Dataset Compendium* (Kopf et al., 2022; Schulder et al., 2025) excluded this dataset from its selection, possibly because it does not fulfill the curation criterion of containing natural non-interpreted sign language.

Prior SLT experiments run on both PHOENIX and the Public-DGS corpus (e.g. Angelova et al., 2022; Zhu et al., 2023) also indicate a large discrepancy between the automatic evaluation scores and the perceived translation quality, for models trained and tested on the two corpora, with the scores for Public-DGS being much lower than the ones of PHOENIX. This might point to the differences in language complexity and domain spread between the corpora, but also motivates a further inspection of PHOENIX.

## 3. Methods

In this study, we proceed with the following research questions:

**General quality issues of the corpus.** Whereas several quality issues have been highlighted in previous work, we aim to obtain systematic quantitative evidence about the quality issues of the corpus. Because the German text derives from broadcast subtitles/ASR text rather than a direct translation of the signing, we treat it as weak supervision and perform an **error annotation of the training set** based on a comparison of the glosses with the original German transcription (Sections 4.1 and 5.1). We focus on the training partition, since our goal is to identify corpus-level quality issues rather than properties specific to the test set alone. To keep the annotation effort manageable while still covering the corpus broadly, we perform the annotation on a structured sample. The annotation is inspired by the Translation Error Rate (TER; Snover et al., 2009, 2006) and uses the following error categories:

- information missing in the glosses,
- information missing in the German text,
- lexical errors, and
- minor differences that do not significantly affect translation adequacy.

The "reordering" error label of TER was excluded due to the structural difference in word order between spoken German and DGS.

**Quality of the test set.** Given the central role of test set quality in automatic evaluation, we additionally examine to what extent the signed content is understandable to a fluent DGS signer working independently of the corpus annotations. For this purpose, they produced a back translation of the test set by watching the videos without access to either the glosses or the German transcribed text (Sections 4.2 and 5.2). The aim is to approximate what a competent human viewer can recover from the signed videos alone, which is also the basic expectation placed on automatic systems evaluated on this material. To obtain an additional indicator of interpretability, the signer recorded their translation confidence for each item. Obvious technical problems in the video material were also flagged, since such issues may directly affect comprehensibility.

**Divergence between back translation and German transcription under automatic metrics.** We want to assess semantic preservation under the assumed conditions of noise and weak parallelism. Additionally, we want to know to what extent the automatic metrics can

measure this semantic preservation. We therefore perform **automatic scoring of the human translations using the original German text as a reference** (Sections 4.3 and 5.3). The automatic scoring is performed with the lexical state-of-the-art machine translation automatic metric BLEU (Papineni et al., 2002), commonly reported in SLT experiments. Additionally, we report ChrF++ (Popović, 2017), TER (Snover et al., 2009), and BLEURT (Zan et al., 2024) scores.

**Effect of sign-to-text back translation on scoring existing systems.** We conduct an additional evaluation of previously reported SLT system outputs under different reference conditions. While the previous question focuses on how a human signer's back translations compare to the official PHOENIX references, the present experiment asks a complementary question: whether comparative system evaluation remains stable when uncertain evaluation items are removed, and the reference side is grounded more directly on the signed videos.

## 4. Experiment Setup

For this evaluation, both the test set and the training set were manually analysed by a human annotator through error annotation and back translation, respectively.[4]

### 4.1. Error Annotation of the Training Set

The training set of PHOENIX consists of 7096 pairs of gloss sequences and corresponding German sentences. A structured sample of 307 sentence pairs was annotated by one deaf fluent L2 DGS signer, requiring 25 person-hours in total.

The annotated sentence pairs were selected by starting at the beginning of the training set and sampling blocks of ten to eleven consecutive items at regular intervals, first in steps of 100 up to sentence 1500, and then in steps of 500 until the end of the corpus. This sampling strategy was chosen to obtain a structured overview of gloss and text quality throughout the full training set while covering material associated with each of the nine signers.

### 4.2. Sign-to-Text Back Translation of the Test Set

The test set consists of 642 video files of nine sign language interpreters. As part of our analysis, all of these files were back translated, either fully or partially, by one deaf fluent L2 DGS signer. The signer spent 55 person-hours on this task. For each item, they recorded a translation confidence score of 0, 0.5, or 1, where 1 indicates high confidence, 0.5 indicates partial or medium confidence, and 0 indicates that no reliable back translation could be produced. In addition, each video was marked for the presence or absence of noticeable technical quality problems using a binary flag (0, 1).

### 4.3. Scoring of Human Translation

We compute the metrics BLEU, chrF++ and TER to score the human back translation against the original German text reference. Scores were computed with SACREBLEU (Post, 2018). We repeat the scoring for 3 different versions of the back translation: translations for which the signer reported high confidence, translations for which they reported at least medium confidence, and the full set of back translations (including cases that could only be translated incompletely or not at all).

### 4.4. Benchmark Evaluation with Original and Sign-to-Text References

We evaluated one gloss-based and three gloss-free systems spanning a broader range of quality. The evaluated systems were TwoStream-SLT (Chen et al., 2022), SpaMo (Hwang et al., 2025), SEM-SLT (Hamidullah et al., 2024), and Signformer (Yang, 2024). The original PHOENIX test set scores serve as a baseline over the full test set of 642 items.

In addition, we evaluated the subset of 462 items for which a back translation with confidence score 1 was available. For this high-confidence subset, system outputs were scored under three reference conditions: using the original PHOENIX references only, using the back-translated references only, and

[4]The material is publicly available at https://github.com/DFKI-SignLanguage/sacre-bird-phoenix

| Error type | sentences | percentage |
|---|---|---|
| Information missing in glosses | 97 | 31.6% |
| Information missing in text | 13 | 4.2% |
| Lexical errors | 33 | 10.7% |
| Minor differences | 165 | 53.7% |
| Total annotated parallel sentences | 307 | 100% |

Table 1: Error annotation of the training set

using both jointly in a two-reference setting.[5] This design allows us to separate three effects: the difference between the full benchmark and its subset having uncertain evaluation items removed, the effect of replacing the official reference with a sign-to-text back translation on the same items, and the effect of allowing both reference formulations simultaneously.

Before scoring, outputs and references were normalised in the same way in order to reduce superficial mismatches caused by formatting conventions. Specifically, all text was lowercased, a full stop was added at the end of each sentence, and numbers were written out in words following the conventions of the official PHOENIX German references. Automatic evaluation was then carried out with BLEU for all conditions and with BLEURT for the single-reference conditions. We focus on BLEU because it is the dominant metric in published PHOENIX-based work and therefore most directly reflects prevailing benchmark practice. BLEURT is included as a complementary semantics-oriented metric that is less dependent on exact lexical overlap. The two-reference condition was evaluated with BLEU only.

## 5. Results

We report findings from the corpus audit, including the manual annotation and back translation, automatic scoring, and benchmark results under different reference conditions.

### 5.1. General Quality Issues of the Corpus

The results of the error annotation on the training set can be seen in Table 1. Out of the 307 annotated pairs of gloss sequences and corresponding German text, substantial quality issues were found in a considerable proportion of them. The glosses were missing information as opposed to what was provided in the reference text in 31.6% of the pairs. In 4.2%, information present in the glosses was missing from the written German text. Lexical errors, such as mistranslations, could be found in 10.7% of the pairs, some of them as significant as different numbers or months. Only 53.7% of the pairs contained none or only minor differences that did not meaningfully affect translation adequacy.

The lexical errors varied in severity: some involved relatively minor meaning differences, such as good vs. exciting, clouded vs. heavy clouds, evening vs. night, rain vs. storm. Others were more consequential, including mismatches in numbers or place references, e.g. mostly vs. partially, seven vs. minus seven, North vs. Baltic Sea, and February vs. August[6].

Quality issues did not appear to be evenly distributed across the sampled material. Some sampled blocks associated with particular signer material appeared more affected than others, especially in the handling of locations, indices, and references to previous sentences. Among our 27 sampled blocks, there was only one annotated block of eleven consecutive pairs without considerable issues.

### 5.2. Back Translation of the Test Set

The back translation statistics for the test set are shown in Table 2. Out of the 642 video clips of the test set, 72% were back translated

[5]The two-reference setting was done only for BLEU as BLEURT does not support multiple references.

[6]There are signs for February and August that are identical in hand form and movement, but have different mouthings. Still, in PHOENIX, glosses and reference text refer to different months.

| Back translation status | sentences | percentage |
|---|---|---|
| Successful | 462 | 72.0% |
| Issues due to video quality | 78 | 11.2% |
| Not translated because of lexical issues | 17 | 2.6% |
| Incomplete due to unclear location signs | 26 | 4.1% |
| Total number of videos | 642 | 100% |

Table 2: Statistics on the back translation of the test set

with high confidence by the annotator. 11.2% of the videos were found to have at least one noticeable quality problem, such as signs being cut off at the beginning or end, poor video quality, or the signer's face or hand movements not being visible due to overlapping text.

The translator also reported difficulties in the back translation, such as the frequently unclear use of region-specific location signs in 4.1% of the videos, and dialectal differences between signers. An example of this was "Donnerstag" (Thursday) being signed differently by Signer01 and Signer05, and Signer03 and Signer05 using different signs for "Schnee" (snow). While dialectal and regional variation is a normal and expected property of DGS rather than an error or deficiency, this can affect the suitability of PHOENIX as a benchmark.

### 5.3. Automatic Scores of Back Translation against Original Text

The automatic evaluation scores of the back translation scored against the original German text can be seen in Table 3. Notably, scores for all 3 automatic metrics are at the very low-end of the scoring scale (max. 14.8 BLEU), considering that an identical text scores 100 BLEU and that good quality text translations in standard test sets go up to approximately 45 BLEU.

These very low scores indicate that a substantial amount of information is lost during the full translation pipeline. As mentioned, the German text is an automatic and manually corrected transcription of what is said by the speaking TV presenter, which is then interpreted in DGS, which is then edited in the corpus segments, which is then translated by our DGS signer. Even allowing for possible back translation errors by the DGS signer when operating at high confidence, there is likely a sequence of cumulative processing and translationese effects, including:

- possible errors introduced by an early-stage open source ASR system and not fully resolved during manual correction by native German speakers,
- considerable information loss during the interpretation process, as documented by the original authors (high time pressure),
- considerable information loss during the preprocessing/segmenting of the videos and their alignment with the German transcript, and
- considerable paraphrasing that occurred during the interpretation and the back translation process, which cannot be captured by lexical metrics.

Taken together, these factors point to what is perhaps the most important conclusion of this paper, that the practice of hundreds of papers to evaluate against this particular test set using solely lexical metrics should be reconsidered.

It is also worth noting that while the high-confidence human back translation scored 14.8 BLEU, the scores reported for automatic SLT systems are in the range of 17 to 30 BLEU (e.g. Table 4 column PH14T-subset). A superficial reading could misleadingly suggest that the automatic systems outperform the human back translation. However, this is an artefact of the evaluation protocol, as automatic systems may mimic the language style and punctuation of the German text (as seen in parts of our manual analysis), or may simply obtain more coincidental matches with the reference. We may also note that the effort of the translation model developers to increase the BLEU scores of their models, by iteratively

| human translation type | BLEU | chrF++ | TER | BLEURT |
|---|---|---|---|---|
| full back translation | 11.8 | 37.8 | 75.1 | 0.493 |
| mid-confidence back translation | 12.9 | 40.8 | 72.6 | 0.529 |
| high-confidence back translation | 14.8 | 43.6 | 70.4 | 0.573 |

Table 3: The human back translation of PHOENIX scored as hypotheses, using the official test set as a reference. Items below the confidence threshold are removed from both hypothesis and reference.

using PHOENIX as a development set, may result in not only learning how to translate from sign language to text, but also reproducing translation artefacts that are not relevant to translation quality.

As it can capture non-trivial semantic similarities, BLEURT appears to be less misleading than lexical-based metrics. However, it also produces scores indicating that automatic SLT systems perform equally well or better than human back translations.

### 5.4. SLT Benchmark Results across Reference Conditions

Results are shown in Table 4. First, the confidence=1 subset is systematically easier than the full PHOENIX test set. When evaluation is restricted to the 462 items with confidence=1 back translations while retaining the original PHOENIX references, all systems obtain higher BLEU and BLEURT than on the full 642-item benchmark. This is consistent with the fact that the subset excludes test items that a human annotator could not translate with high confidence and that SLT systems would likely also struggle to translate reliably. The subset differs in difficulty from the full test set and thus, the effects of subset selection and reference condition must be distinguished.

When the same 462 items are scored against the back-translated references instead of the original PHOENIX references, BLEU drops sharply for all systems, whereas BLEURT decreases far less. For example, TwoStream-SLT drops from 30.26 BLEU on the original reference subset to 13.39 BLEU on the back-translated reference subset, while BLEURT declines only from 0.626 to 0.587. The same pattern holds for the other systems. This suggests that the change in reference condition has a much stronger effect on lexical overlap than on semantic similarity. Put differently, system outputs remain substantially closer in meaning to the back translations than the BLEU drop alone would suggest. In addition, the confidence intervals are generally narrower under the back translation condition, most clearly for BLEU, suggesting more consistent score estimates across the evaluated items.

Overall, the confidence intervals of BLEU are so broad that no confident distinction can be made between some of the middle-ranked systems, whereas the confidence intervals of BLEURT are much narrower, indicating that relative system comparisons (ranking) made using this metric are valid. Additionally, system ranking based on BLEU is not fully stable across reference conditions. If one ignores the lack of statistical significance, on the original PHOENIX benchmark, and also on the confidence=1 subset scored with the original references, the scores suggest that SEM-SLT outperforms SpaMo in BLEU, whereas the order gets reversed once the same subset is scored against the back-translated references. Meanwhile, BLEURT appears to be more consistent across different settings and produces the same rankings as BLEU evaluated against back-translated references.

Finally, the two-reference condition yields the highest BLEU scores for all systems on the confidence=1 subset. This suggests that the back translations capture additional acceptable target realizations that are not adequately rewarded under a single-reference protocol. When both the official PHOENIX references and the back translations are treated as valid references, systems receive credit for outputs that would otherwise be penalised. This supports the interpretation that current PHOENIX evaluation is constrained both by the narrow-

| Model | BLEU | | | | BLEURT | | |
|---|---|---|---|---|---|---|---|
| | PH14T full | PH14T subset | Back Transl. subset | Multi-ref subset | PH14T full | PH14T subset | Back Transl. subset |
| TwoStream-SLT (gloss-based) | 28.2 ±2.1 | 30.26 ±2.39 | 13.39 ±1.12 | 34.89 ±2.22 | 0.597 ±0.013 | 0.626 ±0.010 | 0.587 ±0.010 |
| SpaMo | 22.2 ±2.0 | 24.87 ±2.64 | 10.59 ±1.43 | 27.91 ±3.04 | 0.542 ±0.015 | 0.574 ±0.013 | 0.543 ±0.013 |
| SEM-SLT | 23.7 ±2.1 | 27.27 ±2.32 | 8.89 ±1.47 | 31.01 ±2.39 | 0.484 ±0.012 | 0.511 ±0.016 | 0.448 ±0.013 |
| Signformer (S3D features) | 14.8 ±1.4 | 17.33 ±2.11 | 6.72 ±1.06 | 19.60 ±2.01 | 0.425 ±0.016 | 0.452 ±0.018 | 0.435 ±0.016 |

Table 4: Benchmark results across reference conditions. *PH14T full* denotes the full original PHOENIX test set (642 samples). All *subset* columns refer to the same 462-sample subset consisting of items for which a back translation with confidence=1 was available. *Multi-ref* uses both original and back-translated references.

ness of the reference side and by the sensitivity of lexical metrics to the exact wording of a single reference. However, the confidence intervals still do not allow a significant comparison between the two middle systems, and the observed order is different from the one suggested by the more robust BLEURT.

Taken together, these findings strengthen our main argument. The earlier human back translation experiment already showed that high-confidence human back translations receive unexpectedly low lexical scores when compared against the official PHOENIX references. The present system-level experiment extends that observation: benchmark results shift when systems are evaluated against back translations rather than against the original PHOENIX references, and they shift again when both are available. This suggests that current PHOENIX-based evaluation does not simply reflect how well systems recover the content of the signed videos, but is also influenced by the fact that the official references are tied to the original spoken-language source rather than to a direct translation of the signed material. Secondly, the lack of robustness of BLEU as a lexical metric, as compared to the semantics-aware BLEURT, is emphasised.

## 6. Conclusion

Our audit of the RWTH-PHOENIX-2014T corpus reveals substantial quality issues in both its training and test partitions, with systematic information loss, lexical inconsistencies, and signer-dependent variability that compromise its suitability as a benchmark for sign language recognition and translation. The low automatic scores obtained when comparing human sign-to-text back translations to the official speech-transcribed test references reveal that the test set suffers from strong translationese effects and other pipeline artefacts, and demonstrate that it does not provide a stable or linguistically faithful target for automatically evaluating systems, particularly when such evaluations rely solely on lexical metrics. The drop in scores observed when replacing the original speech-transcribed references with the human sign-to-text back translation indicates a lack of robustness of the lexical metrics and underlines that current PHOENIX-based evaluations may be too tied to the original spoken-language text, becoming unreliable in the presence of paraphrasing or reference errors.

Given the widespread use of the dataset and the risk of misleading system comparisons, we argue that future research should adopt more robust evaluation practices, incorporate human-centric and context-aware assessments, and prioritise datasets that better reflect natural sign language usage. We hope this work encourages the community to critically reassess the role of PHOENIX and to invest in the development of transparent, high-quality resources that support reliable scientific progress.

## 7. Acknowledgements

The research reported in this paper was supported by BMBF (now BMFTR, German Federal Ministry of Education and Research) through projects SocialWear (grant number 01IW20002) and BIGEKO (grant number 16SV9093), and by the European Union via the project SignReality, as part of financial support to third parties by the UTTER project (Horizon Europe, GA: 101070631).

## 8. Limitations

It should be acknowledged that while the back translations were produced by one fluent L2 signer, a stronger design would involve multiple native DGS signers in order to assess the robustness of the back translations. Also, our benchmark experiments cover only a small but deliberately heterogeneous set of systems, meaning that the observed ranking shifts should be read as evidence of sensitivity rather than as an exhaustive account of benchmark instability.

## 9. Ethical Considerations

We would like to note that this paper is by no means intended to be a criticism of the work of the original authors of the PHOENIX corpus, or the output produced by the interpreters. We strongly believe that the corpus has been a valuable resource for the community and has made an important contribution to scientific progress in our field. Additionally, we believe that live TV interpretation is necessary and valuable for the sign language community in terms of access to information, and the level of quality achievable under live on-air time pressure is widely accepted as fit for purpose. Here, we are focusing on the suitability of this content as part of strict evaluation protocols.

## 10. Bibliographical References

Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. 2022. Using Neural Machine Translation Methods for Sign Language Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 273–284, Dublin, Ireland. Association for Computational Linguistics.

Sylvain Arlot and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Eleftherios Avramidis, Vera Czehmann, Fabian Deckert, Lorenz Hufe, Aljoscha Lipski, Yuni Amaloa Quintero Villalobos, Tae Kwon Rhee, Mengqian Shi, Lennart Stölting, Fabrizio Nunnari, and Sebastian Möller. 2025. The TUB sign language corpus collection. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA. ACM.

Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. 2008. The ATIS sign language corpus. In *Proceedings of LREC 2008*, pages 2943–2946, Marrakech, Morocco. ELRA.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh

Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie LIU, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. In *Advances in Neural Information Processing Systems 35, (NeurIPS 2022)*, volume 35, pages 17043–17056. Curran Associates, Inc.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with Sign Language Datasets for Sign Language Recognition and Translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France. European Language Resources Association.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3785–3789, Istanbul, Turkey. European Language Resources Association (ELRA).

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be Guilty but References are not Innocent. In *Proceedings of EMNLP 2020*, pages 61–71, Online. Association for Computational Linguistics.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2024. Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–434, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in Size and Depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer, New York.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico. Association for Computational Linguistics.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of IJCAI 1995*, pages 1137–1143, Montreal, Canada.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. The Sign Language Dataset

Compendium: Creating an Overview of Digital Linguistic Resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association.

Frank J. Landy, Janet L. Barnes, and Kevin R. Murphy. 1978. Correlates of perceived fairness and accuracy of performance evaluation. *Journal of Applied psychology*, 63(6):751.

Yann LeCun, Koray Kavukcuoglu, and Clement Farabet. 2010. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of ACL 2020*, pages 4984–4997, Online. Association for Computational Linguistics.

Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. 2012. Dicta-Sign – building a multilingual sign language corpus. In *5th Workshop on the Representation and Processing of Sign Languages (LREC 2012 Satellite Workshop)*, pages 117–123, Istanbul, Turkey.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Fabrizio Nunnari, Eleftherios Avramidis, Cristina España-Bonet, Marco González, Anna Hennes, and Patrick Gebhard. 2024. DGS-Fabeln-1: A Multi-Angle Parallel Corpus of Fairy Tales between German Sign Language and German Text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4847–4857, Torino, Italia. ELRA and ICCL.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL 2003*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of EMNLP 2020*, pages 2685–2702, Online. Association for Computational Linguistics.

Marc Schulder, Thomas Hanke, and Maria Kopf. 2025. Making Sign Language Research Findable: The sign-lang@LREC Anthology and the Sign Language Dataset Compendium. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 277–288, Naples, Italy. Unior Press.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul, and Ralph Weischedel. 2006. A Study of Translation Error Rate with Targeted Human

Annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA. International Association for Machine Translation.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER?: Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece. Association for Computational Linguistics.

Daniel Stein, Jens Forster, Uwe Zelle, Philippe Dreuw, and Hermann Ney. 2010. RWTH-Phoenix: Analysis of the German Sign Language Weather Forecast Corpus. In *SignLang@ LREC 2010*, pages 225–230. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Eta Yang. 2024. Signformer is all you need: Towards Edge AI for Sign Language. Version Number: 1.

Shakib Yazdani, Yasser Hamidullah, Cristina España-Bonet, Eleftherios Avramidis, and Josef van Genabith. 2026. A critical study of automatic evaluation in sign language translation. In *Proceedings of the 15th Edition of the Language Resources and Evaluation Conference (LREC-2026)*, volume -. European Language Resources Association.

Changtong Zan, Liang Ding, Li Shen, Yibing Zhen, Weifeng Liu, and Dacheng Tao. 2024. Building Accurate Translation-Tailored LLMs with Language Aware Instruction Tuning.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural machine translation methods for translating text to sign language glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.

## 11. Language Resource References

von Agris, Ulrich and Kraiss, Karl-Friedrich. 2010. *SIGNUM Database: Video Corpus for Signer-Independent Continuous Sign Language Recognition*. ELRA, ISLRN 083-847-814-698-3.

# Diffusion-Based 3D Sign Language Motion Anonymization: A Feasibility Study on Balancing Identity Confusion and Semantic Preservation

**Zixuan Dai, Shinji Sako**
Graduate School of Engineering, Nagoya Institute of Technology
Nagoya, Aichi, Japan
cnq14068@ict.nitech.ac.jp, s.sako@nitech.ac.jp

**Abstract**

Sign language motions contain individual-specific kinematic features. As the engineering applications of sign language become more widespread, privacy protection of sign language data has emerged as a new challenge. This paper proposes a diffusion model-based approach for sign language motion anonymization. The proposed framework combines conditional diffusion processes with adversarial training to transform identity features while preserving semantic information. For the design and preliminary validation of the proposed model, we conduct a proof-of-concept experiment using a subset of 22 signers from the ASL100 dataset of WLASL, which demonstrates the feasibility of the proposed approach for sign language anonymization.

**Keywords:** Sign Language Anonymization, Diffusion Model, Privacy Protection, Identity Confusion, Transformer

## 1. Introduction

Sign language serves as a vital means of communication for deaf and hard-of-hearing communities worldwide, constituting an independent visual-spatial language system with its own grammatical structures and expressive modalities. In building a truly equitable and inclusive digital society, the needs and rights of the deaf community cannot be overlooked. While recent advances in deep learning have substantially improved the performance of sign language recognition systems, they have simultaneously given rise to new privacy risks. Sign language, as a complex visual-spatial language, not only conveys semantic information through handshapes, positions, and movements, but also unconsciously transmits rich personal behavioral characteristics. These dynamic features, including hand movement velocity patterns, trajectory curvature, joint coordination, and temporal rhythm, form distinctive "motion signatures" unique to each individual. At the same time, these individual motion characteristics constitute an important aspect of each signer's personal identity and self-expression.



Figure 1: Motion style variation across signers. Two different signers performing the sign "AIM" at the same relative frame position.

Respecting such individuality is essential. However, in certain contexts, such as public dataset sharing or anonymous communication, the ability to conceal these identifying features becomes equally important. Balancing respect for personal signing identity with the need for privacy protection presents a unique challenge, one that has become increasingly urgent in the current technological landscape.

In today's highly digitized and intelligent era, sign language users face unprecedented privacy risks. Modern computer vision models can extract and analyze subtle motion features from seemingly innocuous videos, raising significant privacy concerns for sign language communication. As the social demand for sign language grows, engineering applications related to sign language are expected to become increasingly active. Accordingly, sign language datasets are likely to become larger in scale and closer to real-life conditions, which in turn amplifies the urgency of privacy protection.

Several application scenarios illustrate the severity of this concern. In educational settings, student practice videos collected by online sign language teaching platforms may inadvertently expose learners' identities, learning progress, and common error patterns. Similarly, sign language communication in telemedicine consultations involves highly sensitive information, including health conditions and psychological states, with serious consequences if personal identities are revealed. Furthermore, training sign language AI systems requires large-scale datasets, yet using raw video data without adequate consent mechanisms may conflict with privacy protection regulations across countries. Grow-



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 93–99
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ing awareness of privacy rights and biometric data protection regulations necessitates that sign language AI research balance technological advancement with privacy protection. These challenges highlight the urgent need for effective sign language motion anonymization techniques.

Several studies have demonstrated that signers can be identified from their motion patterns alone. (Bigand et al., 2021) showed that kinematic features serve as reliable cues for signer identification in French Sign Language. (Dai and Sako, 2025) analyzed motion-based individualization patterns in Japanese Sign Language, confirming the cross-linguistic nature of kinematic identity features. Furthermore, human perception experiments have demonstrated that observers can identify signers at significantly above-chance accuracy even when facial information is completely concealed (Bigand et al., 2020). Figure 1 illustrates this phenomenon using samples from our dataset: two different signers performing the same sign exhibit visibly distinct motion characteristics in hand height, arm extension, and overall posture, despite conveying identical semantic content.

Conventional video anonymization techniques have primarily focused on facial processing (Perea-Trigo et al., 2025). While approaches using face-swapping techniques and virtual avatars have been proposed, these methods suffer from two fundamental limitations. First, as the findings of (Bigand et al., 2021, 2020) and (Dai and Sako, 2025) indicate, simply concealing the face cannot remove the motion signatures embedded in body movements. These identity-revealing features manifest through individual-specific motion patterns such as arm swing amplitude, wrist rotation patterns, shoulder tilt angles, and subtle differences in finger curvature. Second, facial expressions themselves constitute a crucial component of sign language grammar, conveying essential linguistic information through non-manual signals such as eyebrow movements, mouth patterns, and head tilts. Removing or replacing facial information therefore risks degrading the semantic content of the signed message. More recent work has further confirmed that signer identity can be reliably inferred from pose-level skeletal data alone (Battisti et al., 2024), and that pose-based appearance transfer methods can modify visual identity while attempting to preserve signing content (Moryossef et al., 2025). Taken together, these findings demonstrate that facial concealment alone cannot achieve true anonymization in sign language, and may even provide users with a false sense of security while simultaneously compromising linguistic integrity. This situation calls for a new technical approach capable of effectively removing identity features at the motion level while preserving the semantic content of the movements.

Recent advances in human motion anonymization have leveraged foundation motion models such as VPoser (Pavlakos et al., 2019) and HuMoR (Rempe et al., 2021), which learn latent representations of everyday movements like walking and sitting. By operating in these learned latent spaces, such approaches can generate biomechanically plausible anonymized motions. However, sign language presents fundamentally different challenges. Unlike general body movements, sign language functions as a linguistic system where fine-grained finger configurations and hand shapes directly determine semantic meaning. Moreover, no foundation motion model currently exists for sign language, making it infeasible to directly apply these latent-space anonymization techniques. This motivates our approach of designing constraint mechanisms tailored to sign language characteristics, focusing on transforming existing motions rather than generating entirely new ones from a general prior.

Diffusion models offer a promising framework for such controlled transformation. In recent years, they have achieved remarkable success in image and video generation tasks (Ho et al., 2020; Tevet et al., 2023), and their application to sign language video anonymization has also been reported (Xia et al., 2024). Their conditional generation capability suggests the potential to selectively remove identity-specific features while preserving semantic information. Motivated by these observations, this paper presents a feasibility study on diffusion-based 3D sign language motion anonymization, proposing a framework built upon a Transformer-based diffusion architecture. The proposed method consists of three main components: a multi-modal encoder for joint representation learning that integrates joint coordinates with semantic information, a Transformer-based diffusion module for feature transformation in latent space, and a decoder for reconstructing anonymized 3D motion sequences. In particular, we introduce a semantic conditioning mechanism for preserving linguistic content and an adversarial identity confusion mechanism for actively disrupting identity information, thereby seeking a balance between privacy protection and semantic preservation. Through experiments using a subset of the WLASL dataset involving 22 signers, we demonstrate that the proposed approach can achieve a meaningful degree of identity confusion while maintaining high semantic fidelity, confirming the feasibility of diffusion-based sign language motion anonymization.

The remainder of this paper is organized as follows. Section 2 describes the dataset and the proposed framework. Section 3 presents the experimental setup and results. Section 4 discusses limitations and future directions.

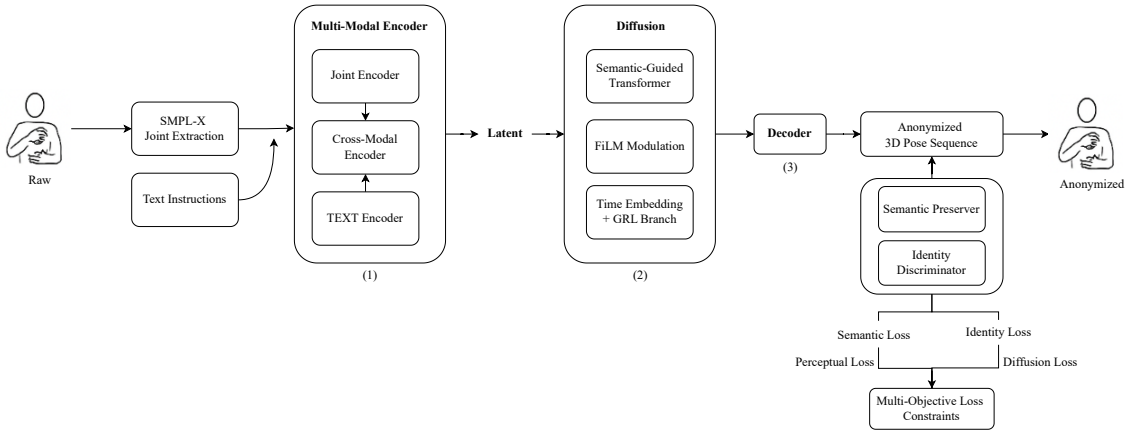



Figure 2: Overall Framework.

## 2. Methodology

### 2.1. Dataset

This study utilizes 3D joint coordinate sequences derived from the SignAvatars dataset (Yu et al., 2024), which applies 3D human body reconstruction techniques to video data from WLASL (Word-Level American Sign Language) (Li et al., 2020). WLASL is currently the largest word-level ASL recognition dataset, encompassing 2,000 sign language glosses recorded by over 100 signers in diverse environments. SignAvatars extends WLASL by reconstructing 3D human body models from the original video data through an automated annotation pipeline that performs hierarchical initialization followed by optimization with temporal smoothing and biomechanical constraints, yielding SMPL-X parameters from which sequences of 65 3D joint coordinates per frame (25 body, 20 left-hand, and 20 right-hand joints) are extracted. Since WLASL is a word-level dataset, each sample corresponds to an isolated sign rather than a continuous utterance. This 3D representation offers richer spatial information than 2D pose estimation, making it particularly suitable for analyzing the fine-grained motion characteristics relevant to both sign language understanding and anonymization.

While large-scale sign language datasets have been developed primarily for recognition and translation tasks, they present a structural challenge for anonymization research. In WLASL, although a large number of signers are represented, the distribution is highly imbalanced: a few signers contribute a large proportion of samples, while many signers appear only a handful of times with minimal overlap across glosses. This distribution is well-suited for vocabulary-level recognition tasks but makes it difficult to model individual-specific motion characteristics, which is essential for both identity modeling and anonymization evaluation. Moreover, no existing 3D sign language motion dataset has been specifically designed for anonymization research, and available resources remain far scarcer than their 2D video counterparts. Given these constraints, we select 22 signers from the ASL100 subset, each contributing a sufficient number of samples to enable meaningful identity modeling.

### 2.2. Framework Overview

Figure 2 illustrates the overall architecture of the proposed framework, which consists of three components: a multimodal encoder, a Transformer-based diffusion module, and a decoder.

The multimodal encoder takes as input 3D joint coordinate sequences (65 joints $\times$ 3 dimensions per frame) together with textual semantic information, and produces fused latent representations. Skeletal motion is encoded through a Transformer encoder that attends over all frames in the sequence to capture spatiotemporal dependencies, while semantic information is embedded via a pre-trained Sentence-BERT encoder (Reimers and Gurevych, 2019). In the current proof-of-concept implementation, the text input provides a fixed semantic category descriptor; extending this to per-sample gloss-level conditioning is planned for future work. A cross-modal attention mechanism integrates these heterogeneous modalities, producing latent representations that are both motion-aware and semantically informed. This joint encoding is essential because the downstream diffusion process must be able to distinguish between what a signer is saying and how they are saying it.

The core of the framework is the Transformer-based diffusion module, which operates on the encoded latent representations. A key observation motivating our design is that conventional anonymization techniques (such as geometric transformations or noise injection applied directly to joint coordinates) treat all motion features uniformly.

They lack the ability to selectively target identity-specific characteristics while leaving semantic content intact. Diffusion models offer a more principled alternative: by learning to iteratively denoise from a corrupted signal, they can be trained to generate motion sequences that satisfy specific constraints.

However, a standard unconditional diffusion process would modify all latent features indiscriminately, potentially degrading the linguistic content of the sign. To address this, we introduce a semantic conditioning mechanism that dynamically guides the denoising process based on the semantic information extracted by the encoder. This allows the model to be aware of which aspects of the motion carry linguistic meaning and should therefore be preserved throughout the transformation.

Semantic preservation alone, however, is not sufficient for effective anonymization. The model must also actively disrupt identity-specific features rather than merely hoping they are incidentally removed during generation. To this end, we incorporate an adversarial identity confusion mechanism using a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015). An identity classifier is attached to the diffusion module's output features; during backpropagation, the GRL reverses the gradient signal, encouraging the feature representations to become maximally uninformative with respect to signer identity. This creates an explicit adversarial objective: the classifier attempts to recover the original signer's identity, while the diffusion module learns to produce representations that confuse this classifier. Together, the semantic conditioning and adversarial confusion mechanisms establish a dual optimization objective: preserving what is said while obscuring who is saying it. The decoder reconstructs anonymized 3D joint coordinate sequences from the transformed latent representations, producing motion outputs that maintain the spatiotemporal structure of natural sign language.

The model is trained with a multi-objective loss function that combines diffusion denoising loss, reconstruction loss, semantic preservation loss, and adversarial identity confusion loss, with loss weights adjusted through a curriculum learning strategy to ensure stable convergence.

## 3. Experiments

To assess the feasibility of the proposed approach, we conduct preliminary experiments on the 22-signer subset described in Section 2.1. The signers are split into 15 for training, 4 for validation, and 3 for testing, yielding approximately 700 samples across 100 glosses. We evaluate the anonymization performance along three dimensions: identity protection (measured by the drop in identity classification accuracy after anonymization), semantic preservation (measured by cosine similarity between original and anonymized latent representations), and reconstruction quality (measured by root mean square error between joint coordinates).

While prior work has developed identity recognition models based on 2D pose or kinematic statistics (Bigand et al., 2021), no established model exists specifically for 3D sign language skeletal motion data. We therefore train a lightweight classifier based purely on skeletal dynamics features to capture each signer's unique motion patterns. This model achieves 81.7% accuracy on the validation set, substantially exceeding random chance, confirming that identity information is clearly encoded in the raw motion data. The same model serves dual purposes in our framework: providing quantitative evaluation of anonymization effectiveness and supplying adversarial training signals through the gradient reversal mechanism.

| Category | Metric | Before | After |
|---|---|---|---|
| Privacy | Identity Acc. | 81.7% | 38.0% |
| Utility | Semantic Sim. | 1.000 | 0.994 |
| Quality | RMSE | – | 0.064 |

Table 1: Evaluation results before and after anonymization.

Table 1 summarizes the evaluation results. After anonymization, identity classification accuracy drops from 81.7% to 38.0%, a reduction of 43.7 percentage points, demonstrating that the proposed method effectively disrupts identity-specific features. To evaluate semantic preservation, we employ a pre-trained graph convolutional network (GCN) for sign language recognition to extract motion features from both original and anonymized sequences, then compute their cosine similarity. The observed value of 0.994 suggests that 99.4% of the semantic information, as captured by the recognition model, is retained after anonymization. This high preservation rate indicates that the linguistic content of the signs remains largely intact despite the identity-related modifications. The RMSE of 0.064 measures the geometric deviation between original and anonymized joint coordinates, confirming that the overall motion structure is maintained.

Figure 3 visualizes the identity feature distributions using t-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction. In the original data (left), samples show a general tendency to form clusters by signer identity, with each color representing a distinct signer, though some inter-cluster overlap is present. After anonymization (right), the clusters significantly overlap, confirming that identity-specific features have been effectively obfuscated while the overall structure of

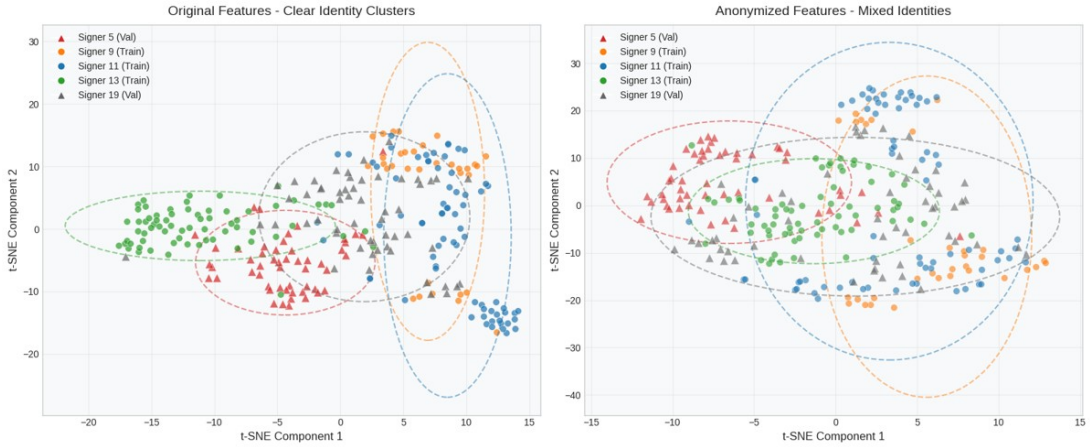



Figure 3: t-SNE visualization of identity features.

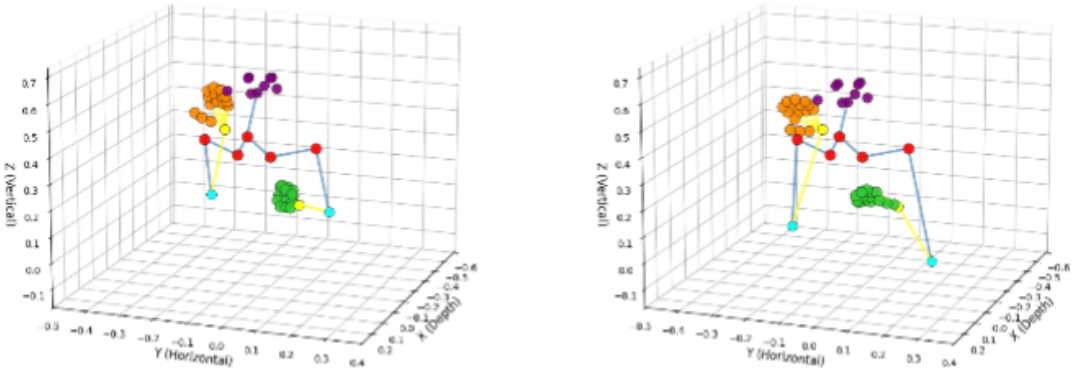



Figure 4: Visual comparison before and after anonymization.

the motion data is retained.

Figure 4 presents a qualitative comparison between original and anonymized motion for the same sign. The overall signing posture, hand configuration, and trajectory are preserved, maintaining semantic interpretability. At the same time, noticeable differences can be observed in hand elevation, arm extension angles, shoulder orientation, and subtle postural characteristics. These are precisely the regions where individual motion styles typically manifest. These modifications suggest that the model has learned to target identity-revealing features while leaving semantically relevant aspects largely intact. However, we note that formal validation by sign language experts remains necessary to confirm full semantic preservation from a linguistic perspective, which we leave for future work.

## 4. Conclusion and Future Work

This study presents a preliminary exploration of diffusion-based 3D sign language motion anonymization. While the results demonstrate the feasibility of balancing identity confusion with semantic preservation, several limitations should be acknowledged. The current evaluation is constrained by data scale. Our experiments involve only 22 signers. This limited scope may not fully capture the diversity of signing styles across broader populations. Furthermore, 3D sign language motion datasets remain scarce compared to 2D video resources, restricting opportunities for large-scale validation. Additionally, while quantitative metrics suggest high semantic preservation, formal validation by sign language experts has not yet been conducted, leaving open the question of whether linguistic meaning is fully retained from a native signer's perspective.

Several directions warrant further investigation.

Extending the framework to larger, more diverse datasets would enable better modeling of cross-demographic variations in signing styles. Adapting the approach to continuous sign language dialogue presents additional challenges, requiring preservation of paralinguistic features such as emotional expression and conversational dynamics. Extending evaluation to sign languages beyond ASL, such as Japanese Sign Language or Chinese Sign Language, could validate cross-linguistic generalizability and reveal universal patterns in gestural identity characteristics.

From a practical perspective, developing user-controllable anonymization levels would enable adaptive privacy protection based on specific use cases and risk profiles. Investigating the perceptual boundaries of identity recognition in sign language through human subject studies could further inform targeted anonymization strategies that minimize semantic disruption while maximizing privacy protection.

## 5. Acknowledgements

This research was supported by JSPS KAKENHI Grant Numbers: 23K11197, 23K17511, 25H00473, and 26K14865.

## 6. Bibliographical References

Alessia Battisti, Eveline van den Bold, Anja Göhring, Franziska Holzknecht, and Sarah Ebling. 2024. Person identification from pose estimates in sign language. In *11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 17–23. ELRA and ICCL.

Félix Bigand, Elise Prigent, and Annelies Braffort. 2020. Person identification based on sign language motion: Insights from human perception and computational modeling. In *Proceedings of the 7th International Conference on Movement and Computing (MoCo)*, pages 1–7. ACM.

Félix Bigand, Elise Prigent, and Annelies Braffort. 2021. Machine learning of motion statistics reveals the kinematic signature of the identity of a person in sign language. *Frontiers in Bioengineering and Biotechnology*, 9.

Zixuan Dai and Shinji Sako. 2025. Motion-based analysis of personalization and kinematic features in Japanese Sign Language video data. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, New York, NY, USA. Association for Computing Machinery.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1180–1189. PMLR.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Amit Moryossef, Gal Sant, and Zifan Jiang. 2025. Pose-based sign language appearance transfer. In *Proceedings of the Third International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–6. European Association for Machine Translation.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985.

Marina Perea-Trigo, Manuel Vázquez-Enríquez, Jose C. Benjumea-Bellot, Jose L. Alba-Castro, and Juan A. Álvarez-García. 2025. Sign language anonymization: Face swapping versus avatars. *Electronics*, 14(12).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. 2021. HuMoR: 3D human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11488–11499.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023.

Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*.

Zhaoyang Xia, Yang Zhou, Ligong Han, Carol Neidle, and Dimitris N. Metaxas. 2024. DiffSLVA: Harnessing diffusion models for sign language video anonymization. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages*, pages 395–407. ELRA and ICCL.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. 2024. SignAvatars: A large-scale 3D sign language holistic motion dataset and benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–19.

# GeoQuery-LSFB: A French Belgian Sign Language Corpus with Procedural Semantic Annotations

**Liesbet De Vos**[*], **Laurence Meurant**[‡], **Paul Van Eecke**[†◇], **Katrien Beuls**[*◇]

[*]Faculté d'Informatique, Université de Namur, Namur, Belgium
{liesbet.devos, katrien.beuls}@unamur.be

[†]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Belgium
paul@ai.vub.ac.be

[‡]NaLTT & LSFB-Lab,Université de Namur, Namur, Belgium
laurence.meurant@unamur.be

**Abstract**

Procedural semantic representations describe the meaning of natural language expressions in terms of computer programs that can be evaluated against images, databases, knowledge graphs or other external resources. While resources annotated with procedural semantic representations already exist for a variety of spoken languages, such resources are still lacking entirely for signed languages. In this paper, we introduce GeoQuery-LSFB as a signed language extension to the multilingual GeoQuery corpus. Concretely, we have complemented each procedural semantic annotation from the original corpus with a corresponding French Belgian Sign Language (LSFB) expression that was phonetically transcribed from video recordings following the HamNoSys convention and annotated with French ID-glosses. The GeoQuery-LSFB corpus constitutes a new resource for a low-resource language and offers for the first time the possibility to study, from an onomasialogical perspective, a signed language along a diverse variety of spoken languages.

**Keywords:** datasets for low resource languages, French Belgian Sign Language, multilingual corpora, NLP datasets, procedural semantics

## 1. Introduction

Procedural semantic representations describe the meaning of natural language expressions in terms of computer programs that are compositionally structured and can be evaluated by a machine (Woods, 1967; Johnson-Laird, 1977; Woods, 1981; Winograd, 1972; Woods, 2010). A defining property of procedural semantics is that the evaluation of semantic representations involves their grounding in some kind of 'world model'. Such a model can range from a database or knowledge graph, through a quantitative or qualitative simulation, to the real world as perceived through a robot's sensory system. The grounded and compositional nature of procedural semantic representations is of great interest to a variety of natural language understanding tasks, including database querying (Zelle and Mooney, 1996; Kwiatkowski et al., 2010; Berant et al., 2013; Liang, 2016; Dong and Lapata, 2016; Cheng et al., 2019), visual question answering and dialogue (Andreas et al., 2016; Johnson et al., 2017b; Hudson and Manning, 2019; Verheyen et al., 2023), and robot instruction (Bollini et al., 2013; Misra et al., 2016; van Trijp et al., 2024).

While a number of resources that annotate natural language expressions with procedural semantic representations already exist today, most of these resources exclusively include English data (Hemphill et al., 1990; Zelle and Mooney, 1996; Kuhlmann et al., 2004; Zettlemoyer and Collins, 2005; Chen and Mooney, 2008; Tasse and Smith, 2008; Johnson et al., 2017a; Hudson and Manning, 2019; Nevens et al., 2024). For other spoken languages, resources that come with procedural semantic annotations exist much more scarcely, but some corpora are available for German (Jones et al., 2012; Gross et al., 2018), Chinese (Lu and Ng, 2011), Spanish, Japanese and Turkish (Wong and Mooney, 2006), Greek and Thai (Jones et al., 2012), and Indonesian, Swedish and Farsi (Susanto and Lu, 2017). All non-English resources, with the exception of Gross et al. (2018), extend the original English GeoQuery corpus (Zelle and Mooney, 1996) with translations, resulting in a parallel corpus that includes a typologically rather diverse selection of languages. When it comes to signed languages however, no corpora annotated with procedural semantic representations exist to date.

In this paper, we introduce the GeoQuery-LSFB corpus as a signed language extension to the multilingual GeoQuery corpus. Concretely, this new resource complements each procedural semantic annotation from the original corpus with a correspond-

[◇]Joint last authors.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 100–112
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ing French Belgian Sign Language (LSFB) expression. Based on video recordings that feature a native LSFB signer, the expressions were annotated with French ID-glosses (Johnston, 2008, 2010) and phonetically transcribed following a time-aligned, multilinear extension to the Hamburg Notation System (HamNoSys) convention (Hanke, 2004). Apart from the 250 utterances in the original parallel corpus, we also present an augmented version of the corpus that covers 4519 utterances and thereby better fits today's data-intensive processing methods.

The GeoQuery-LSFB corpus constitutes a significant contribution to the linguistic resource landscape in three main respects: first of all, it is the first resource that aligns signed language expressions with procedural semantic representations. The corpus thereby facilitates research into language processing technologies that could so far only be developed for spoken languages. Second, it adds valuable corpus data to the limited resources that are currently available to support the study of LSFB. Finally, the parallel and semantically annotated nature of the corpus offers for the first time the possibility to study, from an onomasiological perspective, a signed language along a diverse variety of spoken languages.

The GeoQuery-LSFB corpus was released under the GNU General Public License 2.0 and is available for download at https://doi.org/10.5281/zenodo.19220833.

## 2. Background and Related Work

### 2.1. French Belgian Sign Language Resources

LSFB (*Langue des signes de Belgique francophone*) is used by the deaf and hard-of-hearing community within the French-speaking Community of Belgium. The total number of LSFB users (including deaf, hearing and hard-of-hearing users) is estimated at 20,000 of which 4,600 are first language users[1].

As a sign language, LSFB is produced and comprehended through the visual-gestural modality. This means forms are produced using articulators of the upper body (e.g. hands, head, face, shoulders) and comprehended using the visual perception system (Meier, 2002, 2012). This contrasts with spoken languages, which are produced in the aural-oral modality. Although both types of languages share structural properties, they also differ in some regards. A first point of divergence is the degree of iconicity between linguistic forms and their meanings. Iconicity is present when the phonetic form of a sign or word is motivated by a perceptible property of its referent (e.g. what it looks or sounds like) (Taub, 2012). While all human languages include iconic forms, the resources provided by the visual-gestural modality for iconicity are considered to be more abundant than those offered by the aural-oral modality. Another difference is the temporal organisation of linguistic units. Spoken language is primarily organised sequentially, meaning that linguistic units follow each other in a linear order. In contrast, sign languages contain more multilinear structures in which information is organised simultaneously. Finally, sign languages make extensive use of the signing space (i.e. the space surrounding the upper body of the signer) to manage referents and their relationships.

Phylogenetically, LSFB is closely related to Vlaamse Gebarentaal (VGT), the sign language of the Flemish Community of Belgium. Both languages are historically rooted in the Old French Sign Language (VLSF) and were considered under the umbrella of Belgian Sign Language until the 1970s. LSFB received recognition as an official language in the French-speaking Community of Belgium in 2003.

While significant efforts in creating linguistic resources that can be used for LSFB research, education and language technology development are currently ongoing, LSFB for now remains a low-resource language. The most extensive resource today is the LSFB Corpus (Meurant, 2015; Meurant et al., 2016), which consists of 90 hours of unscripted yet task-moderated video-recorded conversations between pairs of signers. French ID-gloss annotations (Johnston, 2008) have been made available for 25 hours of the corpus material. An overview of all ID-glosses that were used to annotate the corpus, along with French translations and isolated video recordings of the corresponding LSFB signs, has been released as the lex-LSFB lexical database (Meurant et al., 2015).

The LSFB online dictionary (https://dico.lsfb.be) provides LSFB video translations for over 5000 French words and a tool for translating LSFB signs into French words was made available by Fink et al. (2022). The LSFB Corpus was made available to the machine learning community through the release of two datasets, one of which wraps the original corpus data (LSFB-CONT) while the other groups occurrences of the most frequently used signs (LSFB-ISOL) (Fink et al., 2021). A spoken Belgian-French counterpart to the LSFB Corpus, where participants were asked to carry out the same conversational tasks, is currently under construction (Lepeut et al., 2024).

[1]The overall number of LSFB users is estimated at 0.44% of the total population following Haeusler et al. (2014). The number of first language users is estimated at 0.1% of the total population following Pasikowska-Schnass (2018).

## 2.2. Sign Language Transcription

On a high level, sign language transcription systems can be grouped into two categories. The first category transcribes each sign within a signed utterance using a gloss, i.e. an identifier for the sign that essentially corresponds to one or more words from the ambient spoken language[2], or to a numerical code. The use of descriptive spoken language words within the label (e.g. THANK-YOU for a sign that expresses gratitude) is generally preferred over opaque numerical labels, as it helps annotators remember the correspondences between signs and their labels. It is important to note, however, that sequences of glosses should not be confused with spoken language translations, since they retain the grammatical structure of the signed language. Of particular interest are ID-glosses (Johnston, 2008, 2010), where contextually different variants of a sign, i.e. variants with the same form but a different meaning, are grouped under the same gloss. (ID-)glosses do not contain any information about the form of a sign itself and are heavily dependent on a spoken glossing language.

The second category of writing systems transcribe signed utterances phonetically or phonologically. Systems in this category do not rely on any external glossing language, but transcribe the form components of signs directly. Phonetic/phonological transcription systems for signed languages were pioneered by Stokoe (1960) for American Sign Language (ASL). The Sutton SignWriting (Sutton, 1995) and HamNoSys (Prillwitz et al., 1987; Hanke, 2004) systems are currently the most widely used language-agnostic transcription systems.

SignWriting represents manual and non-manual components of signs in 2D space using an alphabet of 652 base symbols[3]. HamNoSys provides a linear, phonetic representation of sign components using a more restricted inventory of just over 200 base symbols (Smith, 2013). More recently, the Typannot system for handshape annotation was introduced with the goal of reaching a HamNoSys-like phonetic system while achieving a SignWriting-like readability (Rébulard et al., 2018), but is still to find a more widespread adoption.

Gloss-based transcriptions and phonetic transcriptions are very different in nature, have different properties and serve different purposes. Gloss-based transcriptions are, relatively speaking, less time-consuming to annotate. Their grounding in an ambient spoken language naturally supports their use in domains where the relationship between the signed and the spoken language is of significant interest, such as (machine) translation and different forms of sign language education. Phonetic transcriptions on the other hand are a valuable resource for linguistic research as they allow for a representation of signed forms that do not need to be cast into linguistic theories that were developed for spoken languages. As noted by Hodge and Crasborn (2022), it can be considered good practice in corpus development to incorporate different levels and types of annotation. In this way, the corpus becomes multi-purpose and is thereby likely to better stand the test of time.

[2]The spoken language predominantly used within the same geographic region as the relevant signed language.

[3]information retrieved from the Sutton SignWriting computer manual: https://www.signwriting.org/archive/docs7/sw0636_SignWriting_Alphabet_Manual_2010.pdf

## 2.3. Procedural Semantics

The procedural approach to semantics was introduced in the field of artificial intelligence in the 1960s and 1970s by Woods (1967), Winograd (1972) and Johnson-Laird (1977). The innovative idea behind this approach was that the meaning of a natural language expression can take the form of a program that can be evaluated by a machine. The evaluation of such a program involves its grounding in a 'world model', i.e. an external resource such as a database, knowledge graph, sensory observation, or simulation. Let us consider an example drawn from the GeoQuery resource, which will be described in more detail in Section 2.4. The world model consists of a database of geographical facts, including the area and population of cities and states, the height of mountains, the length of rivers, and the states and cities rivers run through. Procedurally, the meaning of a question like "*How big is Texas?*" could be represented through a query that, when evaluated against the database, would return the area of the state of Texas. Importantly, the answer itself is irrelevant when it comes to representing the meaning of the question. If a new treaty is signed and the world model is consequently updated, a query that accurately captured the meaning of the question would still return the correct answer, even if the answer is no longer the same. Likewise, a program that adequately represents the meaning of the statement "*The Mississippi river runs through Iowa*" should evaluate to true in every world where Iowa is among the states the Mississippi river runs through, and evaluate to false in any other world.

The procedural approach to representing meaning subscribes to the Davidsonian view that the meaning of an utterance corresponds to its truth conditions (Davidson, 1967). A complete understanding of an utterance amounts to being able to determine its truth value in every possible world, a task that requires an exact knowledge of the conditions under which the utterance is true. In terms of procedural semantics, the programs that represent

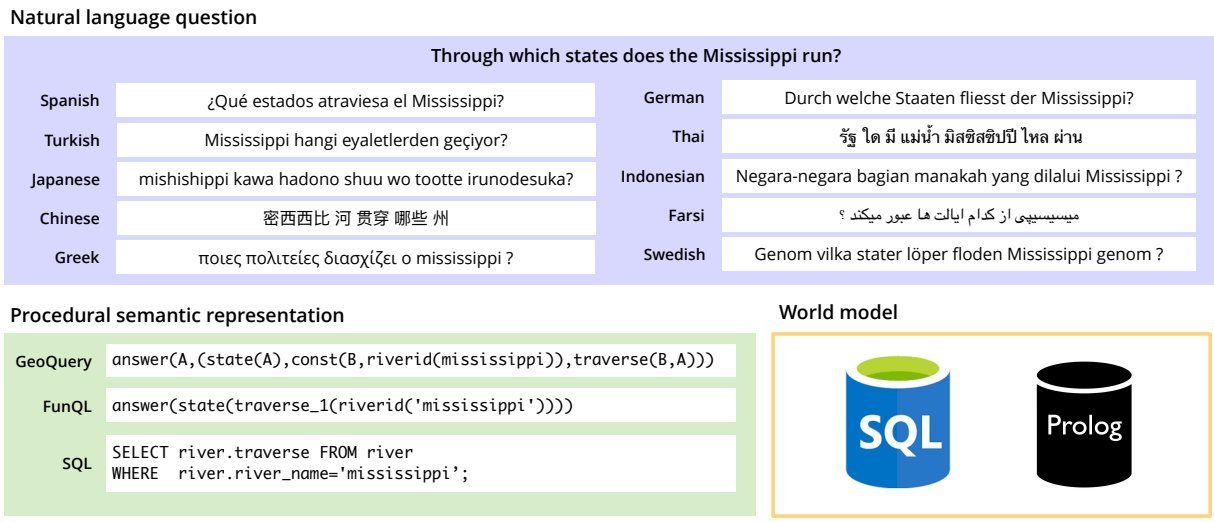



Figure 1: Schematic representation of a single entry in the multilingual GeoQuery corpus, featuring the original English question, its translation into ten other spoken languages, its procedural semantic representation in the GeoQuery, FunQL and SQL languages, and the relational and logic databases that constitute the world model

the meaning of natural language utterances thus correspond to their truth conditions. The process of evaluation then corresponds to the task of determining the truth of an utterance in a particular world[4]. In general, there are many ways in which procedural semantic programs can be formalised and in which their evaluation can be implemented, and the choice will typically depend on the task at hand and the nature of the world model that is provided (see e.g. Verheyen et al., 2023).

Despite the terminology that is used, procedural semantic programs almost never take the form of a sequence of operations to execute. Instead, they state the logic underlying the computation they are supposed to represent, for example through a conjunction of predicates. As long as a procedural interpretation can be linked to their logic interpretation, the programs are considered procedural semantic representations. In this paper and the associated GeoQuery-LSFB corpus, all procedural semantic representations will take the form of logic expressions that correspond to machine-evaluatable database queries.

The ability to evaluate procedural semantic representations against an external source sets them apart from the declarative, non-executable representations used in other sign language applications, such as the Discourse Representation Structures (DRS) used by Marshall and Sáfár (2004), Elliott et al. (2008) and Sáfár and Glauert (2010) or the logical structures from Role and Reference grammar (RRG), used by Murtagh (2011) and Murtagh et al. (2022).

## 2.4. The GeoQuery Corpus

The GeoQuery corpus was originally introduced by Zelle (1995) and Zelle and Mooney (1996) as a benchmark dataset for the supervised learning of semantic parsers. In its original form, it consisted of 250 English questions[5] about US geography that were collected from undergraduate students at the Department of Computer Sciences of the University of Texas at Austin. Each question was annotated by a human expert with a logical query that would answer the question upon evaluation against the GeoBase database of US state geography (Borland International, 1988). The corpus was extended to 880 question-query pairs by Tang and Mooney (2001) and Tang (2003) using data collected from users of the GeoQuery web interface.

The logical queries annotated in the original dataset straightforwardly integrate with logic programming languages such as Prolog. In order to better accommodate semantic parsers written in other programming languages, Kate et al. (2005) introduced FunQL, a functional, variable-free query language for the GeoQuery domain. A set of scripts allows for automatic conversion between the annotated logical queries and their FunQL counterparts (Kate et al., 2005). Popescu et al. (2003) and Iyer et al. (2017) introduced a relational database

[4] In Fregean terminology, the programs would correspond to the *Sinn* of an utterance, while the evaluation process would reveal its *Bedeuting* (Frege, 1892).

[5] We refer to the GeoQuery utterances as *questions* as they constitute requests for information to be answered. Grammatically, these requests can be expressed through interrogative sentences (e.g. '*What is the city with the smallest population?*') but also through imperative sentences (e.g. '*Give me the biggest city in Wisconsin*').

schema for the Geobase database and manually annotated all GeoQuery questions with SQL queries.

The first multilingual version of the GeoQuery corpus was introduced by Wong and Mooney (2006), who provided Spanish, Japanese and Turkish translations of the original GeoQuery questions. Later, translations of all questions in the extended GeoQuery dataset were added for many other languages, in particular for Chinese by Lu and Ng (2011), for German, Greek and Thai by Jones et al. (2012) and for Indonesian, Farsi and Swedish by Susanto and Lu (2017).

Figure 1 shows a single entry in the multilingual GeoQuery corpus. The original English question "*Through which states does the Mississippi run?*" is shown on top along with its translation into the 10 other languages of the dataset. The procedural semantic representations that correspond to the questions are shown in the bottom-left part of the figure. Note that the logical query and the FunQL query are equivalent, but that the SQL query was constructed independently. A schematic representation of the 'world models' is shown in the bottom-right part of the figure. The SQL meaning representation can be evaluated against the relational database, while the GeoQuery and FunQL meaning representations can be evaluated against the Prolog factbase.

## 3. Data Collection and Transcription

The overall corpus creation task consisted of annotating the 250 utterances of the original GeoQuery corpus with corresponding LSFB expressions. Given that LSFB is not a written language, one cannot simply ask a native signer to write down LSFB translations of the English sentences. Our native LSFB informant was however experienced in signing LSFB expressions that semantically correspond to utterances written down in the ambient French language. The first stage in the corpus creation process therefore consisted of making video recordings of LSFB expressions. These recordings could then be transcribed both phonetically and in terms of ID-glosses by an expert transcriber in a second stage.

In order to optimise the workload of the LSFB informant and transcriber, the first challenge was to select the minimal number of expressions that needed to be signed, video-recorded and transcribed. We started from the 221 unique meaning representations in the corpus and reduced them to 95 schemata by replacing named entities by their higher-level entity type (<state>, <river>, <city> or <capital>). For each of these schemata, the first corresponding utterance from the GeoQuery dataset was translated into French and provided to the signer. The signer was instructed to interpret the meaning of the French utterance and sign a corresponding LSFB expression. For meaning representations that occurred multiple times in the corpus with the exact same named entities, the signer was asked to sign the same number of variants. The signed utterances were video-recorded, yielding a total of 124 recordings.

The expert transcriber annotated the recordings first with French ID-glosses and then with HamNoSys-based phonetic transcriptions. ID-gloss annotation layers were created for the left and right hands using glosses from the lex-LSFB lexical database (Meurant, 2015). During the annotation process, five new ID-glosses were created for namesigns that do not occur in the database (NS:CALIFORNIE, NS:FLORIDE, NS:TEXAS, NS:DES-MOINES, and NS:MEXIQUE), as well as four new regular ID-glosses (KILOMETRE-CARRE, TAILLE, ME-DIRE and SERRER). Depicting and pointing signs were annotated following the guidelines established by Johnston (2010, 2016). Fingerspelled named entities were transcribed as such and variants of existing ID-glosses were tagged following Johnston (2010)'s handshape categorisation. Phonetic transcriptions were made using a time-aligned, multilinear extension to the HamNoSys convention. Concretely, a separate HamNoSys annotation layer was added for each hand. The segmentation of the layers was copied from the corresponding ID-gloss layers, temporally aligning the HamNoSys and ID-gloss transcriptions. Two-handed signs, which HamNoSys transcribes using a single linear expression, are annotated in the layer for the dominant hand, in which case the corresponding segment in the layer for the other hand is left empty.

An example transcription is shown in Figure 2 for the original utterance "*What are the high points of the states surrounding Mississippi?*". The annotation layers associated to the signer's dominant right hand are divided into eight time-aligned segments, one for each ID-gloss that was identified. The signer opens with the PALM-UP sign, a discourse marker that indicates that the speaker is thinking (Gabarró-López, 2017). This sign is followed by the DANS sign (English: IN) and the fingerspelled proper name Mississippi (FS:MISSISSIPPI, where FS stands for 'fingerspelling'). Then, the IL-Y-A sign is annotated (English: THERE-IS), followed by a depicting sign DS(BENT5):ETAT+, where DS stands for 'depicting sign', BENT5 for the handshape categorised according to Johnston (2016) and + signals that the sign is repeated multiple times. Finally, the pointing sign PT:DET/LOC(1)+ is annotated, where PT stands for 'pointing sign', DET/LOC for 'determiner or location', 1 for the used handshape and + for repetition, followed by the ID-glosses HAUT (English: HIGH) and QUOI (English: WHAT). Note that

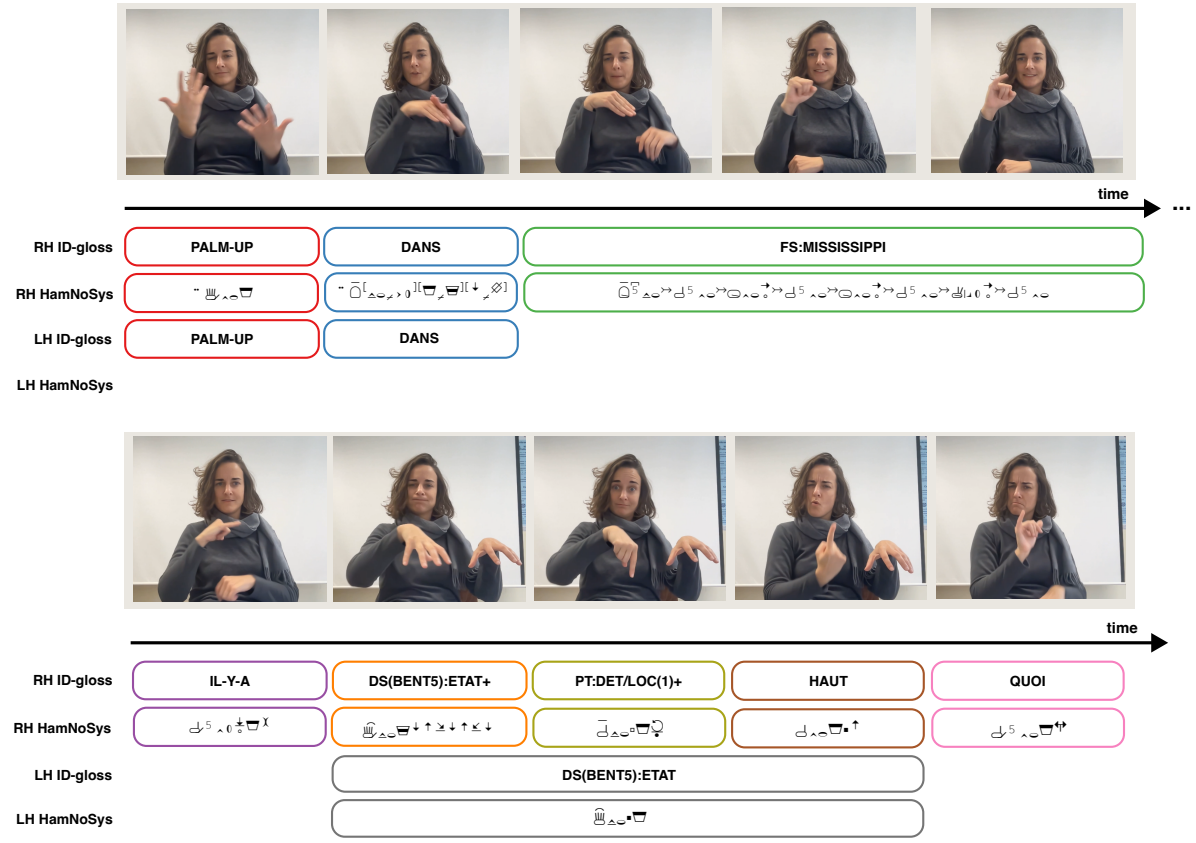



Figure 2: An example transcription of a video fragment in which the LSFB expression that corresponds to the English utterance *What are the high points of the states surrounding Mississippi?* is signed. Tiers for right-hand ID-gloss, right-hand HamNoSys transcription, left-hand ID-gloss and left-hand HamNoSys transcription are shown

the circular motion of the pointing sign is included in its HamNoSys transcription, but is reduced to 'repetition' in the ID-gloss annotation.

The annotation layers for the left hand hold three segments, two of which are ID-gloss annotations for two-handed signs (PALM-UP and DANS). For these segments, the HamNoSys transcription layer for the left hand remains empty, as the symmetry operator ¨ is used in the corresponding right-hand HamNoSys segment. The third left-hand segment concerns a depicting sign referring to a state-like entity that is again signed using Johnston (2016)'s handshape BENT5 and which coincides with three right-hand segments.

After the transcription of the 124 video recordings was completed, the resulting annotation layers were extrapolated to the 250 utterances of the original corpus. This could be done straightforwardly by substituting the named entity segments that were annotated in the recordings with segments containing the namesigns or finger-spelled transliterations of the named entities that occur at the same location in the remaining utterances. A separate, larger corpus of 4519 annotated utterances was also created by extrapolating the transcriptions of the video recordings to all named entities in the GeoBase database, enforcing the constraint that states, rivers, cities and capital cities can only be substituted with an entity of the same type.

As indicated above, the French ID-gloss annotations and HamNoSys transcriptions were contributed by a single expert transcriber. As a result, no extensive inter-annotator agreement study could be performed. While we acknowledge this as a limitation of the resource that we created, we would very much like to emphasize the difficulty of recruiting and training annotators for a low-resource sign language like LSFB. The annotator that was recruited is to the best of our knowledge the only person sufficiently proficient in both LSFB and HamNoSys to perform this task reliably. When it comes to the French ID-gloss annotations, a small-scale inter-annotator agreement study was performed in the pilot phase of the corpus creation process in order to validate its feasibility. This pilot study yielded a Cohen's kappa of 0.83 on 10 corpus entries selected to be maximally different. This result led to the green light needed to initiate the corpus creation process, but cannot be reported as such as a measure of the quality of the final resource.

All annotations were created using the ELAN software toolkit (Crasborn and Sloetjes, 2008)[6], the HamNoSys input palette[7] and the CWA Signing Avatars SiGML Player[8].

## 4. The GeoQuery-LSFB Corpus Release

The GeoQuery-LSFB corpus was released under the GNU General Public License 2.0 and can be downloaded from https://doi.org/10.5281/zenodo.19220833. The release consists of the following parts:

**Documentation** A user guide that documents the contents of the release, especially with respect to using the corpus for research purposes. The documentation provides a detailed description of all aspects of the corpus, including the required installation procedures, the labels used during the transcription process and the data structures that were used.

**Video recordings** All 124 video recordings, encoded using MPEG-4 compression. The videos amount to a total length of 21 minutes and 44 seconds, which corresponds to an average of 10.35 seconds per video.

**HamNoSys font** A version of the Menlo font that was extended with the HamNoSys character set[9].

**Annotation files** The 124 raw ELAN annotation files in `.eaf` format, as well as their corresponding configuration files in `.psfx` format.

**Corpus files** The actual corpus files, which hold for each entry:

- a unique identifier,
- the original English utterance,
- a reference to the higher-level semantic schema of which the meaning of the utterance is an instantiation,
- a reference to the corresponding ELAN file and video recording (where applicable),
- translations of the utterance into Spanish, Turkish, Japanese (Romaji), German, Greek, Thai, Indonesian, Swedish, Chinese and Farsi,
- the French translation of the utterance that was provided to the LSFB signer,
- procedural semantic representations of the utterance in the GeoQuery, FunQL and SQL formats,
- our integrated representation of the ID-gloss and HamNoSys annotations of the corresponding LSFB expression.

Apart from the unique identifiers and references to the higher-level semantic schemata, ELAN files and video recordings, we contributed the French and LSFB annotations and adopted all other annotations from prior versions of the corpus as documented in Section 2.4 above. The corpus files are provided in both the JSON and XML formats, with both versions containing the exact same information.

**Database files** Scripts for executing and evaluating the procedural semantic representations of the utterances in GeoQuery, FunQL and SQL formats, including Iyer et al. (2017)'s script for creating and populating a relational database. Instructions on how to use the execution and evaluation scripts are included in the user guide.

Overall, the transcription layers of the GeoQuery-LSFB corpus for the dominant right hand contain 2681 segments, while those for the left hand hold 1258 segments. 2794 segments (70.93%) were annotated with lexicalised ID-glosses. 143 distinct lexicalised ID-glosses occur, with the most frequent ones being either generally frequent in LSFB, such as UN (English: A/ONE) and IL-Y-A (English: THERE-IS), or being related to the domain of the dataset (e.g.VILLE (English: CITY) and RIVIERE (English: RIVER)). 595 segments (15.11%) were annotated with pointing sign glosses and 380 segments (9.65%) were annotated with depicting sign glosses. The remaining 170 segments (4.32%) were annotated as finger-spelled proper names.

## 5. Potential Uses of the Corpus

The GeoQuery-LSFB corpus can serve various use cases within the fields of NLP and linguistics.

In NLP, the corpus facilitates a range of language processing tasks, such as semantic parsing, database information retrieval, sign language recognition and machine translation. Currently, the performance of multilingual semantic parsers has exclusively been tested on spoken language data.

[6] Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands: https://archive.mpi.nl/tla/elan

[7] https://www.sign-lang.uni-hamburg.de/hamnosys/input/

[8] https://vhg.cmp.uea.ac.uk/tech/jas/std/

[9] Original HamNoSys character set was retrieved from: https://doi.org/10.25592/uhhfdm.9725

The GeoQuery-LSFB corpus provides the first semantic parsing benchmark which combines a variety of spoken languages with a signed language, allowing multilingual parsers to be trained and tested on a set of languages with heterogeneously diverse forms. The developed semantic parsers can be used within natural language database interfaces, which allow users to retrieve information from a database using natural language, increasing accessibility to these databases (Copestake and Jones, 1990). Given the ID-gloss and HamNoSys annotations provided, the corpus can also be used to train sign language recognition models. Such models typically map a sign language video to a single ID-gloss or series of ID-glosses (Camgoz et al., 2020). The inclusion of HamNoSys also allows the development of models that map video to HamNoSys representations. Finally, the parallel nature of the corpus is ideal for machine translation between a signed language and multiple spoken languages.

Within the field of linguistics, the corpus's main use is the onomasiological study of a signed language along with a diverse set of spoken languages. Onomasiology is a branch from linguistics which analyses how different languages express the same concept (Fernández-Domínguez, 2019). The semantic representations and multiple languages in the corpus are ideal for this, as they allow researchers to study how one component of the semantic representation is expressed in each of the corpus' languages. This can give insight into the influence of the visual-gestural versus aural-oral modality on language variation. Additionally, the combination of signed forms and semantic annotations in GeoQuery-LSFB increases the limited amount of semantically annotated data that is available for computational construction grammar development. Computational construction grammar operationalises the core ideas of construction grammar and analyses language through its form-meaning mappings (e.g. Steels 2011; Beuls and Van Eecke 2023, Van Eecke and Beuls 2025). While several computational construction grammars have been engineered (e.g. Nevens et al., 2019; Verheyen et al., 2023) and learned (Nevens et al., 2022; Doumen et al., 2023, 2024) for spoken languages, similar resources for signed languages are currently limited (e.g. van Trijp 2015; De Vos et al. 2025). GeoQuery-LSFB can help fill this gap by providing the first semantically annotated corpus for a signed language.

## 6. Discussion and Conclusions

We have presented the GeoQuery-LSFB corpus as a linguistic resource for the low-resource French Belgian Sign Language (LSFB). Concretely, we have extended the multilingual and semantically annotated GeoQuery corpus with transcriptions of corresponding LSFB expressions. Based on video recordings featuring a native LSFB signer, an expert transcriber has annotated each expression from the original corpus with French ID-glosses (following Meurant et al., 2015; Johnston, 2016) and phonetic HamNoSys transcriptions (following Hanke, 2004). The corpus was released under the GNU General Public License 2.0 and includes (i) an extensive user guide, (ii) the video recordings that were made, (iii) the original English corpus data and procedural semantic annotations, (iv) the multilingual versions of the corpus and FunQL and SQL annotations that were previously contributed by different researchers, (v) our LSFB annotations in terms of French ID-glosses and HamNoSys transcriptions, and (vi) scripts for executing and evaluating the procedural semantic representations. A synthetically augmented version of the corpus was also included in order to meet user demand.

The GeoQuery-LSFB corpus constitutes a contribution to the linguistic resource landscape in three main respects. First of all, the corpus contributes valuable data to the limited pool of resources that are currently available for the linguistic study of LSFB, as well as for the development of LSFB language technologies. Such data is challenging to create, not only for financial reasons, but also given the limited number of people that both sufficiently master LSFB and have experience or interest in the creation and management of linguistic resources. Second, the GeoQuery-LSFB corpus is, to the best of our knowledge, the first corpus that aligns signed utterances with procedural semantic annotations. The corpus thereby facilitates research into language processing technologies, such as data-efficient semantic parsers or natural language interfaces, that could so far only be developed for spoken languages. Finally, the parallel and semantically annotated nature of the corpus offers for the first time the possibility to study, from an onomasiological perspective, a signed language along a diverse variety of spoken languages.

## 7. Limitations

**Single annotator** The LSFB transcriptions and annotations were performed by a single expert transcriber. We acknowledge this as a limitation of the resource that we created. Yet, we would like to emphasize the difficulty of recruiting and training transcribers for a low-resource sign language like LSFB. We were not able to identify a second potential transcriber sufficiently proficient in both LSFB and HamNoSys to perform the transcription task. Regarding the French ID-gloss annotations, a small-scale inter-annotator agreement study was performed in the pilot phase of the corpus creation

process in order to validate its feasibility. This pilot study yielded a Cohen's kappa of 0.83 on 10 corpus entries selected to be maximally different. This result led to the green light needed to initiate the corpus creation process, but the final corpus was annotated by only one of the transcribers that was part of the pilot due to a lack of available human resources.

**Limited size and domain of the corpus** With its 250 and 4519 utterances, the resource that we introduce is limited in volume and restricted to a single domain. As such, it might not be suitable for many data-intensive processing methods and should not be taken as a representative sample of language use. At the same time, it is the first resource of its kind for LSFB, and for any signed language for that matter. Given the extremely limited resources available for the low-resource LSFB language, the corpus will find immediate use in LSFB-related research and to the best of our expectations in sign language research more generally.

## 8. Statements and Declarations

**Funding** This research was partially supported by the ARIAC project (No. 2010235), funded by the Service Public de Wallonie (SPW Recherche).

**Ethics statement** The French Belgian Sign Language (LSFB) informant consented to the publication and distribution of the video recordings that are part of the GeoQuery-LSFB resource. Both the LSFB informant and the expert transcriber were remunerated for their work according to the official pay scales for scientific personnel employed by universities of the French community of Belgium.

**Competing interests** The authors have no competing interests to declare that are relevant to the content of this article.

**Author Contributions** All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Liesbet De Vos and Laurance Meurant. The original manuscript was written by Katrien Beuls and Paul Van Eecke, with contributions by Liesbet De Vos. All authors commented on previous versions of the manuscript and read and approved the final manuscript.

**Use of Generative AI** The authors declare that this article was conceived and written without the use of generative AI tools.

## 9. Bibliographical References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.

Katrien Beuls and Paul Van Eecke. 2023. Fluid Construction Grammar: State of the art and future outlook. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 41–50. Association for Computational Linguistics.

Mario Bollini, Stefanie Tellex, Tyler Thompson, Nicholas Roy, and Daniela Rus. 2013. Interpreting and executing recipes with a cooking robot. In *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, pages 481–495. Springer.

Borland International. 1988. *Turbo Prolog 2.0 Reference Guide*. Borland International, Scotts Valley, CA, USA.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine Learning*, pages 128–135.

Jianpeng Cheng, Siva Reddy, Vijay Saraswat, and Mirella Lapata. 2019. Learning an executable neural semantic parser. *Computational Linguistics*, 45(1):59–94.

Ann Copestake and Karen Sparck Jones. 1990. Natural language interfaces to databases. *The Knowledge Engineering Review*, 5(4):225–249.

Onno Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 39–43. European Language Resources Association (ELRA).

Donald Davidson. 1967. Truth and meaning. In Jack Kulas, James H. Fetzer, and Terry L. Rankin, editors, *Philosophy, Language, and Artificial Intelligence: Resources for Processing Natural Language*, pages 93–111. Springer.

Liesbet De Vos, Paul Van Eecke, and Katrien Beuls. 2025. A computational construction grammar framework for modelling signed languages. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 1–12, Düsseldorf, Germany. Association for Computational Linguistics.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33–43. Association for Computational Linguistics.

Jonas Doumen, Katrien Beuls, and Paul Van Eecke. 2023. Modelling language acquisition through syntactico-semantic pattern finding. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1317–1327.

Jonas Doumen, Katrien Beuls, and Paul Van Eecke. 2024. Modelling constructivist language acquisition through syntactico-semantic pattern finding. *Royal Society Open Science*, 11(7):231998.

Ralph Elliott, John RW Glauert, JR Kennaway, Ian Marshall, and Eva Safar. 2008. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Universal Access in the Information Society*, 6:375–391.

Jesús Fernández-Domínguez. 2019. The onomasiological approach. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, Oxford, UK.

Jérôme Fink, Pierre Poitier, Maxime André, Laurence Meurant, Benoît Frénay, and Anthony Cleve. 2022. Dictionnaire contextuel langue des signes belge francophone vers français.

Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. 2021. LSFB-CONT and LSFB-ISOL: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Gottlob Frege. 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50.

Sílvia Gabarró-López. 2017. *Discourse Markers in French Belgian Sign Language (LSFB) and Catalan Sign Language (LSC): BUOYS, PALM-UP and SAME*. Ph.D. thesis, Faculty of Arts, Université de Namur, Namur, Belgium.

Stephanie Gross, Matthias Hirschmanner, Brigitte Krenn, Friedrich Neubarth, and Michael Zillich. 2018. Action Verb Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Laurence Haeusler, Thibaud de Laval, and Charlotte Millot. 2014. Etude quantitative sur le handicap auditif à partir de l'enquête "handicap-santé". Technical Report 131, Direction de la recherche, des études, de l'évaluation et des statistiques (DREES), Paris, France.

Thomas Hanke. 2004. HamNoSys: representing sign language data in language resources and language processing contexts. In *Proceedings of the fourth international Conference on Language Resources and Evaluation (LREC)*, volume 4, pages 1–6. European Language Resources Association (ELRA).

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop held at Hidden Valley*. Morgan Kaufmann.

Gabrielle Hodge and Onno Crasborn. 2022. Good practices in annotation. In Trevor Johnston, Julie A. Hochgesang, and Jordan Fenlon, editors, *Signed Language Corpora*, pages 46–89. Gallaudet University Press, Washington, D.C., USA.

Drew A Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017b. Inferring and executing programs for visual reasoning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2989–2998, Washington, D.C., USA. IEEE Computer Society.

Philip N. Johnson-Laird. 1977. Procedural semantics. *Cognition*, 5(3):189–214.

Trevor Johnston. 2008. Corpus linguistics and signed languages: no lemmata, no corpus. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 82–87. European Language Resources Association (ELRA).

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Trevor Johnston. 2016. *Auslan Corpus Annotation Guidelines*.

Bevan Jones, Mark Johnson, and Sharon Goldwater. 2012. Semantic parsing with Bayesian tree transducers. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 488–496. Association for Computational Linguistics.

Rohit J. Kate, Yuk Wah Wong, and Raymond J. Mooney. 2005. Learning to transform natural to formal languages. In *Proceedings of the 20th National Conference on Artificial Intelligence-Volume 3*, pages 1062–1068.

Gregory Kuhlmann, Peter Stone, Raymond Mooney, and Jude Shavlik. 2004. Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer. In *The AAAI-2004 Workshop on Supervisory Control of Learning and Adaptive Systems*. San Jose, CA.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1223–1233. Association for Computational Linguistics.

Alysson Lepeut, Clara Lombart, Sébastien Vandenitte, and Laurence Meurant. 2024. Spoken and signed languages hand in hand: parallel and directly comparable corpora of French Belgian Sign Language (LSFB) and French. *Corpora*, 19(2):241–253.

Percy Liang. 2016. Learning executable semantic parsers for natural language understanding. *Communications of the ACM*, 59(9):68–76.

Wei Lu and Hwee Tou Ng. 2011. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1611–1622. Association for Computational Linguistics.

Ian Marshall and Éva Sáfár. 2004. Sign language generation in an ALE HPSG. In *11th International Conference on Head-Driven Phrase Structure Grammar*, pages 189–201, Stanford, CA. CSLI Publications.

Richard P. Meier. 2002. Why different, why the same? Explaining effects and non-effects of modality upon linguistic structure in sign and speech. In Richard P. Meier, Cormier Kearsy, and Quinto-Pozos David, editors, *Modality and structure in signed and spoken languages*, pages 1–26. Cambridge University Press, Cambridge, United Kingdom.

Richard P. Meier. 2012. Language and modality. In Roland Pfau, Marcus Steinbach, and Bencie Woll, editors, *Sign Language: An International Handbook*, volume 37 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 574–601. De Gruyter Mouton, Berlin, Germany.

Laurence Meurant. 2015. *Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB)*. LSFB-Lab, University of Namur.

Laurence Meurant, Christophe De Clerck, Sibylle Fonze, Gauthier Raes, Susana Sanchez, and Éric Bernagou. 2015. Lex-LSFB: online lexical database from and linked to the 'Corpus LSFB'.

Laurence Meurant, Aurelie Sinte, and Eric Bernagou. 2016. The French Belgian Sign Language corpus a user-friendly searchable online corpus. In *7th workshop on the Representation and Processing of Sign Languages: Corpus Mining (LREC 2016)*, pages 167–174.

Dipendra K. Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. 2016. Tell me Dave: Context-sensitive grounding of natural language to ma-

nipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300.

Irene Murtagh. 2011. Developing a linguistically motivated avatar for Irish Sign Language visualisation. In *2nd International Workshop on Sign Language Translation and Avatar Technology*.

Irene Murtagh, Víctor Ubieto Nogales, and Josep Blat. 2022. Sign language machine translation and the sign language lexicon: A linguistically informed approach. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 240–251.

Jens Nevens, Robin De Haes, Rachel Ringe, Mihai Pomarlan, Robert Porzel, Katrien Beuls, and Paul Van Eecke. 2024. A benchmark for recipe understanding in artificial agents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 22–42.

Jens Nevens, Jonas Doumen, Paul Van Eecke, and Katrien Beuls. 2022. Language acquisition through intention reading and pattern finding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 15–25.

Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2019. Computational construction grammar for visual question answering. *Linguistics Vanguard*, 5(1):20180070.

Magdalena Pasikowska-Schnass. 2018. Sign languages in the EU. Technical Report PE 625.196, European Parliamentary Research Service (EPRS), Brussels, Belgium.

Ana-Maria Popescu, Oren Etzioni, and Henry Kautz. 2003. Towards a theory of natural language interfaces to databases. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 149–157.

Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, Jan Henning, et al. 1987. HamNoSys. Hamburg Notation System for sign language. An introduction. *Zentrum für Deutsche Gebärdensprache*.

Morgane Rébulard, Claire Danet, Adrien Contesse, Claudia S. Bianchini, Jean-François Dauphin, Léa Chèvrefils, Patrick Doan, and Dominique Boutet. 2018. TYPANNOT, a new glyphic system to annotate handshapes in any sign languages. In *8th Conference of the International Society of Gesture Studies "Gesture and diversity" (ISGS8)*, pages 139–140, Cape Town, South Africa. Heather Brooks.

Éva Sáfár and John Glauert. 2010. Sign language HPSG. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 204–207. European Language Resources Association (ELRA).

Robert Smith. 2013. HamNoSys 4.0 user guide. Technical report, Institute of Technology Blanchardstown, Blanchardstown, Ireland.

Luc Steels. 2011. Introducing Fluid Construction Grammar. In Luc Steels, editor, *Design Patterns in Fluid Construction Grammar*, pages 3–30. John Benjamins, Amsterdam, Netherlands.

William C. Stokoe. 1960. Sign language structure: An outline of the visual communication system of the American deaf. *Studies in Linguistics: Occasional Papers*.

Raymond Hendy Susanto and Wei Lu. 2017. Semantic parsing with neural hybrid trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3309–3315, Washington, DC, USA. AAAI Press.

Valerie Sutton. 1995. *Lessons in SignWriting: textbook & workbook.* Deaf Action Committee for Sign Writing, La Jolla, CA.

Lappoon R. Tang. 2003. *Integrating Top-down and Bottom-up Approaches in Inductive Logic Programming: Applications in Natural Language Processing and Relational Data Mining*. Ph.D. thesis, Department of Computer Sciences, University of Texas, Austin, TX.

Lappoon R. Tang and Raymond J. Mooney. 2001. Using multiple clause constructors in inductive logic programming for semantic parsing. In *European Conference on Machine Learning*, pages 466–477. Springer.

Dan Tasse and Noah A. Smith. 2008. SOUR CREAM: Toward semantic processing of recipes. Technical Report CMU-LTI-08-005, Carnegie Mellon University, Pittsburgh, PA, USA.

Sarah F. Taub. 2012. Iconicity and metaphor. In Roland Pfau, Marcus Steinbach, and Bencie Woll, editors, *Sign Language: An International Handbook*, volume 37 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 574–601. De Gruyter Mouton, Berlin, Germany.

Paul Van Eecke and Katrien Beuls. 2025. PyFCG: Fluid Construction Grammar in Python. *arXiv preprint arXiv:2505.12920*.

Remi van Trijp. 2015. Towards bidirectional processing models of sign language: A constructional approach in fluid construction grammar. In *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science*, pages 668–673. University of Torino.

Remi van Trijp, Katrien Beuls, and Paul Van Eecke. 2024. The proof is in the almond cookies. In Luc Steels and Robert Porzel, editors, *Narrative-based Understanding of Everyday Activities: A Cookbook*, pages 59–77. Venice International University, Venice, Italy.

Lara Verheyen, Jérôme Botoko Ekila, Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2023. Neuro-symbolic procedural semantics for reasoning-intensive visual dialogue tasks. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023)*, pages 2419–2426.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Yuk Wah Wong and Raymond J. Mooney. 2006. Learning for semantic parsing with statistical machine translation. In *Proceedings of the 2006 main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics.

William A. Woods. 1967. *Semantics for a question-answering system*. Ph.D. thesis, Harvard University, Cambridge, MA, USA.

William A. Woods. 1981. Procedural semantics as a theory of meaning. In Aravind K. Joshi, Bruce L. Webber, and Ivan A. Sag, editors, *Elements of Discourse Understanding*, pages 300–334. Cambridge University Press.

William A. Woods. 2010. ACL lifetime achievement award: The right tools: Reflections on computation and language. *Computational Linguistics*, 36(4):601–630.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, pages 1050–1055, Washington, D.C., USA. AAAI Press.

John Marvin Zelle. 1995. *Using inductive logic programming to automate the construction of natural language parsers*. Ph.D. thesis, The University of Texas at Austin.

Luke Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 658–666. AUAI Press.

# The De-Sign Platform: An Online Psychometric Tool for Dementia Screening of Deaf Older Adults in two Sign Languages, GSL and ÖGS

**Athanasia-Lida Dimou[1], Theodore Goulas[1], Marianna Tsatali[2,3], Tarsitsa Ntova[3,4], Doris Hoffmann-Lamplmair[5], Stavroula-Evita Fotinea[1], Eleni Efthimiou[1], Birgit Teichmann[3], Magda Tsolaki[2], Joanna Atkinson[6], Bencie Woll[6]**

[1]ATHENA – Research & Innovation Information Technologies, Athens, Greece,
[2]Greek Association of Alzheimer's Disease and Related Disorders,
[3] Network Aging Research, Heidelberg University, Heidelberg, Germany
[4]Association of the Deaf of Northern Greece
[5]Krankenhaus Barmherzige Brüder, Wien, Austria
[6]Deafness, Cognition and Language Research Centre, University College London, DCAL

[1]Artemidos 6 & Epidavrou, 151 25 Maroussi, Athens, Greece,
[2]Syndika 13, Thessaloniki 546 43, Greece
[3]Network Aging Research, Heidelberg University, Bergheimer Str. 20, 69115 Heidelberg, Germany
[4]3 Kallitheas Ave., Ampelokipoi Thessaloniki, Greece
[5]Johannes von Gott Platz 1, 1020 Wien, Austria
[6]49 Gordon Square, London, WC1H 0PD, United Kingdom
ndimou@athenarc.gr, tgoulas@athenarc.gr, mtsatali@yahoo.gr, psy.tarsitsa@gmail.com, Doris.Hoffmann-Lamplmair@bbwien.at, evita@athenarc.gr, eleni_e@athenarc.gr, teichmann@nar.uni-heidelberg.de, tsolakim1@gmail.com, joanna.atkinson@ucl.ac.uk, b.woll@ucl.ac.uk

**Abstract**

This article presents the De-Sign platform, a web-based psychometric tool specifically designed for screening dementia in Deaf older adults (50+) who use Austrian Sign Language and Greek Sign Language hereinafter ÖGS and GSL respectively. The limited access to dementia services for these populations is primarily attributed to a scarcity of healthcare professionals fluent in sign language. Hence, enhancing access to relevant diagnostic services has become a priority. Currently, there is a significant lack of screening tools specifically developed to identify early signs of dementia that are compatible with national sign languages. To address this issue, the De-Sign Erasmus+ (2022-2025) project has employed suitable psychometric instruments that are adapted to the cultural contexts and linguistic norms of Deaf communities in Austria and Greece. The only existing Cognitive Screening Test (CST) for British Sign Language (BSL), used for diagnosing dementia in Deaf older adults, was initially adapted from English by Atkinson et al. (2015). The De-Sign platform hosts a cognitive screening test in ÖGS and GSL. Both were linguistically and culturally adapted from the BSL-CST test, providing two web-based versions of a psychometric tool that enables dementia screening within these populations.

**Keywords:** Dementia screening, Deaf populations, digital platform, psychometric tool, GSL, ÖGS.

## 1. Introduction

The De-Sign project "Raising Awareness for Dementia in Deaf Older Adults in Europe" (ERASMUS+, Project N° 2022-1-DE02-KA220-ADU-000088591) (https://de-sign-project.eu/index.php/el/ ) was an initiative aiming to meet an existing substantial lack in the field of mental health services, and particularly in the early detection of Dementia in Deaf and Hard-of-Hearing adult sign language users. Project partners included Germany, Greece, Austria and Italy, A German Sign Language (DGS) dementia screen had already been developed (Stockleben et al., 2025) and the Italian partners were not involved in the dementia screening part of the project. So, the study described here reports on the development of dementia screening tools for two of the partners' languages: Greek Sign Language (GSL) and the Austrian Sign Language (ÖGS) SLs. The De-Sign platform (https://sign.ilsp.gr/de_sign/) was one of the outcomes of the of the Erasmus+ funded project "De-Sign: Raising Awareness for Dementia in Deaf Older Adults in Europe".

### 1.1 Dementia Screening in Deaf Populations

Dementia is not a single disease entity but rather "a syndrome resulting from diseases of the brain, usually of a chronic or progressive nature, in which there is disturbance of multiple higher cortical functions, including memory, attention, thinking, orientation, comprehension, calculation, learning capacity, language, and judgement. The impairments of cognitive function are commonly accompanied, and occasionally preceded, by deterioration in emotional control, social behavior, or motivation. This syndrome occurs in Alzheimer disease, in cerebrovascular disease, and in other conditions primarily or secondarily affecting the brain" (WHO, 2019).

Alzheimer's disease is the most common form of dementia and accounts for over 50% of the cases (Alzheimer Europe, 2023).



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 113–119
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Timely and accurate diagnosis of dementia is a health priority in many countries, as it enables access to a range of services that support a better quality of life, such as access to appropriate medication, time to plan future care preferences, and opportunities to express needs, wishes, and goals (NICE, 2018). Also, early diagnosis enables individuals to receive appropriate symptomatic treatment (e.g., cholinesterase inhibitors where indicated), engage in advance care planning, and participate in shared decision-making regarding future care preferences while decisional capacity is still preserved (Dubois et al., 2016). Recognition of early signs of dementia by family, friends, and communities plays a crucial role in facilitating help-seeking and referral to specialized assessment services (Alzheimer's Disease International, 2024).

However, there are few dementia awareness programs in sign languages worldwide (but see, e.g. https://bda.org.uk/dementia/. This The absence of such programs contributes to low awareness of the early signs of dementia and underdiagnosis in Deaf communities around the world (Flower, Heferman & Dening, 2024; Ferguson-Coleman, Keady & Young, 2014). The aim of the De-Sign project was also to promote equal access of Deaf people to dementia information and diagnosis by raising awareness about dementia within Deaf communities in Europe.

The main innovation of this project was the creation of a digital platform for dementia screening, the De-Sign platform. The platform delivers an online digital administration of a neuropsychological test that screens for early signs of dementia in Deaf populations. The developed psychometric tool allows Deaf participants to be tested in their native language without the mediation of interpreters. A built-in evaluation section allows the health care expert to review the participants' answers. The De-Sign platform was developed in two different SLs, GSL & ÖGS.

## 2. Methodology

For the development and implementation of the De-Sign platform, Deaf communities, healthcare professionals, and academic experts were brought together, resulting in practical and accessible tools that directly address the needs of Deaf people.

At the time of the start of the De-Sign Project (2022-2025), except for the British Sign Language Cognitive Screening Test, hereinafter BSL-CST, there were no screening tools for detecting dementia in older Deaf adults, while none have been implemented as a fully technology-assisted application.

The creation of a fully web-based screening test filled a gap that existed in the administration of such tests not only for the homogeneous parametrization of audiovisual materials but also in terms of the methodology of answer reception (answers are directly video-recorded and stored in a safe database), data protection, and classification as well as with respect to its potential remote use.

### 2.1.1 The BSL-CST Test

The British Sign Language Cognitive Screening Test (BSL-CST) is a screening tool for cognition assessment created and delivered in British Sign Language that is used to identify dementia and acquired cognitive impairment in older deaf individuals (Atkinson et al. 2015). It can be used to detect neurodegeneration or to provide a clinical baseline for a follow-up assessment. It is designed to screen all areas of cognition, with items relating to memory, language (BSL and British 2-handed fingerspelling), executive function, visuospatial ability, orientation, and attention

It uses standardized on-line video administration to screen cognition, using signed, rather than spoken or written, instructions and was normed on a sample of 226 deaf people aged 50-90o. This innovative testing approach transforms the ability to detect dementia in deaf people, avoids the difficulties of using an interpreter, and enables culturally and linguistically sensitive assessment of deaf signers, with international potential for adaptation into other signed languages. (Atkinson et al. 2015).

### 2.1.2 The Adaptation of the BSL-CST Into GSL and ÖGS

To obtain a highly efficient psychometric tool in the two SLs that are currently hosted on the platform, a series of methodological steps were followed for linguistic and cultural adaptation.

Two multidisciplinary teams of Deaf and hearing experts were brought together, one for each SL. In Austria this role was assumed by the Krankenhaus Barmherzige Brüder Wien[1]. The expert group in Austria consisted of a Deaf social worker and a Deaf nurse, as well as a physician and four clinical psychologists with expertise in ÖGS.

In Greece two institutions were involved in the process, the Association of the Deaf Northern Greece (EKVE)[2] and the Greek Association for Alzheimer and Related Disorders (GAARD)[3]. The expert group in Greece consisted of two Deaf teachers of GSL with extensive teaching

[1]Krankenhaus Barmherzige Brüder Wien (https://www.barmherzige-brueder.at/portal/wien/home)

[2] EKVE (https://www.ekbe.gr )

[3] Greek Alzheimer Association and Related Disorders (https://www.alzheimer-hellas.gr/index.php/en)

experience and Deaf family backgrounds; a psychologist fluent in GSL with expertise in neuropsychological assessment; a linguist and GSL interpreter; a neuropsychologist specialized in dementia assessment; and a linguist specializing in SL research.

Athena RC[4] was the institution that undertook the development of the platform and had regular communication with both teams.

The test contains 24 tasks, and each task contains several items. All test items of the BSL-CST were initially translated into the Austrian and Greek spoken/written language; they were then translated and adapted into the respective SLs, including thorough review to identify items requiring linguistic and/or cultural adaptation and to make decisions regarding the insertion of new items that were necessary for the efficacy of the test.

The main cultural adaptations were applied to items involving vocabulary, to ensure linguistic and cultural appropriateness. In contrast, visuospatial tasks did not require any modifications, as they were considered culturally neutral.

No new subtests were required. In some cases, existing items were modified to align with Austrian and Greek Deaf cultures. Variants of some tasks needed to be created to allow for fingerspelling particularities in both languages; in ÖGS, signers use either one- or two-handed fingerspelling, so two sets of elicitation material were created. The choice of which to use depended on information provided by the participants before the beginning of the test.

Similarly, for GSL, a limited number of alternative items were introduced in specific tasks. It had been expected that some elderly Greek Deaf participants might not know or use fingerspelling. Although in the course of testing it was found that most participants in the 50+ and 60+ groups could fingerspell, the platform provides an alternative option for those participants who cannot fingerspell: the Color Trails Test was administered in place of the Alternating Trails task to reduce language-related bias in executive function assessment.

Once agreement was reached on all items, a Deaf native signer of each of the SLs was video recorded signing all test instructions to create the pilot version of the test.

During the pilot phase, five Deaf experts provided detailed feedback. Their comments were discussed within each expert group, and appropriate amendments were implemented. Guidance and supervision on procedures for adaptation of the BSL-CST were provided by the original authors of the BSL-CST.

Following finalization of the above-described procedures, each team recorded the final elicitation videos for all items in the test. The Austrian team recorded the videos with the support and equipment of the hospital's communication department, while the recording of the videos for GSL took place in the premises of ATHENA RC using high-quality audiovisual equipment.

A common methodology for creating the video materials was followed by both teams, and each elicitation item was filmed multiple times (2-3 times each). Afterwards, the videos were reviewed jointly by Deaf and hearing experts, and the best takes were selected.

## 3. The De-Sign Platform: Design and Implementation

Simultaneously with the procedures outlined above, the De-Sign platform was developed from the ground up. The platform currently hosts the adapted version of the BSL-CST test in GSL and ÖGS (https://sign.ilsp.gr/de_sign/). Both tests are constructed on a unified framework and design, incorporating parallel tasks to eliminate biases, ensure homogeneity in test administration and data acquisition across the two countries.

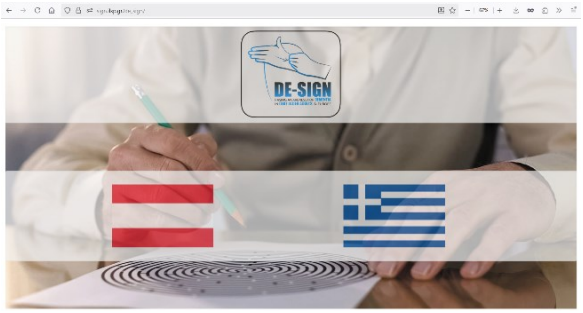



Figure 1: Initial page of the De-Sign platform

To facilitate the design of the platform, an initial blueprint was created to outline the methodological framework required for implementation. This step was crucial, as it enabled an examination of the technical specifications and accessibility standards that needed to be integrated into the platform.

The layout of the platform was designed with a primary focus on ensuring optimal test performance by incorporating adequate accessibility features for Deaf participants. These features include a comprehensive and user-friendly menu, instructions available in both video and written formats, a limited number and clearly identifiable buttons for advancing to the next task, pausing, and returning to the main menu, as well

[4] ATHENA RC (https://www.athenarc.gr/en)

as options to skip tasks and navigate between them, along with the capability to record participants' responses.

A circular feedback strategy was implemented during the development of the platform, enabling the incorporation of insights from health and special needs experts. This approach facilitated the adaptation of the platform's functionalities to meet their expectations. As a result of this constructive dialogue, modifications and enhancements were made to the platform's features, overall performance, technical specifications, and testing flow. These improvements are intended to facilitate the administration of the tests and increase the usability of the tool itself.

Consequently, only essential navigation buttons were included to allow health experts to navigate between tasks.

The second pillar of the platform's architectural design was based on security and storage issues. Data storage and security issues are bound by local regulations in each country. To address this, ATHENA R.C collaborated closely with Krankenhaus Barmherzige Brüder Wien for the ÖGS and the EKVE and the GAARD for the GSL. They developed a suitable scheme for data transfer, storage, and security protocols tailored to each language set.

For the integration of video recordings, including both general and task-specific instructions, a

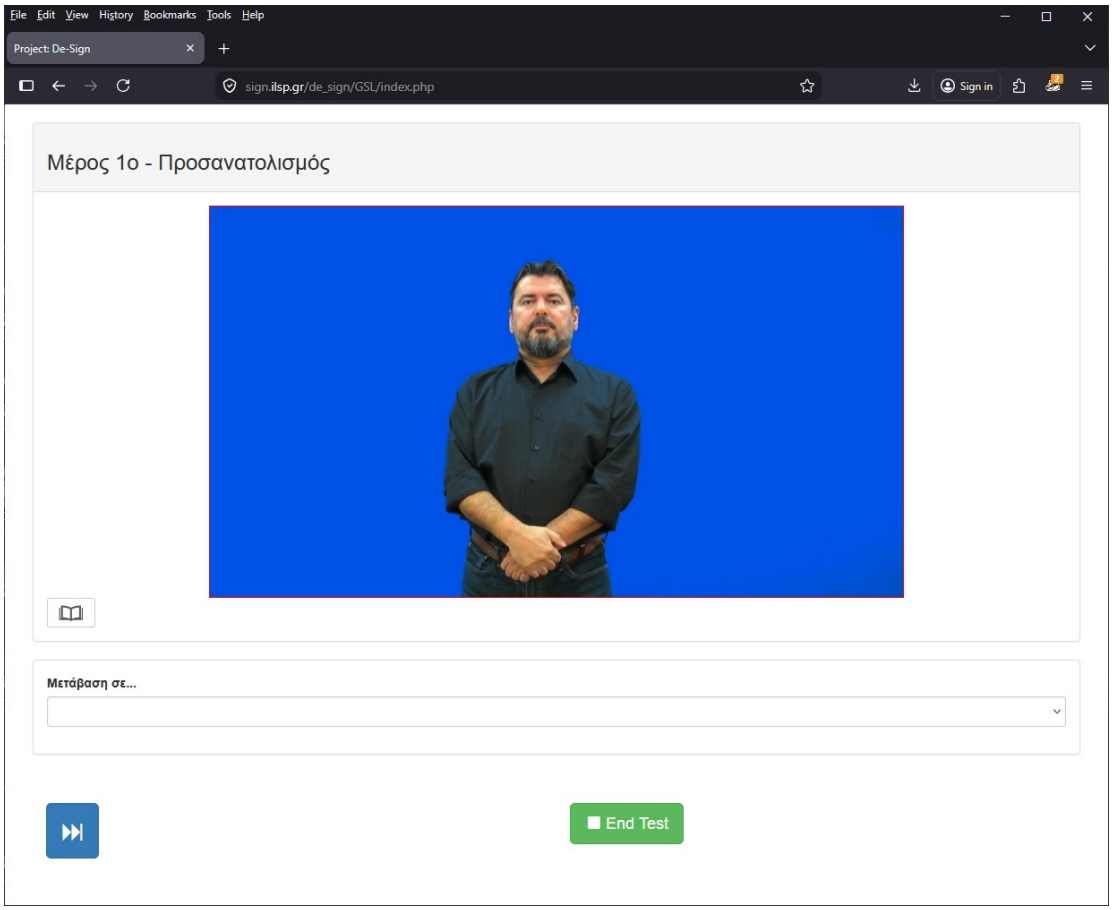



Figure 2: View of the layout pages on the De-Sign platform; the elicitation video of the orientation task on the Greek version.

Figure 2 illustrates the layout of the platform's pages which maintain a consistent design across both tests. The central area of each page features an elicitation video that provides instructions for each task, while a text version of the instructions is accessible beneath the video screen via a designated button.

The layout was intentionally designed to be minimalist, following recommendations that excessive visual stimuli could negatively impact on the efficacy of the tests, particularly for participants who may be prone to dementia.

flexible development and management approach (agile methodology) was adopted. This was necessary for the quick design, production, and delivery of subsystems within the project timeline. The use of proven technologies (JS, JSON, web services) allows usability, interoperability, and indexing requirements of audiovisual content. Content integration on the platform was carried out over the three years of the De-Sign project, with the content of both tests was available in both sign languages (ÖGS & GSL).

Testing and trial phases validated each development stage of the platform and confirmed its efficient performance.

### 3.1 Scoring, Result Reviewing and Final Evaluation Sheet

A significant aspect of the platform is the evaluation page, which serves as a distinct section within each language set. This feature enables health experts to simultaneously view video productions submitted for each participant and assess them on the platform. The final score for each task, along with the overall score for each participant, is calculated automatically, and a report is generated in .pdf format. Access to this section is restricted to the health expert responsible for each participant.

The result of the assessment is a total score, calculated by summing the scores across all items. Inter-rater reliability is typically examined in rating scales rather than in cognitive screening tools, as this type of reliability is less relevant when multiple cognitive domains are assessed rather than a single underlying construct.

Raters in Austria were not provided with additional training, as there was already a sufficient number of experienced health professionals, both Deaf and hearing, working in the field of dementia.

A comprehensive clinician User Manual has been developed and is available on the platform, outlining procedures and scoring guidelines. This manual is designed to assist new users in becoming acquainted with the tasks, the test flow, and the specificities of the digital test administration. It offers detailed instructions on biases to avoid during test administration and provides all necessary information regarding the use and scoring of the evaluation section.

### 3.2 Ethics and GDPR

The De-Sign platform is intended as a screening tool rather than a diagnostic instrument; it identifies early signs of dementia and facilitates referrals for further examinations by health experts. Although the data collected on the platform does not include medical information, it remains sensitive and personal, originating from individuals who may be at risk of dementia.

The platform adheres to strict GDPR regulations, and informed consent is obtained from every participant. Specifically, upon entering the platform for each language set, participants are presented with a general video outlining the project's objectives and the platform's purpose. Subsequently, each participant views two detailed videos: the first addresses the aims of the psychometric test, the research processes, and the management of their video recordings, while the second discusses GDPR considerations regarding their personal information on the platform. These videos are also available in written form. Participants have the option to either read the informed consent documents or view them in sign language. The documents must be signed, and the corresponding box checked on the platform before the test can commence. Failure to accept the specified clauses constitutes an exclusion criterion, resulting in the cessation of the test.

Following this, each health expert must log into the platform using their personal credentials. The health expert is responsible for inputting the minimum required information for each participant, which includes their name, surname, date of birth, and the date of test administration. This information is utilized to create a unique identification sequence for each participant, which is subsequently employed to access the evaluation page containing performance scores for all tasks.

## 4. Results

Both the Greek and Austrian pilot versions of the psychometric test on the digital platform were administered to 20 Deaf individuals aged 50 or older who met stringent inclusion criteria. In Austria, there were no significant comprehension or technical issues, which allowed the pilot data to be incorporated into the final analysis. Conversely, minor adjustments were made in Greece.

For the purposes of developing test norms, 100 Deaf older adults from Austria and 70 Deaf older adults from Greece (with ongoing recruitment) have participated in the official administration of the test thus far. Normative samples from Austria and Greece were recruited from diverse regions to ensure representation of regional lexical variation. Participants provided informed consent and underwent screening for inclusion and exclusion criteria prior to engaging in one hour of video-based neurocognitive testing. As of the date of publication of the present paper, results for ÖGS are undergoing preparation for publication (Hoffmann-Lamplmair et al., in preparation); data are still being collected for GSL.

In Austria, four clinical psychologists fluent in ÖGS administered and scored the assessments using a structured scoring protocol. Approximately half of the scores were assessed through double rating to evaluate inter-rater reliability. In Greece, two clinical psychologists fluent in GSL administered and scored the assessments in accordance with protocol guidelines. Test administrators were members of the Deaf community who had received formal training in test administration under the supervision of one of the two clinical psychologists. They were health professionals, professional interpreters, or teachers. All were university graduates, used Sign Language as their primary means of communication, and self-identified as members of the Deaf community.

The outcome of this endeavor was the creation of a standardized web-based adaptation of the BSL-CST in two distinct SLs, which is characterized by high quality, linguistic precision, and cultural relevance.

## 5. Conclusion

The main target group of the De-Sign platform were Deaf older adults who use ÖGS and GSL.

This population has historically faced significant barriers in accessing linguistically and culturally appropriate cognitive screening tools, often resulting in underdiagnosis or misdiagnosis of dementia. By developing an adapted and web-based version of the BSL-CST, De-Sign platform effectively addressed this gap, offering an accessible, reliable, and standardized assessment mechanism tailored for these communities.

Consequently, the Deaf populations in Austria and Greece have been afforded access – for the first time – to a scientifically validatable cognitive screening tool that meets their linguistic and cultural requirements. This initiative marks a significant advancement toward achieving equity in neuropsychological assessment for Deaf individuals.

## 6. Acknowledgments

This article was the outcome of the De-Sign Project (KA220-ADU-000088591), co-funded by the Erasmus+ Program of the European Union.

## 7. Bibliographical References

Alzheimer Europe. (2023). *Dementia in Europe yearbook*. https://www.alzheimer-europe.org

Alzheimer's Disease International. (2024). *World Alzheimer Report 2024: Global changes in attitudes to dementia. London*, England: Alzheimer's Disease International. https://www.alzint.org/resource/world-alzheimer-report-2024/

Joanna Atkinson, Tanya Denmark, Jane Marshall, Cath Mummery, Bencie Woll. (2015). Detecting Cognitive Impairment and Dementia in Deaf People: The British Sign Language Cognitive Screening Test, *Archives of Clinical Neuropsychology*, 30(7):694–711, https://doi.org/10.1093/arclin/acv042

Athanasia – Lida Dimou, Theodore Goulas, Eleni Efthimiou, Stavroula – Evita Fotinea, Panagiotis Karioris, Kyriaki Vasilaki. (2025). The De-Sign platform: design and development of the adapted screening test for detecting early onset of dementia in deaf older adults in two Sign Languages - GSL and ÖGS. In *14th Pan-Hellenic Conference on AD & Related Disorders and 6th Mediterranean Conference on Neurodegenerative Diseases,* Thessaloniki, Greece.

Bruno Dubois, Harald Hampel, Howard Feldman, Philip Scheltens, Paul Aisen, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Kaj Blennow, Karl Broich, Enrica Cavedo, Sebastian Crutch, Jean-François Dartigues, Charles Duyckaerts, Stéphane Epelbaum, Giovanni Frisoni, Serge Gauthier, Remy Genthon, Alida Gouw, Marie-Odile Habert, David Holtzman, Miia Kivipelto, Simone Lista, José-Luis Molinuevo, Sid O'Bryant, Gil Rabinovici, Christopher Rowe, Stephen Salloway, Lon Schneider, Reisa Sperling, Marc Teichmann, Maria Carrillo, Jeffrey Cummings, Cliff Jack Jr. (2016). Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia, 12*(3):292–323. https://doi.org/10.1016/j.jalz.2016.02.002

Emma Ferguson-Coleman, John Keady and Alys Young. (2014). Dementia and the Deaf community: Knowledge and service access. *Aging & Mental Health*, *18*(6):674–682. https://doi.org/10.1080/13607863.2014.880405

Isabelle Flower, Eithne Heffernan and Tom Dening. (2024). Dementia and the Deaf community: prevalence, assessment and management in people with hearing loss since childhood, *Aging & Mental Health*, DOI:10.1080/13607863.2024.2430533

Doris Hoffmann-Lamplmair, Partick Martinetz, Marie-Luise Burgstaller, Romeo Seifert, Lucas Nerat, Sabrina Cernek, Manuela Schwanzer, Athanasia-Lida Dimou, Theodore Goulas, Birgit Teichmann, Bencie Woll, Joanna Atikinson, Marianna Tsatali, Ege Karar, Tarsita Ntova, Thomas Ströbele. (2025). Videsta: A tailored dementia screening tool for the Austrian Deaf sign language community. In 3*5th Alzheimer Europe conference,* Bologna, Italy.

Gill Livingston, Jonathan Huntley, Andrew Sommerlad, David Ames, Clive Ballard, Sube Banerjee, Carol Brayne, Alistair Burns, Jiska Cohen-Mansfield, Claudia Cooper, Sergi Costafreda, Amit Dias, Nick Fox, Laura Gitlin, Robert Howard, Helen Kales, Mika Kivimäki, Eric Larson, Adesola Ogunniyi, Vasiliki Orgeta, Karen Ritchie, Kenneth Rockwood, Elizabeth L Sampson, Quincy Samus, Lon Schneider, Geir Selbæk, Linda Teri, Naaheed Mukadam. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet*, *396*(10248): 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6

National Institute for Health and Care Excellence. (2018). *Dementia: Assessment, management and support for people living with dementia and their carers (NG97).* NICE. https://www.nice.org.uk/guidance/ng97

Lisa Stockleben, Bencie Woll and Joanna Atkinson. (2026). Screening for Acquired Cognitive Impairment and Dementia in Deaf Users of German Sign Language: The

Cognition Test German Sign Language. *Archives of Clinical Neuropsychology*, *41*(2), acaf094.

Marianna Tsatali, Sabrina Cernek, Athanasia-Lida Dimou, Erico Dolza, Doris Hoffmann-Lamplmair, Nicola Della Maggiora, Partick Martinetz, Tarsitsa Ntova, Evanthia Plachoura, Romeo Seifert, Thomas Ströbele, Magda Tsolaki and Birgit Teichmann. (2024). Raising Awareness for Dementia in Deaf Older Adults in Austria, Germany, Greece, and Italy: The De-Sign Erasmus+ project. In *13th Pan-Hellenic Conference on AD & Related Disorders and 5th Mediterranean Conference on Neurodegenerative Diseases*, Thessaloniki, Greece.

Marianna Tsatali, Sabrina Cernek, Athanasia-Lida Dimou, Erico Dolza, Doris Hoffmann-Lamplmair, Nicola Della Maggiora, Partick Martinetz, Tarsitsa Ntova, Romeo Seifert, Thomas Ströbele, Magda Tsolaki and Birgit Teichmann. (2023). A Cognitive Screening Test for detecting Alzheimer's Disease Dementia in the Deaf older adults in Austria, and Greece: The De-Sign Erasmus+ project. In *33rd Alzheimer Europe Conference*, Helsinki, Finland.

Marianna Tsatali, Athanasia-Lida Dimou, Erico Dolza, Doris Hoffmann-Lamplmair, Tarsitsa Ntova, Ioanna-Antigoni Angelidou, Marie-Luise Burgstaller, Kiriaki Chatziavannidou, Theodore Goulas, Nicola Della Maggiora, Romeo Seifert, Thomas Ströbele, Magda Tsolaki, Birgit Teichmann. (2024). Raising awareness about dementia in Austrian, German, Greek and Italian Deaf communities: the De-Sign Erasmus+ project. In *34th Alzheimer Europe conference,* Geneva, Switzerland.

Marianna Tsatali, Athanasia-Lida Dimou, Tarsitsa Ntova, Ioanna-Antigoni Angelidou, Doris Hoffmann-Lamplmair, Marie-Luise Burgstaller, Ege Karar, Nicola Della Maggiora, Romeo Seifert, Erico Dolza, Magda Tsolaki, Birgit Teichmann. (2025). The De-Sign Project: Identifying Dementia Knowledge in the Deaf Community and Findings from an Adapted Cognitive Screening Test for Older Deaf Adults in Austria and Greece. In *35th Alzheimer Europe conference*, Bologna, Italy.

Marianna Tsatali, Sabrina Cernek, Athanasia-Lida Dimou, Erico Dolza, Doris Hoffmann-Lamplmair, Nicola Della Maggiora, Partick Martinetz, Tarsitsa Ntova, Evanthia Plachoura, Romeo Seifert. (2023). A Cognitive Screening Test for Detecting Alzheimer's Disease Dementia in the Deaf Older Adults in Austria, and Greece: The De-Sign Erasmus+ Project. Alzheimer's & Dementia 19:e072396. https://doi.org/10.1002/alz.072396

World Health Organization. (2019). *International statistical classification of diseases and related health problems (10th rev.)*. https://icd.who.int/browse10/2019/en

# Feature Analysis of MoCap Data for Optimised Sign Language Processing

**Yves A. Duppen, Mirella De Sisto, Ifigeneia Mavridou, Phillip Brown, Lisa Lepp, Dimitar Shterionov**
{Y.A.Duppen, M.DeSisto, I.Mavridou, P.C.Brown, L.B.Lepp, D.Shterionov}@tilburguniversity.edu
Tilburg University
Warandelaan 1, 5037AB Tilburg, The Netherlands

**Abstract**

Despite the rapid advances in AI and its impact on machine translation (MT), when it comes to sign language (SL) processing and MT, there is a big bottleneck – the lack of substantial quantities of quality signed data suitable for developing SLMT models. Marker-based motion capturing (MoCap) is a technique for tracing and recording the body movements (including hands and figures) in 3D space with high precision and has been widely used in SL research. MoCap data is of high representative accuracy, making it very suitable for analysing movement patterns and articulatory features. However, it is also very complex – a recording of a single sign may contain more than 240 entries over 156 features making it difficult for processing. In this paper we analyse MoCap data aiming to understand which captured features are of high importance. Consecutively, we optimise the MoCap data representation, reducing the number of features, and assess how this feature-reduced data impacts sign classification task. We organise MoCap features based on their importance and show how models trained on feature-reduced representations outperform those developed on the complete feature set.

**Keywords:** sign language NLP, motion capturing, data optimisation, sign classification

## 1. Introduction and background

The rapid technological and methodological advances in deep learning, and in AI in general, have not only improved language modeling and machine translation (MT), recognition of image, video and audio signals, the synthesis of life-like 3D avatars, etc., but have also led to the fusion of interdisciplinary research innovations. These advancements have created a fruitful environment for impactful research in the field of sign language processing and translation. Initiatives such as the SignON [1] and EASIER [2] consortia, the sign language MT shared tasks in the WMT conference (Müller et al., 2022; Müller et al., 2023) and, in general, the growth of number of publications in the field testify for this trend. Despite this spur of activities, one significant bottleneck remains. In particular, this bottleneck is data – high volumes quality data are required for training models of sufficient performance. Unfortunately, sign language (SL) data are scarce and scattered making training MT systems for SLs impractical. Furthermore, SL data have not been collected with MT (and the underlying pipeline components) in mind (Vandeghinste et al., 2024). This challenge can be addressed by either collecting new data and / or processing existing data to fit the requirements of an MT (or LLM) framework.

When processing existing video data, it is common to use a pose estimator, such as OpenPose (Cao et al., 2018) or MediaPipe (Lugaresi et al., 2019) for feature extraction. This facilitates the processing of input videos by reducing the complexity in the video stream, retaining information relevant to the signer (Rastgoo et al., 2021). However, these methods suffer from some limitations when it comes to SL processing. First, they have limited capacities with estimating depth and with occlusion. Second, they are not specifically trained on SL data or developed for SL research use-cases. As De Coster et al. (2023) shows, the lack of domain-specificity leads to errors in the (recognition). Existing SL specific methods (e.g. Forte et al. (2023); Ivashechkin et al. (2023)) have a better ability to process SL data, but still suffer from limitations related to depth accuracy and occlusions (Andersen et al., 2025). Pose estimators are an efficient solution to processing existing data and despite their limitations they do provide a mechanism to convert a signed video into a numerical format that can be processed with existing ML and DL tools.

However, when it comes to collecting new data, an alternative to first recording video and then processing the video data (with pose estimators) is to use *motion capturing*. In recent years, data collection via marker-based motion capture techniques has gained increasing attention for SL linguistics and SL technology purposes (Andersen et al., 2025). Given their ability to provide fine-grained 3D information, marker-based motion capture methods generate highly accurate representations of what is being signed. In addition, collecting

[1] https://signon-project.eu
[2] www.project-easier.eu



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 120–128
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

data in a 3D setup prevents a number of limitations, such as those concerning depth accuracy or occlusions previously mentioned, especially when this setup is tailored towards sign language production. The recent work by Andersen et al. (2025) and earlier work by Jedlička et al. (2020) present such tailored setups outlining alternatives and highlighting limitations and decision points.

Similar to those of pose estimators, MoCap data are very elaborate – a sequence of spatial data in 3D, i.e. three numerical values for a keypoint, and a number of keypoints for every captured frame, typically in a 120 frame-per-second recording. These data are quite useful when it comes to sign language synthesis, i.e. the task of generating an artificial 3D signer (often referred to as avatar) that can convey a signed utterance (Bernhard et al., 2022; Jedlička et al., 2020; Naert et al., 2020).

While the majority of work using MoCap data focuses on data-driven synthesis (Naert et al., 2020) and (Bernhard et al., 2022; Jedlička et al., 2020; Naert et al., 2020), in this work we analyse the MoCap data through the lens of sign language recognition (SLR), the task of processing an input video (containing a sign or a signed utterance) and generating a format suitable for a specific downstream task. SLR is typically the first step in the sign language machine translation pipeline (Shterionov et al., 2024). We explore the structure of MoCap data with the goal of identifying important features and reducing complexity. MoCap generates highly detailed and close to realistic motion data of a signer's movements and articulations. By analysing MoCap data, in contrast to video recordings (potentially processed with a pose estimator) aiming at a better understand of which human features have highest impact on classification of signs.

We analyse a dataset of 44 isolated signs, each produced by 3 signers (i.e. a total of 132 signs). The language is Sign Language of the Netherlands (*Nederlands Gebarentaal* – NGT). These 44 signs were selected because they capture the whole phonological inventory of NGT. Our investigation involves feature importance analysis and classification of individual signs based on BioVision Hierarchy (BVH) files. This format is widely used for motion capturing data and for its usability with avatar-generating softwares (e.g. Animics[3]).

Feature importance analysis for sign language has been previously conducted in the context of signer identification (among others (Bigand et al., 2020, 2021b)). Bigand et al. (2020) and 2021b also employ classification (and use dimensionality reduction) however their task differs from ours – we aim at classification of signs rather than of sing**ers**. Dimensionality reduction has also been used in the context of spontaneous sign language decomposition into elementary movements Bigand et al. (2021a); however, our work specifically looks into the features native for MoCap and MediaPipe (expressed in BVH format) rather then of their principle components.

[3] https://github.com/upf-gti/animics

## 2. Dataset Creation

The dataset we used is a newly recorded set of 44 signs from the Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT). The data collection took place as part of the CoCoS project – small-scale project funded by Tilburg University and executed in collaboration with the Nederlands Gebarencentrum (NGT). Within this project, we organised a co-creative workshop which, among others, allowed us to record MoCap data.[4] Three people participated in the data collection – two female and one male participants; one of them with a dominant left hand.

We used a setup with six OptiTrack Flex 13 motion capture cameras[5] to record full-body motion data. All cameras were synchronised with the OptiTrack Motive software[6] (version. 3.3). System calibration was performed prior to the capture sessions achieving <0.3 mm mean residual error. Reflective markers (14 mm diameter) were placed on an OptiTrack suit; these markers were distributed on the anatomical regions of the upper body in accordance to a standard rigidbody marker set, with particular emphasis on wrists, elbows, shoulders, and torso. Hand marker placement was in accordance with the Manus Quantum Metaglove alignment.[7]

The captured data-three signers signing 44 individual signs-was exported in BVH format for post-processing and analysis. Our data processing code is available at: https://github.com/dimitarsh1/CoCoS_dataprocessing.

## 3. Experiments and results

We used the CoCoS MoCap data (Sec. 2) to train various classifiers and to assess their performance.

### 3.1. Objectives

Our experiments aimed to assess the impact of various features on the performance of various temporal-spacial architectures on the task of classification of signs based on inputs in BVH format. The following model families were trained:

[4] In a separate session we also recorded videos of the same 44 signs to be used with a body / pose estimation framework. For the scope of this paper, we only focus on the MoCap data.

[5] https://optitrack.com

[6] https://optitrack.com/software/motive

[7] https://manus-meta.com

LSTM, BiLSTM, CNN1D, CNN+LSTM hybrid, Transformer Encoder, Temporal Convolutional Network (TCN), and Graph Convolutional Network (GCN). In particular, we aim to answer the following research question:

*Which features are of high importance (for MoCap-based inputs) and how they impact the predictive performance of the models and how they impact model performance?*

### 3.2. Data preprocessing

The MoCap data from the three participants were concatenated together to form a larger data set. This raw data was organised at frame level as a tabular feature matrix $X$, and a frame-wise label vector $y$. To enable sequence-based models, the frame stream was converted into fixed-length sequences using a sliding-window segmentation approach. Sequences were generated using a window length of 60 frames and a stride of 20 frames. Formally: $X[t_k : t_k + 60, :]$ where $t_k \in \{0, 20, 40, \ldots\}$.

Importantly, individual annotated sign segments were not treated as single input sequences. Instead, the fixed-length sliding window was applied over the continuous frame stream irrespective of sign boundaries. Consequently, a single annotated gesture could contribute to multiple overlapping windows, and sliding windows may cross sign-segment boundaries, including transitions among different gestures or between gesture and non-gesture frames. Therefore, for a given window $X[t_k : t_k + 60, :]$, the corresponding frame-level labels $y_t$ may contain multiple class identities present in the same window.

Because labels were available per frame, a single class label had to be assigned to each 60-frame window. For each window, the final sequence label was computed using a majority vote over the 60 constituent frame labels: $\hat{y}^{(k)} = \arg\max_c \sum_{t=t_k}^{t_k+59} \mathbf{1}[y_t = c]$. This choice gives precedence to the class most dominant within the window and decreases sensitivity to short label noise or transitional frames.

The resulting dataset of sequences $(2248 \times 60 \times 156)$ was split into train ($72.25\%$ of all data), validation ($12.75\%$ of the data), and test ($15\%$) segments. This validation split was used for early stopping and model selection.

For the dataset format, sequences were wrapped in a PyTorch Dataset and loaded using DataLoaders with batch size 64. Shuffling was only enabled for training to reduce ordering bias.

For graph-based models, the feature vector at each time step was assumed to consist of concatenated joint coordinates, to ensure structural compatibility, the feature dimension $F$ was required to be divisible by the number of coordinate channels per joint $C$, allowing inference of the number of joints $J = F/C$. The input was therefore interpreted as a structured tensor of shape $T \times J \times C$.

A chain-structured adjacency matrix was constructed to define local spatial connectivity between neighboring joints. This simplified topology yields a computationally streamlined and reproducible baseline graph setup.

### 3.3. Models and model setup

We experimented with the following architectures and hyperparameters using the aforementioned same data preprocessing (see Section 3.2) pipeline:

- **LSTM:** with $input\ size = F$, $hidden\ size = 128$, $num\ layers = 1$, $cell = "lstm"$, and $bidirectional = False$.
- **BiLSTM:** identical to the LSTM configuration, but with $bidirectional = True$.
- **CNN1D:** instantiated with $n\ features = F$ and $num\ classes = C$ (convolutional block details follow the repository implementation).
- **TCN:** instantiated with $n\ features = F$ and $num\ classes = C$ (dilation schedule, kernel sizes, and residual block settings follow the repository implementation).
- **Transformer encoder:** instantiated with $num\ classes = C$ (number of heads, encoder depth, and feed-forward dimensions follow the repository implementation).
- **CNN+LSTM hybrid:** with $input\ size = F$, $cnn\ channels = (64, 128)$, $lstm\ hidden = 128$, $lstm\ layers = 1$, and $bidirectional = False$.
- **GCN:** with $num\ joints$, $coords\ per\ joint$, $num\ classes = C$, and $adj\ matrix = adjacency_full$ (graph layer configuration follows the repository implementation).

Training and evaluation settings are as follows: the "Adam" optimizer from PyTorch, 5-fold cross-validation, maximum epochs of 100, learning rate $10^{-3}$, and early stopping patience of 3 epochs without validation improvement (Bergman et al., 2024; Goodfellow et al., 2016). Cross-validation was used to obtain a solid estimate of model performance and reduce dependence on a single train/validation split, given the small dataset. Early stopping was applied to limit overfitting, given that overlapping windows may increase sample similarity.

## 4. Results

### 4.1. Model performance

Table 1 shows the mean F1 scores on the validation set across the seven models, averaged over all

cross-validation folds, For completeness, the post-hoc test F1-scores on the held-out test set are also reported for comparison, although model selection was based exclusively on validation performance. The TCN model achieved the highest performance (best validation F1 = 0.605), followed by the Transformer model (best validation F1 = 0.558). The recurrent models (LSTM, BiLSTM) and the convolution–recurrent hybrids performed significantly worse.

| Model | Val F1 | Test F1 |
|---|---|---|
| TCN | 0.605 | 0.632 |
| Transformer | 0.558 | 0.568 |
| CNN1D | 0.558 | 0.580 |
| GCN | 0.271 | 0.273 |
| BiLSTM | 0.268 | 0.270 |
| CNN+LSTM | 0.268 | 0.270 |
| LSTM | 0.268 | 0.270 |

Table 1: *Validation F1 (mean over folds) and Test F1. Test score is reported post-hoc and was not used for model selection.*

In the final evaluation on the held-out test set, the TCN achieved a test F1-score of 0.57 on the tenth epoch, confirming its robustness and generalizability of the model Figure 1.

To demonstrate the stability of the TCN model, we also show the TCN's training F1 score during the different epochs in Figure 1. The model exhibited consistent improvement in discriminative performance (train F1 from 0.264 to 0.681 over 10 epochs), and the steady increase in training F1-score indicates stable optimisation behaviour across epochs.

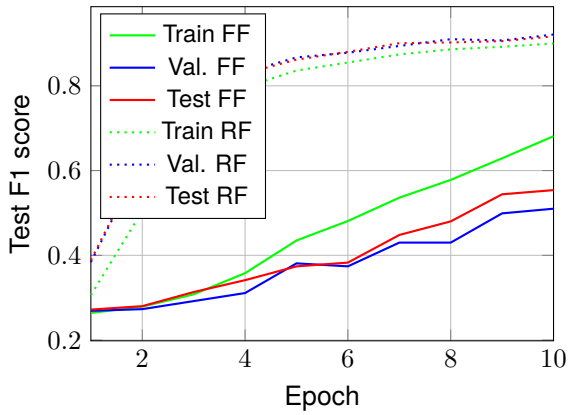



Figure 1: Post-hoc test F1 scores for the best model trained on the full-feature (FF) set, and the reduced-feature (RF) set.

### 4.2. Feature importance

According to Tsinganos et al. (2022), gesture discrimination is mainly influenced by features related to the forearm and hand. This is also shown in our feature importance analysis of the TCN model (Table 2). The 30 features with highest F1-score drop are mainly associated with right-hand articulators, particularly the right forearm and right hand. Given that most of our participants has primarily a dominant right hand, we assert that these results should reflect this feature (hand dominance) rather than be interpreted as absolute.

We assessed feature importance through a model-agnostic permutation importance process. After training, the baseline weighted F1-Score was computed on a held-out set. For each feature, its values were randomly permuted across samples while preserving the temporal structure within each sequence. The model was re-evaluated on the permuted data, and feature importance was defined as the decrease in F1-score relative to the baseline.

This method measures the extent to which the trained models rely on each feature for classification. Because permutation was applied across samples, the analysis captures sequence-level discriminative contribution rather than time-step-specific saliency. To control computational cost, evaluation was performed on a subset of the validation data.

Table 2 shows an F1 drop of 0.212 for the `RightForeArm_Zrotation` (most significant absolute reduction) to 0.100 for the `LeftForeArm_Yrotation` (list significant absolute reduction). This reduction in performance signals a notable degradation in classification performance when temporal patterns were disrupted. Left-sided joints (e.g., `LeftThumb3`, `LeftForeArm`, `LeftPinky2`) also contribute, but overall feature importance is clearly dominated by right-sided joints.[8]

## 5. Unsupervised Structure: Clustering in PCA Space

To further examine the latent structure of the motion feature space, K-means clustering was performed on PCA-reduced representations of the motion features. The optimal number of clusters was determined using silhouette analysis. The silhouette analysis indicated $k = 3$, which produced the highest average silhouette coefficient of 0.558 (Figure 2). Although this value reflects only moderate cluster separability, these results show that signs retain a low-dimensional structure that remains separable even in the absence of label supervision (Rousseeuw, 1987).

[8] As noted above, $\frac{2}{3}$ of our participants use their right hand as dominant. We stress that as hand dominance impacts signing, so would these results be impacted in case the majority of our participants were left-hand-dominant signers. This falls under our limitations and opens the potential for future work.

| Feature idx | Feature name | F1 drop | F1 perm mean | F1 base |
|---|---|---|---|---|
| 81 | RightForeArm_Zrotation | 0.212 | 0.504 | 0.716 |
| 84 | RightHand_Zrotation | 0.189 | 0.527 | 0.716 |
| 27 | LeftHand_Zrotation | 0.170 | 0.545 | 0.716 |
| 86 | RightHand_Yrotation | 0.154 | 0.562 | 0.716 |
| 82 | RightForeArm_Xrotation | 0.148 | 0.567 | 0.716 |
| 36 | LeftHandThumb3_Zrotation | 0.141 | 0.575 | 0.716 |
| 29 | LeftHand_Yrotation | 0.117 | 0.599 | 0.716 |
| 83 | RightForeArm_Yrotation | 0.116 | 0.600 | 0.716 |
| 24 | LeftForeArm_Zrotation | 0.102 | 0.614 | 0.716 |
| 26 | LeftForeArm_Yrotation | 0.100 | 0.616 | 0.716 |

Table 2: *Top 10 Permutation-based feature importance ranked by F1 score drop for the TCN model.*

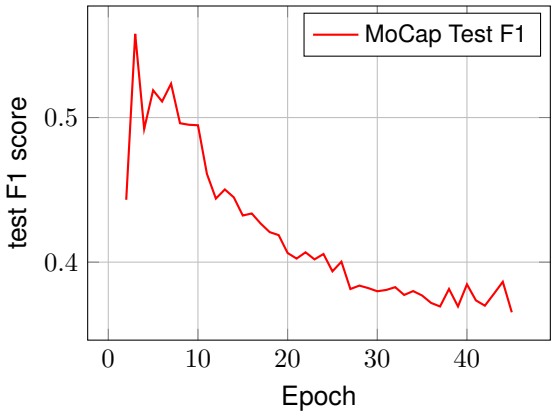



Figure 2: Silhouette score

Correlation matrices revealed very high correlations (> 0.95) between anatomically adjacent joints, including: `RightHandRing2 X_rotation` and `RightHandRing3_Xrotation`; `LeftHandMiddle2_Xrotation` and `LeftHandMiddle3_Xrotation` as well as `Spine_Zrotation` and `Spine1_Zrotation`.

This is expected in articulated-body motion capture: rotations of physically connected joints propagate through the kinematic chain, creating mechanical interdependence. This redundancy partially explains why convolutional models, which leverage local temporal and spatial coherence, outperform recurrent models on this task.

## 6. Models with reduced feature set

By modifying the experimental arrangement to reduce the number of input features, the models were retrained using only the 30 most influential features identified through permutation importance. As reported in Table 3, this feature reduction led to a marked increase in validation F1-scores across all evaluated architectures. For the TCN model, the validation F1-score improved from 0.605 (see Table 1) with the full 156-feature representation to 0.936 using the reduced feature set (see Table 3). This suggests that removing redundant or less significant features may facilitate more successful learning.

Because the feature ranking was derived prior to splitting, these results may be optimistic owing to potential information leakage; we therefore interpret them as indicative rather than conclusive; we leave the detailed analysis for future work.

| Model | Val F1 | Test F1 |
|---|---|---|
| CNN+LSTM | 0.936 | 0.933 |
| TCN | 0.931 | 0.925 |
| CNN1D | 0.925 | 0.921 |
| BiLSTM | 0.920 | 0.918 |
| Transformer | 0.913 | 0.911 |
| LSTM | 0.906 | 0.904 |
| GCN | 0.823 | 0.807 |

Table 3: *Validation F1-scores and Test F1. cross deep learning architectures using top 30 features. Test score is reported post-hoc and was not used for model selection.*

The superior performance of the Temporal Convolutional Network (TCN) compared to the other evaluated architectures can be interpreted in light of the structural properties of the motion capture data and the modeling assumptions underlying each architecture.

### 6.1. Temporal Convolutions and Local Motion Structure

TCNs employ dilated temporal convolutions, which allow the model to detect both short-range and long-distance dependencies within sequential data (Bai et al., 2018). Through stacked dilation layers, the receptive field increases exponentially, empowering the model to integrate information across multiple temporal scales. In the context of sign languages, this design is particularly relevant. Individual signs consist of fine-grained articulatory movements, such as finger and wrist rotations, embedded inside broader motion trajectories of the arm and upper body. The dilated convolutional

structure may therefore facilitate the simultaneous modeling of micro-level kinematic variations and macro-level movement patterns.

Given the large dimensionality and arranged nature of the BVH feature encoding, the convolutional approach appears well-suited to capturing locally coherent movement sequences while maintaining sensitivity to longer temporal connections.

### 6.2. Limitations of Recurrent Models for High-Dimensional Data

In contrast, recurrent models such as LSTMs and BiLSTMs process each time step sequentially and operate on flattened feature vectors. While recurrent models are intended to capture temporal connections, they do not clearly incorporate spatial relationships among joints unless additional framework constraints are introduced. Although CNN+LSTM combines spatial feature extraction with temporal recurrence, the added architectural complexity does not necessarily yield better generalization when the temporal structure is already well captured by expanded convolutions (Jafari and Jafari, 2026). Prior work has shown that explicitly modeling skeletal topology can improve performance in motion recognition tasks (Yan et al., 2018).

Furthermore, the comparatively lower performance of recurrent models may be attributed to their limited ability to exploit spatially structured correlations among joint rotations. Given the strong interdependencies observed between anatomically adjacent joints, architectures that leverage local feature coherence appear better suited to this representation.

### 6.3. Transformer Performance and Data Requirements

Transformer-based models rely on self-attention modules to model long-range dependencies across sequences (Vaswani et al., 2017). In principle, this allows for flexible modeling of long-range interactions between time steps. However, Transformers benefit from large-scale training data to obtain robust focus patterns and minimise overfitting.

As shown in both Table 1 and Table 3, the Transformer achieved competitive but slightly lower performance compared to the TCN. This may reflect insufficient data for the attention mechanism to fully exploit its modeling capacity; we ought to stress that our dataset is relatively small and acknowledge the need for further experimentation with a much larger dataset. Nevertheless, the strong performance of the Transformer suggests that attention-based architectures remain an encouraging path for larger motion capture datasets.

## 7. Analysis

Permutation importance evaluation showed that perturbing right-hand rotational channels yielded the largest decreases in weighted F1-score, indicating their central role in classification performance. If we consider that two of the signers who were recorded had their right hand as dominant, we might see this finding as consistent with phonological descriptions of sign languages, in which the dominant hand typically carries the primary articulatory load (Brentari, 1998; Sandler and Lillo-Martin, 2012). However, at the current stage we cannot be certain of the role played by signer's handedness. Further data collection and experiments should be performed to test this hypothesis.While the present study does not aim to validate linguistic theory, the correspondence between model-derived feature significance and phonological structure strengthens the understandability of the results.

Unsupervised clustering further supports this observation. The emergence of moderately separable clusters in PCA-reduced space suggests that sign instances retain structured organization even within lower-dimensional representations. The fact that the most influential PCA components correspond predominantly to right-hand rotations reinforces the importance of distal articulators in discriminating signs.

### 7.1. Redundancy and correlated joint movements

Correlation analysis showed very high correlations between anatomically adjacent joints, commonly exceeding 0.95. This superfluity is consistent with the mechanical dependencies intrinsic in articulated-body kinematics. Rotational changes in one segment of the kinematic chain naturally propagate to connected segments, resulting in highly correlated feature channels (Morasso, 2025).

From a machine learning perspective, such redundancy increases feature dimensionality without proportionally increasing discriminative information. This phenomenon likely contributes to the comparatively lower performance recorded when models are trained on the full 156-feature dataset. The substantial performance increases observed in the reduced-feature experiment suggest that excluding redundant or weakly informative channels can simplify the learning task and improve generalization.

### 7.2. Architectural implications

The comparative performance of the evaluated models provides insight into how temporal-spatial inductive biases interact with skeletal motion data.

The excellent performance of the TCN suggests that convolutional modelling of time sequences is

notably well-suited to motion-capture–based sign classification. Dilated temporal convolutions allow the model to integrate fine-grained articulatory movements with wider motion trajectories, while preserving local integration in feature space. This inductive bias appears advantageous in representations characterised by strong local correlations and structured redundancy.

In contrast, recurrent models process flattened feature vectors sequentially and do not explicitly encode spatial structure unless further mechanisms are introduced. Given the high degree of inter-joint correlation observed in the dataset, this lack of overt spatial modelling may limit the effectiveness of these models. Furthermore, LSTM (and BiLSTM) suffer from forgetting issues, especially with long sequences. Recurrent neural networks (RNNs) suffer from the vanishing gradient descent problem; Long-short term memory (LSTM) units were proposed to address the aforementioned forgetting issue, however they still struggle with long sequences. The work of (Bahdanau et al., 2015; Luong et al., 2015) proposes the attention mechanism for RNNs / LSMTs, reducing significantly the forgetting issue in such networks. The observed results suggest the presence of the forgetting issue and, consecutively, that the encoded sequences perhaps contain significant spatio-temporal variations.

Transformer-based models achieved competitive performance, indicating that attention components can successfully capture long-range dependencies in motion sequences. Nevertheless, their slightly lower performance relative to the TCN may reflect the limited dataset size, as attention-based architectures typically benefit from larger training corpora (Abdullayeva and Alishzade, 2025).

## 8. Conclusions

This work investigated the role of feature importance in motion capture representations for deep learning-based sign classification. Specifically, BioVision Hierarchy (BVH) motion-capture data where evaluated across multiple temporal models, including recurrent, convolutional, transformer-based, as well as graph-based models. The objective was to examine how feature saliency patterns relate to model performance and to assess whether high-dimensional skeletal representations are fully required for accurate sign recognition.

Taken together, the results show that motion-capture representations of isolated signs contain substantial structured redundancy, with distinctive information concentrated in a relatively small subset of distal articulatory features. This observation suggests that extremely high-dimensional skeletal representations may not be strictly indispensable for effective sign classification. Instead, targeted feature selection or structured dimensionality minimization may improve computational effectiveness while preserving, or even enhancing, predictive performance. Furthermore, the comparative results across architectures indicate that models that exploit local temporal–spatial coherence are particularly well suited to BVH-based sign modeling.

**Limitations and future work** At the same time, the aforementioned conclusions must be interpreted in light of the study's limitations. The small dataset consists (three participants signed 44 signs),[9], a limited number of participants (only 3) and was collected under monitored conditions, which restricts the generalizability of the findings. Validation on larger and more diverse sign-language corpora would be necessary to determine whether the observed architectural advantages and feature-importance patterns extend beyond the present dataset. Forthcoming experiments should includes a larger group of left-hand- and right-hand-dominant signers to allow for the thorough investigation of handedness. Furthermore, future studies should include analyses of by-signer attributes to get a more fine-grained understanding of feature importance linked to individual signers.

## Bibliographical References


Gulchin Abdullayeva and Nigar Alishzade. 2025. A study on recurrent, attention-based, and hybrid neural architectures for sign language recognition. *Informatics and Control Problems*.

Jari Ivar Andersen, Gomer Otterspeer, Robert Belleman, and Floris Roelofsen. 2025. Designing a marker based motion capture setup for sign language research. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv*.


[9] Ideally this would imply 132 sign MoCap recordings; however, as not all signs were signed by all participants, our dataset includes 116 sign MoCap recordings

Edward Bergman, Lennart Purucker, and Frank Hutter. 2024. Don't waste your time: Early stopping cross-validation. *arXiv*.

Lucas Bernhard, Fabrizio Nunnari, Amelie Unger, Judith Bauerdiek, Christian Dold, Marcel Hauck, Alexander Stricker, Tobias Baur, Alexander Heimerl, Elisabeth André, Melissa Reinecker, Cristina España Bonet, Yasser Hamidullah, Stephan Busemann, Patrick Gebhard, Corinna Jäger, Sonja Wecker, Yvonne Kossel, Henrik Müller, Kristoffer Waldow, Arnulph Fuhrmann, Martin Misiak, and Dieter Wallach. 2022. Towards automated sign language production: A pipeline for creating inclusive virtual humans. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '22, page 260–268, New York, NY, USA. Association for Computing Machinery.

Félix Bigand, Elise Prigent, and Annelies Braffort. 2020. Person identification based on sign language motion: Insights from human perception and computational modeling. In *Proceedings of the 7th International Conference on Movement and Computing*, MOCO '20, New York, NY, USA. Association for Computing Machinery.

Félix Bigand, Elise Prigent, Bastien Berret, and Annelies Braffort. 2021a. Decomposing spontaneous sign language into elementary movements: A principal component analysis-based approach. *PLOS ONE*, 16(10):1–18.

Félix Bigand, Elise Prigent, Bastien Berret, and Annelies Braffort. 2021b. Machine learning of motion statistics reveals the kinematic signature of the identity of a person in sign language. *Frontiers in Bioengineering and Biotechnology*, Volume 9 - 2021.

Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. Mit Press.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv*, abs/1812.08008.

Mathieu De Coster, Ellen Rushe, Ruth Holmes, Anthony Ventresque, and Joni Dambre. 2023. Towards the extraction of robust sign embeddings for low resource sign language recognition.

Maria-Paola Forte, Peter Kulits, Chun-Hao Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. 2023. Reconstructing signing avatars from video using linguistic priors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12791–12801.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. 2023. Improving 3d pose estimation for sign language. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Alireza Jafari and Fatemeh Jafari. 2026. How much temporal modeling is enough? a systematic study of hybrid cnn-rnn architectures for multi-label ecg classification. *arXiv*.

Pavel Jedlička, Zdeněk Krňoul, Jakub Kanis, and Miloš Železný. 2020. Sign language motion capture dataset for data-driven synthesis. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 101–106, Marseille, France. European Language Resources Association (ELRA).

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Pietro Morasso. 2025. Coordinating the redundant dofs of humanoid robots. *Actuators*, 14(7). 354.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 744–772. Association for Computational Linguistics.

Lucie Naert, Caroline Larboulette, and Sylvie Gibet. 2020. LSF-ANIMAL: A motion capture corpus in French Sign Language designed for the animation of signing avatars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6008–6017, Marseille, France. European Language Resources Association.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Syst. Appl.*, 164:113794.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Wendy Sandler and Diane Lillo-Martin. 2012. *Sign Language and Linguistic Universals*. Cambridge University Press.

Dimitar Shterionov, Lorraine Leeson, and Andy Way. 2024. The pipeline of sign language machine translation. In Andy Way, Lorraine Leeson, and Dimitar Shterionov, editors, *Sign Language Machine Translation*, pages 1–25. Springer Nature Switzerland, Cham.

Panagiotis Tsinganos, Bart Jansen, Jan Cornelis, and Athanassios Skodras. 2022. Real-time analysis of hand gesture recognition with temporal convolutional networks. *Sensors*, 22(5). 1694.

Vincent Vandeghinste, Mirella De Sisto, Santiago Egea Gómez, and Mathieu De Coster. 2024. Challenges with sign language datasets. In Andy Way, Lorraine Leeson, and Dimitar Shterionov, editors, *Sign Language Machine Translation*, pages 117–139. Springer Nature Switzerland, Cham.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Red Hook, NY, USA.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-based Action Recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

# Leveraging Text-side Augmentation For Sign Language Translation

**Diandra Fabre[1], Julie Lascar[2], Julie Halbout[2], Markarit Vartampetian[1]**
[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
[2]Univ. Paris-Saclay, CNRS, LISN
{diandra.fabre, markarit.vartampetian}@univ-grenoble-alpes.fr
{julie.lascar, julie.halbout}@universite-paris-saclay.fr

**Abstract**

Sign language translation faces significant challenges due to the scarcity of annotated data and the inherent complexity of sign languages. This paper presents a method to improve sign-to-text translation models by augmenting data on the text side. We conduct experiments using two state-of-the-art models on two publicly available datasets: PHOENIX-2014T for German Sign Language and Mediapi-RGB for French Sign Language. Our main contributions are : (1) augmenting the training sets of both datasets on the text side using a generative model, (2) evaluating the impact of paraphrasing on BLEU and BLEURT scores, and (3) analyzing the impact of paraphrasing on translation outputs. We observed a significant improvement in translation for both languages. This suggests that adding variability to the training dataset through paraphrasing can lead to better generalization of the models. These results are comparable to state-of-the-art methods that use more complex approaches, such as Visual-Language fine-tuning, to improve translation.

**Keywords:** low-resources, sign language, data augmentation, translation

## 1. Introduction

Sign languages (SL) are visual languages used within Deaf communities. While spoken and written languages encode information sequentially, sign languages simultaneously use multiple articulators (hands, face, and body) to convey meaning (Sandler and Lillo-Martin, 2006). A single signed utterance can express information that would require multiple sequential units in a spoken language. Sign languages also make grammatical use of the signing space, encoding syntactic and referential relations spatially rather than through word order. As a result, sign and written languages differ fundamentally in their grammatical organization, and direct translation is not always possible. Speech-to-text or text-to-text translation likely requires less data than sign-to-text translation to reach the same performance. In the machine translation (MT) domain, a dataset of 6000 standard-length sentences is already considered tiny (De Coster et al., 2021; Gu et al., 2018). Given the additional complexity of sign languages described above, we hypothesize that reaching equivalent performance would require substantially more data than for sequential language pairs.

SL datasets are hard to collect and are several scales smaller than those used for speech or text translation [1]. However, text data are relatively easy to acquire as shown by the development of Large Language Models (LLM). These LLM are trained on an extensive amount of data, aggregated from open resources on the Internet, such as Wikipedia data or publicly available books. Millions of sentences are fed into these models, which learn language representations and perform on various tasks, such as summarizing, translation or generation. Using LLM knowledge to guide our translation system on the text side should improve translation quality and generate syntactically correct sentences.

In this paper, we relied on the paraphrasing capability of generative models. Today, human interpreters remain the most accurate translators between sign and spoken languages. A SL interpreter is a certified professional who facilitates communication between Deaf and hearing individuals, translating bidirectionally between spoken and signed language, as well as between written text and sign language. They act as cultural mediators, promoting accessibility in professional, social, and civic life. Their role requires bilingualism and deep cultural awareness, ensuring effective communication across both languages and communities.

Interpretation refers to real-time translation to or from signing, while translation is done offline with the possibility of correcting the proposed results. We will use the term "translation" in the following sections, and differentiate between the two terms if necessary. As with spoken language translation, variations in translation between different humans might be observed. Liu and Dou (2023) experienced the lexical diversity on simultaneous interpretation between different languages, and shows that interpreters used simplification techniques such as paraphrasing or summarizing, but are also required to add some explaining parts depending

[1]for more details see the "List of datasets": https://research.sign.mt/



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 129–139
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

on the specificities of the languages. Translating between any sign language in a 3D configuration, and a sequential language should require the same techniques.

When translating from spoken to signed language, interpreters must hear the speaker, meaning they usually work into a non-native language unless they are hCODAs (hearing Children of Deaf Adults). Emerging practices introduce Deaf interpreters as intermediaries between hearing interpreters and the audience, resulting in more natural signed expression. In different contexts, in a similar situation, hearing and deaf interpreters produce different signed language (Stone and Russell, 2011). Interference from the source language can cause the signed output to mirror spoken grammar rather than follow natural signed language conventions (Dayter, 2019). Halley (2020) also showed the contextual and personal influence of the interpreter on the translation. The work of Archambeaud (2022) also provides several potential influences of the interpretation. The media used for conveying the interpretation influence the sign space and thus, the way the interpreter signs. The physical space and work conditions can be factors of variability. The origin, political beliefs, or discrimination that could exist against the interpreter can also influence the way some specific subjects would be signed, even if the interpreter mission implies total neutrality.

We aim to reproduce this translation variability using generative models to artificially increase available datasets. These models, built on the Transformer architecture, are pretrained on large corpora, and have proven effective in diverse text generation tasks, including paraphrasing. By doing that, we hope to reduce model overfitting and increase its understanding of sign language by providing a bigger dataset. To our knowledge, this is the first systematic study of text-side paraphrasing across two SL datasets. While prior work explored paraphrasing in a gloss-to-text setting and reported no significant improvements (Angelova et al., 2022), we address the video-to-text translation task. Our main contributions are the following: (1) Augmenting the training set of two datasets (French and German SL) using a generative model on the text side; (2) Comparing two state-of-the-art architectures for small datasets with two different SL, and observing the impact of language-dependent pretrained models on the results ; (3) Evaluating the impact of paraphrasing on the scores and the translation output.

## 2. Related Works

Most research on Sign Language Translation (SLT) follows methodologies from text-to-text or speech-to-text translation. Recent works have focused on Transformer-based encoder-decoder architectures, originally introduced by Vaswani (2017).

The most straightforward application of Transformer-based encoder-decoder architecture for SLT was first proposed by Camgoz et al. (2020) and later optimized by Sincan et al. (2023), both on PHOENIX-2014T dataset (Forster et al., 2012). The encoder output was modified to simultaneously perform classification via a Connectionist Temporal Classification (CTC) decoder in parallel with the decoder used for next token prediction. This architecture enables both direct sign-to-text translation and translation guided by sign classification. These models show that gloss annotations can be useful, but glosses remain costly to annotate and are not always available. We use glosses in our study when provided, but our augmentation only increases the text side.

Recent research has explored integrating Large Language Models (LLMs) into Sign Language Translation (SLT). LLMs, based on the Transformer architecture, include encoder-only models for language representation, decoder-only models for language generation, and encoder–decoder models. In De Coster et al. (2021), a frozen BERT model was used to enhance SLT performance. Their architecture is an extension of the model introduced by Camgoz et al. (2020). Weights of either the decoder or both the encoder and the decoder were replaced by those of a BERT model: only the cross-attention and the fully connected layer on the decoder side were updated during training. They used initial BERT model trained on English. In our study, we evaluated the impact of a language-specialized BERT depending on our dataset.

Other works have investigated fine-tuning LLMs for SLT. In the works of Uthus et al. (2024) and Tarrés et al. (2023), a T5 encoder-decoder model was fine-tuned on an American Sign Language dataset using Mediapipe features (Lugaresi et al., 2019) extracted from YouTube videos, improving translation accuracy. Meanwhile, Sincan et al. (2024) trained a sign-spotting model on a large dataset, passing the sequence of detected signs to ChatGPT for sentence generation. Similarly, Wong et al. (2024) proposed Sign2GPT, a method using a frozen GPT decoder to enhance text generation. The authors reported a slight improvement on the BLEU score compared to previous methods, as we can observe from Table 1 for PHOENIX-2014T dataset if compared to the baseline Sign2gls2txt. Moreover, fine-tuning such large models requires a huge computational capability and also represents an ecological cost to take into account.

Recent studies have leveraged architectures that integrate both visual and language modalities to address Sign Language Translation (SLT). For example, Hwang et al. (2025) introduced a gloss-free

| Model | Dataset | Glosses | BLEU-4 | BLEURT |
|---|---|---|---|---|
| SLRT (Camgoz et al., 2020) | PHOENIX-2014T | yes | 20.85 | 49.02 |
| BERT2RND (De Coster et al., 2021) | PHOENIX-2014T | yes | 22.47 | x |
| BERT2BERT (De Coster et al., 2021) | PHOENIX-2014T | yes | 21.29 | 49.42 |
| Sign2GPT (Wong et al., 2024) | PHOENIX-2014T | no | 22.52 | x |
| DVE-SLT (Sincan and Bowden, 2025) | PHOENIX-2014T | no | 23.81 | 53.59 |
| SpaMo (Hwang et al., 2025) | PHOENIX-2014T | no | 24.32 | x |
| Youtube-ASL (Uthus et al., 2024) | How2sign, Youtube-ASL | no | 12.39 | 46.63 |
| SpaMo (Hwang et al., 2025) | How2Sign | no | 10.11 | 42.23 |

Table 1: Selected state-of-the-art models for SLT and their reported BLEU scores and BLEURT scores when available.

SLT framework that achieves state-of-the-art performance on Phoenix14T without relying on gloss annotations. Their approach employed off-the-shelf visual and video encoders (e.g., ViT and VideoMAE) combined with fine-tuned LLMs (e.g., mBART-L, mT0-XL, Flan-T5-XL, and Llama-2). Similarly, Jang et al. (2025) used BLIP-2 (Language–Image Pre-training with Frozen Image Encoders and LLMs) to extract visual cues from the video background, alongside a Video Swin Transformer pre-trained on a sign recognition task to encode the signs. Translation was then performed using a LLaMA model, resulting in improved SLT performance. Finally, the work of Sincan and Bowden (2025) presented DVE-SLT, a two-phase, dual visual encoder framework for gloss-free SLT, leveraging contrastive visual–language pretraining to align complementary visual features with sentence-level text embeddings. Most of these experiments results are reported in Table 1, on different datasets: PHOENIX-2014T (Forster et al., 2012), How2Sign (Duarte et al., 2021), Youtube-ASL (Uthus et al., 2024). When available, we reported both BLEU-4 and BLEURT scores (defined in Section 3.3). These approaches demonstrate the potential of LLMs in SLT, but they require large datasets for fine-tuning, and thus heavy computational resources and ecological impact.

In this paper, we focus on lightweight architectures, as opposed to the finetuning of LLMs. We conduct experiments on two datasets: German Sign Language dataset PHOENIX-2014T (Forster et al., 2012), and French Sign Language dataset Mediapi-RGB (Bull et al., 2024). We explore the data augmentation on the text side to improve translation performances, by taking advantage of generative models knowledge on languages in their text format. Data augmentation for low-resource machine translation has been widely explored for spoken languages (Fadaee et al., 2017), and more specifically applied to speech-to-text on low-resourced languages (Mi et al., 2022). For sign language, Jang et al. (2022) applied text-side augmentation with back-translation and paraphrasing on Korean Sign Language gloss data with modest improvements, and Angelova et al. (2022) experimented with paraphrasing on PHOENIX 2014T in a gloss-to-text setting. In contrast, our work applies LLM-based paraphrasing to the direct video-to-text translation task and obtains interesting improvements across two SL datasets.

## 3. Methodology

### 3.1. Datasets

**PHOENIX-2014T** is a dataset of interpreted German Sign Language (Forster et al., 2012). It consists of daily weather forecast translated in sign language. It is composed of 9 signers, with 1081 gloss vocabulary, and 7096 sentences for the training dataset. The video embedding is based on 2D CNN model fine-tuned on the same sign language dataset (Momeni et al., 2022), with a sliding window of 16 frames resulting in 1024-dimensional feature vectors (embeddings).

**Mediapi-RGB** is a dataset of news and reports in LSF, produced by deaf journalists between 2018 and 2022 (Bull et al., 2024). These videos were subtitled in French after recording. It is composed of 10 main signers, with a 445 sparse gloss vocabulary. These gloss annotations have been shown to improve translation performance (Fabre et al., 2025). Mediapi-RGB is composed 27 840 sentences after deduplication for the training dataset. The video embedding is based on SWIN model architecture from Liu et al. (2022) fine-tuned on a large-scale dataset of isolated signs in British Sign Language (Momeni et al., 2022), with a sliding window of 16 frames resulting in 768-dimensional embedding features. The authors of the Mediapi-RGB dataset provided SWIN features along with the dataset, we decided to redo the fine-tuning to have full control of our experiment, starting from the videos. Thus, baseline results may differ from other contributions on the Mediapi-RGB dataset. The main difference from the German Sign Language dataset is that the task is translation from sign to text, rather than interpretation from speech

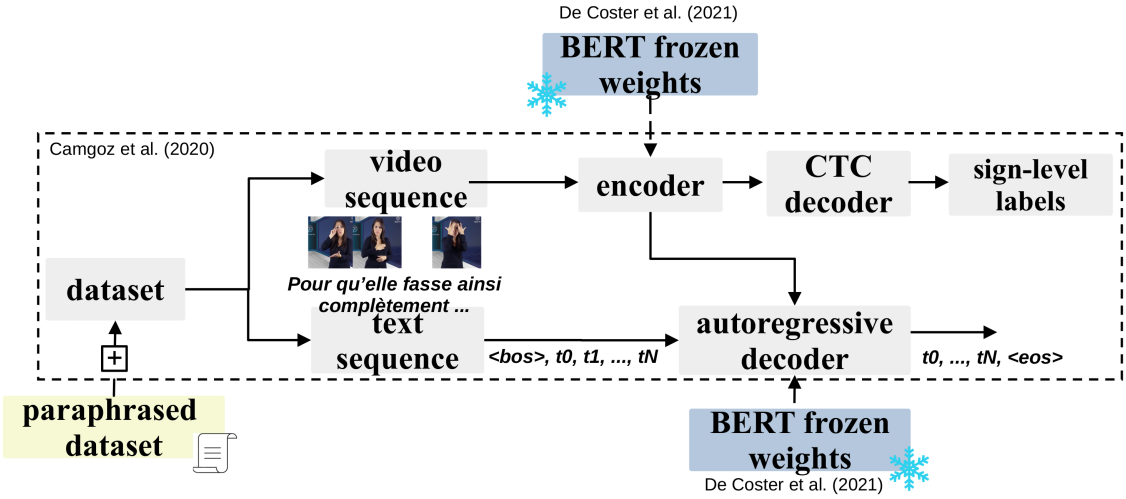



Figure 1: Simplified illustration of the architecture proposed by Camgoz et al. (2020) with our different proposed modifications: a language-specialized BERT model using De Coster et al. (2021) frozen-weights approach, and a data augmentation on the text side to increase dataset.

to sign. While the German SL dataset covers only one topic, the French SL dataset covers multiple topics accross multiple years.

## 3.2. Model Architectures

In this work, two architectures were used. The first architecture tested is the architecture proposed in Camgoz et al. (2020) and will be called *Baseline*. We used the configuration proposed by the authors, which consists of an encoder-decoder Transformer architecture with a CTC layer at the end of the encoder for classification purposes. This architecture provides both Word Error Rate (WER) for sign recognition, and BLEU score (defined in Section 3.3) for sign translation. The loss is computed as the weighted sum of CTC loss and cross-entropy loss. The second architecture, proposed by De Coster et al. (2021), consists of a Frozen Pre-trained Transformer. We will use the acronym *FPT* throughout this paper. It consists of using BERT pre-trained encoder-only model as initialization for some of the weights of both the encoder and the decoder. Only the cross-attention and the fully connected layer were updated during training. Figure 1 schematizes the architecture of the model and the two main modifications: the FPT approach using frozen weights, and the data augmentation on the text side. The encoder (resp. decoder) of the model is made of 4 (resp. 3) transformer blocks with 8 heads used in each self-attention layer. As Camgoz et al. (2020), we applied a weight of 5 for the CTC (classification) loss and 1 for the translation loss. For the remaining parameters, we kept the configuration proposed by Camgoz et al. (2020).

### 3.2.1. BERT Models

BERT-based models can also provide a structured representation space that captures linguistic proximities, and be used to tokenize the text before translation for the Baseline architecture. For *FPT* architecture, a BERT model is required to initialize some weights of the models.

The model initially used for FPT initialization was the initial `bert-base-uncased`. As we wanted to see the influence of language-specific model on performances, we changed this depending on the target language. For the German SL dataset, we chose the German BERT model `dbmdz/bert-base-german-uncased` . This is a German-only encoder, and it is uncased, as is the PHOENIX-2014T dataset. As for the model proposed by Chan et al. (2020), this model is trained on German Wikipedia dump, OpenLegalData dump and news articles. For the French SL dataset, we chose the French model CamemBERT [2] (Martin et al., 2020), trained on 138GB of French text.

## 3.3. Metrics

The BLEU metric is widely used in SL translation papers. It relies on the match of a chain of n-words between reference and estimated sequence Papineni et al. (2002). This metric suffers from its lack of adaptability regarding the possible variability when one sentence can have multiple translations. The BERT space can be used to evaluate the translation performances of a model. While BLEU score relies on a sequence of exactly matching words,

[2] https://huggingface.co/docs/transformers/en/model_doc/camembert

BLEURT score will rely on BERT space to determine the proximity between the predicted sequence and the reference sentence (Sellam et al., 2020). BLEURT score is based on a multilingual BERT space tested on 13 languages (including German, English and French for instance). It has been shown to be closer to human judgment than BLEU score for translation evaluation (Sellam et al., 2020). The BLEURT score range is between 0 and 100 (or 0 and 1) but may output values outside this range.

### 3.4. Text-Side Data Augmentation

Data augmentation is a way of artificially increasing the size of a given dataset to improve performance and reduce the risk of overfitting. In this work, we propose a dataset augmentation on the text side by doubling the training dataset. For this paraphrasing task, we used the GPT API (Paredes et al., 2023). We chose GPT-4o-mini for its strong instruction-following capabilities and paraphrase quality. While we acknowledge that using a closed-source model is not ideal for reproducibility, recent benchmarking from Beauchemin et al. (2025) shows that the best open-source alternatives (Qwen2.5-32b, Llama-3.1-8B-Instruct) perform approximately 20 points below GPT-4o-mini on French paraphrase quality metrics. Larger open-source models like Pixtral-Large (123B) approach comparable performance but require substantially more resources.

Using GPT-4o-mini allowed us to focus on data augmentation rather than model-specific tuning, as our pipeline is model-agnostic with minimal prompting (limited inference steps, low engineering overhead). Future work will systematically compare open-source alternatives to assess whether the performance gap justifies the trade-offs in cost and openness.

**German data augmentation** was performed with the following prompt: "Paraphrasiere einfach den folgenden Satz auf Deutsch:" (en: *Simply paraphrase the following sentence in German*).We calculated BLEURT and sacreBLEU scores as similarity metrics between the original training dataset and the paraphrased training dataset. We obtained a BLEURT score of `71.97` and a BLEU score of `17.82`. The training dataset was doubled using the proposed paraphrasing approach. All texts were provided in lowercase, as in the original version of dataset. This resulted in an increase in the size of the dataset vocabulary from 3000 to 4623 words. Table 2 shows some examples of paraphrasing.

**French data augmentation** Since the dataset is composed of parts of sentences, we used the contextual parts to paraphrase simultaneously all parts of a sentence. The following prompt was used : messages=[ "role": "system", "content": "Tu es un assistant linguistique expert en français." (en: *You are a language assistant specialising in French.*), "role": "user", "content": "Ta tâche est de paraphraser la phrase suivante en gardant la segmentation et le format. Portion à paraphraser :" (en: *Your task is to paraphrase the following sentence whilst retaining the line breaks and formatting. Portion to be paraphrased:*) ]

We obtained a BLEURT score of `72.78` and a BLEU score of `23.89`. This resulted in an increase in the size of the dataset vocabulary from 36160 to 37169 words. Table 3 shows some examples of paraphrases.

## 4. Experiments and Results

Each run of a model was estimated to produce about 262.88 gCO2e on an NVIDIA GTX 1080 Ti [3]. Each request on GPT API is estimated to produce 0.38 gCO2e / query [4], so a total of 2696.48 gCO2e for paraphrasing the German SL training dataset, and a total of 10579.58 gCO2e for the French SL training dataset. This last part could be reduced by running generative models locally.

### 4.1. German Dataset

As shown in Table 4, doubling the training dataset with paraphrases increases the overall BLEURT score on the test set, but results in a decrease of the BLEU score. This aligns with the observations made in Section 3.3. BLEU score is sensitive to word substitutions and perform poorly when synonyms are used, as it evaluates precise n-grams sequences. We can make the hypothesis that the BLEURT metric recognize paraphrasing as a different version of the same sentence, and can give a high score to a sequence that differs from the reference, whereas the BLEU score will assign a low score. Table 5 shows examples of the results on the test set after data augmentation.

We observed an improvement of more than 20 points of BLEURT for 113 sentences, and an improvement of more than 30 points for 48 of the test sentences (out of 642). Augmenting the dataset on the text side by paraphrasing the train dataset definitively improves the performances of the model. Table 5 shows examples of sentences where the BLEURT score increases when using data augmentation.

All pairwise differences between experiments were evaluated for statistical significance using a t-test ($p < 0.001$). We observed no significant difference between the Frozen Pre-trained Transformer and the random weight initialization, either with or without data augmentation. These results partially replicate those of De Coster et al. (2021) on

[3] https://calculator.green-algorithms.org/

[4] https://www.climateqa.com/docs/carbon-footprint/

| | |
|---|---|
| Sentence | **heute nacht kühlt die luft ab auf vierzehn bis acht grad.** *(tonight the air will cool down to fourteen to eight degrees.)* |
| Paraphrase | in der nacht wird es auf 14 bis 8 grad abkühlen. *(at night it will cool down to 14 to 8 degrees.)* |
| Sentence | **im süden kaum schauer.** *(hardly any showers in the south.)* |
| Paraphrase | es gibt im süden fast keine regenfälle. *(there is almost no rainfall in the south.)* |

Table 2: Examples of sentences and their paraphrases in German as generated by GPT API, along their translation in English.

| | |
|---|---|
| Sentence | **mais son apport à l'équipe n'avait pas été notable.** *(but his input to the team had not been noteworthy.)* |
| Paraphrase | mais sa contribution à l'équipe n'avait pas été significative. *(but his contribution to the team had not been significant.)* |
| Sentence | **handicap : mobilisation prévue le 5 mars.** *(handicap : mobilization scheduled for march 5.)* |
| Paraphrase | handicap : rassemblement programmé le 5 mars. *(handicap: rally scheduled for march 5.)* |

Table 3: Examples of sentences and their paraphrases in French as generated by GPT API, along their translation in English.

| German Sign Language (PHOENIX-2014T) | | | French Sign Language (Mediapi-RGB) | | |
|---|---|---|---|---|---|
| **Configuration** | **BLEURT** | **BLEU** | **Configuration** | **BLEURT** | **BLEU** |
| Baseline | 49.02 | 20.85 | Baseline | 10.55 | 4.14 |
| Bas. + Data aug. | **55.01** | 17.35 | Bas. + Data aug. | **11.56** | **4.71** |
| Bas. + Data aug. + BERT tok. | 51.70 | 20.71 | Bas. + Data duplication | 10.72 | 4.18 |
| FPT | 49.42 | **21.29** | FPT | 9.68 | 3.70 |
| FPT + Data aug. | 54.71 | 19.19 | FPT + Data aug. | 10.17 | 3.91 |

Table 4: BLEURT and BLEU scores for different configurations: baseline from Camgoz et al. (2020), BERT tokenizer and FPT (De Coster et al., 2021), adding paraphrasing.

PHOENIX-2014T. Data augmentation, in both configurations, surpasses the baseline dataset with significant difference. We conducted two supplementary experiments exploiting BERT space. We applied tokenization to PHOENIX-2014T text data. Adding a BERT tokenizer to the text increases drastically the size of the vocabulary (from 4623 words to 31106 tokens) and results in a slight drop of BLEURT score for our model.

## 4.2. French Dataset

Based on the results obtained on the German SL dataset, we reduced the number of experiments in French SL to four, as shown in Table 4: *Baseline* with and without data augmentation, and *FPT* with and without data augmentation. As for German SL dataset, paraphrase augmentation significantly improves the results. All experiments show significant differences (t-test with $p < 0.001$). However, unlike on PHOENIX-2014T, we observe a loss in performance when using the *FPT* architecture on Mediapi-RGB (scores of 10.55 to 9.68 on BLEURT). We attribute this to the dataset's larger size, multi-topic diversity, and vocabulary exceeding CamemBERT's token capacity, which frozen weights cannot accommodate. We observed an improvement of more than 20 points in BLEURT for 289 sentences, and an improvement of more than 30 points for 125 test sentences (out of 8060). We also compared the results between doubling the dataset by repeating each sentence twice, instead of using paraphrases. Dataset duplication resulted in a score close to the baseline (10.72 in BLEURT against 10.55), with no statistical difference between the two experiments.

From the examples displayed in Table 6, we observe the relevance of augmented datasets for the model. For the first example "mais le ministre français de l'économie bruno le maire", the sequence of sign produced in the video is as follows: " mais | ministre | économie | France | nom | POINTAGE | dire ", the last element "dire" ('to say') does not appear in the reference [5] but it ap-

[5]This last word belongs to the next subtitle, but since videos are overlapping for practical reasons, some signs appears in one video and are not translated in the subtitle.

| Model | Sentence | BLEURT |
|---|---|---|
| Ref. | **später ist es meist trocken.** *(later it is mostly dry.)* | |
| Baseline | ansonsten aber trocken ist es trocken. *(but otherwise dry it is dry.)* | 20.62 |
| Augm. | später wird es dort trocken sein. *(later it will be dry there.)* | 72.61 |
| Ref. | **morgen schwacher bis mäßiger ostwind.** *(tomorrow there will be a light to moderate easterly wind.)* | |
| Baseline | dort weht morgen meist schwach bis mäßig. *(tomorrow it will be mostly light to moderate there.)* | 40.07 |
| Augm. | morgen weht ein schwacher bis mäßiger wind aus östlichen richtungen. *(tomorrow there will be a light to moderate wind from the east.)* | 78.01 |
| Ref. | **in den alpen wird es dann sogar schneien.** *(it will even snow in the Alps.)* | |
| Baseline | am alpenrand da regen. *(on the edge of the Alps there will be rain.)* | 22.31 |
| Augm. | an den alpen fällt noch schnee. *(on the Alps there will still be snow.)* | 59.67 |
| Ref. | **an den temperaturen ändert sich wenig.** *(there is little change in the temperatures.)* | |
| Baseline | das ändert sich wenig. *(that changes little.)* | 39.58 |
| Augm. | die temperaturen ändert sich wenig. *(the temperatures changes little.)* | 79.10 |

Table 5: Examples of German Sign Language to text translation with an improvement of more than 20 points in BLEURT score between baseline and augmented dataset. *Ref.* is the reference, *Baseline* refers to Baseline model without augmentation, and *Augm.* refers to baseline + data augmentation.

| Model | Sentence | BLEURT |
|---|---|---|
| Ref. | **mais le ministre français de l'économie bruno le maire,** *(but french economy minister bruno le maire,)* | |
| Baseline | le ministre de la santé olivier véran a indiqué que le ministre de la france. *(health minister olivier véran said the french minister.)* | 11.10 |
| Augm. | le ministre de l'économie bruno le maire, a déclaré que *(economy minister bruno le maire said that)* | 43.81 |
| Ref. | **les six accusés encourent trois ans d'emprisonnement.** *(the six defendants face up to three years in prison)* | |
| Baseline | ces deux ans de prison. *(these two years in prison.)* | 23.22 |
| Augm. | ils encourent une peine de trois ans d'emprisonnement. *(they incur a three-year prison sentence.)* | 68.05 |
| Ref. | **23 titres du grand chelem à son actif,** *(23 grand slam titles to his credit,)* | |
| Baseline | il a été condamné à 23 ans de prison. *(he was sentenced to 23 years in prison.)* | 4.88 |
| Augm. | et a remporté 23 titres au grand chelem, *(and won 23 grand slam titles, )* | 55.57 |
| Ref. | **pour quelle raison ?** *(for what reason?)* | |
| Baseline | en allemagne, *(in germany,)* | 0.00 |
| Augm. | quelles sont les raisons ? *(what are the reasons?)* | 76.26 |

Table 6: Examples of French Sign Language to text translation with an improvement of more than 20 points in BLEURT score between baseline and augmented dataset. *Ref.* is the reference, *Baseline* refers to Baseline model without augmentation, and *Augm.* refers to baseline + data augmentation.

pears rightly in the prediction of the augmented model. For the second sentence, "les six accusés encourent trois ans d'emprisonnement.", the corresponding sequence of signs is: "eux | risquer | 3 ans | prison" (en: *"them | risk | 3 years | prison"*) the notion of "six defendants" is implied by the space position of the group "them", meaning they were previously mentioned. The sentence proposed by the augmented model is a good translation of this out-of-context isolated signed sentence.

As stated earlier, the Mediapi-RGB dataset is complex and includes multiple news topics that evolve over time. Let us take the example of three persons: Edouard Philippe (16 occurrences in training set) / Jean Castex (36) / Boris Johnson (10). All three of them are Prime ministers, and two of them were consecutive Prime ministers of France. We observed a confusion on model predictions between all of them. In the last example in Table 7, the signed sequence is: " venir | premier ministre |

| Model | Sentence | BLEURT |
|---|---|---|
| Ref. | **à l'époque des dinosaures,** *(during the time of the dinosaurs,)* | |
| Baseline | jean-luc mélenchon *(jean-luc mélenchon)* | -1.26 |
| Augm. | | |
| Ref. | **l'isnar-img reconnait qu'il est difficile de résister à la puissance de l'industrie pharmaceutique** *(the isnar-img acknowledges that it is difficult to resist the power of the pharmaceutical industry.)* | |
| Baseline | chine : un changement de jeu vidéo.*(china: a change in video games.)* | 1.21 |
| Augm. | ukraine : le grand public ? *(ukraine: the general public?)* | 1.39 |
| Ref. | **notamment édouard philippe et plusieurs de ses ministres,** *(including édouard philippe and several of his ministers)* | |
| Baseline | le premier ministre britannique boris johnson a donc été un premier ministre. *(british prime minister boris johnson was therefore a prime minister.)* | 5.58 |
| Augm. | le premier ministre, boris johnson, *(prime minister boris johnson,)* | 4.61 |

Table 7: Examples of French Sign Language to text translation where less than 1 point difference in BLEURT score was observed between baseline and augmented dataset. *Ref.* is the reference, *Baseline* refers to Baseline model without augmentation, and *Augm.* refers to baseline + data augmentation.

le-grand | plusieurs | ministre | aussi " (en: *" come | prime | minister | several | minister | too "*). Prime minister is followed by a sign supposed to represent the name of Édouard Philippe. However, at that time (before July 2020), there was no consensus on the signed name for him, and multiple propositions coexisted in the dataset. Both our models predicted "Boris Johnson" instead of "Édouard Philippe", ignoring this sign. Following the COVID lockdown and its numerous communications interpreted into French Sign Language, we observed a progressive convergence and standardization of many signed names. Rare vocabulary tends to translate randomly as opposed to frequent themes in the news report such as covid (793 out of 27841 sentences when counting only original training data) or deaf (949).

1390 sentences got the same score ( $\pm 1$ point) before and after augmentation. From this extract, we analyzed some sentences with lowest BLEURT scores. We picked three examples displayed in Table 7. The word "dinosaure" only appears twice in training dataset."Jean-Luc Mélenchon" sign is a two-hands signs where hand configuration is similar to the one chosen for "dinosaur", however placed differently (above the head versus shoulder height). The name "isnar-img" does not have a specific sign and is spelled using dactylology, currently lacking in translation system. As a consequence, output of baseline and augmented models seems random.

## 5. Discussion and Conclusion

### 5.1. Random Initialization Versus FPT

Our results partially replicate those of De Coster et al. (2021) on German SL. However, they reveal a critical limitation on Mediapi-RGB, where FPT seems to degrade performance on larger, multi-domain datasets. Our results suggest that the FPT approach is more suitable for small, narrow-domain datasets (7K sentences, weather-only, 3K vocab), where frozen weights help prevent overfitting. On larger, multi-domain datasets however, frozen weights appear to limit the model's capacity to adapt, leading to underfitting. The tokenizer might not match the requirements: 36K words vocabulary exceed CamemBERT's 32K token capacity. FPT seems more suitable to small and narrow-domain datasets. This highlights the importance of matching architectural choices to dataset characteristics, independently of data augmentation.

### 5.2. Paraphrasing Contribution

These experiments on two sign languages show the efficiency of data augmentation on the text side to improve SL translation models. In the French dataset, paraphrasing helps understanding of the main theme of a sentence. The examples in Table 6 confirm this tendency, showing that paraphrasing improves predictions on a specific theme, sometimes with better precision, such as "economy minister" versus "health minister."

Paraphrasing methods show promise for augmenting text data, but they also present limitations. The German weather forecast dataset is well-defined and scientifically constrained, reducing the likelihood of bias. In other domains, however, LLM-generated paraphrases may introduce distortions, altering the original meaning of sentences due to societal or stereotypical influences, as shown in prior work such as Ducel et al. (2024). Next steps would include a bias analysis of the French SL dataset paraphrasing. We evaluated paraphrasing using the open-source model Qwen2.5-32b mentioned previously, obtaining a BLEURT score of

75.23 and a sacreBLEU of 31.17, both higher than those achieved with GPT-4o-mini. However, a comparative qualitative analysis between Qwen-32b and GPT-4o-mini demonstrated that Qwen-32b revealed more gaps in its knowledge of Deaf culture. For example, it confuses key terms such as “Deaf” and “hCODA” and incorrectly uses *entendeuse* instead of the correct term *entendant* for "hearing person". Precise terminology is particularly critical when translating or paraphrasing Deaf-specific content. A second consideration arises in the German SL dataset, which involves interpreted data, moving from spoken language to sign language. We kept the sign as ground truth and paraphrased the speech. We have to verify if paraphrases would be interpreted in the same way as the original signed sequences.

Doubling the dataset through paraphrasing helps improve the translation task. We questioned the relevance of variability in translation between oral language and SL, or between SL and text. As most of the literature focuses on oral-to-sign language interpretation, the observations largely apply to that context, but they are also valid for sign-to-text translation. This proves the importance of adding this variability to our translation models, as there is not one single way to translate from French Sign Language to French.

For completing this evaluation, we will include in the next steps a human-in-the-loop. We would provide human translators, interpreters and also hCODA and Deaf participants with our out-of-context signed sentences and ask them to produce precise translations. In parallel, we would ask them to validate proposed automatic paraphrases and, when possible, generate new ones that accurately reflect the signed sequence. By doing so, we should reduce noise in our data and obtain more reliable sign-to-text sequences with multiple paraphrases.

### 5.3. Conclusion

In this study, we explored how generative models can help improve sign language translation through lightweight text-side data augmentation. We showed that paraphrasing the target text improves translation performance on two datasets (German and French SL), even with limited training resources. Next steps will include a human-in-the-loop for expert paraphrasing validation and generation, and a thorough analysis of the effect of associating two text translations to a single signed production.

## 6. Ethical Considerations

None of the authors is Deaf. As previously mentioned there might be biases in LLM-generated paraphrases alongside a risk of hallucinations. The dataset used for French Sign Language is sparse and covers multiple subjects with variable occurrences.

## 7. Bibliographical References


Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. 2022. Using neural machine translation methods for sign language translation. In *Proceedings of the 60th annual meeting of the association for computational linguistics: Student research workshop*, pages 273–284.

Florine Archambeaud. 2022. *Vers un modèle des espaces en interprétation du Français vers la Langue des Signes Française*. Theses, Université de la Sorbonne nouvelle - Paris III.

David Beauchemin, Yan Tremblay, Mohamed Amine Youssef, and Richard Khoury. 2025. COLE: a comprehensive benchmark for French language understanding evaluation. *arXiv preprint arXiv:2510.05046*.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proc. of CVPR*, pages 10023–10033.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German’s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Daria Dayter. 2019. *Collocations in non-interpreted and simultaneously interpreted English: a corpus study*. Routledge.

Mathieu De Coster, Karel D’Oosterlinck, Marija Pizurica, Paloma Rabaey, et al. 2021. Frozen pretrained transformers for neural sign language translation. In *18th Biennial Machine Translation Summit (MT Summit 2021)*, pages 88–97. Association for Machine Translation in the Americas.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2024. “You’ll be a nurse, my son!” Automatically assessing gender biases in autoregressive language models in French and Italian. *Language Resources and Evaluation*, pages 1–29.

Diandra Fabre, Julie Lascar, Julie Halbout, Yanis Ouakrim, Annelies Braffort, Thomas Hueber, et al. 2025. Exploring sign-level strategies to enhance automatic translation of French Sign Language. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, pages 1–7.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. LLMs are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18362–18372.

Jiatao Gu, Hany Hassan Awadalla, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354.

Mark Halley. 2020. Rendering depiction: a case study of an American Sign Language/English interpreter. *Journal of Interpretation*, 28(2):3.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico. Association for Computational Linguistics.

Jin Yea Jang, Han-Mu Park, Saim Shin, Suna Shin, Byungcheon Yoon, and Gahgene Gweon. 2022. Automatic gloss-level data augmentation for sign language translation. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6808–6813.

Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025. Lost in translation, found in context: Sign language translation with contextual cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8742–8752.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211.

Zhibo Liu and Juhua Dou. 2023. Lexical density, lexical diversity, and lexical sophistication in simultaneously interpreted texts: a cognitive perspective. *Frontiers in psychology*, 14:1276705.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2022. Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing. *Neural Networks*, 148:194–205.

Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *European Conference on Computer Vision*, pages 671–690. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Cristian Mauricio Gallardo Paredes, Cristian Machuca, and Yadira Maricela Semblantes Claudio. 2023. ChatGPT API: Brief overview and integration in software development. *International Journal of Engineering Insights*, 1(1):25–29.

Wendy Sandler and Diane Carolyn Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Ozge Mercanoglu Sincan and Richard Bowden. 2025. Contrastive pretraining with dual visual encoders for gloss-free sign language translation. *arXiv preprint arXiv:2507.10306*.

Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. 2023. Is context all you need? Scaling neural sign language translation to large domains of discourse. In *Proc. of ICCV*, pages 1955–1965.

Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. 2024. Using an LLM to turn sign spottings into spoken language sentences. *arXiv preprint arXiv:2403.10434*.

Christopher Stone and Debra Russell. 2011. Interpreting in International Sign: Decisions of Deaf and non-deaf interpreters.

Laia Tarrés, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. 2023. Sign language translation from instructional videos. In *Proc. of CVPR*, pages 5625–5635.

Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. 2021. BERTTune: Fine-tuning neural machine translation with BERTScore. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 915–924.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2024. Youtube-ASL: a large-scale, open-domain American Sign Language-English parallel corpus. *Advances in Neural Information Processing Systems*, 36.

Ashish Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations*.

## 8. Language Resource References

Hannah Bull, Yanis Ouakrim, Julie Lascar, Annelies Braffort, and Michèle Gouiffès. 2024. Mediapi-RGB corpus.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, et al. 2021. How2sign: A large-scale multimodal dataset for continuous American Sign Language.

Jens Forster, Christoph Andreas Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H. Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A large vocabulary Sign Language recognition and translation corpus.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2024. YouTube-ASL: A large-scale, open-domain American Sign Language-English parallel corpus. volume 36.

# The Construction of the CORALSE Corpus, Now and Beyond: A Tool for Documenting Spanish Sign Language

**Ana Fernández Soneira, María C. Bao-Fente, Rayco H. González-Montesino, Inmaculada C. Báez Montero**

Universidade de Vigo, Universidade de A Coruña, Universidad Rey Juan Carlos, Universidad de Vigo
anafe@uvigo.gal, maria.bao@udc.es, raycoh.gonzalez@urjc.es, cbaez@uvigo.gal

**Abstract**

The main objective of this paper is to present the experience of building the CORALSE corpus and to discuss the challenges that arise when attempting to provide a comprehensive description of a sign language.

To this end, we address the following questions, drawing on the data obtained in the completed phases of the CORALSE project as well as on the foundational principles guiding the project's third phase.

THE CORALSE CORPUS TODAY: TODAY How have we developed a linguistic corpus of sign language?, What steps have we taken in developing the CORALSE corpus?, Which informants have we recorded and what criteria have guided their selection?

THE CORALSE CORPUS IN THE FUTURE: Which (native) languages do we prioritise when selecting informants?, How do the perspectives of reference signers, interpreters, educators, and psycholinguists contribute to a more complete understanding of a sign language?

Corpus linguistics is understood as a set of methodologies designed to study language through collections of digitised texts. Its development over recent decades—initially driven by advances in computing and, subsequently, by the emergence of the internet—represents one of the most significant transformations in contemporary linguistic research.

The projects CORALSE: Annotated Inter-university Corpus of Spanish Sign Language and Textual Typology, Registers and Styles in Spanish Sign Language: New Data for the Expansion of the CORALSE Corpus adopt a corpus linguistics approach to collect, analyse and describe a representative sample of Spanish Sign Language (LSE). We also reflect on the types of linguistic data that are truly necessary to document the actual use of Spanish Sign Language.

**Keywords**: CORALSE Corpus, Spanish Sign Language (LSE), linguistic description, translation, acquisition.

## 1. The CORALSE Corpus Today

The need to produce a comprehensive linguistic description of Spanish Sign Language led us to ask ourselves how we could obtain primary linguistic and cultural materials that were representative of the linguistic behaviour of the deaf community, including different registers and varieties, as well as their metalinguistic knowledge. The search for an answer to this question led to the creation of the CORALSE corpus. This project began in 2013, with the initial objective of documenting the current state of LSE through a broad and representative sample of different types of signed discourse and different areas.

In the first phase of the project (2013-2016)[1], we designed a visogestural corpus, the first reference corpus of Spanish Sign Language, CORALSE (http://www.coralse.org/). The main achievement is the work on linguistic documentation, reflected in the collection of real speech samples from deaf[2] informants from different parts of Spain, which has allowed us to begin documenting the language from a synchronic perspective, collecting linguistic variation and exploring the sociolinguistic situation of Spanish signers based on real communication situations.

We pay particular attention to the influence of corpus linguistics on our approach to the study of sign languages in terms of their standardisation, translation and acquisition. As a research methodology, corpus linguistics is compatible with any theoretical model; however, the use of textual samples produced in natural contexts makes it particularly well-suited to use-based functional and cognitive models

In the field of linguistic documentation, the CORALSE corpus is an open visual archive that aims to play a key role as a contextual source by enabling the reconstruction of sociolinguistic and cultural practices that accompany and enrich the linguistic data itself. We intend it to be a linguistic tool composed of structured and citable data and metadata that complies with the FAIR principles (Findable, Accessible, Interoperable, Reusable), i.e. accessible without technical restrictions,

[1] All phases of the project have been funded by the Spanish government.

[2] In this article, we use the term 'deaf' to refer to individuals who identify as members of a sign language linguistic minority and participate in the cultural practices of sign language communities, in line with established sociolinguistic perspectives in Deaf Studies.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 140–147
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

interoperable with other heritage repositories and suitable for scientific and educational reuse.

To achieve this, we first developed seven signed tasks (see Table 1), with linguistic, sociolinguistic and cultural content, recorded a pilot test, modified some items[3] and finalised the task list design before beginning the recordings (Báez Montero et al., 2016):

| CORALSE CORPUS | | Stimulus | Objective/ Function | Time |
|---|---|---|---|---|
| Presentation | | - | 1st contact | 5 min. |
| Questionnaire | | Q. in LSE | socio-ling information | 15 min. |
| Description-narration | Historical fact | 2 images | Narration in the past tense | 5 min. |
| | Map | 2 images | Sequencing | 5 min. |
| | Illustrated history | 2 images | Description and Pragmatic information | 5 min. |
| | Tom and Jerry | 2 videos | Synthesis | 5 min. |
| Free conversation | | -- | Spontaneous discourse | 15 min. |
| Naming | | 130 images | Lexicon, Sociolectal variation | 10 min. |
| Questionnaire | | Q. in LSE | Diachronic variation | 10 min. |

Table 1. CORALSE corpus task list

Questionnaires are more structured and controlled activities, designed to ascertain informants' views on issues relating to the social and linguistic situation of the deaf community and to linguistic variation (in the latter case). They include clear instructions and a defined format, explained in a signed video, to ensure that all informants receive the same information. To obtain more natural data without losing focus on certain linguistic phenomena, we also develop various elicitation tasks (picture description, naming, etc.), designed to elicit the use of certain linguistic forms (classifiers, time markers, etc.).

For the selection of participants, in keeping with the tradition of sociolinguistic studies and other sign language corpora (Schembri, 2013), we were interested in a series of demographic variables, including gender, age, place of residence, socioeconomic status, the school they attended, and the age at which they acquired sign language. Participants were selected according to some of these variables (see Table 2); we established three age groups as priorities (age will determine whether they attended a school for the deaf or an integrated school), gender, place of residence, and whether they were deaf people who used LSE.

| Demographic characteristics | | N=81 | % |
|---|---|---|---|
| | | | |
| Sex | Men | 38 | 46.9 |
| | Women | 43 | 53.1 |
| | | | |
| Age range | 18<35 | 29 | 35.8 |
| | 35<65 | 37 | 45.7 |
| | >65 | 15 | 18.5 |
| | | | |
| Region | Galicia | 16 | 19.8 |
| | Basque Country | 8 | 9.9 |
| | Madrid | 8 | 9.9 |
| | Andalusia | 20 | 24.7 |
| | Canary Islands | 6 | 7.4 |
| | Valencia | 14 | 17.3 |
| | Extremadura | 9 | 11.1 |
| | | | |
| LSE acquisition | Native | 15 | 18.5 |
| | Early <6 | 32 | 39.5 |
| | Late prelocutive 6<18 | 30 | 37 |
| | Late postlocutive 6<18 | 4 | 4.9 |
| | | | |
| Type of school | For the Deaf | 32 | 39.5 |
| | Mainstream | 27 | 33.3 |
| | Both | 21 | 25.9 |
| | None | 1 | 1.2 |
| | | | |
| Higher Education University or Professional training | Yes | 38 | 46.9 |
| | No | 43 | 53.1 |
| | | | |

Table 2. Sociodemographic data of the informants in the CORALSE corpus

Once the above decisions had been made, we began the recordings. The participants were always filmed in pairs belonging, whenever possible, to the same age group and of opposite sexes. We gave the informants we contacted the opportunity to choose their own partner for the recording. The participants were filmed with one

[3] For example, we had to clarify the instructions in the spatial descriptions in the map task, as the suggested routes were rather vague and caused confusion among the informants.

camera focused on each of them and another central camera that recorded both of their sign language production. We used a blue or green background chroma key to ensure a clear view of hand movements, and three LED spotlights (see Figure 2).

During the recordings, the informants were supported by a deaf moderator to answer any questions that might arise. In addition to the moderator, all task are recorded and projected to guide the recording and ensure that all informants receive the same input.




Figure 1. Map of filming locations

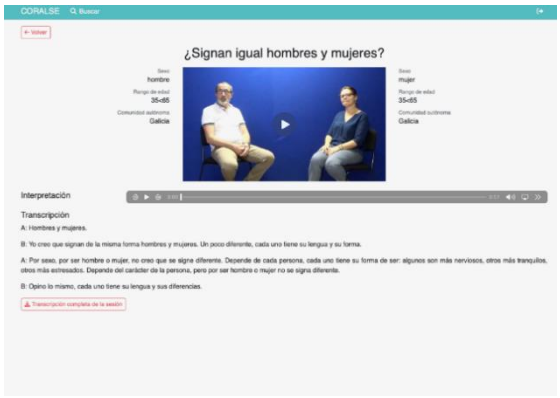



Figure 2. Example of a coded recording

In the second phase of this project (2018-2021), in addition to enriching the corpus with samples of diatopic and diachronic variants, we began the description and analysis of this sign language, with the aim of providing a more up-to-date picture of the linguistic reality of LSE (see Table 3).

| Measure | Galicia | Basque Country | Community of Madrid | Andalusia | Canary Islands | Valencian Community | Extremadura | Total |
|---|---|---|---|---|---|---|---|---|
| Sessions | 8 | 4 | 4 | 10 | 3 | 7 | 5 | 41 |
| Recording Hours-minutes | 8.1 | 4.1 | 3.1 | 14.29 | 3.44 | 8.41 | 7.37 | 49.44 |
| Tasks | 72 | 36 | 36 | 86 | 27 | 63 | 45 | 365 |
| Participants | 16 | 8 | 8 | 20 | 6 | 14 | 9 | 81 |
| Videos | 400 | 98 | 200 | 450 | 150 | 175 | 140 | 1613 |
| Interpretation | 175 | 98 | 100 | 282 | 76 | 87 | 78 | 896 |
| Transcription | 148 | 75 | 75 | 180 | 131 | 134 | 93 | 836 |

Table 3. Corpus production data.

The analysis of the variants we obtained allowed us to begin describing LSE based on the sociolinguistic data collected. We also began preparing the corpus so that it could be annotated and consulted collaboratively, thereby bringing LSE closer to new researchers who, using this tool, could create new research projects with FAIR data (Báez et al., 2020).

On the one hand, we carried out oral interpretation into Spanish, transcription, coding and analysis of the two questionnaires designed to analyse regional and generational variation in LSE (tasks 2 and 6 in table 2). The data obtained has provided us with an initial approximation of our informants' beliefs and attitudes towards two fundamental issues in the study of this language: generational variation and contact with Spanish (Báez and Bao-Fente, 2023), as well as its influence in the school/educational context (Fernández Soneira and Bao-Fente, 2021). We have also gathered their opinions on variation in Spanish sign languages (Báez and Bao-Fente, 2024). As the focus of these studies was more socially oriented, the responses were not transcribed into ELAN for these studies; instead, they were interpreted into spoken Spanish by the interpreters involved in the CORALSE project and were also translated into written Spanish. Furthermore, we have begun the grammatical description by analysing the Naming test, a lexical task with which we aim to establish the lexical variants and the most common signs in different semantic fields of LSE. The initial work has focused on the analysis of colours and numbers (Báez et al., 2020). The data analysed has also allowed us to begin comparative work with other sign languages (Müller de Quadros et al., 2024).

We are currently focused on tagging the corpus; we intend to annotate the recorded samples using ELAN (see Figure 3), translate them from sign language into Spanish, and catalogue the samples so that they can be made available on the consultation website that is currently under development.

The samples will be annotated employing the following lines or tiers in a hierarchical structure to facilitate data readability once exported: for each informant, one line for the translation, two for the glossing (dominant hand and non-dominant hand), and one general line for comments.

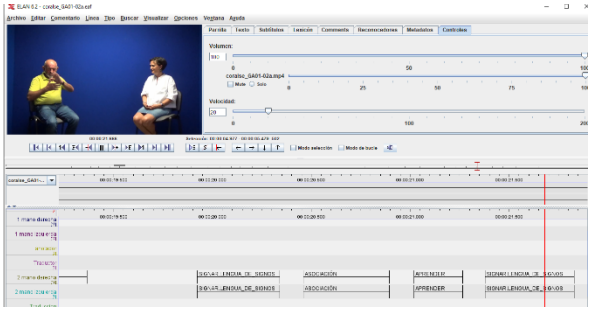


Figure 3. Test annotated with ELAN

## 2. The Corpus CORALSE in the Future

To complete the description we set out to create in the early stages of the CORALSE project and to address the questions and requirements involved in continuing to expand the corpus, we are tackling new projects that will provide us with new perspectives from informants (interpreters, educators, deaf reference persons), registers (more formal), and communicative contexts (academic, scientific, etc.).

In 2023, we began the third phase of the project, in which we aim to expand the number of potential users of the corpus so that it is useful not only for linguists but also for translators, interpreters, and specialists in specialised languages and in the acquisition and development of spoken and sign languages.

To this end, we are developing the project *Textual typology, registers and styles in LSE: new data for the expansion of the CORALSE corpus (TR3/CORALSE)*, whose three sub-projects share the objective of completing the description of Spanish Sign Language begun with the annotated CORALSE corpus and organising the data to make it more accessible to users. We will collect and insert into CORALSE the new language samples from the three subcorpora alongside the samples already collected of colloquial language in interaction: 1.- Subcorpus of register and style of reference signers of LSE in the CORALSE corpus; 2.- Subcorpus of children's language and bilingual acquisition in spoken and sign languages AdCORALSE; 3.- Parallel multilingual subcorpus of translations and interpretations of specialised languages to/from LSE. In the following lines, we will briefly describe each subproject.

### 2.1 Reference Signers

This subcorpus aims to incorporate examples of sign language from the educated register into the CORALSE corpus, characterised using an extensive, complex and precise vocabulary, which forms the basis of our corpus of reference signers. The figure of the reference deaf signer, although little studied and discussed, plays a fundamental role both in the deaf community and in the standardisation and dissemination of sign language. These linguistic models are recognised for their high linguistic and cultural competence and act as key references in the transmission of language and the construction of the collective identity of deaf people. However, there are still few studies that delve into their specific characteristics or the criteria necessary for their identification.

For signer selection, we will not use the exact same variables as in the CORALSE corpus. Instead, our criteria will focus on sociolinguistic and language-related factors. These include family background, age, stage of LSE acquisition, type of basic schooling, higher education completion, specific LSE training, and inference ability. Here, educational background is considered not as a direct indicator of linguistic competence or proficiency, but as one element of the signer's broader sociolinguistic profile (see table 4).

So far, we have analyzed the Galician group in the CORALSE corpus (CORALSE_GA), which includes 16 participants. These data helped us identify reference signers for initial recordings in Galicia. Our broader goal is to collect data from 20 to 30 signers across Spain. The corpus shows that informants distinguish between two language models: native signers (deaf children of deaf parents) and teachers of Spanish Sign Language. Native signers acquire the language at home from birth, while teachers often learn Spanish Sign Language in adulthood and may not be native users. This distinction is based on participant reports, not a predetermined hierarchy.

These evaluations show that, in the corpus, linguistic competence is seen through acquisition history and communicative expertise. This aligns with the communicative goals of the CEFR, which defines proficiency as the ability to use language in real-life situations. However, the CEFR and its Companion Volume (2020) serve only as broad reference frameworks, not for direct LSE application without adaptation (see Table 4). LSE signer profiles are complex and dynamic, reflecting factors present in contacts between majority and minoritised languages, regardless of modality.

| Value | Sociodemographic characteristics of the reference signer |
|---|---|
| 4 (Excelent) | Early acquisition in childhood or school + training in LS + higher education |
| 3 (Good) | Early acquisition in childhood or school + training in LS |
| 2 (Acceptable) | Early acquisition in childhood or school or training in LS (+0.5 higher education) |
| 1 (Not acceptable) | Late acquisition of LS, no LS training (+0.5 higher education) |

Table 4. Analysis of sociodemographic characteristics of the reference signers.

No single variable identifies advanced proficiency. Instead, the chosen criteria combine acquisition-related variables and pragmatic skills, like inference, relevant to advanced communicative use. Higher education, included here only alongside LSE training, is not an independent factor.

We have developed an initial set of questions for the recordings of the reference signers ("What does Spanish Sign Language mean to you?"; "Have you ever experienced any form of discrimination because you are deaf?"), which meet the following criteria: questions with

inferential content, questions that require the use of formal language and specialised terminology. The skills of what is considered an advanced user according to the CEFR will be taken into account, i.e. the skills of levels C1 and C2 (Council of Europe, 2020).

### 2.2 Translators and Interpreters

CoMPARTiR is conceived as a multilingual and multimodal parallel corpus of translations and interpretations between Spanish Sign Language (LSE) and various spoken languages (Spanish, Galician, Catalan, English, French, etc.), with special attention to specialised languages and professional registers. Its objective is twofold: to systematically document real interlinguistic mediation products produced by deaf and hearing professionals; and to provide empirical evidence for terminological and discursive standardisation processes in LSE, integrating transparent criteria of representativeness, annotation and FAIR access within the TR3/CORALSE framework.

The relevance of this approach is based on three reasons. First, the methodology based on parallel corpora allows for the rigorous description of communication mediated by sign language translators and interpreters, the identification of patterns of equivalence and linguistic variation, and the support of the training and work of these professionals (Treviño et al., 2020; Albanie et al., 2021; Duarte et al., 2021; Uthus et al., 2023, Jiang et al., 2024).

Secondly, interpreters and translators (both deaf and hearing) act as agents of normalisation (Nogueira et al., 2012) and standardisation: their decisions regarding terminology, translation and bilingual/bimodal communication management shape acceptable usage in academic, healthcare, legal and media contexts. Capturing these decisions in aligned corpora allows such conventions of use to be evidenced and transferred to lexicographical and didactic materials, reinforcing standardisation through practice.

Thirdly, in the Spanish ecosystem, the need for reliable data is evident: the deaf community has expressed dissatisfaction with the quality and availability of interpreting services, and the lack of training and research on LSE translation and interpreting is a reality. CoMPARTiR responds with a curated and ethical infrastructure, based on informed consent and the use of open-licensed audiovisual materials (which guarantees their responsible reuse in accordance with current regulations), institutional agreements and an annotation template harmonised with TR3/CORALSE.

Methodologically, CoMPARTiR is structured around three pillars: the collection of real samples of translation and interpreting to/from LSE in defined fields; the construction of a parallel corpus through LSE↔spoken language alignment enriched with metadata; and analytical triangulation that integrates the analysis of the translation product, the interlinguistic mediation process and its educational implications. To date, the project has achieved the following objectives:

- Conducting a bibliographic and documentary review of existing corpora in the field of specialised translation/interpreting and sign languages, confirming the scarcity of specific references in this field.
- Develop a preliminary typology of texts that could be included in the corpus, grouped by specialised fields (education, health, law, science, technology, economics, religion, etc.), and define the priority language combinations (LSE ↔ Spanish, LSE ↔ Galician, LSE ↔ Catalan, LSE ↔ Basque, LSE ↔ English, LSE ↔ international signs, among others), taking into account the availability of samples and research relevance.
- Establish an initial classification of criteria that allows for systematic and consistent classification of samples. This classification includes issues such as:
  - the type of procedure (translation or interpreting);
  - the typology, either according to context (for public services, television, business, conferences, audiovisual, etc.) or according to the type of interaction (dialogical or monological); and
  - the modalities, such as linguistic (spoken/signed), mode of work (simultaneous/consecutive/sight), directionality (unidirectional/bidirectional), use of technology (remote/automatic), professional status (professional/natural), etc.
- We have received a favourable opinion from the Research Ethics Committee of Rey Juan Carlos University, after analysing the complexity involved in collecting and using audiovisual samples that include the image (and in many cases the voice) of interpreters and deaf or hearing people, which raises significant ethical implications.
- Begin collecting samples: approximately 20 hours of recordings have been collected in the fields of education, healthcare and conferences (see Table 5), all of which are licensed under Creative Commons or with the informed consent of the participants. Specifically, the following table shows the number of samples collected and the total time for each field of specialisation:

| | N.º of vídeos | Recording time |
|---|---|---|
| Educational field | 80 | 6:53:41 |
| Healthcare | 71 | 2:13:59 |
| Academic (conferences) | 25 | 13:41:21 |
| TOTAL | 176 | 22:49:01 |

Table 5. Samples collected from CoMPARTiR

The next steps in this project include: increasing the number of samples in different areas of specialisation; training the members of the working team in the use of annotation and corpus management tools; generating a template for annotation and transcription from LSE to glosses; establishing a common protocol to ensure consistency in annotation; and initiating preliminary analysis of the annotated data, according to the interests and needs of the research team.

### 2.3. Population under 18 Years of Age

The purpose of AdCORALSE is to design a corpus that allows for the study of language development in bilingual and intermodal acquisition contexts, in which sign language and spoken language are in contact, regardless of which is the mother tongue, first or dominant language (L1) and second language (L2). The aim is therefore to document: a) the reality of bilingualism between the two modalities, b) the situation in which LSE education is taking place, and c) language learning among the signing population. Consequently, the purpose of the corpus will not be limited to recording linguistic uses, but will also include conditions of exposure, acquisition trajectories and educational practices that directly influence the communicative development of the signing population.

In the case of LSE, the documentation of these factors is particularly relevant because the data obtained in CORALSE confirm that the majority of adult signers who use this language regularly (N=81) did not acquire it early on, as their mother tongue or first language (n=15), but rather through schooling and at a later age, after the age of 3-4 (n=66).

In order to document the acquisition or learning of a language in the process of development (L1/L2), the evolutionary and dynamic nature of the object of study must be taken into account (Fernández-Pérez, 2020). The variability of children's language implies the collection of data in development; therefore, the aim is to design a tool that allows for the collection of significant samples in this context of acquisition, both transversally and longitudinally. The aim is to build a corpus that records language development in the sign language population, specifically the acquisition of communicative competences in LSE and Spanish, either in the same subjects through different tests or tasks according to their stage of development, or by comparing subjects at a given moment and developmental milestone.

The design of AdCORALSE requires a flexible structure that allows samples to be collected from different stages of development and incorporates linguistic properties that are still emerging in both L1 and L2, but which are particularly relevant in sign languages given the limitations of existing documentation and low intergenerational transmission. The scarcity of linguistic descriptions of sign languages also means that an acquisition corpus must comply with international standards, as proposed in the design of CORALSE (Báez et al., 2020).

We are developing the protocol for the child language corpus with a focus on protecting children's rights and privacy. The ethical standards for research with minors guide all procedures. We will collect two types of data. First, we will gather language samples through child-friendly tasks tailored for young participants. Second, we will use test-based data to document linguistic abilities at the phonological, lexical, grammatical, pragmatic, and discourse levels. Data sharing will comply with the FAIR principles and ethical standards for research with minors. Only anonymised, quantifiable data from assessment instruments will be shared with external researchers through controlled access, and requests will be reviewed for ethical compliance. Raw audio, video recordings, and personally identifiable information will not be shared.

Given the scarcity of tools for Spanish Sign Language, the corpus also aims to advance the development of resources for assessing sign language in children and adolescents. The aim is also to ensure its subsequent usefulness not only for other linguistic studies and analyses, but also for future applications, for example, in the clinical or educational fields.

Table 6 shows the main design criteria for the CORPUS.

| Design criteria | |
|---|---|
| Significance of samples | Native deaf subjects in LSE (L1), non-native subjects (L2) and CODA subjects for each age group, according to the time of acquisition and developmental stage<br>Balance of samples according to gender, family status and linguistic varieties |
| Functional balance between modalities | Tasks for each level of linguistic development, in LSE and Spanish:<br>Skills for language and communication development<br>Phonological and lexical-semantic development<br>Grammatical (morphosyntactic) development<br>Pragmatic and discursive development<br>Tasks for each developmental stage, in LSE and Spanish<br>Functional parallelism between LSE and LO/written tasks: they record the same skills, processes or constructs in each language. |

| | |
|---|---|
| Equivalence of data in both languages | Records per subject, timed according to developmental period; monthly, quarterly, half-yearly or yearly (0-3, 3-6, 6-9, 9-12, 12-15 and 15-18 years)<br>For each subject and time record:<br>Samples of spontaneous and (semi)structured language<br>Samples elicited through standardised (LO) and non-standardised tests (LS)[4].<br>Audiovisual records, with linguistic and technical quality |
| Contribution from/to the community | Developmental records provided periodically by families, teachers and speech therapists<br>Records provided by informants under the age of 18, with prior authorisation and informed consent from both the family and the minor. |

Table 6. Design criteria for the AdCORALSE corpus

The AdCORALSE corpus is thus conceived as an open, dynamic, longitudinal and multimodal repertoire; a systematic set of data suitable for documenting the emergence and development of language in deaf populations exposed, simultaneously or sequentially, to a sign language and an oral or written language. We intend for this corpus to specifically promote research on the acquisition of LSE, which has been less studied than other sign languages (Sánchez-Amat and Quer, 2018), to provide data and evidence that will help guide methodological decisions, teaching resources and educational practices with deaf children and adolescents.

## 3. Conclusion

The purpose of the CORALSE corpus is to provide the linguistic and sociolinguistic data necessary to study and analyse the linguistic situation of the deaf community and thus draw conclusions about sign language. Through this linguistic corpus, which is accessible, technological, representative and portable, i.e. it will contribute to the survival of the language, we will be able to offer future users all the necessary information on linguistic, sociolinguistic, cultural and linguistic variation issues. In addition, the corpus will have an impact on the scientific community, as it will make data available for comparative and typological studies.

The data from the CORALSE corpus will enable two key outcomes: drawing technical conclusions, thanks to its design for analysis with future artificial intelligence tools and participating in world-renowned projects such as The Language Archive or CHILDES, using tools like Catcher, Sketch Engine, or sign language macro-corpora.

The availability of the CORALSE corpus, not only in its public section but also in its specific (private) section, will also make it possible to offer research and application material to researchers, interpreters, teachers, etc.

In short, the corpus will continue to be enriched over time, not only with new data but also with analyses by different researchers, publications derived from research, and the application of analysis to language teaching, psycholinguistics, computational linguistics, translation, and interpreting.

From an open heritage perspective, the availability of CORALSE not only guarantees the preservation of the linguistic heritage of the Spanish deaf community but will also facilitate its integration into other digital corpora, historical ethnography projects and interdisciplinary analyses, reinforcing the concept of heritage as a common infrastructure for research, teaching and society.

In conclusion, this corpus is not conceived merely as a project for documenting and compiling linguistic data on Spanish Sign Language (LSE), but as a key linguistic tool—that is, a comprehensive scientific framework designed to generate linguistic resources. The project is grounded in an empirical linguistic perspective, based on the systematic analysis of real-world usage data, ensuring the validity and representativeness of the data, as well as the replicability of analytical procedures through the explicit specification of protocols, annotation criteria and data processing methodologies.

## 4. Acknowledgments

This project is funded by the Spanish Ministry of Science, Innovation and Universities (Subproject references: RefCORALSE, PID2022-139084NB-C31; AdCORALSE, PID2022-139084NA-C33; CoMPARTIR, PID2022-139084NA-C32).

## 5. Bibliographical References

Albanie, S., Varol, G., Momeni, L., Bull, H., Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R. Mc-Parland, A. and Zisserman, A. (2021). *BOBSL: BBC-Oxford British Sign Language Dataset*. https://doi.org/10.48550/arXiv.2111.03635

Báez Montero, I. C., Fernández Soneira, A. and Freijeiro Ocampo, E. (2016). "CORALSE: diseño de un corpus de lengua de signos española". In A. Moreno Ortiz and C. Pérez-Hernández (eds.), CILC2016. *EPiC Series in Language and Linguistics*, vol. 1, pp. 111-120.

[4] The Spanish abbreviation LO refers to spoken language and LS to sign language.

Báez Montero, I. C. and Bao-Fente, M. C. (2024). Identity and Sign Language Varieties in Spain: Attitudes and Beliefs. In Mara Barbosa and Talia Bugel (Eds.), *Language Attitudes and the Pursuit of Social Justice: Identity, Prejudice, and Education*. Taylor and Francis, pp. 208-227.

Báez Montero, I. C. and Bao-Fente, M.C. (2023). Actitudes e ideologías lingüísticas en la Lengua de Signos Española: creencias de las personas sordas ante la variación en su lengua. *Revista de Estudos da Linguagem*, 31:2 (thematic issue), 947-980.

Consejo de Europa (2020). Consejo de Europa (2020), *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación. Volumen complementario*. Servicio de publicaciones del Consejo de Europa: Estrasburgo. www.coe.int/lang-cefr.

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J. and Giro-i-Nieto, X. (2021). How2sign: a large-scale multimodal dataset for continuous American Sign Language. In L. O'Conner (Ed.), *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2735-2744). The Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.48550/arXiv.2008.08143

Fernández-Pérez, M. (2020). Corpus lingüísticos de habla infantil y representatividad: el valor de los datos en repertorios de habla en desarrollo. *Revista De Filología Hispánica*, 36(2), 651-73. https://doi.org/10.15581/008.36.2.651-73

Fernández-Soneira, A. and Bao-Fente, M. C. (2021). ¿Qué supone ser sordo a nivel escolar? Reflexiones sobre educación inclusiva y bilingüe a partir del corpus CORALSE. *Revista de estilos de aprendizaje*, 14, 46-61. doi: 10.55777/rea.v14i27.2819

Jiang, Z., Göhring, A., Moryossef, A., Sennrich, R. and Ebling, S. (2024). SwissSLi: the Multi-parallel Sign Language Corpus for Switzerland. In N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti and N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 15448–15456). ELRA y ICCL. https://aclanthology.org/2024.lrec-main.1342/

Müller de Quadros, R., Báez Montero, I. C., Fernández Soneira, A. and da Silva Reis, L. (2024). Verbos copulativos, existenciais e de posse na LIBRAS e na LSE: Un estudo morfossintático-semántico. *INTERLETRAS*, v. 11 (39), pp. 1-17.

Nogueira, R., Villameriel, S., Costello, B., Barberá, G. and Mosella, M. (2012). Efectos de la presencia de intérpretes en el aula para la normalización de la lengua de signos española. In CNSE & UNED (Eds.) *Estudios sobre la lengua de signos española* (pp. 401-415). Editorial UNED.

Sánchez-Amat, J. and Quer, J. (2018). El desarrollo de las lenguas de signos. In Melina Aparici and Alfonso Igualada (Eds.), *El desarrollo del lenguaje y la comunicación en la infancia* (pp. 225-241). Editorial UOC.

Schembri, A., Fenlon, J. and Cormier, K. (2013). Building the British Sign Language Corpus. *Language Documentation & Conservation* (7), 136-154. http://hdl.handle.net/10125/4592

Treviño, R., Hochgesang, J. A., Shaw, E. P. and Willow, N. (2020). One Side of the Coin: Development of an ASL-English Parallel Corpus by Leveraging SRT Files. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, and J. Mesch (Eds.), *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives* (pp. 224-30). European Language Resources Association (ELRA). https://aclanthology.org/2020.signlang-1.36.pdf

Uthus, D., Tanzer, G. and Georg, M. (2023). *YouTube-ASL: A large-scale, open-domain American Sign Language-English parallel corpus*. https://doi.org/10.48550/arXiv.2306.15162

# The Community and Ethics Shaping the Norwegian Sign Language Corpus

**Lindsay Ferrara**

NTNU, Norwegian University of Science and Technology
Edvard Bulls veg 1, 7491 Trondheim, Norway
lindsay.n.ferrara@ntnu.no

**Abstract**

Recently, the Norwegian Sign Language Corpus has been published, and it includes language data from over 100 signers from around Norway. Collecting and building such multimodal signed language corpora have important implications for both research and deaf communities. However, consideration is needed to protect the personal nature of signed language data, while also making a long-term resource that is as accessible as possible to various community, research, and professional stakeholders. In addition, the potential exploitation of corpus resources by commercial and other interests, which are not necessarily aligned with the deaf community itself, must also be deliberated. Here, these seemingly opposing issues and the ethics that surround them are discussed. Current best practices in Open Science (including FAIR and CARE data principles), along with ethical discussions raised by scholars working, for example, in Deaf Studies, are shown to be important in navigating this complex research data landscape.

**Keywords:** Norwegian Sign Language Corpus, Open Science, data sharing, ethics

## 1. Introduction to the Norwegian Sign Language Corpus

In 2015, the long-term work of building a linguistic corpus for Norwegian Sign Language began. As is well known, the primary language data of signed language corpora entail video-recordings of the language. These data are also accompanied by (more or less) extensive background information on the signers, annotation files of the recordings, documentation of the corpus, and other supplementary materials. Since many of these data contain personal data and are not anonymized, creating signed language corpora and subsequently publishing them require researchers to engage with a range of legal and ethical concerns. In this paper, an exploration of the specific issues relevant to the creation and publication of the Norwegian Sign Language Corpus is presented. First, a brief overview of the corpus is given. Then the main legal framework along with best practices in Open Science is described. Additional ethical issues relevant in today's research and technology landscape are also raised, before concluding with how the corpus has been published and/or made available to the research community, the Norwegian deaf community, and other stakeholders.

Ten years after the initial work in 2015, the first iteration of the Norwegian Sign Language Corpus has been completed and published in CLARINO, the Norwegian branch of CLARIN.[1] It is comprised of four main datasets, which were collected at different times and under different circumstances. Even so, all the datasets were collected with the intention of creating a long-term resource that could contribute to the documentation of Norwegian Sign Language. These datasets are:

- Data collected as part of Rolf P. Halvorsen's doctoral research (Ferrara and Halvorsen, 2021): collected in 2007, recordings of four signers (two younger and older men and two younger and older women)[2], re-telling of "Frog, Where Are You?" (Mayer, 1969) and personal experience narratives of the 9/11 attacks,
- Pilot project (Ferrara and Bø, 2022): collected in 2015, recordings of seven elderly signers from Trondheim, Oslo, and Bergen (six women and three men, aged between 61-74) having conversations (dyad, triad, and multiparty),
- Depicting Perspective in Norwegian Sign Language (Ferrara and Ringsø, 2021): collected in 2017-2018, recordings of 21 deaf signers (13 women and eight men, aged between 22-57) in conversation (dyad and triad), and
- A Norwegian Sign Language Ecology (Ferrara, 2024a; Ferrara, 2024b): collected in 2019-2025, recordings of 87 signers of all ages (46 women and 40 men, aged 23-91) in elicited and (semi-)spontaneous interactions (dyad and triad).

[1] CLARIN is a European digital infrastructure that supports research based on language resources, https://repo.clarino.uib.no/xmlui/

[2] Exact ages of participants were not archived.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 148–154
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Over 100 signers[3] participated in these projects, and they vary by gender, age, education, where they live, and socio-economic background. Some have deaf parents; some have hearing parents. However, they all report being active members of the deaf community and using Norwegian Sign Language in their everyday lives. These signers were recorded engaging in a variety of language-based tasks, for example: informal conversation; personal experience narratives; re-tellings of picture books; interviews focusing on growing up deaf in Norway; as well as discussions on a range of topics relevant to the deaf community.

In total, the Norwegian Sign Language Corpus contains approximately 60 hours of video-recordings, edited into approximately 650 video files. These video files are being annotated in ELAN (Crasborn and Sloetjes, 2008), across approximately 410 ELAN files. The corpus also contains extensive background information on the signers, consent forms, the Norwegian Signbank (Ferrara, 2020), stimuli materials, and other supplementary files that organize and document the corpus (see Ferrara, 2025 for access to some of these non-video materials).

## 2. The General Data Protection Regulation, Open Science, and research ethics in a digitized world

In the early stages of all signed language corpus projects, a decision must be made on how open and accessible the corpus will be, and what its future uses can entail. As is well known, the primary data of a signed language corpora are video recordings of people using the language—as such, they contain personal data which are legally protected in many places. Relevant to the Norwegian context is the Norwegian Personal Data Act, which, in part, enacts the European Union's General Data Protection Regulation (GDPR). These legal protections work to give individuals more control and agency over their own data and how it is used in today's digitized world.

The enactment of GDPR in Norway in 2018 led to uncertainty in academia regarding (personal) research data. Plans to collect data for the Norwegian Sign Language Corpus was met with high levels of scrutiny by Sikt, The Norwegian Agency for Shared Services in Education and Research, who approve the collection of personal data in research contexts. While this scrutiny was warranted and welcomed, original demands that the data not be stored long-term and restricted to only the purview of the original project went against the goals of collecting a linguistic corpus in the first place.

These initially proposed restrictions were a large contrast to the aims of Open Science, which were also gaining more traction in Norway: "…[O]pen science is defined as an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community» (UNESCO, 2021, p. 7).

One important facet of Open Science is making the data science is based on available to others. Open research data "can be openly used, reused, retained and redistributed by anyone, subject to acknowledgement. Open research data are available in a timely and user-friendly, human- and machine-readable and actionable format, in accordance with principles of good data governance and stewardship...» (UNESCO, 2021, pp. 9-10). Best practices in creating open research data, especially for open signed language data, include the adoption of both FAIR and CARE principles. The FAIR Guiding Principles (Wilkinson et al., 2016; summarized by GO FAIR, no date) recommend that research data (or even non-data objects) be:

- Findable: data can be found by machines and humans, largely through metadata,
- Accessible: information about how the resource can be accessed is provided, with possibly additional authentication and authorization,
- Interoperable: "The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.» (GO FAIR, no date), and
- Reusable: data and metadata need good documentation so that they can be reused in other contexts.

FAIR data supports Open Science goals by facilitating the long-term care of digital research objects that are open and available to 'down-stream' human- and machine-driven activities. They also encourage transparency, reproducibility and re-usability in science. However, while FAIR data should be as open as possible, it should also be as closed as necessary

[3] While most of these signers are deaf, due to the nature of the interaction in some of the recordings, the Corpus also includes two hearing signers.

(UNESCO, 2021, p. 11) to respect, for example, the rights of research participants and their personal data. This tension between protecting personal data while creating research data that was as open as possible guided the initial consideration of how data in the Norwegian Sign Language Corpus was to be collected, processed, and further shared.

Nuance was then added to these seemingly opposing forces (GDPR and data protection on one side, and Open Science on the other) by socio-political realities and ethical considerations. First, as a foreign, hearing researcher it was important to consider how a Norwegian Sign Language Corpus (and subsequent linguistic research) could contribute and align with the goals of the Norwegian deaf community. How would such a long-term resource support and promote this minority language? How could members of the deaf community become involved and hopefully engage in linguistic research using the corpus in the future?

These questions were notably relevant because at the time a new Norwegian Language Act was being drafted and ultimately passed (Språklova 2022). The act, among other things, officially recognizes Norwegian Sign Language as the natural signed language of the deaf community in Norway and stipulates that the Norwegian state has the responsibility to promote and strengthen this language. In response to this law, the government established a committee to survey the status of Norwegian Sign Language in society and to draw up recommendations which would enact the responsibilities outlined in the Language Act (NOU 2023:20). One of the recommendations put forth is the need for more linguistic research and documentation of Norwegian Sign Language. Research work on Norwegian Sign Language is also an important part of the other six recommendations laid out in the report. The Norwegian Sign Language Corpus as a resource along with research resulting can be seen to contribute to this national responsibility of promoting and strengthening Norwegian Sign Language.

Ethical considerations also played a prominent role in decisions around the collection and sharing of the Norwegian Sign Language Corpus. For example, the SLLS Ethics Statement for Sign Language Research[4] raised specific issues relating to research with deaf communities and gives advice about how to respect signers' personal data as well as train, promote, and include deaf people in the research process.

It was also vital to consider how the Corpus might be used (and/or exploited) by technology and AI projects, e.g., machine learning and automatic signed language recognition and translation. There has been a large increase in the number of projects aiming, for example, to develop "assistive technologies" for deaf people, with generally, very little engagement or input from deaf people (De Meulder, 2021; Angelini et al., 2024). Some of these projects have hoped to leverage signed language corpora in their workflows, even with the serious limitations and biases that come with that (De Meulder, 2021; Jantunen et al., 2021; Schembri & Cormier, 2022). While the Norwegian Sign Language Corpus was not created to support developments in machine learning or AI, the high-paced developments across those fields and the ethics surrounding it cannot be downplayed. The European Union of the Deaf's Ethical Framework On AI And Sign Language is relevant here, which demands adherence to *Deaf Digital Law* "a developing legal framework aimed at regulating and safeguarding the equitable and ethical use of digital technologies by and for deaf people» (Venade de Sousa, 2025, p. 13).

## 3. CARE data principles in the context of deaf signing communities

In addition to the ethical guidelines mentioned in the previous section, I argue here that the CARE data principles can help signed language researchers consider and navigate a complex research data landscape (see also Schulder and Hanke, 2022 on how CARE principles played a role in the creation of the DGS, German Sign Language, Corpus). Originally designed to promote Indigenous Peoples' data sovereignty, the goals of the principles included, "fostering Indigenous self-determination by enhancing Indigenous use of data for Indigenous pursuits» (Carroll et al., 2020, p. 3), and honoring the FAIR data principles. The principles recognize that Indigenous Peoples historically have been subjected to "data inequities and data exploitation" (Carroll et al., 2020, p. 2), which includes the fact that most indigenous data is not held by indigenous people. By increasing control over their own data Indigenous People can reposition themselves from being "subjects of data that perpetuate unequal power distributions to self-determining users of data for development and wellbeing" (Carroll et al., 2020, p. 2). The CARE principles are summarized in Figure 1.

The spirit behind the CARE principles equally apply to deaf signing communities. There should be an aim to foster deaf peoples' self-determination by enhancing deaf people's use of data for deaf pursuits. One can equally recognize that, similar to Indigenous Peoples, deaf people have also been subjected to data inequity and exploitation. We can probably also assume that even today data collected about deaf people are

[4] https://slls.eu/slls-ethics-statement/

still largely controlled by hearing people, perpetuating unequal power distributions.

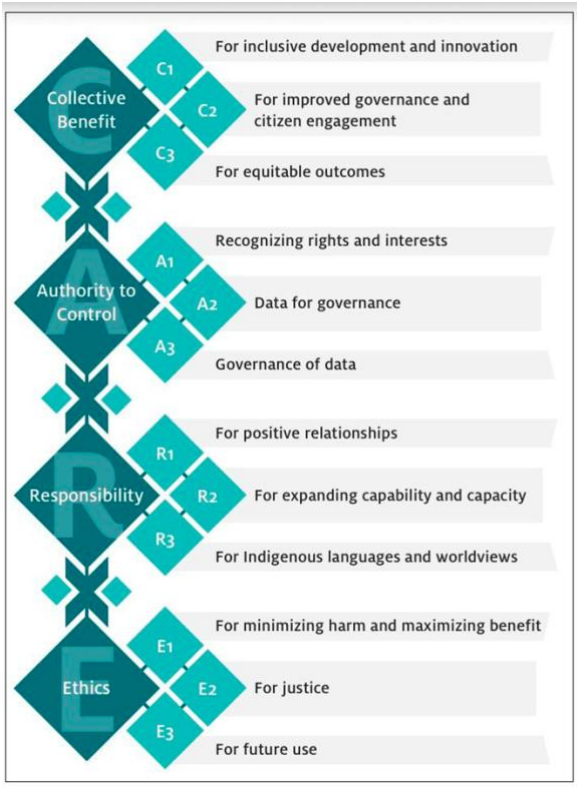



Figure 1: The CARE Principles for Indigenous Data Governance (reproduced from Carroll et al., 2020, p. 5).

The CARE principles stipulate that data should "facilitate the collective benefit" for the community, that communities have an "authority to control and govern" their data including stewardship decisions, that the people working with the data have a responsibility to the communities (e.g., respect, capacity building), and that ethical use of the data should reduce harm and maximize benefit (Carroll et al., 2020, p. 6). The ethical use of data can be further elaborated: "Indigenous Peoples' rights and wellbeing should be the focus across data ecosystems and throughout data lifecycles in order to minimize harm, maximize benefits, promote justice, and allow for future use. Paramount to ethics in data practices is representation and participation of Indigenous Peoples, who must be the ones to assess benefits, harms, and potential future uses based on community values and ethics" (Carroll et al., 2020, p. 6).

Again, it is not a big leap to see how these principles are equally applicable to the world's diverse deaf communities. Researchers working with deaf research data, such as signed language corpora, should facilitate benefit for the deaf community (see also Hochgesang and Palfreyman, 2022 about linguists working with deaf communities). And deaf communities themselves should have authority in the control of such data and how it is used. The ultimate guiding principle should be respect and supporting deaf people's rights and wellbeing.

## 4. Conclusions: accessing the Norwegian Sign Language Corpus

Before decisions were made about the collection and archival of the Norwegian Sign Language Corpus, careful deliberation over best practices in Open Science along with the relevant legal and ethical obligations, mentioned above, was carried out in collaboration with research and community partners (e.g., discussions were had with international colleagues who have experience building signed language corpora, a focus group of deaf people from the community, the Norwegian Sign Language representative at the Norwegian Language Council). Beginning with the aim that the data be 'as open as possible, and closed as necessary,' the first set of considerations revolved around legal obligations. Ensuring compliance with the EU's GDPR was requisite, so information and consent documents were drafted and revised in collaboration with the Norwegian national service Sikt. The consent documents in particular included specific formulations surrounding the use of personal data within and external to the EU. Also, consent was requested for the long-term archival and use of the data for teaching and research purposes.

In the end, it was decided that the video-recorded data be licensed under two different conditions. First, a Creative Commons license (CC BY-NC-SA 4.0) was applied to data considered non-sensitive. These data include recordings of signers re-telling picture books, a set of personal experience narratives, and public events. In this way, a portion of the corpus data could be used by the (non-academic) deaf community and relevant stakeholders. For example, it was envisioned that the recordings could be used in the future to make teaching materials.

The recordings of conversations, private meetings or events, and discussions were licensed under a more restricted academic license. Among other things this license requires a (research) plan to be submitted to the Norwegian Sign Language Corpus manager detailing how the data will be used.[5] This restricted license was chosen for several reasons. First, it works to protect the signers' and any third persons' privacy (aligning with GDRP). Second, it helps measure

[5] Currently the role of Norwegian Sign Language Corpus manager is held by the author. This role, however, will need to be passed on over time.

engagement with the corpus and keep track of how it is being leveraged by the deaf and academic communities (which may help with future development work). It is important to note that both licenses are non-commercial. More information about the two licenses can be found on the landing pages to the specific datasets in CLARINO.

As a note on timing of the corpus collection: three of the corpus datasets were collected before GDPR came into force. Although participants in these projects did sign consent for their data to be used long-term, in research and teaching, special consideration related to GDPR and the current technological landscape was missing. Therefore, it was decided to reach out to the signers from those previous projects and ask if they would be willing to consent to their data being archived in a (modern) Norwegian Sign Language Corpus, online, in line with GDPR. This process was time-consuming, but worthwhile, because all data now archived in the corpus are made available under the same conditions. The re-consent procedure also gave participants an ability to update their consent (or not) over the use of their data in the current internet age (e.g., data from Ferrara and Halvorsen, 2021, was collected in 2007, well before the age of social media and AI).

Next, also with privacy considerations in mind, it was decided to publish only some aspects of the signer's background information with the video recordings: summarized information about gender, age, where the signer lives, and when they acquired Norwegian Sign Language. These variables are often important for sociolinguistic studies. However, additional details about the signers' background (e.g., language practices in the family and school, work background, etc.) can be made available by contacting the Norwegian Sign Language Corpus manager. Decisions are made case by case.

A final decision was made to not publish the annotation files along with the video data. Although publishing annotation files and video files together are what make a signed language video collection into a corpus (Johnston, 2010), several reasons motivated this decision. First, for a very practical reason, if the annotation files were indeed to be published with the video data, it would mean the video data would not be published for many years. The project team in Norway is very small, with only one primary person working on the data. Thus, it will take many years to complete basic gloss annotation for the corpus videos. In addition, only very little translation work has been carried out, which is also considered requisite initial work (see Johnston, 2024). Out of a perceived need to publish the video data as soon as possible, so that it can be used by the community, it was decided to publish the video data without annotation files. Perhaps, in the future, annotation files can be published, and then versioned, as more work is completed. In the meantime, the annotation files can be shared by contacting the Norwegian Sign Language Corpus manager.

A second reason to not publish the annotation files was made in light of the quickly shifting technological landscape relating to big data, machine learning, and AI. Again, these technologies (in the context of deafness and signed language) often use signed language corpora to train computer models, and this process involves using annotations of the corpus data. Since, for the time being, it seems these projects often are not situated within the deaf community and are not aligned with community values or wishes, it was decided to limit access to the annotation files. As a hearing, foreign researcher, I, as the Norwegian Sign Language Corpus manager, decided it was not my place or role to facilitate such projects. However, there is nothing barring the sharing of the annotation files. Thus, in the future, if a technology project arising from within the deaf community is to be carried out, it is possible to facilitate access to the corpus' annotation files, with the caveat that use aligns with the other license restrictions.

Creating a signed language corpus in the modern age comes with a diverse set of legal and ethical considerations. For the Norwegian Sign Language Corpus, decisions about how to make the corpus accessible to the deaf community was as important as ensuring that the personal data of signers was respected. FAIR and CARE principles helped frame considerations, encouraging a discussion of how the deaf community could be empowered and respected by the project and the resulting corpus. While it was decided that the corpus should be available to as many relevant stakeholders as possible, it was important to consider downstream use of the corpus, especially in regard to engagement and/or exploitation by new technologies. These concerns, along with respect for the personal data collected in the corpus, warranted the implementation of some more conservative protections. It is hoped that now the corpus can be used by the deaf community and researchers to promote, strengthen, and underscore the value and importance of Norwegian Sign Language to the Norway's linguistic landscape.

## 5. Acknowledgments

I would like to give a heartfelt thanks to all the signers who participated in data collection for the Norwegian Sign Language Corpus, for their time and sharing their language. I would also like to thank all the people who have helped on the different projects as well as the corpus work itself (for a complete list, Ferrara, 2026).

This work has been financially supported by Norwegian Ministry of Culture, NTNU, and the

Norwegian Research Council under project #287067 'Language Use in the Norwegian Deaf Community: Reflections of a Signed Language Ecology.'

## 6. Bibliographical References


Angelini, R., Spiel, K., and de Meulder, M. (2024). Bridging the gap: Understanding the intersection of deaf and technical perspectives on signing avatars. In A. Way, L. Leeson, and D. Shterionov (Eds.), *Sign language machine translation*. Springer, pp. 291-308. https://doi.org/10.1007/978-3-031-47362-3_12

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., and Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal, 19*(43):1-12. https://doi.org/10.5334/dsj-2020-043

Crasborn, O., and Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In Onno Crasborn, Eleni Efthimiou, Thomas Hanke, Ernst D Thoutenhoofd, and Inge Zwitserlood (Eds.), *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora.* Paris: European Language Resources Association, pp. 39-43. Retrieved from https://www.sign-lang.uni-hamburg.de/lrec/pub/08022.html.

De Meulder, M. (2021). Is "good enough" good enough? Ethical and responsible development of sign language technologies. In Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), Virtual. Association for Machine Translation in the Americas, pp. 12–22

GO FAIR. (no date). *FAIR principles*. Retrieved 9 February 2026 from https://www.go-fair.org/fair-principles/

Hochgesang, J. A., and Palfreyman, N. (2022). Signed language corpora and the ethics of working with signed language communities. In J. Fenlon and J. A. Hochgesang (Eds.), *Signed language corpora*. Gallaudet University Press, pp. 159-195.

Jantunen, T., Rousi, R., Rainò, P., Turunen, M., Valipoor, M. M., and García, N. (2021). Is there any hope for developing automated translation technology for sign languages? In M. Mämäläinen, N. Partanen, and K. Alnajjar (Eds.), *Multilingual facilitation*. University of Helsinki Library, pp. 61-73. https://doi.org/10.31885/9789515150257

Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, *15*(1):104-129. https://doi.org/10.1075/ijcl.15.1.05joh

Johnston, T. (2024, November). Auslan corpus annotation guidelines. Manuscript. Sydney: Macquarie University. Retrieved from http://auslan.org.au/about/annotations/

Mayer, M. (1969). *Frog, where are you?* New York: Dial Press.

Schembri, A., and Cormier, K. (2022). Signed language corpora: Future directions. In J. Fenlon and J. A. Hochgesang (Eds.), *Signed language corpora*. Gallaudet University Press, pp. 196-220.

Språklova. (2021). *Lov om språk [The language act]* (LOV-2021-05-21-42). Lovdata. https://lovdata.no/dokument/LTI/lov/2021-05-21-42

NOU 2023:20. (2023). Tegnspråk for livet: Forslag til en helhetlig politikk for norsk tegnspråk [Signed language for life: A proposal for a holistic treatment of Norwegian Sign Language]. Kultur- og likestillingsdepartementet. Retrieved from https://www.regjeringen.no/no/dokumenter/nou-2023-20/id2984187/

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3:160018. https://doi.org/10.1038/sdata.2016.18

Schulder, M., and Hanke, T. (2022). How to be FAIR when you CARE: The DGS Corpus as a case study of open science resources for minority languages. In Nicoletta Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC2022)*. Marseille, France. European Language Resources Association (ELRA), pp. 164-173. https://aclanthology.org/2022.lrec-1.18/

UNESCO. (2021). *UNESCO Recommendation on Open Science*. https://doi.org/10.54677/MNMH8546

Venade de Sousa, F. (2025). *Sign language in the era of artificial intelligence*. European Union of the Deaf.


## 7. Language Resource References


Ferrara, L. (2020). *Norwegian Sign Language dataset in Global Signbank.* Nijmegen: Radboud University, Centre for Language Studies. https://signbank.cls.ru.nl/

Ferrara, L. (2024a). *Norwegian Sign Language Corpus – Language Ecology (Retellings and public events)*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/155.

Ferrara, L. (2024b). *Norwegian Sign Language Corpus – Language Ecology (Conversations*

*and private meetings/events)*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/156.

Ferrara, L. (2025, April 28). Norwegian Sign Language Corpus - Associated files and materials. PID https://doi.org/10.17605/OSF.IO/TYDF4

Ferrara, L. (2026, January). Annotating the Norwegian Sign Language Corpus. Manuscript. Trondheim, Norway: NTNU. Retrieved from https://doi.org/10.17605/OSF.IO/TYDF4

Ferrara, L and Bø, V. (2022). *Norwegian Sign Language Corpus – Pilot Corpus (Conversations)*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/147.

Ferrara, L. and Halvorsen, R. P. (2021). *Norwegian Sign Language Corpus - Halvorsen (2012)*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/141.

Ferrara, L. and Ringsø, T. (2021). *Norwegian Sign Language Corpus – Depicting Perspective*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/144.

# Generations in the DGS Corpus: Evolving Outreach Activities and Cross-Generational Stories on Social Media in a Long-Term Corpus Project

**Anike Fiedler, Marc Schulder, Julian Bleicken, Annika Herrmann**
Institute of German Sign Language and Communication of the Deaf,
University of Hamburg, Germany
{anike.fiedler, marc.schulder, julian.bleicken, annika.herrmann}@uni-hamburg.de

**Abstract**

Social media has become a powerful tool for research projects, community outreach, science communication, and to recruit participants. Due to its differences to traditional media and presentation modes, it provides a particular focus on producing very concise content that is entertaining and accessible while staying informative. In this paper, we describe how the long-term project DGS-Korpus, creators of a corpus and dictionary of German Sign Language, evolved its outreach strategies over time. One unique aspect of its unusually long project run-time of nineteen years is that it has involved several cases of multiple family members participating in the project at different points in time, resulting in cross-generational participation. The paper describes how the project's social media campaign uses these cross-generational connections to illustrate important aspects of the project, such as its relevance for cultural heritage and language identity, the different ways that members of the German deaf communities were and are involved in the project, and its relevance to interpersonal connections.

**Keywords:** sign language corpus, community outreach, science communication, social media dissemination, cross-generational participation

## 1. Introduction

Since the turn of the century, science communication has changed drastically, embracing new paths and formats for reaching its audiences (Kulczycki, 2013). As a long-term project with a run-time of nineteen years, the DGS-Korpus project (Prillwitz et al., 2008) has lived through the better part of these changes. In this article, we describe how it evolves and grows with these developments and how it embraces the unique opportunities of its generation-spanning community involvement.

Social media has been part of the internet since its beginnings as "nothing more than a giant Bulletin Board System" (Kaplan and Haenlein, 2010, p. 60), but its exact nature has evolved repeatedly, from bulletin boards and personal homepages to networking services, micro-blogging and short-form video sharing. Its growing ubiquity in people's lives, being available anywhere and anytime on personal computers and mobile devices, has led to a massive shift from traditional media to social media in terms of advertisement and public relations communication efforts (Allagui and Breslow, 2016). Aside from private use and commercial interests, social media has also become increasingly relevant for the purposes of science communication (Jia et al., 2017).

The DGS-Korpus project has had two phases of data collection, the first running from 2010–2012 and covering adults of all ages, and a second phase from 2024–2026 that focuses on young adults. Public outreach, in particular with the deaf[1] communities, has been a part of the project from the beginning. Aside from in-person activities, the outreach for the first phase focused on providing information via the project's website and on interaction through Facebook. With the shift to a younger generation in the second phase, the focus of online outreach efforts was moved to presenting short-form media on Instagram[2] and Mastodon[3]. A particular opportunity of this second phase lies in the time gap between the two data collection periods. Children and siblings of former participants or moderators who had been too young to take part in the first phase are now old enough to participate themselves.

In the remaining paper, section 2 offers an introduction of the DGS-Korpus project and deaf involvement in it. Section 3 provides an overview of the project's science communication strategy, describing its websites, considerations on accessible social media, and its various types of outreach activities. Section 4 focuses on the relevance of including different generations in the corpus and how this lead to a special social media campaign for the currently ongoing second data collection phase. Our discussion of the project's outreach and social media activities is concluded in section 5.

[1] For an elaboration of the project's decision to use "deaf" rather than "Deaf" or "d/Deaf" see Schulder et al. (2021, p. 19). A thorough discussion of the topic is also provided by Kusters et al. (2017).

[2] https://www.instagram.com/dgskorpusprojekt/

[3] https://fediscience.org/@dgskorpusprojekt/



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 155–163
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

## 2. The DGS-Korpus Project

The DGS-Korpus project (Prillwitz et al., 2008) is a long-term project at the University of Hamburg, funded by the Union of German Academies of Sciences and Humanities and conducted at the Institute of German Sign Language and Communication of the Deaf (IDGS)[4]. It began in 2009 and has a run-time of nineteen years. Its primary goals are the creation of three resources:

- the *DGS Corpus*, a linguistic reference corpus of German Sign Language (DGS, *Deutsche Gebärdensprache*),
- the *Public DGS Corpus*, a fully annotated publicly released subset of the DGS Corpus,
- the *Digitales Wörterbuch DGS (DW-DGS)*, a digital dictionary of DGS primarily based on the DGS Corpus.

These resources were designed to fulfil multiple functions, working as a dataset for linguistic research, resources for sign language education, and as a cultural heritage resource, documenting the life experiences of members of the DGS language communities in Germany.

Data from the Public DGS Corpus are made available through a number of different portals that are optimised for different target groups, such as the deaf communities and researchers (both deaf and hearing) (Jahn et al., 2018) (see section 3.1 for details). Over time, the portals have been revised and expanded based on communities feedback and developments in the academic field (Hanke et al., 2020) and additional portals have been added (Konrad et al., 2024).

Giving back to the communities from which its object of study, German Sign Language, originates has been an integral part of the DGS-Korpus project from its inception. This is why the corpus was designed as a cultural heritage resource and why it has a dedicated communities portal. The project's outreach and science communication activities also reflect this, as has previously been described by Jahn et al. (2022).

### 2.1. Deaf Involvement

Research on sign languages must always involve the deaf communities that are the primary stakeholders of their sign languages (Harris et al., 2009). From the very beginning, the DGS-Korpus project worked together in a team of deaf and hearing staff members and at least for part of the project duration the project was also deaf-led with a deaf principal investigator. Even though, this currently is not the case due to organisational reasons, deaf researchers and annotators are active on all levels: as team members, moderators, focus group members, annotators, student assistants, and participants. We especially invest in the promotion of deaf academics, supporting and accompanying educational training and career building for young deaf researchers.

The main coordinators of participant searches in both data collection phases were well-connected deaf people from the respective regions of Germany. These later also acted as moderators during recording sessions, where their task was to guide participants through the different tasks of the data collection. These recording sessions always took place in a setting with only deaf people present (Schulder et al., 2021). After the first data collection phase, they also continued to work as an advisory board and focus group which provided invaluable feedback on various steps of the endeavour (Jahn et al., 2022).

Annotation of the recordings is performed by a group of deaf and hearing student assistants with sign language competence, supervised by deaf expert annotators who also take care of quality assurance and the anonymization of video data (for further details, see Schulder et al., 2021).

As part of the project's informed consent rules, participants have full control over their recordings and usage licences for the data were designed to protect the participants' rights (Schulder and Hanke, 2022).

### 2.2. Working with Minority Communities

When working with and in minority communities, ethical science practices are essential (Singleton et al., 2015; Harris et al., 2009). In addition to following recommended open science practices such as the FAIR principles (*Findable, Accessible, Interoperable, Reusable*) (Wilkinson et al., 2016), the DGS-Korpus project also aligns with the CARE principles (*Collective benefit, Authority to control, Responsibility, Ethics*) (Carroll et al., 2020) on governance of data from minority communities, as is discussed by Schulder and Hanke (2022).

It is important that minority communities, such as the deaf communities, have control over the data that originated from them (Crasborn, 2010). Aside from ensuring participants' informed consent and appropriate data handling (Hanke et al., 2010), it is also essential to keep communities informed about and involved in research outputs, both in person and online. In-person interactions are facilitated through presentations, publications, exhibitions, events, interviews, etc. and by inviting an exchange of ideas and opinions, for instance at conferences, workshops, practical trainings of language teachers, and in guided tours of the project's

[4] https://www.idgs.uni-hamburg.de

facilities and open exhibitions. For online interaction with our communities, social media posts that focus on video or visual content are particularly helpful and successful.

## 3. Science Communication

In this section, we discuss various aspects of the DGS-Corpus project's science communication strategy. Section 3.1 describes the websites of the project and its resources, with a focus on their considerations regarding community needs. In section 3.2, we discuss how considerations of accessibility drove our choices in social media platform. Section 3.3 provides examples of our various outreach activities, both online and in person events.

### 3.1. Project Websites

The DGS-Korpus project provides a number of websites to disseminate information about the project in general and provide access to the resources it creates. These were designed with the goal in mind of giving back to the communities. See also the related discussions by Jahn et al. (2022) and Schulder and Hanke (2022).

The official project website[5] is presented in four languages: DGS, German, International Sign (IS), and English. We provide videos in DGS and IS for most of the written content. It is regularly updated with reports of events and presentations visited and held by the DGS Corpus team members.

Through the communities portal *MY DGS*[6] the Public DGS Corpus is presented to the deaf communities and all people interested in the cultural heritage and life experiences of participants. Its introduction, the only longer text, is provided in DGS and German. All other German text is kept to the minimum required for navigation. To also provide accessibility for visitors not yet fluent in DGS, videos come with optional German subtitles.

The research portal *MY DGS – annotated*[7] offers access to the recordings of *MY DGS*, enhanced by research materials, such as full sign annotations, translations into German and English, and body pose position data for computer-assisted analyses (Konrad et al., 2022; Schulder and Hanke, 2020). Designed for the international academic communities, the research portal interface was originally provided only in English, but in reaction to feedback from the deaf communities, a German interface was added (Hanke et al., 2020).

Additional portals allow for systematic corpus searches and a comparative linguistics approach via the web-based corpus query tool ANNIS (Isard and Konrad, 2022) and through iLex, the transcription environment used to create the DGS Corpus (Konrad et al., 2024).

Design of the dictionary *DW-DGS*[8] (Langer et al., 2024) also centres around the wealth and complexity of DGS and its position as an independent language, as can be seen in features such as its use of *micons*, moving sign icons designed to fit in with text layouts (Langer et al., 2018). The project website also provides a series of tutorials in DGS to acquaint visitors with the various functionalities of the dictionary.

### 3.2. Accessible Social Media

In science communication in general, but especially when concerning sign languages and deaf audiences, accessibility needs to be considered in a number of ways, such as language barriers, audience-appropriate messaging, but also regarding user privacy protection.

Naturally, deaf audiences should be reached through their preferred language and modality. That means visual materials are prioritised, especially video, as it is the only way to appropriately communicate in sign languages. This matters for project websites (see section 3.1), but also plays a great role in choosing social media platforms through which to communicate. The target community needs to be active on said platform and the platform's capabilities must allow for the intended communication formats (e.g. supporting video posts). As platform capabilities, community preferences and critical discourse on accessibility evolves, these choices also need to be revisited regularly.

During the 2010s, the platform of choice for reaching the German deaf communities was Facebook. Over time this has shifted, especially for younger generations, who are now more active on Instagram due to its focus on visual content.

A downside of these platforms is that they are commercial ventures that, while free to use, finance themselves in part through tracking, harvesting, repurposing and reselling user data. Access to posts on these platforms is often gated, either requiring visitors outright to create a user account and log in (thus allowing them to track users both on the platform and often other websites) or at least strongly urging them to do so by otherwise presenting a degraded user experience. The exact conditions of access can also change at any time, largely without recourse.

In response to this, alternative non-commercial open source platforms have gained popularity in recent years, especially a set of decentralised and cross-compatible services collectively known as

[5] https://www.dgs-korpus.de
[6] https://www.meine-dgs.de
[7] https://ling.meine-dgs.de
[8] https://www.dw-dgs.de

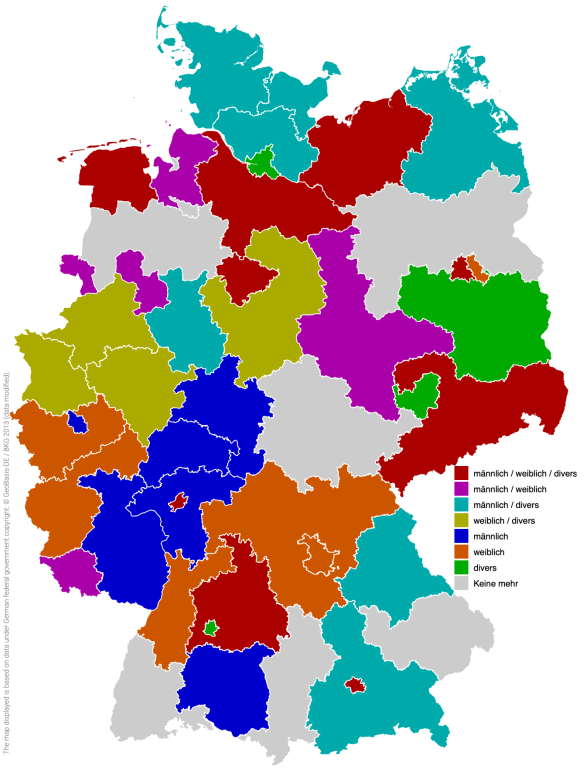



Figure 1: Example of progress map for participant search. Different colours indicate which genders are still missing in the respective region. Grey areas indicate that enough candidates were found.

the *Fediverse*. The most well known of these is *Mastodon*[9], which focusses on the short message *micro-blogging* format previously popularised by Twitter. Instances are hosted on a myriad of private or non-commercial servers in many different countries. These servers are linked to each other through shared open protocols without a centralised controlling authority, following a principle similar to that of email.

While deaf communities presence on the Fediverse is still small, the DGS-Korpus project chose to create an account on `fediscience.org`, a Mastodon instance for scientists hosted in Germany by the non-profit *Fediscience e.V.* The same content is now posted to Instagram and Mastodon, thus creating un-gated versions of all posts.[10]

## 3.3. Outreach Activities

Public relations activities of the DGS-Korpus project are a collaborative team effort. A small designated group of team members and students handle the social media accounts and coordinate other activities as one of their various project duties. All other team members support the effort by contributing to specific events and by providing suitable materials from their other activities (e.g. photos or video messages from public events). In the following, we describe some of our recent efforts. Further activities have also already been described by Jahn et al. (2022).

Aside from keeping people informed about the project's work in general, many outreach activities focus on making them acquainted with the resources the project has created, such as the Public DGS Corpus and the DW-DGS, and to show them how these resources can be of use to them. During data collection phases, an additional focus lies on the recruitment of participants.

### 3.3.1. Social Media

As the second data collection focussed on young signers between 18 and 33 years of age, outreach relied heavily on social media posts on Instagram (and later Mattermost), first to find moderators and then to support them in finding participants from their respective regions (see also section 2.1). For this, moderators would individually record short videos in which they would introduce themselves and the project and describe what region they were searching participants for. These videos were then shared on the project's social media accounts. (A later campaign focussing on cross-generational connections in the corpus is described in more detail in section 4.2.)

Each recruitment video would also show a map of Germany with regions being colour-coded to indicate their respective recruitment progress (see fig. 1). Different colours would indicate whether the project was still looking for male, female or non-binary candidates in that region. An up-to-date version of the map with interactive tooltips was also available on the project website.

### 3.3.2. Exhibitions

Another type of outreach activity is the project's presence in exhibitions. Since 2021, a media station about the project is part of a long-term exhibition at the Humboldt Forum in Berlin, as described already by Jahn et al. (2022).

More recently, the DGS-Korpus project joined seven other long-term projects funded by the *Academy of Sciences and Humanities in Hamburg*[11] in the interactive exhibition *"Notwendig, nützlich, neu – Langzeitforschung in Hamburg"* ("Necessary, useful, new – Long-term research in Ham-

[9] `https://joinmastodon.org/`

[10] We also investigated other Fediverse platforms like *Pixelfed* and *PeerTube* that have a greater focus on image or video materials. Mastodon was chosen based on a number of factors, such as video size limitations, platform maturity, and availability of suitable servers.

[11] `https://www.awhamburg.de/`



Figure 2: The exhibition “Necessary, useful, new – Long-term research in Hamburg” at the Hamburg State and University Library Carl von Ossietzky. Photo © AdWHH / Jann Wilken

burg”).[12] The travelling exhibition started out in Hamburg in late 2025 (see fig. 2) and is currently in Bremen before moving on to other locations.

Our goal for this latest exhibition was to inform the general public about our research outputs and the impact such a project has in terms of preserving the cultural heritage of a minority. At the same time, we also collaborated with the academy to ensure the entire exhibition would provide barrier free access for deaf visitors. In a joint effort between professional deaf interpreters, staff of the IDGS interpreting programme and sign language interpreting students, we created DGS translations for all written and spoken content of the exhibition, available via QR codes at the various exhibits. Aside from ensuring accessibility, this also provided valuable practical experience to the students involved. In addition, we organised tours of the exhibition in DGS, run by a professional deaf museum guide and supported by project members, one tour each in Hamburg and Bremen.

#### 3.3.3. Workshops

The DGS-Korpus project also organises presentations and workshops to promote its resources. While earlier workshops covered the Public DGS Corpus, many of the more recent workshops have focussed on the corpus-based dictionary DW-DGS. For these workshops, project members visit different institutions and universities in Germany to acquaint sign language teachers and students in linguistics or interpreting degrees with the dictionary and corpus, inform them about their development, and teach them how to use them for their teaching, practice of signing and interpreting, and for their own research.

[12] https://www.awhamburg.de/magazin/ausstellung-zur-langzeitforschung-in-hamburg.html

#### 3.3.4. Public Events

Other ways to reach future researchers are taking part in events for pupils organised at the University of Hamburg. The DGS-Korpus project supports the IDGS in taking part in university open house events, classroom visits, school events like *Boys' Day*[13] and other public information events.

## 4. Generations in the DGS Corpus

Due to both its design and long project run-time, the involvement of several generations of signers is a poignant topic of the DGS-Korpus project, which has also been reflected in some of its outreach work. In this section, we describe how it mattered for the design and creation of the corpus itself (section 4.1) and how it was used for a small successful social media campaign during the second data collection phase (section 4.2).

### 4.1. Corpus Design and Creation

The DGS Corpus was designed to provide a diverse and balanced cross-section of DGS communities in Germany. A detailed description of the factors that went into participant selection can be found in Schulder et al. (2021), but the most important ones were region, age, and gender. Aside from ensuring diverse cultural heritage content, this was also done to enable linguistic studies of regional and dialectal differences as well as of language change across generations.

For the first data collection phase beginning in 2010, the participants were divided into four age groups: 18–30, 31–45, 46–60, and 61 or older. This division broadly corresponds to the existing generational cohorts commonly referred to as the millennials, generation X, etc. This division made it feasible to balance age groups during participant recruitment, but also later served as a privacy-retaining way to provide age information in the Public DGS Corpus (Schulder and Hanke, 2022). The majority of participants were partnered with a person of their own age group, but in the interest of also covering cross-generational interactions, some pairs with greater age differences were also selected.

The second data collection phase focuses on a fifth cohort, born between 1993 and 2007. As such it mainly covers people who were not considered during the first data collection period as they were still underage. A small number of participants belonging to the previous cohorts was also recruited to again also include cross-generational pairs.

[13] *Boys' Day*/*Girls' Day* is a national initiative to acquaint pupils with vocations and research fields in which their respective gender is under-represented.



Figure 3: Screenshot of first social media post about family connections in the corpus. The mother (left) was an advisory focus group member during the first data collection phase and the daughter is now moderator in the second phase.



Figure 4: Screenshot of second social media post about family connections in the corpus. The daughter (left) was a participant in the first data collection phase and moderator in the second phase, while her father joined the second phase as a participant.

Since the new data collection focusses on a single cohort of young people, it was decided to also select moderators from the same age group. The responses to the call for new moderators were different to the search for moderators in 2009. This new generation has built strong deaf identities through bimodal-bilingual education and the cultural and legal recognition of sign languages. In contrast to the first phase, we mainly found new moderators that had a deaf family background. This was partly because focus group members of the first phase were asked to help with finding moderators, in some cases resulting in them proposing family members.

In one case, a whole family has now been involved in the project's data collection. Both parents were moderators in the first data collection phase and in the second phase their daughter and son are now moderator and participant, respectively. These inter-familial and cross-generational connections within the corpus in turn inspired the social media campaign described in the following section.

## 4.2. Cross-generational Social Media Campaign

In several instances, families have been involved in both data collection phases, with different family members being moderator or participant in either phase. This provided a unique opportunity for a small social media campaign that highlights these cross-generational connections within the DGS Corpus. The concept of this campaign was to show how participation in the DGS Corpus can connect the different generations, creating a shared experience, and how participation can be of value for the deaf communities. This helped advertise the corpus in general, highlight its relevance to the communities and, most importantly, motivate potential candidates to volunteer for participation.

For the campaign, we asked each pair of family members to record a short video in which they talk about their experiences with the DGS-Korpus project. These were then posted to the project's Instagram and Mastodon accounts. Links to the Mastodon post of each video are provided in the footnotes.

In the remainder of this section, we will present four of the resulting social media posts about cross-generational connections (sections 4.2.1 to 4.2.4) and the impact of this social media campaign (section 4.2.5).

### 4.2.1. Video 1: Heritage

The first video[14] posted to our social media accounts shows mother and daughter (fig. 3). The mother was a moderator in the Leipzig region. She later worked very passionately in the focus group. Her daughter signed up to be moderator in the second data collection phase. Together, they sign about how the first data collection phase was organised a little differently to the second phase, but that together both data collections will build a large archive of German Sign Language utterances and stories and that it is important to support this project.

### 4.2.2. Video 2: Recording Day

The second video[15] features a daughter and her father (fig. 4). She was a participant of the youngest cohort during the first data collection phase. When she learned about the second data collection, she

[14] https://fediscience.org/@dgskorpusprojekt/114986130472182158

[15] https://fediscience.org/@dgskorpusprojekt/114953928082797230



Figure 5: Screenshot of third social media post about family connections in the corpus. Two brothers, one is a moderator in the second data collection phase (left) while the other had already been a participant in the first phase.



Figure 6: Screenshot of fourth social media post about family connections in the corpus. The daughter (left) was a participant in the second data collection phase and her father was a participant in the first phase.

volunteered to participate again, this time as a moderator. Since she represented one of the regions that was selected to have a participant of an older cohort, she chose to ask her father if he wanted to be part of her group. He agreed, and after his data collection session for the corpus they recorded a short video together to promote participation in the ongoing data collection. They sign about how participating is not complicated and that even though the father was initially very nervous because of all the cameras, he soon forgot about them and had a pleasant time.

#### 4.2.3. Video 3: Language Change

In the third video[16], we see two brothers (fig. 5). One of them was a participant in the first data collection. At that time, the other brother was too young to participate, but when offered the chance to be a moderator in the second phase, he accepted to be a part of the project. They recorded a video reflecting on how the language has changed over time and why it is of utmost importance to document these language changes on video.

#### 4.2.4. Video 4: Youth Language

In the fourth video[17], we see father and daughter (fig. 6). The father has been working at the IDGS as a lecturer for sign language and deaf studies for a long time. During the first data collection period he was asked to take part in an equipment testing session. That sparked his interest and he was then selected as a participant. During the second phase, he recommended his daughter as a suitable candidate for the data collection. She was then contacted by the moderator of her region and selected for participation. In their conversation they touch upon language change and youth language among other things.

#### 4.2.5. Impact

All videos were posted to Instagram and Mastodon to promote participation in the data collection and to show that taking part is a unique opportunity that provides value for the deaf communities. On Instagram, these videos were collectively watched 23,300 times with the second one with 8,800 views being the most watched out of these four videos. The posts resulted in 905 likes, 127 new followers and in several people applying to participate in the data collection. In comparison to the engagement with other posts, these were received particularly well in the communities.

In general, we made the observation that posts that feature signed videos are viewed more often than posts that only feature pictures and text, but the cross-generational videos have been the best performing even among other DGS video posts. The small cross-generational campaign might have helped to convey the sense of why we are trying to collect this data and that the DGS Corpus is not only a compilation of signed data for linguistic research but also an archive of deaf cultural heritage and life experiences.

## 5. Conclusion

In today's science communication, social media is a valuable tool, especially to reach younger generations. It has the advantage of directly reaching target groups in a dynamic and open conversational manner (Al-Quran, 2022). Our article showed how the DGS-Korpus project uses social media as well as other outreach formats, such as workshops, exhibitions and public events, to keep the deaf com-

[16] https://fediscience.org/@dgskorpusprojekt/114852199075743085

[17] https://fediscience.org/@dgskorpusprojekt/116198565031730915

munities and the general public informed about the project's work, make them acquainted with its corpus and dictionary, but also to promote interest in sign languages and deaf culture.

Especially when working with minority languages and cultures, it is vital to find a way to reach the target group in their preferred mode of communication. Social media has offered the small public relations team of the DGS-Korpus project the opportunity to serve the necessity of visual communication. Through signed videos and eye-catching posts, they not only keep the communities informed but also engage with them on an equal footing. The amount of user engagement, when gauged by popular metrics such as numbers of views, likes and comments, can naturally not compete with professional accounts with the resources to produce a steadier stream of high quality content. Nevertheless, we could clearly see that posts were rewarded with interactions and increasing numbers of followers, as well as several applications for participation in the DGS Corpus data collection. The cross-generational videos posted on social media, but also the other posts about conferences, papers and exhibitions show how the collection of signed conversations and the research already done on the data of the first data collection of the DGS Corpus impact the different generations of deaf people in Germany. It helps all of us see and understand how unique and worthy of protection sign languages, insider experiences and the deaf cultures are.

## 6. Acknowledgements

This document has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

## 7. Bibliographical References

Marwan Walid Mohammad Al-Quran. 2022. Traditional media versus social media: challenges and opportunities. *Technium*, 4(1):145–160.

Ilhem Allagui and Harris Breslow. 2016. Social media for public relations: Lessons from four effective cases. *Public relations review*, 42(1):20–30.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43.

Onno Crasborn. 2010. What does "informed consent" mean in the internet age?: Publishing sign language corpora as open content. *Sign Language Studies*, 10(2):276–290.

Thomas Hanke, Sung-Eun Hong, Susanne König, Gabriele Langer, Rie Nishio, and Christian Rathmann. 2010. Towards fair licences for data from the DGS Corpus project. Poster presented at the 4th Workshop of the Sign Language Corpus Network (SLCN 4), Berlin, Germany.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in size and depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Raychelle Harris, Heidi M. Holmes, and Donna M. Mertens. 2009. Research ethics in sign language communities. *Sign Language Studies*, 9(2):104–131.

Amy Isard and Reiner Konrad. 2022. MY DGS – ANNIS: ANNIS and the Public DGS Corpus. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 73–79, Marseille, France. European Language Resources Association (ELRA).

Elena Jahn, Calvin Khan, and Annika Herrmann. 2022. Outreach and science communication in the DGS-Korpus project: Accessibility of data and the benefit of interactive exchange between communities. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 80–87, Marseille, France. European Language Resources Association (ELRA).

Elena Jahn, Reiner Konrad, Gabriele Langer, Sven Wagner, and Thomas Hanke. 2018. Publishing DGS Corpus data: Different formats for different needs. In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 83–90, Miyazaki, Japan. European Language Resources Association (ELRA).

Hepeng Jia, Dapeng Wang, Weishan Miao, and Hongjun Zhu. 2017. Encountered but not engaged: Examining the use of social media for science communication by Chinese scientists. *Science Communication*, 39(5):646–672.

Andreas M Kaplan and Michael Haenlein. 2010. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Reiner Konrad, Thomas Hanke, Amy Isard, Marc Schulder, Lutz König, Julian Bleicken, and Oliver Böse. 2024. Corpus à la carte – improving access to the Public DGS Corpus. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 390–399, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. Public DGS Corpus: Annotation conventions. Project Note AP03-2018-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.

Emanuel Kulczycki. 2013. Transformation of science communication in the age of social media. *Teorie vědy / Theory of Science*, 35(1):3–28.

Annelies Maria Jozef Kusters, Maartje De Meulder, and Dai O'Brien. 2017. Innovations in deaf studies: Critically mapping the field. In Annelies Kusters, Maartje De Meulder, and Dai O'Brien, editors, *Innovations in Deaf Studies: The Role of Deaf Scholars*, Perspectives on Deafness, pages 1–53. Oxford University Press, Oxford.

Gabriele Langer, Anke Müller, Sabrina Wähl, and Julian Bleicken. 2018. Authentic examples in a corpus-based sign language dictionary – why and how. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (EURALEX 2018)*, pages 483–497, Ljubljana, Slovenia. Ljubljana University Press.

Gabriele Langer, Anke Müller, Sabrina Wähl, Felicitas Otte, Lea Sepke, and Thomas Hanke. 2024. Introducing the DW-DGS – the digital dictionary of DGS. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 316–325, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and Arvid Schwarz. 2008. DGS Corpus project – development of a corpus based electronic dictionary German Sign Language / German. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 159–164, Marrakech, Morocco. European Language Resources Association (ELRA).

Marc Schulder, Dolly Blanck, Thomas Hanke, Ilona Hofmann, Sung-Eun Hong, Olga Jeziorski, Lutz König, Susanne König, Reiner Konrad, Gabriele Langer, Rie Nishio, and Christian Rathmann. 2021. Data statement for the Public DGS Corpus. Project Note AP06-2020-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.

Marc Schulder and Thomas Hanke. 2020. OpenPose in the Public DGS Corpus. Project Note AP06-2019-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.

Marc Schulder and Thomas Hanke. 2022. How to be FAIR when you CARE: The DGS Corpus as a case study of open science resources for minority languages. In *13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 164–173, Marseille, France. European Language Resources Association (ELRA).

Jenny L. Singleton, Amber J. Martin, and Gary Morgan. 2015. *Ethics, Deaf-Friendly Research, and Good Practice When Studying Sign Languages*, chapter 1. John Wiley & Sons, Ltd.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):9.

# Formalising Sign Language Depiction, Characterising Categories and Measuring Iconicity with AZee

**Michael Filhol, Emmanuella Martinod**
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)
507, rue du Belvédère, 91400, Orsay, France
michael.filhol@cnrs.fr, emmanuella.martinod@gmail.com

**Abstract**

This paper deals with depiction in (French) Sign Language, the formal account AZee can provide, and how it compares, validates or simplifies the linguistic notions of classifiers and iconic structures. It reports on a partial encoding work on *Mocap1*, a corpus with a high density of depicting structures, following the same method that led to the first AZee reference corpus *40 brèves*. The approach does not postulate classifiers or iconic structures as entities separate from lexical signs, and nonetheless manages to model the corpus data. We discuss the entailed possibility to rediscover some of the useful categories, and if so define them from AZee's premises. We also specify how a formal metric can be specified to measure iconicity in signed data. While this paper is of linguistic interest as it compares to existing theories, it also provides a concrete step to covering depicting discourse with AZee, therefore enable automatic SL animation of depiction.

**Keywords:** French Sign Language, Formal Representation, SL Synthesis, AZee, Iconicity, Depicting Signs

## 1. Introduction

Sign Languages (SLs) depend on visual-gestural modality. This modality allows iconicity in the language, where elements resemble what they refer to, even outside of conventionalised signs. In the history of SL study, this non-lexical iconicity has often been overlooked, sometimes even rejected or considered anecdotal (Frishberg, 1975). Although with varying degrees of interest, it is clearly more considered today. This led to the development of a large range of concepts to describe it.

These descriptions are however not formal to the extent that they could be used for automatic generation with avatars for example. In the domain of SL generation, iconicity is indeed still largely set aside because of its complexity. To this end, we have been developing the AZee framework, which provides a formal linguistic description, readable by a computer to fully specify avatar animation. This paper contributes to developing AZee further by confronting a corpus containing heavily iconic signing, a genre still only partially covered.

First, we present the main concepts used in linguistics to describe iconic structures in SLs. After introducing AZee as a theoretical framework and a language, we report on the corpus we encoded and how these structures were handled with AZee. We then propose a comparison to existing linguistic frameworks for their description, attempting to rediscover categories such as "classifiers" which have been essential in linguistic theories.

## 2. State of the Art

### 2.1. Related Work in Linguistics

Iconic signed parts have traditionally been analysed distinguishing low-level combining features, with special interest in handshapes, from the higher level structures in which these are used. We follow the same distinction in the two following sections. Then, we focus on the frequently mentioned links between iconic structures and the lexicon.

#### 2.1.1. Iconic Handshapes

The most widely used concept is that of *classifier*. This terminology comes from an analogy with the description of some spoken languages like Navajo (Supalla, 1978). In those languages, *classifier morphemes* are used to refer to a class of entities. Applied to SL description, *classifier* stands for a meaningful handshape that expresses information about an entity: its size, shape, or location (Emmorey, 2003; Perniss et al., 2007).

Through time, many different types of classifiers have been defined. Today, a salient distinction seems to have emerged between *entity* (or *object) classifiers* and *handling classifiers*: the first concerns the representation of a shape, whereas the second concerns the representation of the handling of a shape. Several phonological differences have been observed between the two groups (Coppola and Brentari, 2014; Brentari and Goldin-Meadow, 2017).

Another term sometimes referring to such iconic representation of a shape in SL is *proform*. Several authors have used this term (Sutton-Spence



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 164–173
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

and Woll, 1999; Johnston and Schembri, 1999), but strongly promoting it is the "Semiological approach", which gives iconicity a crucial role in shaping SL discourse and grammar (Cuxac, 2000; Sallandre, 2014). It proposes a systematic role of eye gaze in the structures the proforms are used in, which we further present in the following section. In general though, we note little difference between the concepts of proform and classifier.

In any case, an iconic handshape is never produced alone: it is necessarily part of a structure that can be analysed as such (Valli and Lucas, 2000). Let us see how these structures are categorised in the literature.

### 2.1.2. Iconic Structures

From the 1990s, iconic structures have been described in SLs analysed separately, but a range of terminology has evolved to refer to those structures everywhere, with various definitions (Emmorey, 2003): classifier signs, classifier verbs, verbs of motion and location, classifier predicates, spatial-locative predicates, polymorphemic predicates, polysynthetic signs, productive signs, polymorphemic signs, etc.

Given the diversity of this terminology, a standardization was attempted by Johnston (2024), in part inspired by the work of Liddell (2003) in a cognitive-functionalist framework, i.e. emphasizing the *function* of linguistic structures. This document, last updated in 2024, serves as a reference for a large part of recent work in linguistics, especially when it comes to annotation of SL corpora. It distinguishes between depicting signs and non-conventional signs.

#### Depicting Signs

Alternatively named *classifier signs*, these structures include movement depictions ("DSM") and location depictions ("DSL"), in which the entity that moves or is located in the signing space is identified "from the context and/or by the handshape". They further include size and shape depictions ("DSS"), which "describe an object by showing a salient physical feature of it and often occur immediately before or after the object is named with a sign". A last possible type is ground depiction ("DSG"), representing the physical background, or a reference point in a situation.

#### Non-conventional Signs

These structures include "enactments", where the signer mimics an action using the whole upper body, without using lexical (i.e. conventional) signs.[1] Enactments can involve handling classifiers thus constitute "handling depiction". The other type is "gestures", which have a role of modulating and expressing the mood or attitude of the signer, or regulating the discourse and interaction.

In parallel, the semiological approach uses another terminology in which the criterion of non-conventionality of some structures is not at the core of the description. They are called *transfers* because of the cognitive process of transferring human experience into visual linguistic structures. The three main types of transfer are: size and shape transfer (SST), to show the size and/or shape of an entity, situational transfer (ST), which shows the movement of an actant in relation to a ground, and personal transfer (PT), where the signer embodies an individual or an entity. This last category has at least five sub-types which we do not develop here. Those three possible transfers require very precise use of manual and non-manual parametric components. In particular, a specific eye gaze pattern is observed: the signer's gaze seems indeed directed towards the hands *versus* towards the addressee for SST and ST, and the signer's gaze is the one of the embodied entity for PT.

| **Johnston (2024)** | **Sem. approach** |
|---|---|
| DSS | SST |
| Combinations of DSG, and/or DSM, and/or DSL | ST |
| Non-conventional Sign: Enactment and/or Gesture | PT + sub-types |

Table 1: Descriptive concepts for iconic structures

We tried to establish correspondences of these main various descriptive concepts in table 1. Yet, the range of descriptive concepts often overlap without matching exactly, and frequently include combinations of sub-types, so the correspondence remains somewhat approximate. Despite standardization attempts, there is currently no official consensus to describe these elements which are nevertheless massively present in SLs.

### 2.1.3. Iconic vs. Lexical Continuum

The complexity of the linguistic literature on iconic handshapes and structures and their difficult categorisation also comes from the possibility to combine those iconic elements with conventional lexical signs. For example one can make use of a conventional sign while in the middle of an enactment. This would correspond to a non-conventional sign for Johnston (2024), and to a sub-type of PT ("semi-PT") in the Semiological approach (Sallandre, 2014). In the classification described by

[1] Enactment is also frequently called *Constructed Action* (CA) in the literature, see Cormier et al. (2015).

Cormier et al. (2015), this case would correspond to a *Reduced CA*.

This kind of combination is part of SL. It is also linked with its diachronic evolution. Indeed, classifiers would be central in the emergence of lexical signs (Brennan, 1990; Fusellier-Souza, 2004; Johnston and Schembri, 1999; Schembri, 2003; Zwitserlood, 2012). This process sometimes makes it difficult to determine with certainty whether one is dealing with a lexicalized sign or not yet. Here too, categorisations diverge depending on the theoretical framework of authors and their respective focus, but they globally tend to overlap. The overall similarity is that even if researchers seem aware of the fact that the frontier between iconic and lexical is not sealed, their description overall include elements of two different natures. In any case, those categorisations are still missing the formal representation that is necessary for the purpose of avatar animation.

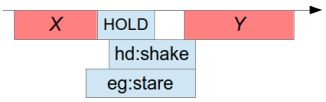
## 2.2. AZee Framework and Corpus Creation

AZee is a formal framework to represent Sign Language utterances and discourse, designed to make both following statements true (Filhol and Hadjadj, 2016):

- compositionality in meaning is accounted for, i.e. if an interpretable piece of signing $S$ is built from a smaller interpretable piece with form $s'$, then the formal representation of $s'$ is visible in that of $S$;
- all body articulations are determined by their representation, in other words automatic avatar animation of any represented utterance is possible, which makes the approach testable.

This is done by encoding linguistic rules, each capturing an association between an identified meaning and form articulations in the target language. Such a rule is labelled according to the attached meaning, and the form specified in the native AZee language, which is able to express body posture constraints, geometric objects in the signing space, and synchronise actions on a timeline. The form–meaning association can be of any language level (including non-manual aspects), carry any semantic granularity, and produce a form constraining any body parts. It can also be parametrised, provided the form can be expressed as a function of the parameters in a stable way. For example:

- a rule "`red`" is specified for the trivial, invariable association between the meaning "red" and the conventional form one finds in a dictionary;




"hd": head; "eg": eye gaze

Figure 1: Form timeline for `not-but`$(X, Y)$

- the more complex, parametrised rule "`not-but`$(X, Y)$" associates the meaning "not $X$ but $Y$" and the timeline shown in fig. 1 (arguments $X$ and $Y$ are open to nested signed content);
- or concerning the matter in this paper, rules making iconic use of space like `place-object`, which captures the common meaning factor between placing a plate on a table, a frame against a wall, etc. and the common form factor: a quick eye gaze to the object location $loc$, immediately followed by a straight movement of body parts jointly configured (denoting a class of object), settling at $loc$.

These rules then serve as the basis for "discourse expressions", i.e. combinations representing signed utterances of arbitrary length. For example, rules `red`() and `with-surprise`$(sig)$ can be combined to mean "really red?!": `with-surprise`$(sig =$ `red`$())$. The examples above would also be useful to compose the meaning: "not [here / at point $A$], but surprisingly [there / at point $B$] instead". Both arithmetically resolve to a timeline containing all of the forms contributed by the respective rules used in the expression.

The approach has been tested both theoretically, through linguistic coverage of the French Sign Language (LSF) videos of the *40 brèves* corpus (LIMSI and LISN, 2022), and technically, through actual avatar synthesis of AZee-encoded expressions (Filhol and McDonald, 2018). The corpus coverage test was published in (Challant and Filhol, 2024), and consists in the AZee encodings of 120 videos of about 30 s each, totalling 1 hour of LSF. It evaluated a coverage of 98%, through timing of the missing pieces, encoded as "ellipses"[2]. It is today a reference corpus for AZee.

## 2.3. What is Missing and New Goal

The genre of the *40 brèves* material has clear benefits to serve as a reference. It comprises

[2]An ellipsis is an expression made of a single dummy rule application, covering a non-encoded part of the signing and placed to keep the overall expression valid and compilable. Its label is not one known and taken from the production set, rather it is chosen ad-hoc to describe the meaning of the omitted part, and enclosed in brackets to make it recognisable. For example:
`:[cars everywhere stuck in traffic]`



Figure 2: Example of elicitation material used in the descriptive task of *Mocap1*



Figure 3: Example of `class-flat-round-large`

monologues facing the camera of error-free SL, not produced by simultaneous/on-the-fly interpretation, and provides a non-restricted semantic field. Besides, since spatial, iconic constructions are ubiquitous in SL, instances can be found in it too. But those are known to reach discourse proportions that are a lot higher in SL, especially when depicting scenes and relative positions of objects or places, which are not strongly exhibited by the *40 brèves*. A goal then was to extend the test above and apply it to a genre of signing containing a high density of these iconic depictions, eventually to evaluate the AZee capability on such data.

## 3. Experiment: *Mocap1* and First Pass

The corpus chosen to represent iconic depiction is *Mocap1*, created through controlled elicitation of image descriptions (LIMSI and CIAMS, 2020). It consists in 187 LSF videos, of 8 signers each freely describing 25 different images such as that in fig. 2, with an average duration of 25 s per video.

We applied the AZee methodology to write the discourse expressions for the contained videos, exactly as was done to reach the *40 brèves* milestone. That is, given the current state of the productions set, either:

- find the combination of rules both accounting for the meaning interpreted of the video and justifying the observable forms;
- or if this is impossible due to missing rules, mine for new rules in SL data to extend or refactor the set, following the published methodology (Hadjadj et al., 2018);
- lastly, if a discourse segment remains difficult to account for with a stable combination of known, new or refactored linguistic rules, capture the whole segment in an ellipsis.

At this point we have reached 60 encoded videos of the corpus, which corresponds to 30 minutes, and one third of the full *Mocap1* material. In total, 22,307 operations and values compose the 60 corresponding expressions, of which 5,471 are rules of the production set, and 35 are ellipses. The remaining 16,801 are native AZee operations (e.g. `translate`) and values (e.g. numericals). No funding is currently supporting further progress, but these numbers to us are enough to create interest in the data. We therefore chose to report on this first pass and look into the way iconic structures are handled.

## 4. Discussion and Comparison to Literature

Section 2.1 presented classifiers as manual forms with attached semantics (usually a category of object, according to its shape), recruitable in structures that fall out of the conventional lexicon, and yet are non-ambiguous between signers. This section studies how these forms and the structures they appear in are accounted for in the AZee material resulting from the encoding of Mocap1.

### 4.1. Iconic Handshapes

In AZee, such a reusable, stable link between a form and an attached semantics precisely makes the case for a linguistic rule in the production set. An example in LSF, and a widely shared one across SLs, is the form–meaning binding between the manual configuration of fig. 3 and a large flat round object.

The form, as any other postural configuration in AZee, is specified as a set of articulatory constraints applying simultaneously. In the case of fig. 3, the form:

- constrains the appropriate finger joints to make the thumb and index bend and the other fingers close;
- orients the index proximal joints along opposite normal vectors to make the hands enclose a disc, in a plane whose orientation might also be relevant (e.g. plate on table vs. on wall);
- places each index base point at a distance from the centre of the disc, that distance giving an idea of its size, fingers in opposite directions.

The location point, the distance from it to each hand, and the disc plane orientation are three variables affecting the form and the meaning jointly. This is captured by arguments for the rule, each named after its interpreted meaning and typed accordingly: the location $loc$ (of AZee native type `POINT`), the orientation of the plane in which it lies, represented by a normal vector $nrmsurface$ (type `VECT`), and a $size$ argument (of type `NUM`, compatible with lengths).

The linguistic rule `class-flat-round-large` is defined as a function of those parameters, yielding the body configuration (native type `POSTURE`) directly oriented and positioned in space according to the given argument values. Other rules accounting for traditional classifiers were of course found useful to encode the full corpus, e.g. `class-flat-surface-narrow` (configures the index and middle fingers straight in the plane of the surface, other fingers closed), or `class-parallel-lines` (four fingers extended and spread, thumb folded in).

We see how no supplementary concept like "classifier" is required to account for such entities. They just fall in the general formal approach to linguistic rule specification in AZee. But after describing *Mocap1* where many are present, an interesting question arises about AZee's account of such documented feature, and whether this category can be rediscovered or characterised in formal terms within the framework, which does not postulate their existence as a distinct case in the first place.

The first thing to notice is that they are modelled as functions yielding instant body postures, i.e. constrain parts of the body into timeless configurations. This is unlike the other linguistic rules defined so far, which yield timelines, i.e. intervals with durations allowing to control posture changes over time.

We also notice that while their arguments can vary, all of them have a location argument of geometric type `POINT`. They often have supplementary spatial orientation or size features, but none that can be made generic in the sense of being applicable to all of them, or interpretable with the same meaning. For example `class-flat-round-large` can be oriented in a plane (e.g. horizontal table on which a plate sits), and `class-person-standing` can be oriented to face a certain direction (index finger pad is understood to be the face of the person), but these are rule-specific interpretations. Besides, `class-flat-round-large` allows a size argument, whereas `class-person-standing` does not. Although a person could semantically be given a size, the form of the latter cannot linguistically depend on such variable, and is only meaningfully variable in location and index facing. The location argument though, is always present. It is in fact a requirement for use in certain structures as we are about to discover. Hence we can tentatively characterise a classifier (or a proform), in formal terms based on AZee specifications, as an AZee function with a signature compatible with `(loc:POINT)` $\rightarrow$ `POSTURE`.

Interestingly, the literature overwhelmingly speaks of *handshapes* to define classifiers, whereas in this framework any set of articulators can be constrained to yield the `POSTURE`. This can involve both hands (e.g. `class-flat-round-large`), more than the hand (e.g. handshape + forearm and its relevant orientation for a tree), or indeed anything else (e.g. the head or an arm alone). It is now worth looking at where these rules are used in the corpus, for insight on the structures that contain them.

## 4.2. Iconic Structures

Of the iconic structures listed in 2.1, we only found 5 role shifts in the corpus. This has to do with its descriptive and non-narrative genre, and leads us to concentrate on the other structures: those categorised by Johnston as "depicting signs".

Let us present examples of AZee code representing such structures. Expression 1 below corresponds to a standing person facing backwards (towards the signer), placed at a point named $P$ defined earlier in the code. Following Johnston's guidelines, such an instance would be annotated "DSL" on the strong hand. Expression 2 shows a similar action on the strong hand, this time performed while the weak hand is holding a fixed anchor, namely a bridge instantiated as a narrow strip (straight index and middle fingers) oriented flat along a left–right axis. The strong hand placed a statue (instantiated as a standing person) facing backwards from behind the bridge. This is another type of classifier predicate, one that would be annotated with a "DSG" on the weak hand, and a "DSL" on the strong. Expression 3 encodes a production where the signer depicts the arm of a crane with a forward and upward movement of the strong hand in a cylindrical configuration ("C"). For Johnston, this would be an instance of a "DSS".

```
(1) :place-object
      'class
      .prf-person-standing
        'facing
        bwd
      'loc
      ^p

(2) :landmark-in-place
      'lm
      :class-flat-surface-narrow
        'loc
        [bridge anchor point]
        'hand
        w
```

```
        'nrmsurface
        dwn  % horizontal surface
        'dirsurface
        rgt  % strip direction
      'sig
      :place-object
        'loc
        translate| ^Mssp| bwd| tiny
        'class
        .class-standing-person
          'facing
          bwd
```

(3)
```
:deploy-shape
  'path
  [fwd & upward from near chest]
  'class
  .class-cylindrical-medium
    'hand
    s
    'nrmsurface
    [fwd & upward]
```

In the encoded data overall, all iconic structures have been covered by combinations of the following rules[3]:

- `place-object`($class$, $loc$, ?$nrmsurface$), mentioned in §2.2;
- `move-object`($class$, $path$), which depicts a movement of an instance of $class$ by making it follow the given $path$;
- `deploy-shape`($class$, $path$), which depicts a fixed line or surface in space, and produces a non-accelerated movement along the given path with a $class$ argument denoting the kind of shape or texture depicted;
- `mult-in-a-row`($item$, $path$), which depicts a set of countable objects of $class$, distributed along the given $path$;
- `mult-around`($item$, $loc$, ?$nrmsurface$, ?$radius$), which depicts a set of countable objects of $class$, distributed on a surface around $loc$;
- `landmark-in-place`($lm$, $sig$), which means $sig$ in a context where a landmark $lm$ is active/present, for example an object in space;
- classifiers as per defined above, i.e. yielding postures from (at least) a location point like `class-person-standing` or `class-flat-round-large`;
- already known rules combining timelines, namely `each-of`/`all-of` (creating sets from parts), `simultaneous` (meaning two pieces $sig1$ and $sig2$ simultaneously true), `side-info` (adding non-focused/incidental information $info$ about a $focus$) and `instance-of` (specifying a $type$ for a focused element $elt$).

We first notice that except for several discovered of the classifier kind, these rules were already known or tentatively proposed in prior work, either to complete the *40 brèves* corpus or to address avatar animation with the Paula system (Filhol and McDonald, 2018). The result is that all of the spatial iconicity covered in the corpus is accounted for through combinations of only very few rules, forming a highly productive system.

Besides, all of the depicting structures encoded involved at least one of the five first items in the list above, which are rules with mandatory arguments of a geometric type (POINT, VECT or PATH). The classifiers, also frequent, verify this condition too given that they have at least a point argument $loc$. The others however, although seen in many iconic structures, are also largely used in non-iconic, non-spatialising contexts. In light of this, we propose the following criterion to qualify a rule as "iconic": having at least one mandatory geometric argument. This outlines a subset of "iconic" rules in the production set. But a noteworthy fact is that structures making use of such an iconic rule almost always contain other, non-geometric ones like `all-of` or `simultaneous`. The next section deals with this typical kind of mix.

## 4.3. Iconic vs. Lexical Continuum

### At the Structures Level

The fact that no a-priori categorisation is assumed between rules makes the definition of lexical vs. iconic units almost redundant on the elementary level in AZee. But this again also consequentially allows to look for a formal definition of the categories accepted in the literature, and perhaps discover an explanation for the borderline cases, such as semi-PT, or reveal a continuum and a way to measure positions thereon.

In the literature, mixes of lexical elements and iconic structures must be analysed as a language phenomenon recruiting elements from different natures to constitute a new kind of structure. In AZee, no such separation exists in the first place, rather rules are applied as arguments for others, regardless of their iconic weight. Those of the so-called lexicon (no mandatory arguments) and those labelled iconic above (with mandatory geometric arguments) may just take part in building discourse expressions. For example, expression 2 in §4.2 contains

[3] Rule arguments prefixed with '?' are optional arguments, i.e. those for which a default value is specified, and which can therefore be left out when the rule is applied in an expression. Non-prefixed arguments are mandatory, i.e. must be given a value when the rule is applied.

4 rules, namely `landmark-in-place`, `class-flat-surface-narrow`, `place-object` and `class-standing-person`. The last three of those are iconic as per our criterion above (they have a mandatory geometric argument in their signature: all of them at least a point named $loc$), whereas `landmark-in-place` is not (its arguments are a timeline and a posture, neither geometric objects).

Now the utterance could typically be signed with the conventional sign for "statue" preceding the strong hand placement (standing person), without breaking the weak hand landmark. This would be encoded by expanding the $sig$ part signed while the $lm$ landmark is held, turning it into another typical pattern in our data:

```
:landmark-in-place
  'lm
  [class-flat-surface-narrow; no change]
  'sig
  :instance-of
    'type
    :statue
    'elt
    [place-object; same as prev. "sig"]
```

This inserts two rules in the expression, one lexical (`statue`, no arguments) and one not (`instance-of`, two timeline arguments). Neither verifies the "iconic" criterion proposed above, so the number of iconic rules is unchanged, while the total rule count went from 4 to 6. The ratio of these two figures $N_{iconic}/N_{total}$ therefore went down from .75 to .5, reflecting a heavier use of non-iconic rules in the mix. We propose that this ratio conveniently measures the iconicity of a signed segment on a $[0..1]$ continuum, through counting rules used in its AZee-encoded expression, instead of defining a-priori labels and creating new categories as new mixes are discovered.

#### At the Discourse and Corpus Level

Intuitively, one would expect to observe very different AZee rules in the encoding of *Mocap1* corpora, compared to that of *40 brèves*. We did the comparison, which showed that among the ten most commonly used rules, three are identical. Those are `info-about` (focused information about a preceding topic), `side-info` (additional information given about a preceding focus), and `in-context` (setting a context for a following process). These three rules therefore appear most useful both in informative and descriptive contexts. For instance, `info-about` is used in *40 brèves* to specify that a man is French, and in *Mocap1* to specify that the lower part of a wardrobe has a sliding door.

However, some of the top 10 rules in *40 brèves* are not among the *Mocap1* top 10. Vice versa, *Mocap1* exhibits many `deploy-shape`, `prf-flat-surface`, `simultaneous`, `place-object`, unlike *40 brèves*. These rules are all present in both, but their ranking is drastically different.

We also observe differences in proportions of lexical signs between the two corpora. Considering an AZee rule "lexical" if it is specified without any mandatory argument, *40 brèves* contains 43% of them, whereas *Mocap1* only holds 10%. This allows to confirm, through AZee representations, that lexical signs are therefore less prevalent in *Mocap1*, and propose a formal metric for iconicity of a corpus.

In summary, the same production rules are used to represent utterances from the two distinct corpora, although these rules are applied in different proportions. First, this shows that discourse is almost never *only* about lexical elements. Second, that iconic structures can be formalised with the same structuring principles than the fully conventional ones. AZee encodings allows us to observe and quantify the proportion of these features.

### 4.4. Limitations and Propositions

A few elements still need improvement or refinement.

Firstly, we have encountered cases where the known production set was insufficient to fully account for an utterance, in other words where more rule discovery is needed. One is the desired semantic difference between actual and potential movement or action. For example, sliding doors of a cupboard were often signed with simultaneous flat surfaces moving, but the AZee expression structure `simultaneous(move-object(...), move-object(...))` reads as doors actually moving whereas the form seems not to allow this interpretation, and rather to impose that the doors *can* move. We hypothesise that a rule is to be discovered, that distinguishes these interpretations. The dynamics of the body and its detachment vs. engagement in the articulation should further be investigated, and may generalise to a "potentiality" rule usable everywhere in LSF, not just for such object movement.

Another case of missing rule was to justify movements repeated at fixed locations, when they are interpreted as continuous processes. While only four cases were encountered in the corpus, to do with flowing or running water, they are enough to reveal that we could not correctly encode such production with the known rule set. This actually takes us back to an unexplored case in our prior study on repeated movements. Two rules had emerged then (`mult-in-a-row` and `mult-around`), in which the countable repetition in the form mapped to a countable set of objects distributed in space. The missing case here means something different because there is no countable object set. Exploring



Figure 4: Example of an iconic structure formalized with a complex `all-of`

this path will probably not only provide a new rule capturing the missing meaning, but also as usual lead to the discovery of some subtle dynamic form difference, useful to refine our movement specifications and produce better animation quality.

Secondly, we observed recurring patterns of descriptions, and one particularly frequent: an `all-of` root applied to a list in which:

- every item applies the `simultaneous` rule, with both of its arguments $sig1$ and $sig2$ applying `deploy-shape`;
- the same rule is used for every $class$ argument contained;
- every $sig1$/$sig2$ pair recruits opposite hands for their deployment;
- every $path$ argument in a $sig1$ is symmetric to that in its sibling $sig2$.

This complex expression pattern is used 167 times to depict things like the multiple sides of a window or a carpet (e.g. fig. 4), the general case being a symmetric shape broken down into edges forming its outline. This is entirely compatible with the semantics of the encoded expression, as `all-of` creates and focuses a set from its parts, namely in such instances an oblong shape from three simultaneous edge deployments. It also produces the correct form, as was demonstrated with the "bedroom walls" example by Filhol and McDonald (2020).

The problem we report here is therefore not a theoretical one, rather to the contrary given the productivity of the system. But such a single piece of signing easily takes up 100 lines of code, with a lot of tedious repetition for the specification of the geometric parts. This to us reveals something impractical if not equipped with at least some form of shorthand. A new function can be introduced for this case, say "`deploy-edges-symmetric-shape`", factoring the few arguments needed to produce the whole structure. This is in a way a new form–meaning binding, but unlike those added to the production set so far, it only wraps an internal complex structure; it is not necessary to encode the utterances it covers. To be faithful to the formal AZee proposition unbroken so far, we therefore propose that such function be introduced to drastically collapse the code and make it more human-readable, but not considered as a linguistic rule of the production set. We will call such functions "wrappers", and encode them separately from the production set, but allow them anywhere in discourse expressions.

Lastly, the case of moving objects has in many places been insufficiently addressed by the only candidate rule at this point: `move-object`. It was introduced with Filhol and Mcdonald (2018) for animating simple object displacements ("DSM"), but it relies on a `PATH` argument, along which a point is dynamically moved for a `POSTURE` to depend on. The problem is that this does not allow things like rotations, e.g. a standing tree falling flat, or even complex compositions of movements, e.g. a car taking a turn while skidding. We understand here that any geometric change in position or rotation can in fact be signed to mean itself, not to mention that any synchronisation between changes can too, if there are several. No linguistic rule or set of rules can therefore cover all cases of object movement. To allow full coverage, some of the geometric specification of what moves and when will ultimately have to be provided in the encoded expression. This means allowing for segments where native AZee objects (points, vectors...) and operators (translations, rotations...) are freely written instead of constraining the expression to rules of the production set. In such segments, the encoding of the dynamic points and vectors transforming over time would be delegated to the author, not the result of linguistic rules.

However, we hypothesise that a generic form–meaning is still to be defined, devoid of geometric handling (factored out), but allowing precisely to interpret that such a segment is being performed, where free-hand geometric changes are happening. Again some subtle body dynamics marking the switch to this depicting mode is probably to be discovered, along with the expected and documented eye gaze detaching from the addressee and looking at the signed scene. We also propose that this switch might be very similar to that happening at the start of role shifts, for which a similar system is needed. Indeed the content of an iconically reported action cannot be the result of a defined linguistic rule; it must be provided by the author as it is the semantic contents itself. But the very fact that we can segment role shifts to interpret them as such may come from some dynamics. This would have to come from a rule enclosing the free-hand content, which itself would have to be provided with native specifications. In our view, this would formally make role shifts and DSMs similar to a wide extent, only one would be enclosing geometric changes and the other posture changes,

perhaps even wrapped under a same rule.

## 5. Conclusion and Future Work

AZee had proved good coverage of news-type discourse, which only contained a few instances of iconic structures. This work has dealt with heavily iconic discourse, to encompass more of what SL produces when not restricted to a genre or use case. While this is often considered as a leap into a new kind of discourse, with new constructs and requiring new descriptive theories, the presented work showed that AZee on the contrary allows to do so without new concepts, and hardly even needs more rules than those already known prior to this work. While this is in itself of theoretical value, we also highlight the concrete benefit of entirely specifying output forms, which enables avatar animation of every encoded expression, including the fine dynamics and crucial eye gaze control where necessary.

In the future, we plan to expand the description in AZee of *Mocap1*. The goal would be to provide an AZee reference corpus for its depicting genre, similarly to what *40 brèves* brought for all-purpose, informative discourse. We should also address yet other discursive genres, in particular narrative discourse containing more role shifts, which were mostly lacking in *Mocap1*.

## 6. Bibliographical References


Mary Brennan. 1990. *Word formation in BSL*. Doctoral dissertation, department of linguistics, Stockholm University.

Diane Brentari and Susan Goldin-Meadow. 2017. Language emergence. *Annu. Rev. Linguist.*, 3:363–388.

Camille Challant and Michael Filhol. 2024. Extending AZee with Non-manual Gesture Rules for French Sign Language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, (LREC-COLING 2024), pages 7007–7016, Turin, Italy. European Language Resources Association.

Marie Coppola and Diane Brentari. 2014. From iconic handshapes to grammatical contrasts: longitudinal evidence from a child homesigner. *Frontiers in psychology*, 5.

Kearsy Cormier, Sandra Smith, and Zed Sevcikova Sehyr. 2015. Rethinking constructed action. *Sign Language & Linguistics*, 18(2):167–204.

Christian Cuxac. 2000. *La langue des signes française (LSF): les voies de l'iconicité*. Faits de Langue. Ophrys, Paris.

Karen Emmorey. 2003. *Perspectives on classifier constructions in sign languages*, lawrence erlbaum associates, publishers edition. Psychology Press, Mahwah, New Jersey, London.

Michael Filhol and Mohamed Nassime Hadjadj. 2016. In *Proceedings of the workshop on Representation and Processing of Sign Languages*.

Michael Filhol and John McDonald. 2018. Extending the azee-paula shortcuts to enable natural proform synthesis. In *Proceedings of the workshop on Representation and Processing of Sign Languages*, pages 45–52, Miyazaki, Japan.

Michael Filhol and John Mcdonald. 2018. Extending the azee-paula shortcuts to enable natural proform synthesis. In *Workshop on the Representation and Processing of Sign Languages*.

Michael Filhol and John McDonald. 2020. The synthesis of complex shape deployments in sign language. In *Proceedings of the workshop on Representation and Processing of Sign Languages*, Marseille, France.

Nancy Frishberg. 1975. Arbitrariness and iconicity: historical change in american sign language. *Language*, page 696–719.

Ivani Fusellier-Souza. 2004. *Sémiogenèse des langues des signes: étude de langues des signes primaires (LSP) pratiquées par des sourds brésiliens*. Ph.D. thesis, Université Paris 8.

Mohamed Hadjadj, Michael Filhol, and Annelies Braffort. 2018. Modeling french sign language: a proposal for a semantically compositional system. In *International Conference on Language Resources and Evaluation*, Miyazaki, Japan. ELRA.

Trevor Johnston. 2024. Auslan corpus annotation guidelines. *Auslan Corpus*.

Trevor Johnston and Adam Schembri. 1999. On defining lexeme in a signed language. *Sign language & linguistics*, 2(2):115–185.

Scott K. Liddell. 2003. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press.

Pamela Perniss, Roland Pfau, and Markus Steinbach. 2007. Can't you see the difference? sources of variation in sign language structure. *Visible Variation: Comparative Studies on Sign Language Structure*, 188:1.

Marie-Anne Sallandre. 2014. *Compositionnalité des unités sémantiques en langues des signes. Perspective typologique et développementale*. Thèse hdr, Université Paris 8.

Adam Schembri. 2003. Rethinking 'classifiers' in signed languages. In *Perspectives on classifier constructions in sign languages*, page 3–34, Mahwah, NJ: Lawrence Erlbaum Associates. Emmorey, K. D.

Ted Supalla. 1978. Morphology of verbs of motion and location in american sign language. *Proceedings of the Second National Symposium of Sign Language Research and Teaching*, page 27–45.

Rachel Sutton-Spence and Bencie Woll. 1999. *The linguistics of British Sign Language: an introduction*. Cambridge University Press.

Clayton Valli and Ceil Lucas. 2000. *Linguistics of American sign language: An introduction*. Gallaudet University Press.

Inge Zwitserlood. 2012. Classifiers. In *Sign language: An international handbook*, page 158–186. Mouton de Gruyter.

## 7. Language Resource References

LIMSI and CIAMS. 2020. *MOCAP1*. ISLRN 502-958-837-267-9. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

LIMSI and LISN. 2022. *40 brèves*. v2, ISLRN 988-557-796-786-3. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

# Extracting Signs from Weakly Aligned Sign Language Corpora: A Study on LSF and LSM

**Lorena de la Garza[1], Julie Halbout[2], Julie Lascar[2], Niels Martinez-Guevara[1], Arturo Curiel[3], Michèle Gouiffès[2], Annelies Braffort[2]**

[1] Centro de Investigación en Matemáticas Aplicadas, Universidad Autonoma de Coahuila, 25280, Saltillo, Mexico

[1]{lorenagarza, niels.martinez}@uadec.edu.mx.edu, [3]me@arturocuriel.com

[2] Univ. Paris-Saclay. CNRS. LISN 91400, Orsay, France

[2]{julie.halbout, julie.lascar, michele.gouiffes, annelies.braffort}@lisn.fr

**Abstract**

This paper presents a framework for the automatic annotation of sign language data across different recording conditions, including original and interpreted content. The proposed approach integrates weak alignment, sign segmentation, and multiple instance learning with a contrastive loss. The resulting annotations are subsequently refined and filtered to enhance their reliability. Our method was applied to two historically related sign languages, French Sign Language (LSF) and Mexican Sign Language (LSM). This lead to the creation of two signaries, comprising approximately 2k categories in LSF (25k occurrences) and 41 categories in LSM (1k occurrences). Both resources provide valuable support for future research in artificial intelligence and linguistics, particularly for comparative analyses between the two languages. A seminal analysis is presented as part of this paper.

**Keywords:** sign languages, automatic annotation, LSM, LSF

## 1. Introduction

Individual sign identification is a core task in sign language processing (Cooper et al., 2011; Rastgoo et al., 2021). It is traditionally divided into isolated recognition, which focuses on single signs, and continuous recognition, which addresses signs within full utterances.

As research moves toward more complex tasks such as machine translation and sign language understanding, effective continuous recognition models have become increasingly important due to the role of context (Hung, 2014). However, training for these languages remains challenging due to the visual nature of sign languages, limited annotated data, and unresolved tokenization issues.

In written languages, modern deep learning methods rely on rich semantic representations learned from large unannotated datasets (Incitti et al., 2023). Attention-based models further enable robust representations, even for Out-of-Vocabulary (OOV) words (Hu, 2020; Galassi et al., 2021; Patel and Domeniconi, 2023).

The success of multimodal models in tasks such as visual question answering and cross-modal information extraction (Wang et al., 2024; Kim et al., 2025) suggests that attention-based methods could benefit NLP sign language research, given sufficient aligned data. In this context, transfer learning offers a promising way to bootstrap aligned corpora for underserved sign languages.

This paper presents a multi-modal sign extraction and representation algorithm, based on transfer learning techniques. The method relies on existing textual and visual embedding models to identify and encode signs in unannotated parallel corpora. In this work, we propose using the subtitles as cues differently. The main idea is to associate sign clips with labels using the textual embeddings along with the contextual information they contain. It is tested in two language pairs in two different setups: French Sign Language (LSF) - French (French subtitling of LSF content) and Mexican Sign Language (LSM) - Spanish (LSM interpretation of Spanish audio). This results in collecting two new corpora that will be made publicly available. Indeed, these language pairs were selected for their typological proximity (French–Spanish and LSF–LSM), under the assumption that their similarities would be reflected in the learned sign representations. To test this assumption, the two sets of multi-modal embeddings were compared against each other to understand whether similar information arising from different Sign Languages (SLs), induced close representations when projected in the same vector space. The results show that some phonetic information is indeed preserved in the resulting embeddings, which can be seen when extracting segments with similar representations from the two corpora.

Our contribution is a segmentation-first pipeline for automatic sign extraction, and a preliminary cross-lingual exploration of LSF and LSM.

The paper is organized as follows. Section 2



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 174–183
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

presents background and related work. Section 3 introduces the dataset used for this study. Section 4 describes the suggested pipeline for sign spotting. Section 5 evaluates the spotted signs within our pipeline and compare the two SL. Finally, Sections 6 and 7 present future work and findings of the study.

## 2. Related work

We review prior work on SL annotation, focusing on the key tasks such as subtitles-to-sign alignment, sign segmentation and spotting.

**Aligning subtitles and video utterances.** Alignment aims to map written-language segments (Spanish or French) to semantically equivalent segments in continuous sign language videos. Early methods for BSL (Bull et al., 2021) were refined with negative alignment losses and self-supervised training (Jang et al., 2025a), but required annotated data. More recently, (Jang et al., 2025b) uses an LLM for alignment, translation, and sentence-boundary prediction.

**Sign segmentation** divides continuous SL videos into meaningful temporal units (Renz et al., 2021). We distinguish sign-level segmentation, which identifies lexical sign boundaries, from utterance-level segmentation such as sentence-like chunks (Jiang et al., 2025). Sign-level segmentation is roughly analogous to word boundaries in spoken language, despite the coarticulated nature of signs (Fenlon et al., 2015), and aligns closely with our goal of comparing lexical units across SLs and spoken languages. Recent approaches treat sign-level using BIO[1]-labels and language-agnostic visual features from pose and motion, being applicable to SLs without lexical annotations (Moryossef et al., 2023). Extensions combine 3D hand reconstruction (HaMeR) with Transformer-based temporal encoders to further improve boundary detection (Low et al., 2025).

**Annotating Sign Language Datasets.** Annotating a SL video dataset consists in labeling frames or video portions, for instance with semantic (such as word, gloss), linguistic or phonologic/formal data. Automatically annotating lexical signs in a SL dataset generally relies on sign spotting. One common method to spot signs relies on a signary composed of isolated signs that are used as queries that are searched by similarity in continuous utterances (Momeni et al., 2022). Additional cues, such as mouthing, help enhance sign spotting accuracy (Varol et al., 2022). In cases where no sign lexicon is available, aligned subtitles can provide weak supervision by indicating which video segments are likely to contain specific signs (Lascar et al., 2024; Momeni et al., 2022). Once a sign has been spotted, its precise temporal window can be obtained using similarity pairing (Lascar et al., 2024).

**Contrastive multimodal learning** has emerged as a dominant paradigm for learning joint representations across modalities and embedding features into a shared vector space (Radford et al., 2021). Recent work has applied contrastive learning to SL tasks under various supervision regimes. SignCLIP aligns dictionary labels with SL videos using large-scale multilingual dictionaries (Jiang et al., 2024), while CVT-SLR (Zheng et al., 2023) and SignCL (Ye et al., 2024) use contrastive signals within pipelines for continuous recognition and gloss-free translation. Among these, CiCo is the most relevant to our work, addressing text–video alignment in continuous signing and learning fine-grained sign-to-word correspondences under weak supervision using MIL-like aggregation (Cheng et al., 2023).

**MIL-NCE** combines Multiple Instance Learning with a contrastive loss with the objective of learning video–language representations from weakly aligned videos. Each text segment is paired with a bag of overlapping visual clips, and softmax-based aggregation highlights the most relevant while treating others as implicit negatives (Miech et al., 2020). MIL-style contrastive methods have been applied to SL under two regimes: utterance-level, aggregating evidence from coarse textual units (Duarte et al., 2022; Cheng et al., 2023), and sign-level, for spotting or dictionary-based retrieval using known sign exemplars (Momeni et al., 2021; Varol et al., 2022). Our approach differs by deriving token-level correspondences from spoken-language subtitles without assuming sign labels and by using a language-agnostic segmentation stage (Moryossef et al., 2023) to reduce temporal redundancy while retaining MIL-style evidence selection.

**Positioning of our approach.** In the context of interpreted SL, subtitle timestamps provide only coarse temporal anchors, and word-to-sign correspondences are often noisy due to interpretation effects such as lag and reformulation. As a result, methods relying on reliable positive pairs or explicit sign identities are not directly applicable, while dense sliding-window approaches generate large and costly candidate sets (Cheng et al., 2023). Although approaches such as the Subtitle Aligned Transformer (Bull et al., 2021) use subtitles for sign-level segmentation, they require domain-specific annotated boundaries, which are often unavailable. We therefore treat subtitles as weak cues rather than precise supervision. Our pipeline first estimates temporal offsets and then constructs candidate units using language-agnostic sign segmentation (Moryossef et al., 2023). This design is related to segmentation-first frameworks such as SEA (Jiang et al., 2025), which decouple segmentation from alignment, though we leave a systematic eval-

[1] Beginning–Inside–Outside

uation of such methods to future work.

## 3. Dataset

For a transversal study on LSM and LSF, the two selected datasets are continuous SL coming from media.

**Mediapi-RGB** (Ouakrim et al., 2024) is a highly qualitative LSF dataset based on broadcasts produced by Deaf journalists[2] and subtitled in French. This dataset consists of 86 hours of LSF programs, with aligned subtitles (obtained in post-production) which cover the period from 2019 to 2023. A bilingual lexicon Mediapi-signary was set up with minimal supervision, and manually checked by an expert (Lascar et al., 2024). Containing 445 classes with 15k occurrences of signs, it constitutes the ground truth for annotation evaluation.

**TV-UNAM** (TV UNAM, 2026) is a dataset of 220 hours of live Spanish-to-LSM interpreted broadcasts produced by the Universidad Nacional Autónoma de México (UNAM). The interpretation is performed simultaneously during the original programs. We cropped the videos to retain only the area where the signer is seen and processed the audio with WhisperX (Bain et al., 2023) to obtain automatic transcripts.

## 4. Method

### 4.1. Weak alignment for the LSM corpus

In the LSF dataset, subtitles are aligned to the videos. In the LSM corpus there is a temporal offset between audio and SL production, due to the interpretation context. A weak alignment algorithm is applied. The goal is not precise word–sign alignment, but to estimate a smooth temporal correction mapping audio timestamps to a video-consistent timeline, forming the basis for subsequent pipeline stages.

*1. Signal Proxy Construction.* To compare text and video modalities, both are converted into one-dimensional event-density signals as structural proxies. For audio, WhisperX timestamps (Bain et al., 2023) are used to build an impulse signal weighted by structural level: words (1.0), subtitle segments (2.0), sentence boundaries (4.0), and speaker changes (8.0), reflecting peaks in visual novelty due to interpreter role-shifting (Quer, 2018). The signal is smoothed with a Gaussian filter. For video, features are extracted using a SlowFast network (Feichtenhofer et al., 2019). Each visual embedding summarizes approximately 1 s of video content and is computed at a temporal stride of about 0.13 s using a overlapping sliding window. Visual novelty is then computed as $n_t = 1 - \text{cos_sim}(\mathbf{v}_t, \mathbf{v}_{t-1})$, where $\mathbf{v}_t$ and $\mathbf{v}_{t-1}$ denote consecutive visual embeddings and $\text{cos_sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$, yielding a one-dimensional signal comparable to the audio proxy.

*2. Offset Estimation via Windowed Cross-Correlation.* Temporal offsets are estimated locally using 30s sliding windows with 15s hops. Within each window, cross-correlation between normalized audio and video signals (Z-score normalization) is computed for non-negative lags up to 5s, reflecting typical interpretation delays (Jiang et al., 2025).

*3. Lag Curve Refinement and Outlier Rejection* Window-based lag estimates may contain outliers due to transcription noise. To stabilize the sequence, a median filter of width 5 windows is applied, suppressing isolated peaks while preserving the global temporal trend.

*4. Time Warping.* Finally, aligned subtitle files (JSON/SRT) are produced using the temporal mapping function $\tau(t) = t + \Delta(t)$. Since $\Delta(t)$ is estimated per window, linear interpolation provides a continuous correction for each timestamp, yielding a weakly aligned corpus for subsequent learning stages.

The followings sections explain our pipeline for the sign spotting. First is the sign segmentation 4.2. Then section 4.3 describes the generation of our data embeddings. The third section introduces our model 4.4 and the last section describes the refinement method of the model prediction 4.5. The pipeline is displayed in the Fig.1.

### 4.2. Pose-based sign segmentation

In this stage, the SL video utterances are temporally segmented into video clips of coherent motion. These clips are approximations for computational purposes, as SL segmentation is inherently ambiguous and variable, even among human annotators (Moryossef et al., 2023).

For this stage, temporal sequences of body poses are extracted using MediaPipe Holistic (Lugaresi et al., 2019), generating `.pose` sequences that are processed using the method proposed by Moryossef et al. (2023). The authors formulate the segmentation problem as a frame-level sequential labeling problem. Thus, the model assigns each frame a label following the *Beginning–Inside–Outside* (BIO) scheme. From this sequence of labels, temporal intervals delimiting segments of continuous body movement are obtained. Both sign and sentence-levels segments are provided. Nonetheless, our work focuses on the sign-level segments.

[2]Médiapi journal https://www.media-pi.fr/

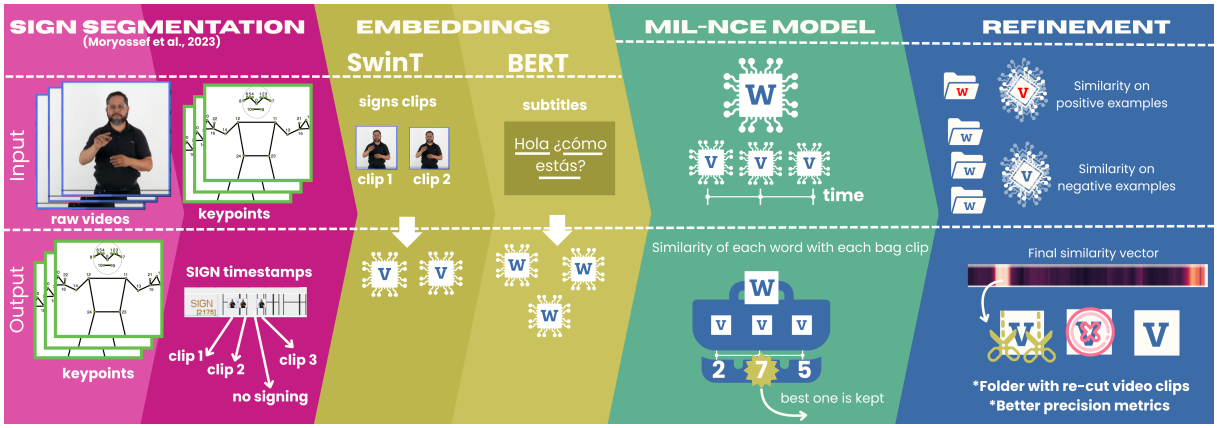



Figure 1: Illustration of the pipeline. Once video and text are *weakly aligned*, we proceeded to the SEGMENTATION. Once videos clips of supposed signs are obtained, we can generate the video EMBEDDINGS for each clip (using Video Swin Transformer trained on BSL) and text EMBEDDINGS (with corresponding LLM for each vocal language). The data pairs are then used for the training of the MIL-NCE MODEL. The predictions are then REFINED to obtain more accurate annotation.

### 4.3. Visual and textual embeddings

**Visual embeddings.** Once temporal video clips are obtained, each of them is decomposed into 16-frames blocks (with one frame overlap). Clips shorter than 16 are discarded to maintain consistency in the model input. The selected frames are resized to a fixed resolution of $224 \times 224$ pixels. The resulting sequence is processed using a Swin Transformer 3D model previously trained on a classification task using British Sign Language (BSL) videos from the BOBSL corpus (Prajwal et al., 2022). For our purposes, the final classification layer is removed, and the model is used solely as a high-level representation (*i.e.* embedding) extractor. Each clip is thus represented by a numerical vector summarizing its visual dynamics, which is stored for use in subsequent stages of the pipeline.

**Textual embeddings.** WhisperX transcriptions (LSM) and subtitles (LSF) are processed too to obtain vector representations at the word level. In particular, an embedding is generated for each occurrence of each word within its original transcription context. For this purpose, each subtitle segment is processed in its entirety using transformer-based language models that produce contextualized text representations. In this way, the representation of each word depends on the surrounding words within the segment, allowing different usages of the same word to be distinguished across contexts. For LSF subtitles, the CamemBERT model (Martin et al., 2020) is used, which has been previously trained on large French text corpora. For LSM transcriptions, a substantial corpus of transcriptions is available. Taking advantage of this data availability, a multilingual model based on XLM-RoBERTa (Conneau et al., 2020) is further fine-tuned using these transcriptions.

### 4.4. MIL-NCE model

Using the constructed pairs, a model is trained with the objective of representing both modalities within a shared vector space, such that corresponding text and video instances exhibit high similarity, whereas incorrect associations remain separated.

**Architecture.** The model is composed of two streams, one for the text (i.e. a sentence-like utterance) and one for the video clips. Each modality is passed through its own linear and normalization layers, and similarities between text embeddings and associated video segments are computed using dot products. Thus, the model learns lightweight projection layers that map both representations into a lower-dimensional shared space while preserving knowledge previously acquired by the base models. Since alignment between words and signs is not exact, a single interval may contain multiple visual segments. Only one, some or none of which may correspond to the target word, while others may represent different words or non-lexical movements. For this reason, training is formulated under a *Multiple Instance Learning* (MIL) framework, where each word is associated with a collection of visual candidates and the model learns to assign higher compatibility to segments that are consistently aligned across many examples. Textual and visual embeddings are projected and normalized into a shared space for cosine similarity. When multiple visual candidates are available, their scores are aggregated using log-sum-exp, emphasizing the most relevant segments.

**Training with contrastive InfoNCE loss.** Training uses a contrastive InfoNCE loss operating on batches of examples processed simultaneously. Within each batch, the model is optimized so that each textual embedding instance is more similar

to its corresponding visual set than to segments belonging to other words in the same batch, thus implicitly generating negative examples through mixing of non-corresponding pairs. The models results on labeling each sign segment with a label from the subtitle.

### 4.5. Annotation Refinement

To remove mislabeled sign clips and obtain finer segmentation of correctly labeled clips, we applied the method proposed by (Lascar et al., 2024; Momeni et al., 2022). It assumes that a common lexical unit is present across video clips that share the same label. Thus, to identify the clips that indeed correspond to a label, pairwise similarity matrices are computed between all video clips associated with a given label (resulting to a $L^+$ vector). To avoid catching signs such as pointing or other similar signs that could appear in the context of our target sign, multiple negative examples are picked (resulting to a $L-$ vector). The final similarity score (vector $L = L^+ - L^-$) are then analyzed at the frame level using two criteria: a similarity threshold and a frequency threshold, defined as the number of frames $n_f$ the similarity exceeds the similarity threshold. Based on these criteria, the method identifies the videos in which the lexical unit is actually present, refines its temporal segmentation, and filters out videos where the unit does not occur.

## 5. Experiments

### 5.1. Training data

We constructed pairs linking contextualized word embeddings with embeddings of candidate visual segments, using pre-extracted representations for both modalities. Different strategies were applied for LSF and LSM: LSF relies on manually aligned subtitles, whereas LSM uses automatic WhisperX transcriptions. For LSF, word intervals are taken directly from subtitle timestamps, and overlapping visual segments are selected. For LSM, timestamps are first refined via weak alignment, and each word is assigned a 2-s window to account for interpretation shifts, after which intersecting visual segments are retrieved. In both settings, each word is paired with one or more visual candidates to train two annotation models.

### 5.2. Implementation details

Regarding MIL, pairwise similarities are aggregated using log-sum-exp pooling and the InfoNCE temperature is fixed to 0.07. Training is run with AdamW using a learning rate of $10^{-4}$, weight decay of $10^{-2}$, gradient clipping at 1.0, and automatic mixed precision when available. We use a micro-batch size of 32 with gradient accumulation of 8 steps, resulting in an effective batch size of 256, which increases the number of in-batch negatives. To improve clip selection within each bag, we anneal the MIL pooling temperature $\tau$ during training (linearly from 0.10 to 0.05 over the first 6k optimizer steps), progressively making the MIL pooling more "max-like" and thus more selective over candidate segments. Additional filtering is applied during data construction and training. For LSM, segments near video boundaries are removed to avoid non-signing content. Proper names are excluded in both corpora, and stop words are filtered out. Training is monitored using a sanity check that compares real batches with shuffled baselines. We report retrieval metrics (top-1/top-5 accuracy), diagonal–off-diagonal similarity gaps, and a hubness indicator. We also track a clip-confidence score, measuring how tilted the within-bag selection becomes as $\tau$ decreases. We additionally track a "clip-confidence" measure, defined as the average maximum softmax probability over candidate segments within the positive bag, to quantify how inclined (clip-seeking) the within-bag selection becomes as $\tau$ decreases. Model selection retains the checkpoint maximizing a composite score combining improvement over the shuffled baseline and clip confidence, with early stopping on stagnation.

### 5.3. Metrics

The model assigns a similarity score, noted $\tau_s$, to each word-clip pair, computed as cosine similarity between the projected contextualized word embedding and the projected video-clip embedding. For evaluation, only the highest-scoring video clip candidate (Top-1) is kept for each word instance. Since a ground truth is available for LSF, we calculate the overlap between each annotation and the corresponding ground truth example. First, the Intersection over Union (IoU) score is computed and the annotation is considered as correct when it exceeds 0.1. Similarly, we used the F-measure and Precision to analyze the model's performance.

### 5.4. Results on LSF

**MIL-NCE outputs.** By applying the MIL-NCE model to the LSF preprocessed data, the 241,430 video clips extracted are annotated across a vocabulary of 22k terms. The 5 most frequent tokens are *été* (summer or been) (2,543), *personnes* (persons) (1,372), *ans* (years) (1,199), *dernier* (last) (1,059) and *sourds* (deaf) (989). The Top-1 similarity scores range from -0.22 to 0.78, with a peak around 0.3.

**MIL-NCE outputs regarding the ground truth (GT).** Mediapi-signary is an expert-validated sign

bank covering 445 lexical categories. The evaluation focuses on the predictions whose labels are covered by this vocabulary. We then report results on 2K automatically evaluated predictions ($\approx 1\%$ of the total outputs) for which a corresponding GT interval is available. It yields an average precision of $P = 0.53$ (with the criterion $\mathrm{IoU} \geq 0.1$), and $P = 0.4$ with $\mathrm{IoU} \geq 0.4$.

**MIL-NCE outputs manually evaluated.** Furthermore, given the small ground truth sample, we also conducted a manual evaluation on a small sample. 11 classes (9 classes$^{signary}$ common to Mediapi-signary, one verb$^{*}$ and two random noun$^{r}$) with varying numbers of occurrences (ranging from 805 to 2) were manually verified, leading to an average precision of 0.52, as shown in Table 2. Keeping the annotations with a minimum similarity value $\tau_s = 0.4$ yields a precision close to 0.8, which corresponds to 59k remaining annotations within an 8k vocabulary size.

Then, the **refinement** procedure (see Section 4.5) is applied, which leads to a higher precision (0.99) and more precise temporal segmentation, as shown in Table 1 for two different $n_f$ values (corresponds to the number of consecutive frames where the similarity have to exceed the threshold to be valid). With $n_f = 4$, we obtain **25,291** occurrences spread in **2,718** classes. It is 6 times mores sign occurrences than Mediapi-signary. These findings should be interpreted with caution, as similar videos may have been mislabeled. A more comprehensive verification process is required to guarantee reliability.

| $n_f$ | C | N | TP | P | R | F1 |
|---|---|---|---|---|---|---|
| Baseline | 672 | – | 480 | 0.81 | – | – |
| 4 | 672 | 287 | 284 | 0.99 | 0.52 | 0.68 |
| 6 | 674 | 247 | 245 | 0.99 | 0.45 | 0.62 |

Table 1: Refinement of the annotation in LSF with $\tau_s = 0.4$. $n_f$ is explained in section 5.3. **C** refers to the number of potential signs to refine.

## 5.5. Results on LSM

**MIL-NCE outputs.** For LSM, the dataset contains 869,629 word instances for which the model produced predictions. The number of unique words (case-insensitive) is 46,844. The most frequent tokens are primarily high-frequency discourse markers and function words as *también* (11,100 occurrences), *más* (9,487) and *muy* (8,270). Institutional and proper names such as *UNAM* (4,527) and *México* (3,347) are also frequent due to the nature of the corpus. The resulting Top-1 similarity scores range from -0.093 to 0.165 (far less than for LSF).

**MIL-NCE outputs manually evaluated.** As in LSF, a small manual evaluation was conducted on a few signs. 4 classes with around 100 occurrences were checked. Performance scores were lower, predictions were less accurate, and overall precision (P) decreased to 0.1 or under.

| label | Count | TP | P |
|---|---|---|---|
| PAYS$^{signary}$ | 805 | 317 | 0.39 |
| VENDREDI$^{signary}$ | 439 | 297 | 0.68 |
| CORONAVIRUS$^{signary}$ | 307 | 186 | 0.61 |
| ENTREPRISE$^{signary}$ | 184 | 104 | 0.57 |
| RÈGLES$^{signary}$ | 94 | 48 | 0.51 |
| CRÉER$^{*}$ | 89 | 40 | 0.46 |
| VÉLO$^{r}$ | 39 | 19 | 0.49 |
| AVORTEMENT$^{r}$ | 8 | 1 | 0.13 |
| COVID$^{signary}$ | 5 | 2 | 0.4 |
| RWANDA$^{signary}$ | 4 | 0 | 0 |
| MALIEN$^{signary}$ | 3 | 0 | 0 |
| IRLANDE$^{signary}$ | 2 | 1 | 0.5 |
| total | 1,977 | 1,015 | 0.51 |
| VACUNA | 129 | 11 | 0.09 |
| DISCAPACIDAD | 112 | 6 | 0.05 |
| OLÍMPICOS | 107 | 9 | 0.08 |
| FRANCIA | 100 | 11 | 0.11 |
| total | 448 | 37 | 0.08 |

Table 2: Resume of LSF and LSM manual evaluation, with the count of occurrences output of the model (Count), the true positives (TP) and the precision score (P)

**Refinement.** Because similarity scores in LSM are lower than in LSF, we adopt a more permissive pre-filtering threshold $\tau_s$= 0.06. For a subset of frequent tokens (at least 100 occurrences) that also appear in the LSF ground-truth vocabulary, we then apply the refinement procedure. The resulting annotations are then manually reviewed, resulting in a final set of 1,311 instances covering 41 distinct lexical items.

## 5.6. LSF and LSM comparison

To analyze the cross-lingual structure of the video embeddings, we visualize word-level centroids using UMAP[3]. Since each lexical item may occur multiple times, we compute a centroid for each label to reduce instance-level noise and highlight stable semantic structure. Figure 2 shows the resulting UMAP projection for LSM and LSF, with lines linking equivalent labels across languages, allowing to qualitatively inspect cross-lingual relationships between LSF and LSM signs. Because UMAP is a nonlinear dimensionality-reduction technique that can distort distances, we additionally computed PCA[4] projections to verify whether apparent similarities were preserved for instance, the embeddings

[3]Uniform Manifold Approximation and Projection for Dimension Reduction

[4]Principal Component Analysis




Figure 2: UMAP visualization of word-level embedding centroids for LSM and LSF. Lines connect equivalent labels across languages.

corresponding to the word *music* appear relatively distant in the UMAP space, yet PCA indicates a higher similarity, suggesting that the UMAP layout may exaggerate certain separations. It is also noteworthy that semantically related items tend to cluster in similar regions of the map. For example, several temporal concepts such as days appear grouped in the upper-left area. This is particularly interesting given that the embeddings are derived solely from visual sign-language data, without any semantic supervision. Building on these observations, we explore the possibility that some of the observed cross-lingual similarities reflect the historical influence of LSF in LSM.

In this visualization, expressions like weekdays appear to show strong proximity. This aligns with the hypothesis that stable lexical items, which are used frequently, are more likely to preserve similarities. However, this pattern does not extend to months that exhibit noticeably lower similarity.

We further hypothesized that other lexical classes might display different behaviors. Neologisms (e.g., *COVID*, not shown in the current map) were expected to exhibit increased variability due to their recent emergence. However, *COVID* doesn't, possibly reflecting rapid worldwide diffusion. Similarly, socially salient terms were hypothesized to show greater divergence. While the sign for *handicap* does, comparable variability is also observed in socially neutral categories such as months.

Taken together, these findings suggest that while we can begin to explore these hypothesis, a larger lexical inventory, along with diachronic data would be required to disentangle historical inheritance.

## 6. Future work

Conducting systematic ablation studies on weak alignment and sign segmentation would help to isolate the contribution of some component of the pipeline. Concerning the alignment, we plan to do a comparative study contrasting the actual method performance with state-of-the-art subtitle-level alignment methods such as SEA (Jiang et al., 2025) , in order to assess trade-offs between accuracy and efficiency. On the sign segmentation side, other methods (such as Varol et al. (2021) or HaMeR) are to be investigated. Indeed, we observed on LSM a oversegmentation of the sign for the creation of the sign clip. On the other hand, evaluating the proposed method on datasets with denser and more reliable annotations, such as BOBSL, would provide a stronger benchmark. Such datasets would allow us to better quantify alignment quality under controlled conditions and to identify failure modes that may be less visible in sparsely annotated data.

## 7. Conclusion

This paper has presented a method for extracting signs in existing media videos, both in LSF and LSM. After a weak alignment between texts and videos for LSM, segmented signs are automatically detected using contrastive learning framework. Since the resulting video clips are under-segmented, the segmentation is further refined using similarity scores between occurrences in a similar category. Around 41 LSM sign categories are collected, and 2k categories in LSF after refinement. The first evaluation of these corpora suggest promising results. The resulting signaries represent precious resources to explore similarities between the two SL, and will allow improving research in SL processing. The results are stronger on the LSF corpus, which can likely be attributed to the fact that the LSF content is original and the subtitles are professionally produced in studio settings. Conse-

quently, the subtitles tend to align more closely with the intended meaning conveyed in LSF. In contrast, interpreted LSM data introduce temporal misalignments and occasional omissions, thereby increasing the divergence between the two modalities and negatively affecting performance.

## 8. Bibliographical References


Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4489–4493.

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. 2021. Aligning subtitles in sign language videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11532–11541.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19016–19026.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Helen Cooper, Brian Holt, and Richard Bowden. 2011. Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. Springer.

Amanda Duarte, Samuel Albanie, Xavier Giró-i Nieto, and Gül Varol. 2022. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14074–14084.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210.

Jordan Fenlon, Kearsy Cormier, and Adam Schembri. 2015. Building bsl signbank: The lemma dilemma revisited. *International Journal of Lexicography*, 28(2):169–206.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2021. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308.

Dichao Hu. 2020. An Introductory Survey on Attention Mechanisms in NLP Problems. In *Intelligent Systems and Applications*, pages 432–448, Cham. Springer International Publishing.

Victor Hung. 2014. Context and NLP. In Patrick Brézillon and Avelino J. Gonzalez, editors, *Context in Computing: A Cross-Disciplinary Approach for Modeling the Real World*, pages 143–154. Springer, New York, NY.

Francesca Incitti, Federico Urli, and Lauro Snidaro. 2023. Beyond word embeddings: A survey. *Information Fusion*, 89:418–436.

Youngjoon Jang, Jeongsoo Choi, Junseok Ahn, and Joon Son Chung. 2025a. Deep understanding of sign language for sign to subtitle alignment. ArXiv preprint.

Youngjoon Jang, Liliane Momeni, Zifan Jiang, Joon Son Chung, Gül Varol, and Andrew Zisserman. 2025b. Lost in translation, found in embeddings: Sign language translation and alignment. ArXiv preprint.

Zifan Jiang, Youngjoon Jang, Liliane Momeni, Gül Varol, Sarah Ebling, and Andrew Zisserman. 2025. Segment, embed, and align: A universal recipe for aligning subtitles to signing. ArXiv preprint.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. Signclip: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193. Association for Computational Linguistics.

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges. *ACM Comput. Surv.*, 57(10):249:1–249:35.

Julie Lascar, Michèle Gouiffès, Annelies Braffort, and Claire Danet. 2024. Annotation of LSF subtitled videos without a pre-existing dictionary. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 204–212, Torino, Italy. ELRA and ICCL.

Low Jian He Low, Harry Walsh, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Hands-on: Segmenting individual signs from continuous sequences. In *Proceedings of the IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG 2025)*, pages 1–5.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. ArXiv preprint.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886.

Liliane Momeni, Hannah Bull, K. R. Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *Computer Vision – ECCV 2022 Workshops*, volume 13695 of *Lecture Notes in Computer Science*, pages 671–690. Springer Nature Switzerland.

Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Watch, read and lookup: Learning to spot signs from multiple supervisors. In *Computer Vision – ACCV 2020: Asian Conference on Computer Vision, Revised Selected Papers, Part VI*, volume 12529 of *Lecture Notes in Computer Science*, pages 291–308. Springer.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724.

Raj Patel and Carlotta Domeniconi. 2023. Enhancing Out-of-Vocabulary Estimation with Subword Attention. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3592–3601, Toronto, Canada. Association for Computational Linguistics.

K. R. Prajwal, Hannah Bull, Liliane Momeni, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Weakly-supervised fingerspelling recognition in british sign language videos. In *Proceedings of the British Machine Vision Conference*, London, United Kingdom.

Josep Quer. 2018. On categorizing types of role shift in sign languages. *Theoretical Linguistics*, 44(3-4):185–196.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign Language Recognition: A Deep Survey. *Expert Systems with Applications*, 164:113794.

Katrin Renz, Nicolaj C. Stache, Samuel Albanie, and Gül Varol. 2021. Sign language segmentation with temporal convolutional networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2022. Scaling up sign spotting through sign language dictionaries. *IJCV*, 130:1416–1439.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16852–16861.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2024. Cross-modal retrieval: A systematic review of methods and future directions. *Proceedings of the IEEE*, 112(11):1716–1754.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving gloss-free sign language translation by reducing representation density. In *Advances in Neural Information Processing Systems 37*, pages 107379–107402.

Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, and Stan Z. Li. 2023. CVT-SLR: Contrastive Visual-Textual Transformation for Sign Language Recognition

With Variational Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23141–23150.

## 9. Language Resource References

Yanis Ouakrim, Hannah Bull, Michèle Gouiffès, Denis Beautemps, Thomas Hueber, and Annelies Braffort. 2024. *Mediapi-RGB: An Extensive Video–Text Corpus for French Sign Language (LSF)*. Language resource available via HAL open archive.

TV UNAM. 2026. *La UNAM responde*. Video series with LSM interpretation.

# An Annotation Formalism for a French–LSF Bilingual Corpus Supporting Sign Language Generation

**Sylvie Gibet, Clément Reverdy, Pierre-François Marteau**
IRISA, Université Bretagne Sud
Vannes, France
{sylvie.gibet,clement.reverdy,pierre-francois.marteau}@irisa.fr

**Abstract**

This paper introduces an annotation formalism for bilingual corpora of written French and French Sign Language (LSF), based on a manually-produced, expert transcription of LSF video data. The formalism captures the grammatical specificities of LSF, including spatial and iconic mechanisms, while explicitly encoding features that support motor programs for animated signing avatars. We propose a parameterized gloss-based approach, called *PGloss-LSF*, which integrates syntactic and semantic structures alongside motion features critical for accurate sign synthesis. We illustrate the framework with examples drawn from our bilingual corpus. The annotation process is incremental, ensuring internal consistency and computational tractability through a two-step evaluation: a qualitative assessment aligning generated signs with the annotation language, and a quantitative evaluation via automatic translation using large language models. By bridging the linguistic specificities of sign language with the computational requirements of sign synthesis, this work advances the integration of sign language corpora into multilingual resources and contributes to the standardization of sign language technologies.

**Keywords:** sign language, annotation, formalism, bilingual corpora, generation

## 1. Introduction

Sign languages are fully-fledged visual-gestural languages with their own grammar and complex linguistic organization. Unlike spoken languages, they rely on spatial structuring, iconic mechanisms, and the simultaneity of manual and non-manual articulators. Each country-specific sign language constitutes an autonomous linguistic system.

In the case of French Sign Language (LSF), most educational, cultural, and administrative information remains primarily available in written or spoken French, limiting accessibility for many deaf signers. Developing computational tools for translation between written French and LSF therefore represents both a societal need and a scientific challenge. A central difficulty lies in representing LSF utterances in a way that preserves their grammatical structure while remaining operational for annotation and computational processing.

This paper focuses on the translation of written French into LSF and introduces a parameterized gloss-based annotation language, hereafter referred to as *PGloss-LSF*, designed as an intermediate representation between French sentences and LSF utterances. Standard gloss sequences are insufficient to capture key properties of LSF, such as simultaneity, spatial anchoring, and the interaction between manual and non-manual components. We therefore extend gloss notation with explicit parameters encoding spatial organization, co-occurrence of both hands, and asynchronous non-manual features (e.g., facial expressions and torso movements). The proposed formalism is grounded in linguistic descriptions of LSF grammar (Cuxac, 2000; Millet, 2019) and is designed to support annotation practices similar to those used in large sign language corpora. Furthermore, while several non-linear or highly expressive formalisms have been proposed, they often remain difficult to apply efficiently in large-scale manual annotation.

Our objective is to propose a structured intermediate representation that facilitates efficient corpus annotation suitable for LLM-based processing, while providing a sufficiently direct and controllable interface for motor-level specification, as required for avatar animation. By balancing linguistic precision and computational tractability, the proposed specification aims to bridge the gap between linear gloss representations and highly detailed scripting systems oriented toward motion generation. The remainder of this paper is organized as follows. We first review related work, then present the annotation formalism illustrated with examples drawn from our corpus, and finally report our initial evaluation results.

## 2. Related Work

The representation of sign languages has been extensively studied from a linguistic perspective. Stokoe's foundational work (Stokoe et al., 1976) introduced a phonological decomposition of signs into minimal parameters, later extended by systems such as *HamNoSys* (Prillwitz, 1989) and SignWriting. Multichannel models, including the



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 184–192
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

PDTS framework (Johnson and Liddell, 2011) and tier-based annotation schemes (Johnston and de Beuzeville, 2009), emphasized the coordination of manual and non-manual articulators. In corpus practice, ELAN has become the standard environment for multi-tier annotation, enabling the alignment of glosses, facial expressions, gaze, and body movements. However, these approaches primarily provide descriptive transcription frameworks rather than structured intermediary representations designed for computational reuse.

From a computational standpoint, scripting and animation languages such as *QualGest* (Lebourque et al., 2001), Partition/Constitution (P/C) (Huenerfauth, 2006), *SiGML* (Elliott et al., 2008), and *AZee* (Filhol et al., 2017) provide highly expressive multichannel specifications of sign articulation. While these systems enable precise modeling of signs and utterances, they require detailed manual encoding and are not well suited to large-scale corpus annotation.

Major corpus efforts, particularly for German Sign Language (DGS), have introduced more efficient annotation schemes (Hanke et al., 2020, 2023). The DGS Corpus annotation framework is primarily designed to support the development of a corpus-based lexical database. It focuses on lexical signs and integrates segmentation cues, lemmatisation rules, and specific annotation features such as double glossing and double-token tags. It also accounts for mouthing, including its coarticulation with spoken-language elements, and captures multi-channel constructions enabling the representation of context-sensitive meanings. Recent work such as Easier Notation (Hanke et al., 2023) further extends gloss-based representations within a linear, human-readable paradigm and provides an editing interface.

Other semi-automatic annotation approaches combine computer vision with linguistic constraints (Momeni et al., 2022; Neidle et al., 2018; Metaxas et al., 2018). More recently, neural end-to-end approaches to sign language translation and production (Kahlon and Singh, 2021; Farooq et al., 2021; Núñez Marcos et al., 2023) have enabled automatic gloss prediction and avatar pose generation (Shalev-Arkushin et al., 2023; Yu et al., 2024; Baltatzis et al., 2024). However, these approaches typically rely on linear gloss sequences or holistic pose representations, which under-specify the structured simultaneity and grammatical parameterization central to sign languages.

In this context, our work proposes a parameterized gloss-based annotation language designed as an operational intermediary layer: it preserves linguistic structure, supports annotation practices, and facilitates both corpus enrichment and downstream generation tasks. While our approach shares similar goals with DGS-based notation approaches, the two can be seen as complementary: the latter focus on scripting and editing based on lexical data, whereas our approach emphasizes syntactic and semantic modeling, as well as scalability for LLM-driven text-to-sign translation. In addition, utterances glossed with *PGloss-LSF* are systematically aligned with produced sign language (video or avatar), and the parameterization explicitly incorporates gestural targets directly usable for synthesis.

Our long-term objective is to develop an end-to-end text-to-sign translation system that maps written input to motor-level representations for sign language generation, while ensuring intelligibility through consistent and linguistically grounded annotation.

## 3. Annotation Formalism

### Motivation and Key Features

The main objectives of our annotation formalism are: (i) to enable in-depth analysis of linguistic processes in French and LSF; (ii) to provide an intermediate language oriented toward motor program specification, thereby facilitating the control of signing avatars; (iii) to enable the text-to-sign translation using large language models by following the approach described in (Reverdy et al., 2025). Our formalism serves as an interface between written French and LSF through a parameterized gloss-based annotation language, called *PGloss-LSF*, which aligns French sentences with actual LSF utterances (video or 3D avatar). It preserves key LSF-specific mechanisms, in particular spatialization and iconicity, by supporting the co-occurrence of manual and non-manual components.

In contrast to highly expressive non-linear representation systems, our approach prioritizes scalability and usability for corpus annotation, while preserving key linguistic structures through explicit parameterization. This makes it possible to construct large annotated datasets while maintaining a representation that remains directly exploitable for motor-level specification and avatar control.

Our annotation approach follows an incremental and controlled design. We begin with simple glosses without arguments and progressively introduce additional parameters (arguments with associated values) to capture specific grammatical targets and increasing sentence complexity, while ensuring that all information required for synthesis is specified. In an initial phase, we annotated utterances containing signs in their citation form, followed by structures for which a syntactic formalization could be explicitly defined, such as directional verbs. Since then, we have extended the

annotation framework to encompass more complex syntactic and semantic phenomena, including pointing gestures, different forms of classifiers (called *proforms*), co-occurring signs enabling the simultaneous use of the left and right hands, non-manual features (e.g., facial expressions), and role shift expressed through torso movements in dialogic contexts.

At each stage of corpus construction and annotation, a two-step validation protocol has been systematically implemented. First, annotations are reviewed by Deaf linguists or experts with advanced proficiency in LSF to ensure linguistic adequacy and theoretical consistency. Second, a computational validation phase is conducted using LLM-based corpus enrichment methods (Reverdy et al., 2025), enabling cross-verification, detection of inconsistencies, and iterative refinement of the formal representation.

### French Word / LSF Gloss Dictionary

The dictionary lists all lexical items in the corpus in the format `<French-word, PGloss-LSF>`. Each gloss is defined by its name and a set of argument–value pairs, represented as:

$$[\texttt{GLOSS-NAME_arg}_1\texttt{:VAL}_1\texttt{:_arg}_n\texttt{:VAL}_n]$$

### Sentence Structure

A sentence is represented as a sequence of glosses, each accompanied by arguments specifying a lexical, syntactic or semantic function: for example the source, destination, and object for an indicating or trajectory sign, the spatial location, the used hand, the facial expression, the torso orientation for role shifts, the size and shape specifier, the classifiers, etc. Each argument is encoded by an identifiers such as `_src`, `_tgt`, `_mpr`, `_nsn`, `_id`, `_loc`, `_hd`, `_trs`, `_pnt`, `_sass`, `_exp`. These (argument, value) pairs enable discourse to be articulated around spatial references (*loci*), geometric processes (trajectory verbs, specific handshapes), and processes that enable simultaneity between several signs (using two hands, facial expression) or narrative processes (torso orientation).

The canonical syntactic structure in LSF typically follows a default-order sequence, encompassing time, location, object, agent, and action. Locative information is usually introduced with a general spatial description and gradually narrowed to a specific locus in signing space.

**French:** *Hier, il neigeait dans la forêt.*
**English:** *Yesterday, it was snowing in the forest.*
**LSF:** `[HIER] [FORÊT] [NEIGER]`
**SL:** `[YESTERDAY] [FOREST] [SNOW]`

In the previous example, the sequence consists of signs in their citation form, each associated with elementary motor programs that are sequentially activated. In the following, examples are presented in English, with corresponding LSF glosses translated into English for clarity.

A key feature of LSF syntax is the semantic distinction between animate (people, animals) and inanimate (objects, places) entities. Gloss ordering reflects this distinction and determines the verb's position relative to its arguments: for inanimate objects, `[OBJECT] [AGENT] [VERB]`; for animate beneficiaries or patients, `[AGENT] [VERB] [BENEFICIARY/PATIENT]`.

*The woman brings food for the picnic.*
`[PICNIC] [FOOD] [WOMAN] [BRING]`

*The teacher asks the student.*
`[TEACHER] [ASK] [STUDENT]`

### Negation

For certain verbs, negation is integrated directly into the sign. Examples of these verbs along with their corresponding negative forms are presented below: `[NEED]` / `[NOT-NEED]`, `[WANT]` / `[NOT-WANT]`, `[CAN]` / `[CANNOT]`, `[BELIEVE]` / `[NOT-BELIEVE]`, `[KNOW]` / `[NOT-KNOW]`, `[HERE]` / `[NOT-HERE]`, `[LOVE]` / `[NOT-LOVE]`

*You do not like strawberries*
`[STRAWBERRY] [PRO2] [NOT-LOVE]`

*He does not want to go to school by bike*
`[SCHOOL] [BIKE] [GO] [PRO3] [NOT-WANT]`

For other verbs, negation is expressed by an additional sign, such as `[NOT]`, `[NEVER]`, `[NOTHING]`, `[NO-MORE]`, or `[THERE-IS-NOT]`, placed after the relevant group of glosses. Note that these signs are typically accompanied by head movements or negative facial expressions.

*I do not understand the lesson*
`[LESSON] [PRO1] [UNDERSTAND] [NOT]`

*They never eat meat*
`[MEAT] [PRO6] [EAT] [NEVER]`

### Punctuation

Punctuation, annotated as `[/]`, is incorporated in *PGloss-LSF* sequence to segment portions of an utterance. These markers correspond to pauses in the flow of signs, which are necessary to preserve the sentence's semantic structure.

### Adjectives

Adjectives follow the words to which they refer.

*This morning, he admires the tired rose. She prepares to open her delicate petals.*
`[MORNING] [ROSE] [TIRED] [PRO3] [ADMIRE] [/] [PETALS] [DELICATE] [PRO3] [OPEN] [PREPARE]`

### Discretization of the signing space: `_loc`

The signing space is divided into discrete zones, following the principles defined in (Lebourque et al., 2001) for motor program specification and in (Millet, 2019) for the pre-semantic space. We discretize this space relative to the signer's body using three biomechanical planes –sagittal, frontal, and transverse, corresponding to forward/backward, left/right, and up/down dimensions. Two additional intermediate planes, oriented at 45° to the right and left, further subdivide the space. Within each plane, directions are coded as follows: planes orthogonal to the transverse plane use left `L`, neutral `N` (default), and `R`; depth (amplitude) is coded as on the body `B`, medium `M` (default), or extended `E`; vertical directions are coded as 0 (horizontal), 1 ($\approx 30°$ from transverse, default), and 2 ($\approx 60°$ upward). These three dimensions uniquely define a location in the signing space, with default values used when information is missing.

Locations are defined relative to an absolute reference linked to the signer's center of gravity. If the torso rotates from the frontal position (e.g., during role shift), the signing space rotates accordingly, and previously defined locations retain their coding. The argument `_loc` specifies the position of an entity within these predefined zones.

`[HOUSE_loc:R1M]`
means the sign `HOUSE` at the location right, medium height, medium amplitude.

### Identity: `_id`, `_nsn`

Individuals are generally identified by their French name, which can be spelled out, and by their signed name `_nsn`. Because the signed name may vary within a narrative while still referring to the same person (e.g., `[CURLED]` or `[LONG] [CURLED]`), an `_id` argument can be used to maintain consistent reference.

John, signed name `CURLED`, is represented as `[nsn:CURLED_id:JOHN]`

### Pre-semantic space, pronouns

We define a pre-semantic space, inspired from (Millet, 2019), as illustrated in Figure 1. This space provides a structured set of predetermined areas that enable shared and consistent referencing within a narrative. These areas are represented by conventionalized spatial loci (rather than discourse-established referential loci), and used as pronominal referents or as starting points or end points of executed trajectories, particularly for directional or trajectory verbs.

In LSF, pronouns exploit the pre-semantic space as deixis elements, whose meaning depends on the discourse context. They are realized through pointing or designation gestures toward specific loci referenced in the discourse. Singular pronouns use a pointing gesture with the index finger extended to a pre-semantic locus. In our annotation, these are represented as `[PROx]`, with x = 1, 2, 3.

Specifically, singular pronouns `[PRO1]` (I/Me), `[PRO2]` (You), and `[PRO3]` (he/she/him/her) are realized by pointing to their reserved loci, while plural pronouns use small gestures within the pre-semantic space: a circular gesture for `[PRO4]` (we) and `PRO6` (they/them), or a sweeping gesture for `[PRO5]` (you).

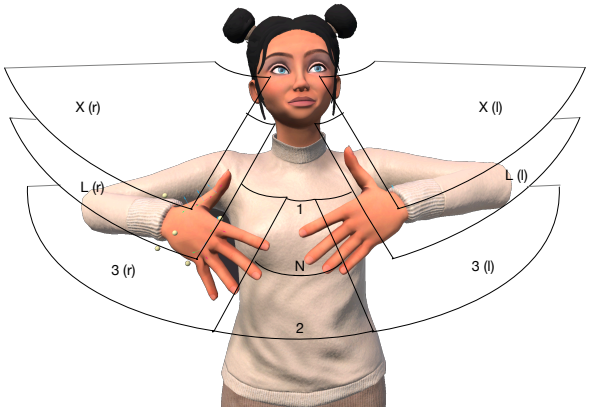



Figure 1: The pre-semantic space, inspired from (Millet, 2019).

Other third-person pronouns vary with animacy and are defined in the pre-semantic space with the prefix `PSS`:
`PSS3`: animate (he/she), in space 3, to the left or right of the signer, in the horizontal plane;
`PSSX`: indeterminate (*someone*), to the left or right of the signer, in the horizontal plane at temple level;
`PSSL`: inanimate (*it*), to the left or right of the signer, referring to a location, between the two previous planes.

Similarly, additional pre-semantic zones are defined as:
`PSS1`: near the signer;
`PSSN`: neutral zone, in front of the signer;
`PSSO`: distant zone, corresponding to the objective.

*Information is sent to me.*
`[INFORMATION] [SEND_src:PSSX_tgt:PSS1]`

*I send information to someone.*
`[INFORMATION] [SEND_src:PSS1_tgt:PSSX]`

*He is going to visit Paris.*
`[PARIS] [PRO3] [VISIT] [GO_tgt:PSSL]`

Possessive pronouns can be represented as:
`TO-PRO1` : *mine*
`TO-PRO2` : *yours*

TO-PRO3 : *his/her*
TO-PRO4 : *ours*
TO-PRO5 : *yours (plural/formal)*
TO-PRO6 : *theirs*

### Pointing: _pnt

Pointing gestures are a core mechanism for spatial reference in LSF. We focus on simple deictic forms with the index finger or hand, leaving other strategies (e.g., gaze or body orientation) aside. The referential target, not the gesture itself, is the key parameter.

Spatial reference is realized via loci in the discretized signing space, corresponding either to entities previously introduced and anchored in the discourse (person, object, place, etc.) or to a predefined zone of the pre-semantic space. In our annotation language, each locus is encoded as a pointing motor program, computed via an inverse kinematics process, with its spatial parameter given by _pnt followed by a value POINTING-TARGET defined in the spatial coding system.

*The house that is there*
[HOUSE] [THERE_pnt:R1E]

The neutral demonstrative pronoun may also be realized by a pointing gesture, with a palm-up handshape, referring to a previously introduced conceptual entity, as illustrated below.

*It is something too often forgotten (Le Petit Prince, Saint Exupéry)*
[THAT_pnt:THING-POINTED] [THING] [FORGET] [OFTEN]

Other deictic forms associated with demonstrative pronouns (e.g., THIS-ONE, THAT-ONE) and demonstrative determiners (e.g., THIS, THAT, THESE) are not addressed here.

### Directional or trajectory verbs in LSF: _src, _tgt, _mpr

In LSF, there exists a set of verbs, called *directional (or indicating) verbs*, which are syntactically structured as follows:

[VERB_src:SOURCE_tgt:TARGET_mpr:OBJECT]

with

- VERB : the action performed
- SOURCE : the agent of the action
- TARGET : the beneficiary of the action
- OBJECT : the entity on which the action is performed (optional)

The arguments _src and _tgt can take predefined location in the pre-semantic space, with values such as _src:PSS1, _tgt:PSS3, or the identifier of an already referenced object, e.g., _id:APPLE1. The third argument is addressed in the paragraph on proforms.

*I tell you a story.*
[STORY] [TELL_src:PSS1_tgt:PSS2]

*He lends me his computer.*
[COMPUTER] [TO-PRO3] [LEND_src:PSS3_tgt:PSS1]

The arguments _src and _tgt can also be used for verbs with trajectories.

*Paris, many tourists go there.*
[PARIS] [THERE_pnt:PSSL] [TOURIST] [MANY]
[GO_src:N_tgt:PSSL]

Note that verbs other than directional or trajectory verbs do not incorporate the agents and beneficiaries of the action in their realization.

*In Brussels, we eat fries.*
[BRUXELlS] [FRIES] [PRO4] [EAT]

### Manual Proforms _mpr

PProforms are linguistic mechanisms grounded in the iconic dynamics of sign languages. Manual proforms act as pronouns, maintaining lexical and referential continuity and supporting syntactic structure. Bodily proforms involve the signer's embodied engagement, whereby the signer adopts the role of a character in the narrative.

From a syntactic perspective, specific hand configurations reproduce or maintain lexical properties within a sign or across discourse. These forms correspond to manual proforms (also called classifiers [1] in English). In our annotation language, manual proforms are specified using the argument _mpr, followed by the name of the proform prNAME (distinct from the citation form of the sign), and optionally by arguments specifying its location and identifier.

We distinguish several types of manual proforms:

1. **Verb-incorporated proforms**, particularly in directional or trajectory verbs involving an object argument; they may also occur in dynamic pointing constructions.

2. **Proforms representing an entity**, annotated as independent glosses. These proformed signs are primarily used to position entities relative to one another and may refer to either animate or inanimate referents.

3. **Quasi-lexicalized dynamic proforms**, including person proforms (with variants such as standing, seated, lying, etc.), and vehicle proforms.

[1] Classifiers include both proforms and size-and-shape specifiers.

Examples of the first category are shown below:

*I give him an apple.*
[APPLE] [GIVE_src:PSS1_tgt:PSS3_mpr:prAPPLE]

*He gives you the book (Renard's book).*
[BOOK] [mpr:prBOOK_id:RENART-BOOK]
[GIVE_src:PSS3_tgt:PSS2_mpr:RENART-BOOK]

In the first example, the proform prAPPLE is integrated into the verb [GIVE], producing a handshape that encodes the object along the verb's trajectory. In the second example, the handshape represents the book identified by RENART-BOOK. The proform may optionally precede the verb to emphasize the object. It should be noted that a list of pre-defined proforms is used in our corpus, each associated with specific hand configurations.

Proforms of the second category frequently involve co-occurring signs, i.e., signs performed simultaneously with both hands. Examples are presented in the following section, which introduces the concept of manual simultaneity. Proforms of the third category are not addressed here due to their high variability and the need for separate analysis.

### Simultaneity of left/right hand: _hd

This mechanism involves producing a sign with one hand while the other maintains an utterance-level element, typically via a proform –for example, executing a sign with the right hand while the left hand holds a proformed sign. The argument _hd specifies the hand (L or R). Once assigned, the hand retains its properties (location, handshape, orientation, etc.) until released or reassigned by a subsequent gloss.

A hand can be released in two ways: (i) implicitly, when a subsequent gloss is produced without the _hd parameter; (ii) explicitly, using the tag [/hd:L] or [/hd:R].

*I am going to the house.*
[HOUSE_loc:L1] [mpr:prHOUSE_loc:L1_id:HOUSE1_hd:L]
[GO_src:PSS1_tgt:HOUSE1_hd:R]

Through spatial anchoring, the sign [HOUSE] is first produced at location L1. The proform prHOUSE is then maintained by the left hand at the same location, while the verb GO is executed by the right hand along a trajectory from PSS1 toward the spatial locus associated with HOUSE1.

*I forgot the keys at the house.*
[HOUSE] [FORGET] [mpr:prHOUSE_id:HOUSE1_hd:L]
[KEY_hd:R] [OPEN-KEY_loc:HOUSE1_hd:R]

In this last example, the left hand maintains the proform prHOUSE (identified as HOUSE1), while the right hand sequentially produces the signs [KEY] and [OPEN-KEY]. The opening action is directed toward the spatial locus associated with HOUSE1, thereby encoding the locative relation through simultaneity.

The spatial relationship of one object relative to another can be encoded using two proforms, each performed by one hand. A non-exhaustive set of relative position vocabulary includes: ABOVE, BELOW, OVER, UNDER, LEFT, RIGHT, NEXT-TO, IN-FRONT-OF, BEHIND, TOP, ON, INSIDE, etc.

*The bowl (proform) is on the table (proform)*
[BOWL] [TABLE] [prm:prTABLE_id:TAB1_hd:L]
[prm:prBOWL_id:BOL1_hd:R_loc:TAB1*ABOVE]

In this example, we first produce the sign [BOWL], followed by [TABLE], and then position the proform prTABLE slightly before prBOWL, applying the modifier ABOVE to indicate the spatial relation.

*There is a tree; at the top of the tree, there is a nest in which there is a bird.*
[TREE_loc:L] [mpr:prTREE_id:ARB1_loc:L_hd:L]
[THERE_pnt:ARB1*TOP_hd:R_id:HA] [NEST_loc:HA]
[mpr:prNEST_id:NEST1_hd:L_loc:HA]
[BIRD_loc:HA*INSIDE_hd:R]

In this example, the sign [TREE] is first produced by the two hands at a left locus (spatialization) and then maintained as a proform with the left hand. The right hand points to the top of the tree, establishing the location labeled HA. Next, the sign [NEST] is produced by the two hands at the left locus identified by HA and maintained as a proform with the left hand. Finally, the sign [BIRD] is executed by the right hand inside the nest.

This utterance can alternatively be produced without pointing, by anchoring the relevant elements with the hands in the signing space:

[TREE] [mpr:prTREE_id:ARB1_hd:R]
[mpr:prNEST_id:NEST1_loc:ARB1*TOP_hd:L]
[BIRD_hd:R_loc:NEST1*INSIDE]

### Role shift (torso orientation): _trs

When a narrative involves multiple characters, the narrator may alternately assume the role of each character, especially in dialogue situations. In such cases, torso rotation from left to right signals the change of character.

Each role shift is annotated with an opening tag indicating torso orientation [trs:VALUE], with VALUE = L, R, or N, and a corresponding closing tag [/trs]. The opening tag can optionally include an _id argument to identify the signing interlocutor.

Because this argument typically applies to a sequence of glosses, [/trs] marks the end of the sequence. An optional _id parameter may also be included to disambiguate the interlocutor when identities change within the narrative. Examples of dialogue with role shifts (from the Tales of the Fox) are provided below. Examples:

Merchant 1: Look! A fox! I think it is dead!
[trs:R] [LOOK] [FOX] [DEAD] [PRO1] [THINK] [/trs]

– Let's stop and take it!
[trs:L] [PRO4] [STOP]
[TAKE_src:R_tgt:L_mpr:prTHING] [/trs]

– Look at the fur! It is beautiful, red with a white neck!
[trs:R] [FUR] [LOOK] [BEAUTIFUL] [RED] [WITH] [NECK] [WHITE] [/trs]

– Merchant 2, angry: Damned fox! We hope the eels will suffocate you!
[trs:L] [DAMN_exp:MEAN] [FOX] [/] [EELS] [PRO2] [SUFFOCATE] [PRO4] [HOPE] [/trs]

**Facial expression: `_exp`**

In LSF, facial expressions are syntactically relevant and fulfill three main functions: (i) marking sentence modality (e.g., interrogative, exclamative, imperative); (ii) expressing adverbial or adjectival meaning; and (iii) conveying the signer's attitude toward the utterance content (e.g., emotion, boredom, concern, contempt).

Facial expressions may co-occur with a single gloss or extend over multiple glosses simultaneously with manual signs. We therefore annotate them using the argument `_exp` in two ways: When the facial expression applies to a single gloss, it is included directly within that gloss, e.g., `[GLOSS_exp:INTENSE]`; When it spans multiple glosses or an entire clause, we use opening and closing tags.

The examples below illustrate these three situation:

```
[exp:QUESTION]    [GLOS1]    [GLOS2]    [GLOS3]
[/exp:QUESTION]
```

*This cat is very cute.*
[CAT_id:CAT1] [THIS_pnt:CAT1] [CUTE_exp:VERY]

*He eats messily.*
[PRO3] [EAT_exp:DISGUST]

*You don't eat soup?*
[exp:DOUBT] [SOUP] [PRO2] [EAT] [NOT] [/exp:DOUBT]

# 4. Resultats and Evaluation

## 4.1. Bilingual Corpus

We rely on a bilingual corpus (French sentences paired with LSF parameterized glosses) which has been created specifically for our *SignToKids* project. It comprises materials produced by Deaf educators from French institutions[2], as well as transcribed children's tales and literary works.

This corpus include several sub-corpora:

- Sentences illustrating grammatical structures such as negation, interrogation, and indicating verbs.
- LSF narratives of the Tales of Renart.
- Short LSF stories demonstrating spatialization mechanisms and proforms (e.g., persons, vehicles).
- Several excerpts from *Le Petit Prince* by Antoine de Saint-Exupéry.

All corpora were annotated and signed by native Deaf signers and recorded using motion capture or video. In total, the dataset comprises approximately two hours of signed LSF utterances.

## 4.2. Evaluation

The evaluation of the proposed formalism follows a twofold approach. First, a qualitative linguistic assessment is conducted in which a Deaf LSF teacher generates signed utterances based on the written transcription. The alignment between these generated productions and the corresponding LSF video recordings is subsequently verified manually, with corrections implemented whenever discrepancies are identified.

Second, we conducted a quantitative evaluation using an automatic text-to-sign translation system. Unseen French sentences were translated into LSF glosses and analyzed for structural consistency with the specification.

The initial evaluation used 165 sentence pairs and was later expanded to 1,650 pairs (Reverdy et al., 2025), covering three-argument glosses with varied spatial loci in the pre-semantic space. Gloss prediction (the semantic label) reached an 18% error rate (Levenshtein distance), while parameters (argument and value) achieved F1-scores of 97% and 98%, respectively.

The annotation scheme was then extended to include additional grammatical mechanisms (pointing, manual proforms for inanimate objects, left/right hand co-occurrence, facial expressions, and role shift), resulting in a dataset of approximately 3,000 sentence pairs. Evaluation of this expanded set is ongoing.

# 5. Conclusion and perspectives

This work has led to the development of a formalism for the annotation of LSF, called *PGloss-LSF*, systematically aligned with LSF video data. The formalism is based on glosses enriched with explicit linguistic and articulatory parameters. Unlike linear gloss sequences, which fail to capture the structural specificities of sign languages, it integrates syntactic and semantic grammatical descriptions

[2]Institut National des Jeunes Sourds and Centre Gabriel Deshayes

that can be dynamically adapted to the discourse context. Its key contribution lies in its ability to explicitly model motor control parameters, thereby bridging linguistic representation and animated sign synthesis.

At this stage, the majority of the formalized grammatical rules of LSF are incorporated into the framework. Future work will extend the annotation scheme to cope with additional sign language mechanisms that exhibit greater variability, such as dynamic proforms and highly iconic constructions. Ultimately, the proposed formalism is designed to be adaptable to other sign languages.

Beyond linguistic documentation, the framework supports computational applications including automatic sign language translation, avatar-based sign generation, and assistive communication technologies. By addressing the challenges of bilingual alignment and semantic interoperability, our *PGloss-LSF* formalism provides a principled and extensible foundation for annotating, integrating, and sharing data across heterogeneous linguistic resources.

## 6. Acknowledgements

The *SignToKids* project is funded by the French National Research Agency (ANR). Several research teams are involved in this project: the IRISA–CNRS laboratory, the MoDyCo–CNRS laboratory, and the National Institute for Deaf Youth (INJS). We would also like to thank the Deaf and signing participants who contributed to the project: Nathalie Irdel, from the Gabriel Deshayes Center, Kristof Gendreau Touchais from the University of Poitiers, and the Deaf/French teaching staff at INJS. They all contributed to the design, recording, and annotation of the corpus.

## 7. Bibliographical References

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024. Neural sign actors: A diffusion model for 3d sign language production from text. *arXiv 2312.02702*.

Christian Cuxac. 2000. *La langue des signes française (LSF) : les voies de l'iconocité*. Faits de langues. Ophrys.

Ralph Elliott, John RW Glauert, JR Kennaway, Ian Marshall, and Eva Safar. 2008. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Universal Access in the Information Society*, 6(4):375–391.

U. Farooq, M. Rahim, N. Sabir, A. Hussain, and A. Abid. 2021. Advances in machine translation for sign language: approaches, limitations, and challenges. *Neural Computing and Applications*, 33:14357–14399.

Michael Filhol, John McDonald, and Rosalee Wolfe. 2017. Synthesizing sign language by connecting linguistically structured descriptions to a multitrack animation system. In *International Conference on Universal Access in Human-Computer Interaction*, pages 27–40. Springer.

Thomas Hanke, Lutz König, Reiner Konrad, Maria Kopf, Marc Schulder, and Rosalee Wolfe. 2023. EASIER notation – a proposal for a gloss-based scripting language for sign language generation based on lexical data. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP 2023 - Workshops, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5, Rhodes, Greece.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the public dgs corpus in size and depth. In *sign-lang@ LREC 2020*, pages 75–82. European Language Resources Association (ELRA).

Matt Huenerfauth. 2006. *Generating American Sign Language classifier predicates for English-to-ASL machine translation*. Ph.D. thesis, University of Pennsylvania.

Robert E Johnson and Scott K Liddell. 2011. A segmental framework for representing signs phonetically. *Sign Language Studies*, 11(3):408–463.

Trevor Johnston and Louise de Beuzeville. 2009. Researching the linguistic use of space in Auslan: Guidelines for annotators using the Auslan corpus. Technical report, Department of Linguistics, Macquarie University, Sydney.

Navroz Kahlon and Williamjeet Singh. 2021. Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society*.

Thierry Lebourque, Sylvie Gibet, and Pierre-François Marteau. 2001. High level specification and animation of communicative gestures. *Journal of Visual Languages and Computing*, 12:657–687.

Dimitris Metaxas, Mark Dilsizian, and Carol Neidle. 2018. Scalable ASL sign recognition using model-based machine learning and linguistically annotated corpora. In *Workshop on the Representation and Processing of Sign Languages (LREC 2018)*, pages 127–132, Miyazaki, Japan.

Agnès Millet. 2019. *Grammaire descriptive de la langue des signes française: dynamiques iconiques et linguistique générale*. UGA Editions.

Liliane Momeni, Hannah Bull, K. R. Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *European Conference on Computer Vision – ECCV 2022*, pages 671–690. Springer.

Carol Neidle, Augustine Opoku, Gregory Dimitriadis, and Dimitris Metaxas. 2018. New shared & interconnected asl resources: Signstream® 3 software; dai 2 for web access to linguistically annotated video corpora; and a sign bank. In *Workshop on the Representation and Processing of Sign Languages (LREC 2018)*, Miyazaki, Japan.

Adrián Núñez Marcos, Olatz Perez-de Viñaspre, and Gorka Labaka. 2023. A survey on sign language machine translation. *Expert Syst. Appl.*, 213.

Siegmund Prillwitz. 1989. *HamNoSys: Version 2.0; Hamburg Notation System for Sign Languages; An Introductory Guide*. Signum-Verlag.

Clément Reverdy, Sylvie Gibet, and Pierre-François Marteau. 2025. Translating sentences from written french into LSF annotation. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents, IVA Adjunct 2025, Berlin, Germany, September 16-19, 2025*, pages 13:1–13:6.

Rotem Shalev-Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2pose: Animating sign language notation into pose sequences. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 21046–21056.

William C Stokoe, Dorothy C Casterline, and Carl G Croneberg. 1976. *A dictionary of American Sign Language on linguistic principles*. Linstok Press.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. 2024. SignAvatars: A Large-scale 3D Sign Language Holistic Motion Dataset and Benchmark. *arXiv 2310.20436*.

# A Pose-Based Pipeline for Annotation of Headshakes in Sign Language Corpora

**Gustaf Gren, Nikolaus Riemer Kankkonen**
Stockholm University
gustaf.gren@ling.su.se, nikolaus.kankkonen@ling.su.se

**Abstract**

This paper introduces a pose-based pipeline designed to support scalable annotation of headshakes in sign language corpora. Motivated by the scarcity of annotated datasets and the need for quantitative typological research, the study evaluates whether automated detection can reduce human annotation effort. The system operates on yaw trajectories extracted with MediaPipe Holistic and uses sliding-window segmentation with neural sequence models (LSTM/CNN) to surface candidate segments for review. Training and evaluation are conducted on a subset of the German Sign Language (DGS) Corpus annotated to target grammatical headshakes functioning as negation rather than for every instance of headshakes. On the DGS dataset the best performing LSTM model achieves an $F_2$-score of 0.45, recall of 0.63. Despite the narrow annotation scope, the pipeline reduces search space: annotators need review only 13% of frames to recover 87% of labeled instances. Error analysis indicates that many false positives correspond to plausible head movements excluded by the annotation criteria. A pilot transfer to Swedish Sign Language shows reduced effectiveness without adaptation, underscoring the need for alignment in cross-lingual transfer scenarios.

**Keywords:** sign language, non-manual signs, non-manual sign detection, computational sign language, pose estimation, sign language resources

## 1. Introduction

Elements that convey linguistic meaning in sign languages without involving the hands are collectively known as nonmanual markers (Pfau and Quer, 2010). These markers can function either affectively (e.g. to express surprise, backchanneling) or grammatically (e.g. as an obligatory negation marker).

With respect to headshakes, Pfau (2015) argues that their grammaticalization as a negation marker originates from their use in spoken language as co-speech gestures expressing negation. Building on cross-linguistic variation, sign languages have been typologically divided into those that are manually dominant and those that are non-manually dominant for headshakes as a negation marker (Zeshan, 2004). Pfau (2015) adopts this distinction in his discussion. However, Johnston (2018) cautions against adopting this manual versus non-manual negation typology, based on the limited quantitative coverage of sign languages to test such typological claims.

Recent work in spoken language typology proposes treating features as gradient rather than binary, modeling them as continuous properties (Kann, 2025; Levshina et al., 2023). In the case of headshakes, this would entail examining the *distribution* of their grammatical use relative to their overall frequency in interaction, yielding a continuous value rather than a binary typological classification. The latter includes a range of discourse and interactional functions (e.g. backchanneling, emphasis, and other pragmatic uses). Investigating this degree of variation quantitatively requires corpus-based studies. Similar gradient-like approaches have been explored in sign language research, for example in corpus-based work on Turkish Sign Language (Makaroğlu, 2021).

Annotated data is scarce, in general, for many sign languages (De Sisto et al., 2022). Public datasets annotated for all instances of headshakes do not, to our knowledge, yet exist for any sign language. In this study, we investigate whether pose-based models can detect headshakes with sufficient recall to assist human annotation for sign language, with the goal of aiding broader data-driven typological investigations.

We frame headshake detection as an annotation-support task rather than a pure classification problem. To evaluate this approach we use German Sign Language (DGS) data annotated for headshakes used as negation. We compare a simple zero-crossing baseline with two neural architectures: a bidirectional LSTM and a 1D CNN. Both models operate on yaw trajectories extracted using MediaPipe Holistic. Model performance is assessed using both aggregate metrics and a targeted error analysis to distinguish between failures arising from representational limitations and those attributable to annotation. Lastly, we perform a pilot study on cross-lingustic transfer. We annotate 50 minutes of Swedish Sign Language data and evaluate the models trained for DGS without additional training.

**Our contributions are:**

- **A practical evaluation framework for non-manual annotation support.** We introduce an annotation-centric evaluation based on re-



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 193–202
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

view rate (the fraction of frames requiring human inspection) and demonstrate how recall trades off against annotation workload.

- **A pose-based pipeline for headshake candidate surfacing.** Using yaw and its temporal derivatives extracted with MediaPipe Holistic, we compare a zero-crossing baseline to BiLSTM and 1D CNN models, establishing performance bounds on DGS headshake-as-negation labels.

- **Analysis of failure modes under label scope and domain shift.** Confidence-stratified manual inspection reveals systematic discrepancies between model predictions and annotation targets, and a zero-adaptation pilot transfer to Swedish Sign Language shows degradation without calibration.

## 2. Background

Human pose estimation models identify human body parts and build representations such as body or hand skeletons, or facial mesh, from images or videos (Zheng et al., 2023). Because sign languages rely on coordinated facial, manual, and bodily movements, these models offer a way around the dimensionality challenges of raw video. They reduce the input from a time series of pixel matrices to a time series of keypoints corresponding to specific body parts (e.g. left pinky knuckle, left elbow). Pose estimation models have been used extensively in sign language research. For example, they have been applied to automatically identify whether someone is signing (Moryossef et al., 2020) and to recognize fingerspelling in ASL (Shin et al., 2021).

The number of keypoints depends on the model. Some pose estimation models have less granularity, like BlazePose (Bazarevsky et al., 2020) which produces 33 body keypoints in total, in order to focus on speed. Others have many more. MediaPipe Holistic (Lugaresi et al., 2019), comparatively, can detect up to 468 keypoints just for the face.

Computational studies of headshakes within sign language have mostly focused on their phonetics and distribution. Chizhikova and Kimmelman (2022) studied the phonetics of headshakes in Russian Sign Language utilizing OpenFace (Baltrusaitis et al., 2018), showing that negative headshakes are optional and relatively rare in Russian Sign Language, appearing in under 30% of negative sentences, and typically consist of one or two small, fast head turns closely aligned with manual negation signs. Kimmelman et al. (2024) also used OpenFace for the Sign Language of the Netherlands, to analyze whether linguistic properties of headshake influence its phonetic form. These studies demonstrate the feasibility of pose-based phonetic analysis but still rely on manual annotation.

The closest related work is Rijbroek (2023), who evaluated automatic headshake detection for the Sign Language of the Netherlands utilizing OpenPose (Cao et al., 2017), achieving the best results using a Hidden Markov Model ($F_1$-score of 0.19, precision of 0.12 and recall of 0.54).

## 3. Data

For training, we base our work on a subset of the DGS Corpus (Konrad et al., 2020), a publicly available collection of German Sign Language recordings from 330 signers. The released videos are provided at 360p resolution and 50 FPS. Headshake annotations were provided by Kimmelman et al. (in preparation) and target grammatical headshake usage in negative clauses rather than all headshakes. Consequently, the labels reflect a narrower linguistic definition than the detection objective used in this study. To better understand the impact of this label mismatch, we conduct an error analysis on model predictions (Section 4.4) to identify systematic inference patterns and assess how annotation discrepancies influences performance.

Data were split at the file level prior to windowing into 72% training, 8% development, and 20% testing sets to ensure that windows from the same recording appeared in only one partition, thereby minimizing temporal leakage from overlapping windows. While a small number of signers contributed multiple recordings, these instances were rare, limiting potential overfitting to signer-specific motion patterns. Because the split was not strictly signer-independent, residual signer effects cannot be ruled out. File-level splitting resulted in a small shift in the global headshake-to-non-headshake ratio (±0.003). The median headshake duration is 0.8 seconds.

To assess performance in transfer scenarios, approximately 50 minutes of video from the Swedish Sign Language Corpus (SSLC) (Öqvist et al., 2020) were manually annotated for headshake events. The annotation was carried out by a single annotator with prior experience working with the corpus. Headshakes were identified based on visible lateral head movements in a frontal view, with a practical threshold of approximately $\pm 10$ degrees from neutral position. Headshake duration was measured from one side of the movement to the other, excluding the lead-in and settling phases. The annotated headshakes have a median duration of 0.4 seconds.

See Table 1 for the full data rundown.

| | Files | Total | HS | non-HS | HS/Total |
|---|---|---|---|---|---|
| **all** | 86 | 2883202 | 37543 | 2845659 | 0.0130 |
| **train** | 61 | 2030449 | 26490 | 2003959 | 0.0130 |
| **dev** | 7 | 297304 | 4915 | 292389 | 0.0165 |
| **test** | 18 | 555449 | 6138 | 549311 | 0.0111 |
| **sslc** | 10 | 82211 | 2841 | 79370 | 0.0345 |

Table 1: File/frame statistics of each split and headshake frames (HS) and non-headshake frames (non-HS), including the domain separate annotated SSLC test-set.

## 4. Methodology

We base our methodology on the pipeline of Pouw (2025), who performed end-to-end multi-gesture detection using a sliding window approach. The approach involves using Mediapipe Holistic (Lugaresi et al., 2019) to extract face, body, and hand kinematic features from video, normalizing the keypoints, and trains a Convolutional Neural Network (CNN) on labeled clips for multigesture detection. Most relevant to headshake detection is the features related to deriving Euler angles using rotation matrix decomposition formulas from the facial keypoints. There are three Euler angles:

- **Pitch**: Head tilting up/down (nodding)
- **Yaw**: Head turning left/right (shaking)
- **Roll**: Head tilting ear-to-shoulder

In practice, deriving Euler angles in this way gives us an array with elements describing where the head is pointing for each given frame where keypoints are found. See Figure 1 for an example, where Euler angles are illustrated using a stick pointing in the direction of the derived angles.

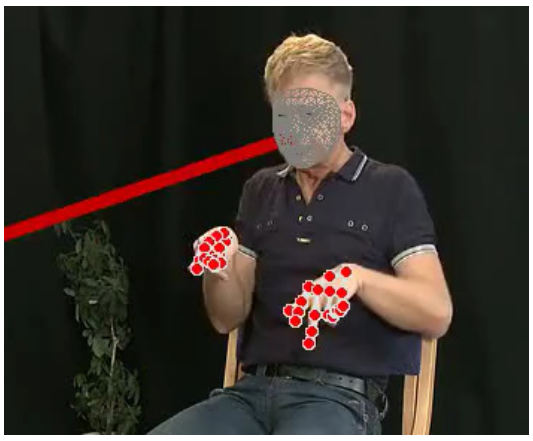


Figure 1: Example frame of video with face mask, hand keypoint estimations, and derived Euler angles from MediaPipe Holistic. Video still from the Swedish Sign Language Corpus.

Since we're doing headshake detection, we reduce dimensionality further, and only use the yaw values because headshakes is primarily horizontal rotation. Initially, we experimented using all three Euler angles, with poor results, with pitch and roll only adding noise. We also resample our video arrays to 25FPS. This is partly to reduce dimensionality further, but primarily to give an idea of future adaptability, given many sign language datasets are recorded at lower speeds, such as 30 or 25FPS. For example, the Swedish Sign Language Corpus (Öqvist et al., 2020) is recorded in 25FPS.

For each video clip, we derive keypoints for every frame. We use MediaPipe Holistic as the pose estimation model for our work because, while it has some limitations in estimating eyebrow movement (Kuznetsova and Kimmelman, 2024), Sargano et al. (2024) found it to be the most accurate among several pose estimation models for yaw estimation. If MediaPipe fails to find keypoints for a frame, then we set it to the same value as the previous frame. We do this instead of more advanced interpolation simply because there are few missing frames, and because it ensures a one-to-one mapping between the yaw array and the frame index. Continuous yaw sequences are segmented into sliding windows with stride. Stride is the number of frames you advance the start of each sliding window (e.g., with a window size of 75 and a stride of 5, windows start every 5 frames, so they overlap by 70 frames). Windows are normalized across clips. Each window $\tilde{y}_t$ is $z$-normalized using the clip's yaw mean $\mu_c$ and standard deviation $\sigma_c$:

$$\tilde{y}_t = \frac{y_t - \mu_c}{\sigma_c + \epsilon}$$

Where $\epsilon$ is a small constant to avoid division by zero.

### 4.1. Zero crossing baseline

As a simple baseline, we classify fixed-length yaw windows using their zero-crossing count. The hypothesis is that headshakes resemble a sinusoidal waveform, and should periodically cross 0. Assuming accurate pose estimation and a stable resting head position, 0 should represent the person looking straight ahead.

Given a windowed yaw signal $y_{1:T}$, we define the number of zero-crossings as

$$Z = \sum_{t=2}^{T} \mathbf{1}[\text{sign}(y_t) \neq \text{sign}(y_{t-1})]$$

A window is labeled positive during tuning if it contains at least one positive frame annotation. On the development set, we compute $Z$ for all positive

windows and set a single decision threshold

$$k = \text{round}\left(\frac{1}{N_+}\sum_{i=1}^{N_+} Z_i\right)$$

where $N_+$ is the number of positive windows. At inference time, a window is classified as a headshake if $Z \geq k$, and negative otherwise.

### 4.2. Neural pipeline

Because this is a single-gesture detection task, we simplify the pipeline in Pouw (2025) and reduce it to use three features: **yaw angle**, **velocity** (derivative of the normalized yaw-array), **acceleration** (second derivative of the normalized yaw-array). Our training data then gets the shape $(B, W, 3)$, with $B$ being our batch size, $W$ our window size. In addition these input vectors are augmented during training, adding small random noise with a probability of $50\%$ during training for each batch.

We test two types of neural models which we implement in `PyTorch`(Paszke et al., 2019)[1]: (i) a bidirectional LSTM with dropout of 0.2 or (ii) a 1D CNN with temporal convolutions. The LSTM was chosen for its ability to model temporal dependencies. It is also a suggestion for future research by Rijbroek (2023), but is to our knowledge still untested for headshake detection in sign languages. The CNN is kept as close to Pouw (2025) as possible. Both models produce per-frame logits for each window, and are trained with a weighted focal loss to address the heavy class imbalance.

A targeted hyperparameter search was performed. Using the LSTM as the target model, the search identified learning rate and window size as the dominant factors; other parameters had limited effect. Therefore we focused our CNN search on learning rate. This resulted in a learning rate of 1e-3 for the LSTM and 1e-4 for the CNN, and a window size of 75 and stride of 5 frames. Other parameters are kept identical between the CNN and LSTM: AdamW optimizer with weight decay set to 1e-5, FocalLoss with $\alpha$=0.25;$\gamma$=2.0, and a ReduceLROnPlateau with a patience of 2 and reduction of 0.5.

We train both models for a maximum of ten epochs, while evaluating against the dev-set three times per epoch. The best checkpoint in regards to $F_2$ against the dev-set is used for the final evaluation against the test-set. Best checkpoint occurred at the third epoch for the LSTM, and fifth for the CNN.

Overlapping windows results in multiple predictions per frame (except a few in the beginning and at the end of a clip). Therefore, we need a temporal aggregation step for inference. We evaluate two strategies during training: mean pooling (mean probability across covering windows for a given frame), and max pooling (maximum probability across covering windows for a given frame).

### 4.3. Evaluation Metrics

We evaluate using precision, recall, $F_1$-scores, and $F_2$-scores. In annotation-support settings, false positives impose a small verification cost, whereas false negatives require full video inspection. Therefore, high recall with moderate precision can still produce efficiency gains. $F_2$-scores reflect this aim, which weighs the score towards recall. During training we set the classification threshold on the dev-set.

Let precision $P$ and recall $R$ be defined as

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

where $TP$, $FP$, and $FN$ denote true positives, false positives, and false negatives. The F$_\beta$ score is given by

$$F_\beta = (1 + \beta^2)\,\frac{PR}{\beta^2 P + R}$$

We use $\beta = 2$, yielding

$$F_2 = 5\,\frac{PR}{4P + R}$$

### 4.4. Error Analysis

We conduct error analysis to assess whether classification performance was driven by discrepancies between model goals and annotation goals, as well as uncover potential inference patterns. Predictions were categorized by model confidence **(high vs. low)** and correctness **(correct vs. incorrect)**.

Using the best-performing model, we extract 25 representative prediction windows from each category of prediction. For example, for the high confidence false positives we sort our false positives and take top $k$ of the predictions by confidence level. We first turn our frame-level ground truth annotations into window-level and sort out windows where headshakes occur <50% of the frames in the window size, in order to avoid windows without much movement information. Otherwise we would have windows where, for example, the last two frames are headshakes. Such windows wouldn't be meaningful to humans. To derive the window-level predictions we use the same frame-level threshold set during training on the dev-set. These filtered window predictions are inspected manually for patterns.

[1]We tested utilizing a Hidden Markov Model similar to Rijbroek (2023), but it was rejected since it was difficult to make it work consistently with `PyTorch`.

### 4.5. Annotation Burden Analysis

To evaluate the feasibility of using the models for annotation work, we calculate the relationship between how many frames the annotator has to go through, compared to the headshakes correctly identified.

For each model we go through each window in the test-set and convert window predictions to per-frame probabilities by *mean* aggregation, taking the mean probability among all windows covering a frame. We then sweep classification thresholds from 0.0 to 1.0, converting probabilities into predicted frames. Finally, we add a dilation of 10 frames before and after predicted ranges to account for context frames that would be added during manual annotation.

Let $F_{pred}$ be the number of predicted headshake frames after dilation, and $F$ the total set of frames. We then calculate Review Rate ($RR$) and compare with recall:

$$RR = \frac{|F_{pred}|}{|F|}$$

For each model we calculate the best weigh-off point as

$$i^* = \arg\max_i (R_i - RR_i)$$

Where $R_i$ is recall and $RR_i$ is review rate at threshold $i$.

### 4.6. Domain Transfer

Finally, we investigate the effect of transferring the best-performing model to another sign language without additional training. We evaluate the annotated SSLC clips described in Section 3 in the same way as for the DGS data, using the models trained for the DGS data. The short length of headshakes in our small SSLC dataset makes a window analysis approach, as in Section 4.4, unfeasible. Instead, we present plots for all frames which aren't true negatives alongside their respective confidence scores in two randomly chosen clips.

## 5. Results

Neural approaches outperformed the baseline for headshake detection in DGS, which achieved very low precision despite moderate recall, resulting in poor overall $F$-scores. Among the tested architectures, LSTM-based models produced the strongest results, with the mean-pooled variant achieving the highest precision (0.21), recall (0.63), and $F_2$ score (0.45), while tying for the best $F_1$ score (0.31) with the max-pooled LSTM. CNN models showed competitive but lower performance, particularly in recall. See Table 2 for full metrics result table.

| | **P** | **R** | $F_1$ | $F_2$ |
|---|---|---|---|---|
| Baseline | 0.02 | 0.59 | 0.04 | 0.08 |
| $\text{LSTM}_{\text{max}}$ | 0.20 | 0.61 | **0.31** | 0.43 |
| $\text{CNN}_{\text{max}}$ | 0.16 | 0.53 | 0.24 | 0.36 |
| $\text{LSTM}_{\text{mean}}$ | **0.21** | **0.63** | **0.31** | **0.45** |
| $\text{CNN}_{\text{mean}}$ | 0.19 | 0.50 | 0.27 | 0.37 |

Table 2: Precision, recall, $F_1$, and $F_2$ for each model and pooling strategy on the DGS data. Threshold set at 0.9, calculated on dev-set.

### 5.1. Error Analysis

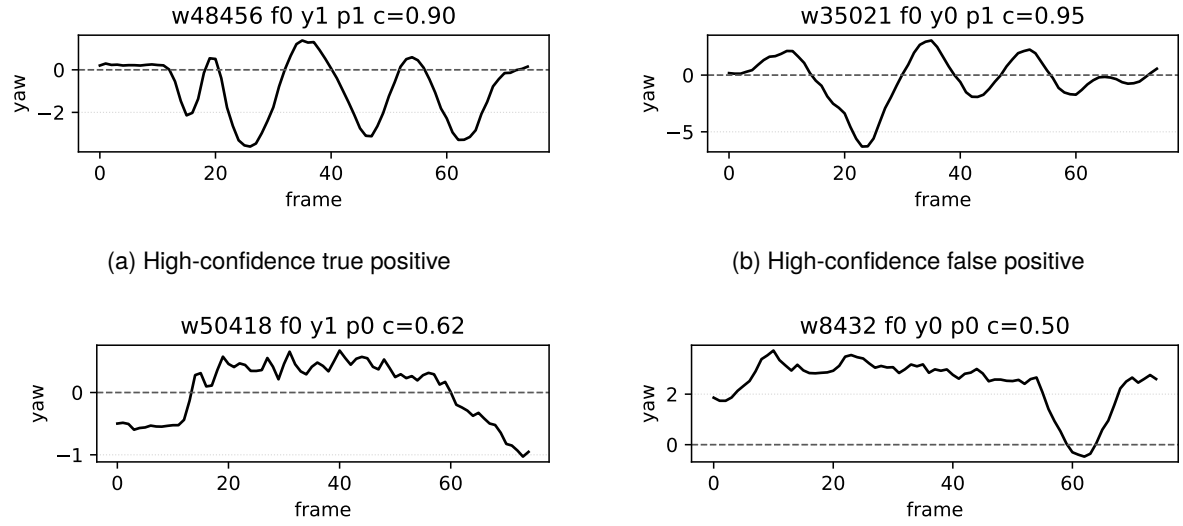
Figure 2 shows example yaw-angle trajectories for the four confidence/outcome categories. For the 25 most representative windows of each category, we generally observe:

- **High-confidence true positives:** Clear periodic waveforms with consistent oscillation between negative and positive yaw angles, matching expected headshake motion across all inspected windows.
- **High-confidence false positives:** Visually plausible headshakes, though the motion is not always centered around a neutral (zero) yaw position.
- **Low-confidence false negatives:** Two distinct patterns emerge in the 25 least confident windows for false negatives:
  1. *Subtle, high-frequency low-amplitude motion*. If you inspect the yaw-array (see Figure 2c) it appears this subtle nature is reflected in the array as well.
  2. *High deviation from resting position*. For example, one instance was observed where the subject had turned just before initiating the headshake. This made the window start highly off-center, even if the pattern was consistent with a headshake.
- **Low-confidence true negatives:** No consistent motion pattern, though they all contain motion. For example, a person with head movement in virtue of swaying left and right with the body.

### 5.2. Annotation Burden

The LSTM and CNN perform similarly when it comes to annotation burden metrics, requiring 13% review rate to achieve 87% recall, and 17% review rate to achieve 89% recall respectively. They both outperform the baseline, which has the best trade-off at a review rate of 36% to achieve 53% coverage.

The full plot is available in Figure 3.




(a) High-confidence true positive

(b) High-confidence false positive

(c) Low-confidence false negative

(d) Low-confidence true negative

Figure 2: Yaw angle trajectories for representative windows across prediction outcomes. .

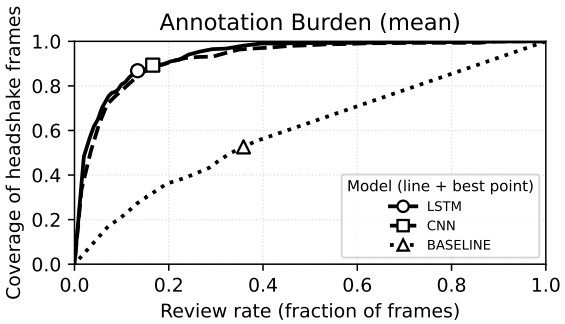



Figure 3: Proportion of required frame review rate of the test-set compared to headshake frame recall, utilizing mean inference pooling across frames.

| | **P** | **R** | $F_1$ | $F_2$ |
|---|---|---|---|---|
| Baseline | 0.05 | **0.77** | 0.09 | 0.18 |
| $\text{LSTM}_{\text{max}}$ | 0.18 | 0.22 | **0.20** | **0.21** |
| $\text{CNN}_{\text{max}}$ | 0.11 | 0.26 | 0.16 | **0.21** |
| $\text{LSTM}_{\text{mean}}$ | **0.36** | 0.05 | 0.09 | 0.06 |
| $\text{CNN}_{\text{mean}}$ | 0.14 | 0.11 | 0.12 | 0.11 |

Table 3: Precision, recall, $F_1$, and $F_2$ for each model and pooling strategy on the SSLC data.

### 5.3. Domain Transfer

The results get worse across all models except the baseline when testing on SSLC data. The baseline gets a recall of 0.77 and precision of 0.05, higher on both than for DGS. While the $\text{LSTM}_{\text{mean}}$ performed best for DGS, it here achieves very poor recall (0.05), though with high precision (0.36). Both neural models with max pooling perform best for $F_2$-scores.

We see this performance difference reflected in Figure 4, where many frames annotated as headshakes exhibit low confidence. This trend is further illustrated in the annotation burden plot for SSLC (Figure 5), which shows a less steep improvement as thresholds become more generous compared to DGS.

## 6. Discussion

Across models for the DGS data, results indicate that headshake detection is achievable with moderate recall but limited precision, suggesting that current pose-based approaches are better suited for assisting human annotation than for fully automated labeling. The most confident false positives during our error analysis correspond to genuine headshakes, in this case pointing to a mismatch between annotation targets (headshakes used as grammatical negation) and what our aim is (general headshake detection). This suggests the metrics above could potentially *under-represent* the models' true performance for general headshake detection, at least for DGS.

Even so, annotation-support systems do not require near-perfect classification performance to be valuable; their utility derives from reliably narrowing the search space for human annotators. The models appear capable of accelerating annotation for headshakes used as negation: the LSTM model requires annotators to review only 13% of the total frames to achieve 87% headshake coverage in DGS. However, this is not consistent with the small

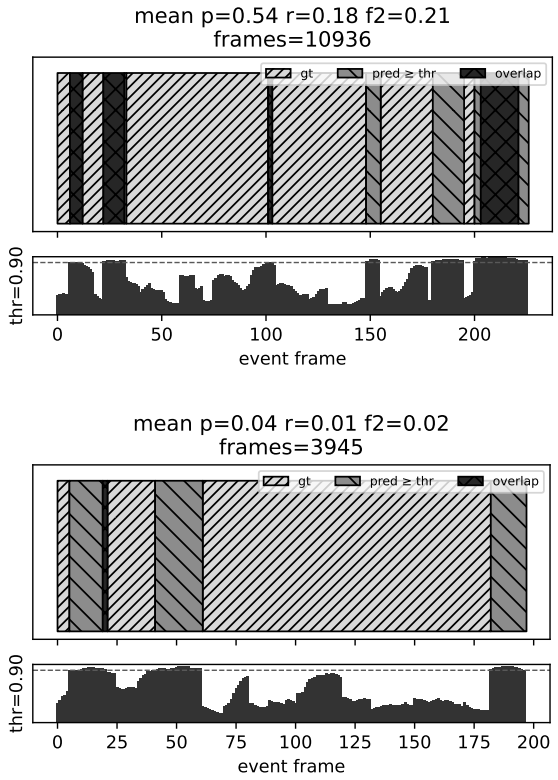



Figure 4: Event-only visualization of frame-level predictions for two SSLC clips using mean probabilities. For each clip, the upper panel shows contiguous segments of frames where the ground truth is positive but the prediction is negative (gt), the prediction exceeds the threshold but the ground truth is negative (pred $\geq$ thr), and frames where prediction and ground truth overlap (overlap). The lower panel shows the corresponding frame-level confidence scores for the same displayed frames, with the decision threshold indicated by the dashed horizontal line. True negatives are omitted; therefore, the x-axis is labeled "event frame" and represents only frames for which either the prediction or the ground-truth label is positive. The title above each plot reports clip-level precision, recall, $F_2$ score, and the total number of frames in the full clip.

transfer scenario we tested, where performance dropped, and therefore also the proposed annotation benefit. While both neural models still were more annotation efficient than the baseline, the gap between them narrowed. This might be because of several factors. Firstly, the length of the median headshake is vastly different from our two datasets. In DGS it is 0.8s, 0.4s for SSLC, which could be cross-lingual difference, or just as likely being due to the difference in annotation goals (i.e. having a different *definition* of a headshake). In a similar line of thought, it could be that the neural models overfit to other aspects of headshakes used as negation

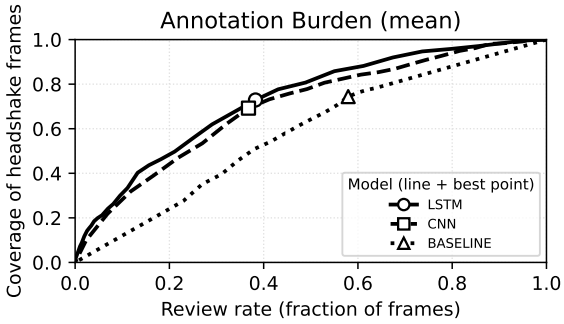



Figure 5: Proportion of required frame review rate of the SSLC data compared to headshake frame recall, utilizing mean inference pooling across frames.

in DGS; such as the amplitude and frequency, for example. One possibility is that headshakes associated with negation in DGS exhibit distinct phonological patterns that are distinct from headshakes more broadly. If so, the models may be capturing these phonological features rather than learning more generalizable representations. Should this interpretation hold, the most direct remedy would be the collection of additional, more diverse training data.

Insufficient tuning to the new domain could also be an explanation for the poor transfer performance. We tested against the best classification threshold from training (0.9). But we see in Figure 4 that many of the ground truth headshake frames would have been captured with a lower threshold. In fact, if we set the threshold to 0.5 we effectively "flip" the precision and recall, and get a precision of 0.11 and recall of 0.36 for the mean-pooled LSTM in the transfer scenario. While we can't tune the threshold like this automatically in a transfer scenario (since we would have no ground truth to tune it to), it does show the effect a little bit of human intervention can have on this task.

Also, a more detailed hyperparameter search could improve performance further; for example utilizing a more complex model design, attention, more layers. We used a simple feature set. Including more hand-crafted features, e.g. from the frequency domain (like Fast Fourier Transform), could be a better fit for modeling headshakes.

The results hint at several future research paths. Firstly, given that the high-confidence false positives predictions contained headshakes, introducing a human in the loop to manually re-label or enhance the DGS dataset for general headshakes seems feasible. A human in the loop could also alleviate the major performance drop seen during our testing of cross-lingual transfer. We could consider an active learning approach to maximize annotation efficiency and increase dataset size: querying the model for least confident negative/positive predic-

tions and manually labeling them. In other annotation workflows, this can drastically lower annotation time (Tharwat and Schenck, 2023).

## 7. Conclusion

This study demonstrates that pose-based models can support the automatic detection of grammatical headshakes in German Sign Language, with neural architectures outperforming a simple baseline and reducing annotation effort. While precision remains limited, error analysis shows that many predicted instances were consistent with headshake movements. The models narrow the search space for human annotators, showing their practical application for corpus construction and typological research. However, the sharp decline in cross-linguistic transfer underscores the sensitivity of such systems, suggesting that adaptation or human-in-the-loop strategies will be necessary for broader applicability in cross-lingual studies. Overall, the findings point to the promise of pose-driven pipelines as scalable tools for investigating headshakes in sign languages.

## Limitations

Several limitations constrain the generalizability of these findings. First, the models rely solely on yaw-based features, potentially overlooking multimodal cues such as facial expressions or body movement that contribute to headshake interpretation. Second, annotations in the DGS corpus target grammatical negation rather than general headshakes, creating a label mismatch that complicates evaluation and may underestimate true performance. The dataset split is not fully signer-independent, leaving residual risk of overfitting to individual motion patterns. Cross-linguistic transfer results are based on a small Swedish dataset with differing headshake durations and definitions, limiting strong conclusions about generalization. Additionally, the models require threshold tuning and hyperparameter exploration that were only partially optimized.

## Acknowledgments

This work has been funded by the Swedish national research infrastructure Språkbanken, in turn jointly funded by its 10 partner institutions and the Swedish Research Council (2025–28, grant 2023-00161). We thank the anonymous reviewers for their insightful comments, which helped improve the paper. Additionally, Marc Schulder for providing helpful comments about formatting for the final paper. We also thank Robert Östling for his input, as well as endless patience during discussions.

## Code Availability

Code is publicly available at:
https://github.com/skogsgren/shakepose

## 8. Bibliographical References

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device real-time body pose tracking.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anastasia Chizhikova and Vadim Kimmelman. 2022. Phonetics of negative headshake in Russian Sign Language: A small-scale corpus study. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 29–36, Marseille, France. European Language Resources Association.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with Sign Language datasets for Sign Language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France. European Language Resources Association.

Trevor Johnston. 2018. A corpus-based study of the role of headshaking in negation in Auslan (Australian Sign Language): Implications for signed language typology. *Linguistic Typology*, 22(2):185–231.

Amanda Kann. 2025. Are translated texts useful for gradient word order extraction? In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 177–182, Vienna, Austria. Association for Computational Linguistics.

Vadim Kimmelman, Anastasia Bauer, Carl Börstell, Jan Bulla, Allah Bux, Laurence Crettenand, Lorena Figueiredo, Anna Kuder, Hannah Lutzenberger, Marloes Oomen, and Roland Pfau. in preparation. Kinematics of negative headshake in seven Sign Languages. Unpublished manuscript, available upon request.

Vadim Kimmelman, Marloes Oomen, and Roland Pfau. 2024. Headshakes in NGT: Relation between phonetic properties & linguistic functions. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 159–167, Torino, Italia. ELRA and ICCL.

Anna Kuznetsova and Vadim Kimmelman. 2024. Testing MediaPipe holistic for linguistic analysis of nonmanual markers in Sign Languages. Preprint. arxiv:2403.10367.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.

Bahtiyar Makaroğlu. 2021. A corpus-based typology of negation strategies in Turkish Sign Language. *Dilbilim Araştırmaları Dergisi*, 32(2):111–147.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020. Real-time Sign Language detection using human pose estimation. In *European Conference on Computer Vision*, pages 237–248. Springer.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Roland Pfau. 2015. The grammaticalization of headshakes: From head movement to negative head. In Andrew D.M. Smith, Graeme Trousdale, and Richard Waltereit, editors, *New Directions in Grammaticalization Research*, pages 9–50. John Benjamins Publishing Company.

Roland Pfau and Josep Quer. 2010. Nonmanuals: their grammatical and prosodic roles. In Diane Brentari, editor, *Sign Languages*, Cambridge Language Surveys, pages 381–402. Cambridge University Press.

Wim Pouw. 2025. EnvisionBox: End-to-end human behavioral classification using Convolutional Neural Networks. Software. GitHub repository. Commit 4f01dd6.

Cas van Rijbroek. 2023. Automatically detecting head-shakes in NGT conversations. Master's thesis, Radboud University Nijmegen.

Allah Bux Sargano, Sébastien Vandenitte, Tommi Jantunen, and Vadim Kimmelman. 2024. Evaluation of head pose estimation algorithms for Sign Language analysis. In *2024 International Conference on IT and Industrial Technologies (ICIT)*, pages 1–6. IEEE.

Jungpil Shin, Akitaka Matsuoka, Md. Al Mehedi Hasan, and Azmain Yakin Srizon. 2021. American Sign Language alphabet recognition by extracting feature from hand pose estimation. *Sensors*, 21(17).

Alaa Tharwat and Wolfram Schenck. 2023. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics*, 11(4):820.

Ulrike Zeshan. 2004. Hand, head, and face: Negative constructions in Sign Languages. *Linguistic Typology*, 8(1):1–58.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*, 56(1).

## 9. Language Resource References

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner,

Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebä rdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release. Dataset. Universität Hamburg.

Zrajm Öqvist, Nikolaus Riemer Kankkonen, and Johanna Mesch. 2020. STS-korpus: A Sign Language web corpus tool for teaching and public use. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 177–180, Marseille, France. European Language Resources Association (ELRA).

# Learning to Spot Signs from Named Entities. A study on French Sign Language.

**Julie Halbout[1] Annelies Braffort[1] Michèle Gouiffès[1] Diandra Fabre[2] Julie Lascar[1]**

[1]Univ. Paris-Saclay. CNRS. LISN, [2]Univ. Grenoble Alpes. CNRS. GIPSA-Lab

[1]91400, Orsay, France, [2]38000, Grenoble, France

[1]{julie.halbout, michele.gouiffes, annelies.braffort, julie.lascar}@lisn.upsaclay.fr

[2]diandra.fabre@gipsa-lab.grenoble-inp.fr

**Abstract**

French Sign Language (LSF) is a low-resourced language, with few available corpora, most of which being only partially annotated. Previous work on other sign languages has explored automatic sign annotation using subtitles as weak supervision, existing signaries, or mouthing cues. This paper focuses on the corpus Matignon-LSF, by first leveraging lexical token spotting then by studying Named Entities (locations, companies, persons). Accounting for the named entities enables the automatic detection of 30% to 100% more signs per class and improves the spotting of rare signs. In addition, this work provides insights into the signing of named entities and contributes resources for improving LSF-to-French translation models.

## 1. Introduction

Sign languages (SLs) are natural languages that rely on visual-gestural modalities rather than vocal or written forms. They are characterized by complex spatial and temporal structures, the use of co-occurring facial expressions and body movements, and non-monolinear syntax that differ significantly from vocal languages. French Sign Language (LSF) is considered as a low-resourced language (lack of data, technical tools, and linguistic resources). Moreover, only a few studies address LSF processing.

A recent study focused on a co-articulated sign spotting task, *i.e.* spotting of a specific sign, in a continuous LSF sequence (Lascar et al., 2024). This task has led to the creation of an initial signary (set of lexical signs) intended for use in automatic processing tasks. However, it remains relatively small, with 445 classes.

In general terms, sign spotting methods rely primarily on a signary when available, or on parallel data composed of videos and subtitles, where subtitles supervise the sign spotting process. One limitation of this second approach is that the correspondence between signs and words in subtitles is not straightforward. Some studies have already addressed this issue by incorporating word synonyms to enrich the possible links between signs and words in subtitles (Momeni et al., 2022).

In this work, we propose a spotting strategy that builds on named entities[1] in the subtitles to increase the number of signs spotted in the corresponding videos.

[1]real-world entities such as persons, locations, organizations, products, etc.

Our main contributions are threefold. (1) We automatically annotate a LSF dataset called Matignon-LSF with lexical signs using the pre-existing LSF-signary, and release these annotations to the community as the Matignon-signary. (2) We extend the dataset with named entity annotations. (3) We propose a new sign spotting approach that enriches existing lexical categories and uncovers novel sign classes, including rare signs.

The remainder of the paper is organized as follows. The section 2 reviews the related work in sign spotting and named entities in Sign Language Processing (SLP). Section 3 presents the preparation of the data used in the study. Section 4 describes the methodology and Section 5 reports the results.

## 2. Related Work

**Sign spotting**: Sign spotting aims to localize a specific sign in a continuous video. One common approach involves using a signary of isolated sign examples as queries. In this method, each sign in the signary, represented by a short video, serves as a reference to search for a similar sign within continuous video sequences. The search can be performed by computing similarity scores between visual representations of the query signs and the target video; the maximum similarity value is used to localize potential occurrences of the sign (Momeni et al., 2022). When multiple occurrences are available for a given sign, a classifier can be trained and used to detect signs directly (Jiang et al., 2021). Additional cues, such as mouthing, can be leveraged to improve spotting performance (Varol et al., 2022). When no signary is available, subtitles can serve as weak



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 203–211
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

supervision by providing textual cues that indicate which video segments are likely to contain a given sign (Momeni et al., 2022; Lascar et al., 2024).

**Named entities recognition**: A named entity (NE) is a sequence of words or signs that refers to an identifiable entity in the real world, such as a person, a location or an organization. Named entity recognition (NER) is the process of detecting and labeling real-world entities in text. Several methods have been proposed in Natural Language Processing (NLP) (Keraghel et al., 2024). Recognizing named entities is crucial for enabling tasks like machine translation, question answering, and information retrieval. Although the field of sign language processing (SLP) has not been explored as extensively as NLP, several authors have nevertheless pointed out that named entities present significant challenges for translation (Camgöz et al., 2020), particularly with regard to localization or anonymization (Bleicken et al., 2016; Yin et al., 2021). Yin et al. (2021) pointed out that, in SLs, named entities may be conveyed either through fingerspelling, lexical sign, or by mouthing the name while pointing the spatialized referent with the index finger. The present research aims to determine how NE are signed in LSF and whether detecting NE in the subtitles provides useful cues for sign spotting.

## 3. Datasets and Data Preparation

### 3.1. Datasets

Two datasets are used in this work, Mediapi-RGB and Matignon-LSF, composed of LSF videos with the corresponding subtitles. They are respectively coming from native LSF (then translated into French subtitles), and interpreted LSF (from a spoken content). They are currently the largest corpora available for SLP in LSF.

The **Mediapi-RGB** (Ouakrim et al., 2024) corpus originates from Médiapi[2], a bilingual online media established in 2018. It contains news and reports in LSF produced by deaf journalists and subtitled in French, covering the period from 2018 to 2022. The corpus contains 86 hours of LSF video data, segmented into 50,084 video excerpts (average duration: five seconds), each aligned with its respective French subtitle (average length: 12.2 words).

Lexical units were extracted from the Mediapi-RGB corpus using the method described by Lascar et al. (2024). This process produced approximately 15k video instances of lexical signs distributed across 445 classes, covering various semantic categories such as locations, numbers, sports, and politics. Each video clip has a resolution of $444 \times 444$ pixels, a frame rate of 25 fps, and contains between 3 and 46 frames (average: 12.5 frames). In this work, we refer to this collection of lexical signs as the *Mediapi-signary*.

**Matignon-LSF** (Halbout et al., 2024) is an interpreted LSF dataset, containing 39 hours of video, and a vocabulary size of 10k subtitles across 18k sentences and 450k tokens. It features 15 LSF interpreters. It comes from the French government's Council of Ministers debriefings (from December 2020 to December 2023) and can be further extended[3]. Videos and subtitles are not aligned, there is a delay of a few seconds. For the sake of annotation, the sentence-like video segments are extracted following the subtitles sentences timestamps. This results in 18,173 videos. For our work, a 5-seconds margin is added at the beginning and at the end of each video to ensure good correspondence with the subtitle. This dataset is used in our study for signs spotting.

### 3.2. Data Preparation

We pre-processed the videos and subtitles, and identified the signers automatically to accommodate signer-dependent variations in signs.

**Video preprocessing.** All videos were encoded using a Video Swin Transformers (VST) model (Liu et al., 2022), initialized with publicly available Kinetics-400 pretrained weights (Kay et al., 2017) and subsequently fine-tuned on BOBSL (Albanie et al., 2021). A sliding window of 16 frames with a stride of 1 was used to encode the videos. Each sign of the Mediapi-signary is represented by a sequence of 16 frames, using zero-padding when the sign is shorter or truncation when it is longer, which occurs rarely. The 16-frames videos are then preprocessed the same way, and we obtain a single vector for each video.

**Subtitles preprocessing.** The Matignon-LSF's subtitles were normalized by lowercasing the text, converting cardinal numbers to digits, and expressing ordinal numbers in words.

**Signer recognition.** Unsupervised clustering was used to identify each signer in the Matignon-LSF dataset. We extracted the first frame from each video and embedded it using FaceNet512, an improved version of FaceNet (Schroff et al., 2015) ranked first in the DeepFace benchmark (Firmansyah et al., 2023). Embeddings were reduced to 100 dimensions via PCA and clustered with HDBSCAN (McInnes et al., 2017). The most relevant clusters were manually selected, and embeddings were classified using a Euclidean distance threshold of 1.3. When new signers appeared among unclassified videos, new classes were manually created. Iterating this process while changing the

[2] https://www.media-pi.fr/

[3] Debriefings of the Council of Ministers take place once a week.

frame representing each video led to the identification of 16 signers.

## 4. Sign Spotting Methods

Three methods have been used for automatically annotating the videos, illustrated by Fig. 1. Lexical signs were annotated using sign spotting from Mediapi-signary examples (Fig. 1 (a,b)), as explained in section 4.1. A second sign spotting technique (Fig. 1 (c)) consisted in pairing features from videos the subtitle of which contain a similar word (section 4.2). This technique was applied for NER, whose annotation is described in section 4.3. Section 4.4 proposes to use NE characteristics to better target the search (Fig. 1 (d)), allowing to collect rare signs.

### 4.1. Signary Spotting

The method is inspired from Momeni et al. (2022) and illustrated by Fig. 1 (a) and 1 (b). As Matignon-LSF does not contain annotation of mouthing cue, we focus on the subtitles as weak supervision for the spotting. From the Mediapi-signary, we select one class composed of a label, that is a French word, e.g. "sept" (*eng: seven*), associated to a collection of videos (e.g. the video A on Fig. 1 (a)). We use the label as the query to search within the subtitles. In Matignon-LSF, we select each video the subtitle of which contains the query word, called token *t* (e.g. "sept"). The hypothesis is that the corresponding lexical sign is present in video B. To find this occurrence, the features of A are compared to each feature of B using a sliding window. To that aim, the cosine similarity is computed between the VST features of A and B, which yields a similarity vector (represented by a heatmap in Fig.1).

A hysteresis thresholding is applied to segment the sign. The query sign A is considered to be found in B when a similarity score exceeds 0.5. Then, to extract the full sign, the segment is expanded to the right and left as long as the neighboring similarity scores exceed 0.45.

### 4.2. Pairing Videos to Annotate New Signs

The second annotation method, illustrated by Fig. 1 (c) comes from (Momeni et al., 2022; Lascar et al., 2024). It was used to create the Mediapi-signary. In that case, no query example is available, only the subtitles are used as weak supervision.

Cosine similarity is computed between videos containing a similar token *t*, and between a set of negative samples. The negative samples are used to unmatched similar frames that are not our target token such as pointing gestures that are common in SL contents. The final similarity is obtained by maximizing the contribution of positive pairs and minimizing that of negative pairs. To handle differences in signing style, videos were clustered by signer prior to similarity computation. This reduces variability in sign forms and leads to more reliable results. At least 9 postive samples (containing the token *t*) and 27 negative samples (not containing the token $t$) are needed to discover new signs. Following recommendations from (Momeni et al., 2022), we remove the 5s video margins (see section 3.1) to have more accurate negative samples. The same hysteresis thresholding as 4.1 is applied. The sign is considered a possible spot when the resulting segment lasts at least 3 frames. The Fig. 1 (c) illustrates the methods for the target sign PAYS.

### 4.3. Named Entities Annotation

Both previous methods rely on the token *t* as weak supervision for the spotting. However, not necessarily all words in spoken (and written) french are signed in LSF. Similarly, some signs have no direct counterparts in the subtitles. We have investigated the named entities by analyzing the annotations performed on the LSF 40 Brèves dataset[4](Challant and Filhol, 2022), and confirmed our findings through qualitative observations on the Matignon-LSF dataset. It appears that NE can be introduced by an hypernym sign (COUNTRY/LOCATION for a country, COMPANY for a company etc.) that are absent of the textual content. In order to confirm and quantify these signing patterns, we carried out the following experiment. We focused on geopolitical named entities (GPE), compagnies (ORG) and person (PERS) entities. GPE offer relevant examples in a constrained domain and allows for rapid and efficient manual annotation in the subtitles. On the other hand, compagnies and person entities requires automatic annotation due to the open vocabularies.

For the sake of clarity in the rest of this article, we introduce the following notations. We use a French word in small capital letters to designate the LSF signs, quotation marks for the words or excerpts from the subtitles, and the word *label* as an index to designate the name of the corresponding class in the Mediapi-signary (e.g. PAYS (*country*) refers to the sign, "pays" to the word, and $\text{pays}_{label}$ to the Mediapi-signary class).

The hypothesis made is that entities may be introduced by specific signs depending on the category forming a specific pattern composed of a minimum of two successive signs and denoted NE-PATTERN: the first sign introduces the NE category and is

[4] 1 hour of journalist content fully annotated, including complex linguistic structures

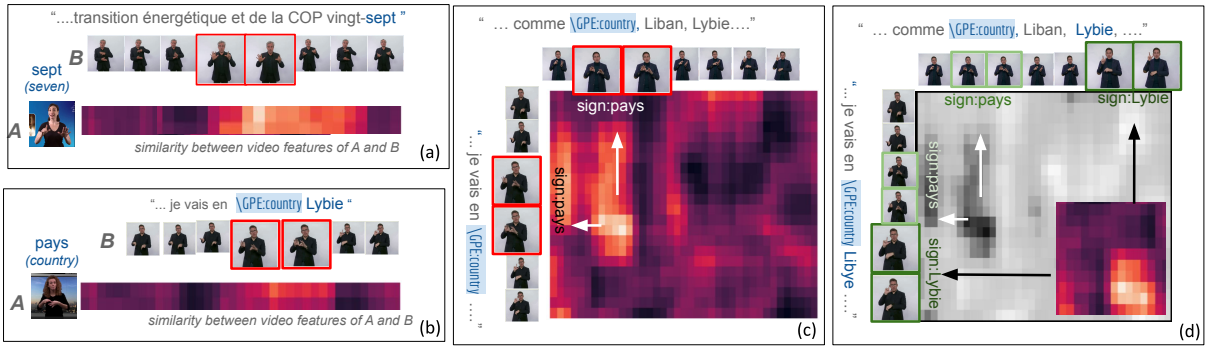



Figure 1: Illustration of the signary spotting (a,b) to find lexical signs associated to a word (a) or to a tag GPE:COUNTRY (b) in subtitles; (c) pairing to discover the signs corresponding to a word / tag present in the two subtitles; (d) targeted pairing where the sign PAYS (country) is detected and the sign corresponding to Lybia is searched locally with a higher precision.

more likely to be an hypernym, and the second refers to the entity. The realization of our example with the Lybia country would be written as follows: PAYS + LIBYE.

In this study, we have selected three GPE categories related to a country, a region and a department and two other categories: company and person. First of all, all the subtitles have been annotated manually with the following tags, written in capital letters: GPE:COUNTRY for a country, GPE:REG for an administrative region, GPE:DEP for an administrative department. For instance, the subtitle (translated in English) *"I am going to Libya."* will become *"I am going to GPE:COUNTRY Libya."*("Je vais en GPE:COUNTRY Lybie."). Concerning the companies (ORG) and persons (PER) annotation, to allow a rapid annotation, we use SpaCy[5] model, followed by a quick manual correction to remove major errors. We specifically labeled organizations (ORG) while restricting them as much as possible to companies[6].

To avoid biasing our study, we excluded the subtitles containing the words "pays", "région", "département", "entreprise" and "personne". For example, the subtitle *"I am leaving our **country** to go to GPE:COUNTRY Libya."* would be excluded ("Je pars de notre pays pour aller en GPE:COUNTRY Lybie."). This exclusion is motivated by the need to avoid duplicate spotting and to verify if the hypernym signs are present in the signing sequence while absent from the subtitles. Thus, 416 GPE:COUNTRY, 99 GPE:DEP, 58 GPE:REG, 2313 PER and 813 ORG were annotated [7] in the subtitles.

[5] https://spacy.io/, $fr_core_news_sm model$

[6] Examples of excluded entities include universities, NGOs, and government organizations, as they do not fall under the hypernym COMPANY

[7] Manual annotation was needed for GPE NE because SpaCy only provides general location NE, without distinguishing countries or departments.

Assuming that named entities can be explicitly introduced in LSF (e.g. PAYS for countries), the pairing method described in section 4.2 is leveraged using named entities tags to localize signs. Then, if these introducing signs belong to the Mediapi-signary, the signary-spotting of section 4.1 is used in a similar way as for lexical signs, but using a tag as query instead of a token. As illustrated by Fig. 1 (b), the lexical sign PAYS is used to find the sign associated to the GPE:COUNTRY tag. In the example, we look for PAYS whereas the word "pays" does not appear in the subtitle.

### 4.4. Targeted Pairing

The observation of the pattern sequence for the named entities inspired us a new method to extract new occurrences of signs, for example signs that are rare in the dataset and unknown in the Mediapi-signary. Indeed, the pairing method of section 4.2 preferably requires a minimum of 9 sign occurrences per class in the signary. Unfortunately LYBIE among others appears only three times in the subtitles. Thus, we propose to apply a *targeted pairing*, illustrated in Fig. 1 (d). The idea is to find the sign associated to the GPE tag (e.g., PAYS for GPE:COUNTRY) and to use it as an anchor. The sign of the entity is then searched by using pairing between two videos containing the sign to be found. Compared to the previous pairing described in section 4.2, the searching area is reduced in size, since the sign is searched right after the anchor. By reducing the area search, the search yields fewer ambiguities compared to the previous pairing, thus only 2 occurrences signed within the NE-PATTERN are enough instead of 9.

## 5. Results and Evaluation

After a description on the evaluation metrics in 5.1, section 5.2 provides the annotation results for to-

kens and NE. The section 5.3 focuses on the evaluation of two spottings methods: signary spotting and pairing videos. Finally, section 5.3 provides insight into the targeted pairing method and then focuses on the named entities pattern evaluation.

## 5.1. Metrics and Evaluation Procedure

First of all, duplicates caused by 5s video margins were automatically removed and visually verified. The number of reference annotations (ground truth **GT**) is obtained in 2 ways. When feasible (i.e. for GPE:REG and GPE:DEP for which the number of occurrences was limited), occurrences have been checked visually: $\mathbf{GT}_{vid}$. Otherwise, it is given by the number of words/tags in the subtitles: $\mathbf{GT}_{sub}$. Among the **N** signs detected, True Positive (**TP**) yields from visual verification by an LSF expert. Thus, the precision **P** is computed as **TP**/**N**. The recall $\mathbf{R}_{vid}$ (respectively $\mathbf{R}_{sub}$) is computed by **TP**/$\mathbf{GT}_{vid}$ (respectively **TP**/$\mathbf{GT}_{sub}$). The F1 score $\mathbf{F1}_{vid}$ is computed when available.

To go further, we also visually analyzed, for each GPE category, whether the signs spotted from the **TP** belong to its associated NE-PATTERN ($\mathbf{TP}_{pattern}$). When feasible (i.e. for GPE:REG and GPE:DEP), we generate a ground truth to indicate the presence of each pattern ($\mathbf{GT}_{pattern}$) in the video.

## 5.2. Quantitative Results

**Token-based sign spotting.** Using the sign spotting method described in section 4.1, *254* classes with **N**=*6,372* signs automatically extracted from Matignon-LSF. Those signs were evaluated by an expert, leading to **TP**=4,545, with a precision **P**=0.71 and $\mathbf{R}_{sub}$ =0.3.

This method does not account for ambiguities arising from the polysemy of certain signs. Consequently, relying solely on subtitles to detect signs is somewhat limited, which may partially explain the low $\mathbf{R}_{sub}$ scores. As an example, in French, the word “nord” (*north*) appears 24 times in the Matignon-LSF subtitles, with different meanings: referring to a French administrative department, a cardinal point, but also a part of the name of a French newspaper (“La voix du Nord”), 4, 11 and 9 times respectively. Using examples of $\text{nord}_{label}$ from Mediapi-signary, only 4 occurrences were retrieved, all related to the cardinal point. The newspaper and the department were signed differently, which illustrates the challenge posed by polysemy in sign detection.

**Named-entities-based sign spotting.** Since the subtitles have been tagged with NE categories, the pairing method of section 4.2 is used to retrieve the corresponding signs. Matching videos containing similar GPE tags results in finding signs corresponding to PAYS (9 occurrences), DÉPARTEMENT (5), RÉGION (2), ENTREPRISE (19) and PERSONNE (2). In a second stage, we use the classes $\text{pays}_{label}$, $\text{departement}_{label}$, $\text{region}_{label}$, $\text{entreprise}_{label}$ and $\text{personne}_{label}$ (present in Mediapi-signary) as queries for spotting (see section 4.1). The table 1 (last line) displays the results for all NE categories. The precision score (0.79) is high in contrast to the low recall $\mathbf{R}_{sub}$ score (0.21). This Table also reports the number of correct occurrences (**TP** = 768). By adding NE tags in the subtitles, the number of annotations is increased. The table also reports the results (indicated by an asterisk) of an alternative annotation approach from (Lascar et al., 2024), based on an MLP. More signs are correctly annotated for the ‘department’ and ‘region’ categories, but fewer signs are detected for the three other categories.

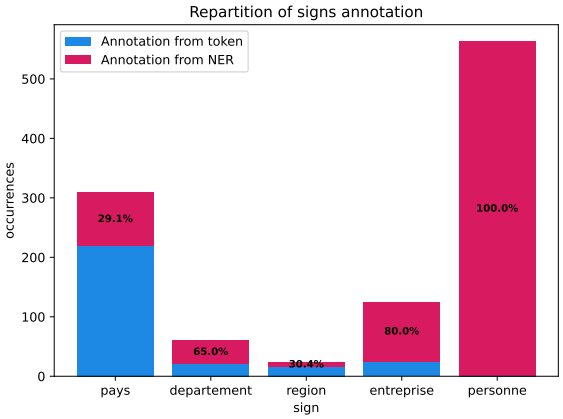



Figure 2: Number of signs after visual inspection (PAYS, RÉGION, DÉPARTMENT, ENTREPRISE and PERSONNE) detected using token (in blue) and NE (in red).

Fig. 2 displays, for each of the five categories, the proportion of signs spotted respectively using the token-based method (blue bar) and NE-based one (red bar). Respectively 29%, 30%, 65%, 80% and 100% more annotations are found for PAYS, RÉGION, DÉPARTEMENT, ENTREPRISE and PERSONNE. The signs RÉGION, DÉPARTEMENT were visually checked ($\mathbf{GT}_{vid}$) to compute F1 (0.78) and $\mathbf{R}_{vid}$ (0.7), which can be considered as solid.

## 5.3. Qualitative Evaluation

**Discovering novel classes.** Inspired by the qualitative analysis, we proposed a method in Section 4.4 to discover signs absent from the Mediapi-signary and thus increase the number of classes. The approach uses category signs (e.g., country, department) as anchors to perform targeted searches for subsequent signs. New classes must meet three criteria: absence from the Mediapi-signary, at least two NE-PATTERN annotations, and fewer than nine subtitle occurrences. We applied

| sign | $\mathbf{GT}_{sub}$ | $\mathbf{GT}_{vid}$ | N | TP | P | $\mathbf{R}_{sub}$ | $\mathbf{R}_{vid}$ | $\mathbf{F1}_{vid}$ | $N^*$ | $TP^*$ |
|---|---|---|---|---|---|---|---|---|---|---|
| PAYS | 416 | NA | 115 | 90 | 0.78 | 0.22 | NA | NA | 13 | 13 |
| RÉGION | 58 | 10 | 8 | 7 | 0.88 | 0.12 | 0.70 | 0.78 | 66 | 56 |
| DÉPARTMENT | 99 | 54 | 45 | 39 | 0.87 | 0.40 | 0.72 | 0.79 | 189 | 158 |
| PERSONNE | 2313 | NA | 647 | 563 | 0.87 | 0.24 | NA | NA | 2 | 0 |
| ENTREPRISE | 813 | NA | 143 | 100 | 0.70 | 0.12 | NA | NA | 108 | 86 |
| total | 3699 | NA | 967 | 798 | 0.79 | 0.21 | NA | NA | 378 | 313 |

Table 1: Evaluation of the spotting of the GPE category signs. $\mathbf{GT}_{sub}$ is the number of tags in the subtitles, $\mathbf{GT}_{vid}$ is the number of GPE category signs in the video (manually annotated), **N** the total number of signs spotted (True Positive + False Positive) and **TP** the true positive. The scores **P** refer to the precision, $\mathbf{R}_{sub}$ to the recall from subtitles ($GT_{sub}$) and $\mathbf{R}_{vid}$ to the recall from the video ($\mathbf{GT}_{vid}$). The $\mathbf{F1}_{vid}$ is the F1 score with **P** and $\mathbf{R}_{vid}$. Results with $^*$ are obtained using an alternative MLP detector.

this method to two countries (Libya, Belarus) and one department (Landes). By computing cosine similarity and locating the nearest similarity blocks after the anchors, we retrieved several occurrences of these signs.

**Country geopolitical entity.** True positives of PAYS were manually analyzed to assess their conformity to the GPE:COUNTRY-PATTERN. Among 90 true positives, 61 matched the pattern, while the remaining 29 require further investigation. Based on 416 GPE:COUNTRY tags in the subtitles, only 14.67% followed this pattern. No full ground truth was established; instead, partial manual checks were conducted. Some countries are signed with specific signs, while others follow the GPE:COUNTRY-PATTERN by introducing the entity by the sign PAYS. Rarely mentioned countries (e.g., Lebanon, Iran, Belarus, Serbia) consistently follow the pattern, whereas frequently referenced countries, such as France's neighbors, influential states, or countries prominent in the news, tend not to. Several limitations were identified, including missed detections, list effects, fingerspelling, absence of signs, distinctive signs, and automatic filtering due to time margins. These factors may explain missing cases in the detected pattern. Overall, well-known countries appear less likely to be introduced by the sign PAYS and therefore less likely to conform to the GPE:COUNTRY-PATTERN. Future work should investigate the factors influencing the use of this pattern.

**Region geopolitical entity.** For GPE:REG, GPE:REG-PATTERNS were manually annotated in the corresponding videos, constituting the ground truth. Among the 7 TP occurrences, 5 of them conform to the GPE:REG-PATTERN, forming the $\mathbf{TP}_{pattern}$. Results are display in Table 2. The two remaining RÉGION are signed to add context in the signing sequences.

**Department geopolitical entity** Here also a GT is created for GPE:DEP-PATTERNS. Table 2 shows that nearly 31 DÉPARTEMENT out of the 99 of the GPE:DEP annotations are introduced by the hypernym sign DÉPARTEMENT. Otherwise, DÉPARTEMENT is used to provide additional context.

According to our observations, departments are signed in three ways: 1) use of its code[8] (70 occurrences among which 31 introduced by the sign DÉPARTEMENT); 2) fingerspelling (2 occurrences systematically introduced by DÉPARTEMENT); 3) a specific sign (12). The absence of the DÉPARTEMENT sign before the department's code is partly explained when several departments are listed, as shown by Table 3.

**Company and person entity** No complete manual observation was made for this two categories. **Company entity**: By inspecting references to SNCF (national railway company), we were able to retrieve the ORG-PATTERN. Out of the 12 mentions in the datasets, seven of them were preceded by the sign ENTREPRISE. Two of the mention in the subtitles are not signed (due to the context where the company has been already mentioned). Finally, three of the mentions in the subtitles are just signed with GARE (station) or fingerspelled, so without the scope of the ORG-PATTERN. Nevertheless, entities annotations remains partial, and the results could be improved through more thorough labeling. **Person entity**: This sign has not been spotted using only the subtitles. However, we were able to annotate more than 500 occurrences within the automatic named entities annotations. Due to the amount of annotation, we only analyze the PER-PATTERN for the entity *French President*: 215 out of the 725 mentions of the french president are preceded by PERSONNE. Further analysis could lead to understand how and when the pattern is used to mention someone.

## 6. Conclusion

This paper investigated three categories of named entities (locations, organizations and persons) and the specificity of their representation in French Sign Language (LSF), in particular the use of a "department" sign followed by a number for most French departments. This study provides several promis-

[8] From 01 to 95 for hexagonal France

| NE-PATTERN | $TP_{pattern}$ | $GT_{pattern}$ | $GT_{sub}$ |
|---|---|---|---|
| GPE:COUNTRY-PATTERN | 61 | NA | 416 |
| GPE:REG-PATTERN | 5 | 7 | 5 |
| GPE:DEP-PATTERN | 31 | 39 | 99 |

Table 2: Number of retrieved patterns ($\mathbf{TP}_{pattern}$) compared with the manually annotated ground truth ($\mathbf{GT}_{pattern}$) for the NE-PATTERN

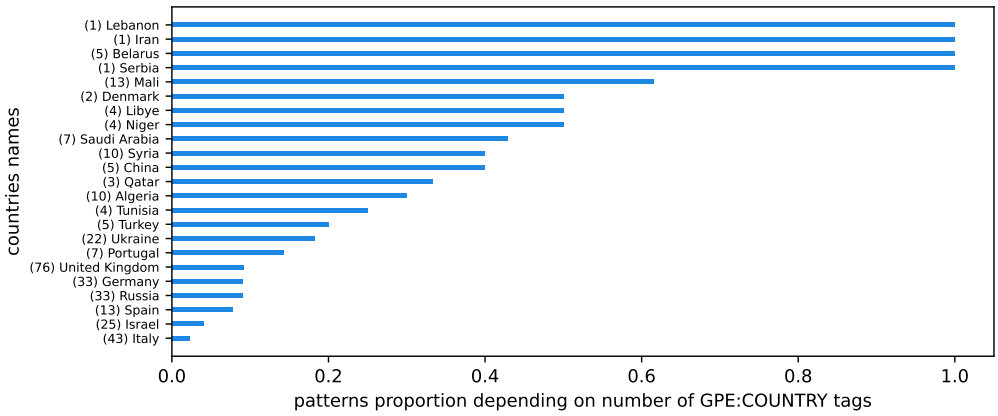



Figure 3: Proportion of GPE:COUNTRY-PATTERN: Number of GPE:COUNTRY-PATTERN instances over the total number of named entities GPE:COUNTRY in the subtitles.

Table 3: A subtitle with a list of departments from the V6VNZD5jLH8_segment_14 video excerpt: green ones conform to the GPE:DEP-PATTERN, blue ones are signed without being preceded by DÉPARTEMENT.

| |
|---|
| L'amélioration est continue, proche de la tendance nationale, dans l'Aube/GPE:DEP/NUM:10/GPE:DEP-PATTERN, dans le Rhône/GPE:DEP/NUM:69, en Île de France/GPE:REG, plus marquée dans la Nièvre/GPE:DEP/NUM:58, dans l'Eure/GPE:DEP/NUM:27, dans les Alpes Maritimes/GPE:DEP/NUM:06, et au contraire un peu plus lente en Seine Maritime/GPE:DEP/NUM:76, dans l'Aisne/GPE:DEP/NUM:02 et dans le Nord/GPE:DEP/NUM:59. |

ing research directions: 1) enriching video annotations by enabling the detection of signs that are not present in subtitles (e.g., "department"). Our experiments showed an increase ranging from 30% to 100% depending on the category; 2) disambiguating certain translations from LSF into French (Ouakrim et al., 2024), where "department 17" can be correctly interpreted as "Charente-Maritime" rather than "17 departments"; 3) targeting the automatic retrieval of rare signs.

Finally, this work enabled a first step toward automatic annotation of the Matignon-LSF corpus, by exploiting both the lexicon present in the subtitles and our analysis of named entities.

## 7. Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011015441).

## 8. Ethical Consideration

Matignon-LSF is an interpreted sign language dataset. Recently, (Desai et al., 2024) pointed out the ethical problem of using interpreted dataset in the development of AI models since they differ from native SL and can lead to inaccuracies. Their use are nonetheless helping in the context of low-resourced data as those data are easy to collect, but the difference as to be acknowledged. Note also that none of the authors is deaf but some of them use LSF on a daily basis with Deaf colleagues, and the team's projects are carried out using an interdisciplinary approach (computer science and linguistics) and in collaboration with deaf organizations.

## 9. Bibliographical References

Julian Bleicken, Thomas Hanke, Uta Salden, and Sven Wagner. 2016. Using a language tech-

nology infrastructure for German in order to anonymize German Sign Language corpus data. In *International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3303–3306, Portorož, Slovenia. European Language Resources Association (ELRA).

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Camille Challant and Michael Filhol. 2022. A first corpus of AZee discourse expressions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1560–1565, Marseille, France. European Language Resources Association.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: state of the art and challenges. *Univers. Access Inf. Soc.*, 23(3):1305–1331.

Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. 2024. Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 366–377, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Diandra Fabre, Julie Lascar, Julie Halbout, Yanis Ouakrim, Annelies Braffort, Thomas Hueber, Michèle Gouiffès, and Denis Beautemps. 2025. Exploring Sign-level Strategies to Enhance Automatic Translation of French Sign Language. In *ACM Int. Conf. on Intelligent Virtual Agents*, Berlin, Germany.

Andrian Firmansyah, Tien Fabrianti Kusumasari, and Ekky Novriza Alam. 2023. Comparison of face recognition accuracy of arcface, facenet and facenet512 models on deepface framework. In *International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pages 535–539.

Tao Jiang, Necati Cihan Camgöz, and Richard Bowden. 2021. Looking for the signs: Identifying isolated sign instances in continuous video footage. In *IEEE Face and Gesture*, page 1–8. IEEE Press.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Beyond words: A comparative analysis of LLM embeddings for effective clustering. In *Advances in Intelligent Data Analysis XXII - 22nd International Symposium on Intelligent Data Analysis, IDA 2024, Stockholm, Sweden, April 24-26, 2024, Proceedings, Part I*, Lecture Notes in Computer Science, pages 205–216. Springer.

Julie Lascar, Michèle Gouiffès, Annelies Braffort, and Claire Danet. 2024. Annotation of LSF subtitled videos without a pre-existing dictionary. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 100–108, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Liliane Momeni, Hannah Bull, K R Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic Dense Annotation of Large-Vocabulary Sign Language Videos. In *IEEE Computer Vision and Pattern Recognition*, volume 13695, pages 671–690. CVPR.

K R Prajwal, Hannah Bull, Liliane Momeni, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Weakly-supervised Fingerspelling Recognition in British Sign Language Videos. In *British Machine Vision Conference*, London, United Kingdom.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering . In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, Los Alamitos, CA, USA. IEEE Computer Society.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2022.

Scaling Up Sign Spotting Through Sign Language Dictionaries. *International Journal of Computer Vision*, 130(6):1416–1439.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *IEEE/CVF Computer Vision and Pattern Recognition*, pages 16852–16861.

Ryan Wong, Necati Cihan Camgöz, and Richard Bowden. 2023. Hierarchical i3d for sign spotting. In *Computer Vision – ECCV 2022 Workshops*, pages 243–255, Cham. Springer Nature Switzerland.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *ACL-International Joint Conference on Natural Language Processing*, pages 7347–7360, Online. Association for Computational Linguistics.

## 10. Language Resource References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv preprint arXiv:2111.03635*.

Julie Halbout, Diandra Fabre, Yanis Ouakrim, Julie Lascar, Annelies Braffort, Michèle Gouiffès, and Denis Beautemps. 2024. Matignon-LSF: a large corpus of interpreted French Sign Language. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 202–208, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Yanis Ouakrim, Hannah Bull, Michèle Gouiffès, Denis Beautemps, Thomas Hueber, and Annelies Braffort. 2024. Mediapi-RGB: Enabling Technological Breakthroughs in French Sign Language (LSF) Research through an Extensive Video-Text Corpus. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: VISAPP*, volume 2, Rome, Italy.

# The Iterative Development and Evaluation Framework for Kazakh-Russian Signing Avatars Targeted to Native Deaf Signers

**Alfarabi Imashev***, **Tohid Alizadeh**†
*National Laboratory Astana, Nazarbayev University,
† Department of Robotics, School of Engineering and Digital Sciences, Nazarbayev University
{alfarabi.imashev, tohid.alizadeh}@nu.edu.kz

**Abstract**

Nowadays, existing research predominantly focuses on already well-researched sign languages. However, the most extensive studies of sign language in Kazakhstan, which adhere to international standards, started about a decade ago. Native deaf signers in Kazakhstan can often suffer from insufficient educational opportunities, which may result in limited reading proficiency too. Sometimes, deaf signers can recognize letters and read words, but they may not fully understand the overall concept and need to break it down into a sequence of simpler ideas to comprehend it better. Consequently, signing avatars have the potential to interpret internet statements, movie subtitles, or YouTube videos, and this sign language production may increase accessibility and improve communication between deaf and hearing individuals, as well as between humans and avatars. An equally critical challenge is how to develop a tool that will help deaf signers evaluate the performance, appearance, and naturalness of signing avatars without relying on written text across all sign languages, particularly in underserved communities. This paper outlines the iterative development of the Kazakh-Russian Sign Language interpreting avatar, ongoing improvements to the evaluation instrument, and a comparative analysis of this instrument with another evaluation method designed to attain the same objective.

**Keywords:** signing avatars, HCI, sign language generation, assessment, questionnaire

## 1. Introduction

Recent advancements in Virtual Avatars (i.e., human-like virtual characters) and testing of their implementations have transitioned these entities from niche gaming studies to mainstream research in education and collaboration (Audrezet et al., 2025; Gasch et al., 2025; Ouhnni et al., 2025). By the same way, all these breakthroughs have the potential to greatly enhance the presence, popularity, and applicability of avatars in the near future for sign languages(Quandt et al., 2022). Generative approaches provide more novel options to enhance the visibility of sign languages, engagement, interactivity, and scalability. Natural language processing implemented for sign languages may allow providing any information to end-users in different languages, including sign languages; and virtual avatars could improve the access to information in sign languages (Naert et al., 2020), especially when there is no physical presence for deaf signers.

Sign language performing characters, or signing avatars, have been a prominent focus in the domains of human-agent interaction (HAI) and human-computer interaction (HCI) research for nearly two decades(Huenerfauth and Hanson, 2009; Kennaway et al., 2007; Lombardo et al., 2010; Kipp et al., 2011b). Due to their promising potential for facilitating communication between deaf and hearing individuals through the translation of written or spoken to sign language and vice versa, they were especially sought after as a valuable tool by mainly hearing researchers.

The research results generally affirm the efficacy and potential of signing avatars in facilitating solving challenges often faced by native deaf signers and improve information accessibility (Kipp et al., 2011a; Adamo-Villani et al., 2016; Bouzid et al., 2016). However, more recent research has identified a problem in the development of suitable avatars that are user-friendly, grammatically accurate, and comprehensible for the deaf (Bragg et al., 2019). Nonetheless, members of Deaf communities often perceive sign language interpreting systems as insufficient (Wolfe et al., 2022).

This counterproductivity, as Wolfe (Wolfe et al., 2022) contends, is likely attributable to the subpar quality of signing, evidenced by the display of robotic gestures that lack both naturalness and clarity. A potential explanation for this may be attributed to the homogeneity of research teams, including mainly hearing researchers and a deficiency of inclusivity, leading to the ongoing challenges faced by deaf sign language users in stating their requirements. So, the involvement of the deaf community is essential in the evaluation of sign language systems, as their insights can address challenges in experimental design (Huenerfauth and Kacorri, 2015; Roelofsen et al., 2021) and identify areas for improvement (Kipp et al., 2011b; Gibet et al., 2011; Schnepp et al., 2012; Roelofsen et al., 2021).

Although quite often bilingual (Gibet et al., 2011),



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 212–225
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

many deaf signers consider the sign language as their first language. Therefore, administering a questionnaire in a written format, as is commonly practiced, is not suitable (Farwell, 1976). Consequently, interaction with the Deaf community must be conducted in their first language (Gibet et al., 2011; Bosch-Baliarda et al., 2019).

Secondly, research indicates that complex and ambiguous concepts present additional challenges for deaf individuals in accurately understanding ideas (Parvez et al., 2019), resulting in a cognitive burden (Bosch-Baliarda et al., 2019). When deaf signers recognize letters and words, they may not fully understand the concept itself; therefore, they have to deconstruct it into several simpler concepts in order to achieve comprehension. According to the experience of local interpreters, some of them are CODA and familiar with deaf culture; instead of utilizing the Internet or encyclopedias to ascertain the meaning of a sign, deaf signers generally seek clarification from their own social group, including friends or relatives. According to (Morere, 2012), the educational level of young adults who graduated high school is reported to be similar to that of fourth-grade students in the USA. Local sign language interpreters and representatives of organizations for the Deaf report a remarkably similar situation in Kazakhstan and the CIS countries.

It is noteworthy that prior research (decades ago) on sign languages was predominantly concentrated on Western sign languages (such as ASL LSF, and DGS), which have been thoroughly investigated, resulting in an abundance of resources. However, comprehensive studies of sign language used in Kazakhstan, which adhere to international standards, started only about a decade ago (Kudubayeva et al., 2016; Imashev, 2017), including data collection (Imashev et al., 2020) and sign language recognition (Kuznetsova et al., 2021; Duisenbek and Zhukabayeva, 2025), and only recently were the first evaluation studies also conducted (Imashev et al., 2022).

One of the primary obstacles to this work is the lack of consensus on the complexities of conducting evaluation studies with the deaf. A major challenge is the development of deaf-friendly survey metrics to maximize researchers' understanding of the responses of deaf participants. Thus, traditional textual questionnaires, commonly used to conduct surveys with hearing individuals, are often inappropriate for user evaluation studies involving the deaf. This approach can undermine the validity of the feedback obtained for several obvious reasons (Bosch-Baliarda et al., 2019).

This work prioritizes both the selection and design of evaluation instruments and the logistics of user studies, which include field studies with deaf participants.

## 2. Avatar Design

This section describes the iterative evolution of data-driven signing avatars designed for Kazakh-Russian Sign Language performance has undergone four significant modifications due to technological advancements, supported by two research-oriented grants over the past six years, with a focus on enhancing performance plausibility and visual representation (see Figure 1). The rationale for all modifications and enhancements was guided by a combination of feedback provided from the participants of each user study and was implemented incrementally from version to version. All user studies described in the subsequent Section 3 and their outputs are associated with each version of avatar.

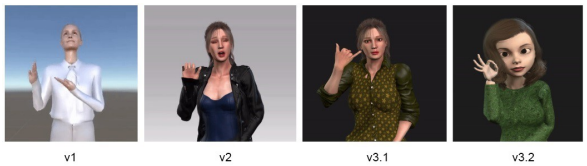



Figure 1: The evolution of Kazakh-Russian Sign Language Avatar development.

v1. The avatar used for the online user study (only flat hand configurations and no facial emotions were available for this version); v2. The avatar used for the first in-person user study (hand configurations, facial mimics, mouthing); v3.1 The updated avatar used for the second user study (more precise finger articulation, less shiver); v3.2 The latest version ended up with cartoon-like appearance because it allowed to obtain more stable motions.

### 2.1. Avatar 1 (v1)

In 2021, it was decided to conduct a user study online due to the COVID restrictions. The approach utilized for the online user study is described in (Martinez et al., 2017). It extracts human body keypoints from 2D videos by OpenPose. Then process depth estimation, which helps to reveal distance from the camera and convert them into 3D domain.

The motion models obtained can be imported into avatars. Unfortunately, it allowed us to use flat-hand configurations only, so it was necessary to construct sentences using flat, open-hand configurations only. Facial expressions are also not involved.

### 2.2. Avatar 2 (v2)

The improved approach used for the first in-person user study. It requires two additional techniques on the human pose sequence stage: namely LightTrack (Ning et al., 2020) and ExPOSE/SIMPL-X (Choutas et al., 2020). LightTrack is a framework

developed by MIT for online human pose tracking, and ExPOSE/SIMPL-X recovers details of hands and faces thanks to its attention model. This improved version mimics the movements in a better and more natural (physiologically believable) way according to the feedback of deaf signers and interpreters.

### 2.3. Avatar 3 (v3)

This is the pipeline of how Agent 3 operates. Basically, it is almost the same as the avatar proposed above. Just another outer appearance, minimal hyperparameter-tuning, minimal fine-tuning, and less body shivering obtained by the iClone (McCallum, 2011; Georgiev et al., 2024) software (see Figure 2).

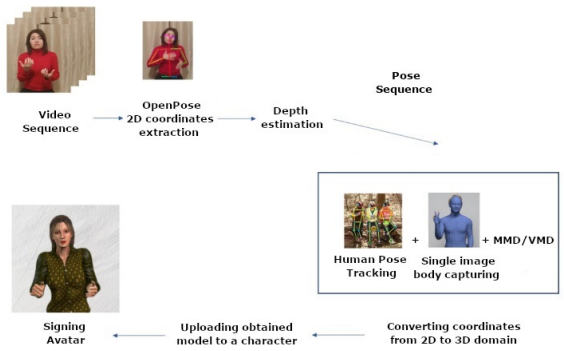



Figure 2: The updated approach used for signing avatar involved in the second user study.

### 2.4. Avatar 4 (v4)

The latest version of the proposed K-RSL signing avatar is depicted in Figure 3. The architecture remains the same as the previous version; however, the body shiver was quite bothersome for participants in the previous user survey. Consequently, experimenting with various body proportions supposedly enhanced motion stability and rendered a more cartoon-like appearance. Despite the fact that (Adamo-Villani and Anasingaraju, 2016) experimentally rejected their hypothesis that avatar stylization could significantly influence participants' perception of signing avatars, they had a relatively small sample size. There is a probability that the stylization of avatars in a manner that is reminiscent of cartoons could be more functional, welcoming, and engaging for the purpose of education and advertising in the context of sign language and avoid potential uncanny valley effect (McDonald, 2025).

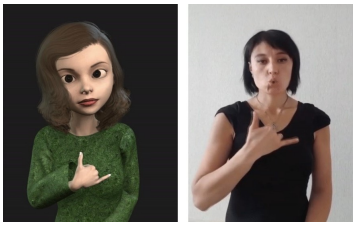


Figure 3: The updated approach (v3.2) used for signing avatar was demonstrated in the second in-person user study compared to a human signer.

## 3. Comparison of the proposed avatar with other existing signing avatars

This section showcases a list of seven well-known state-of-the-art and reputable available 3D signing avatars created for sign language production (generative or pre-recorded) and provides brief comparisons of their merits and demerits:

The first approach is based on the SiGML[1] (Ebling and Glauert, 2016) agents (see Figure 4). It utilizes the HamNoSys (Hanke, 2004) notation system, which encodes the manual features of sign languages. This approach is widely recognized. However, it is rule-based and always requires expert approval.



Figure 4: SiGML avatars

The authors (Yang et al., 2022) of the second signing avatar have used motion capture technologies to record college lessons and import them into mixed reality (see Figure 5 a). This approach is time-consuming and requires expensive technologies with interpreter approval to film each lesson.

The third avatar (see Figure 5 b) is designed to produce German Sign Language (Nguyen et al., 2021). It converts coordinates from 2D videos to 3D domain using Mediapipe and uploads motion models to a Blender avatar. Unfortunately, the eyes of the avatar do not have pupils or iris. As we understood, head position and eye gaze motions are crucial for interaction between deaf signers. In our case, when participants watched the SiGML avatar ANNA, they mentioned that her eyes seemed to look a bit down, and simultaneous head motions from side to side may be perceived as avoiding conversation or even disrespect in local deaf culture.

[1] https://vh.cmp.uea.ac.uk/index.php/SiGML



a b c d e f

Figure 5: a) The mixed-reality avatar of (Yang et al., 2022); b) the 3D avatar of (Nguyen et al., 2021); c) the avatar Paula from DePaul university; d) BuHammad; e) SignMT's stick-figure avatar; f) Project Adaptis.

Next, DePaul University's Paula avatar (see Figure 5 c) demonstrates all ASL linguistic parameters, including hand movements, facial expressions, and body gestures, for clear and natural signing (Davidson, 2006; McDonald et al., 2024). She has been used in practical and instructional settings, including interpreter training at the Jack Mabley Developmental Center in Dixon, Illinois, for the Illinois Department of Human Services.

The avatar named BuHammad (see Figure 5 d) was created for Qatari Sign Language (El Ghoul and Othman, 2022); although it performs convincing signs and facial expressions, the website presents a limited vocabulary for now.

The 6th avatar is cutting-edge technology called SignMT (see Figure 5 e), which supports several sign languages, produces a Skeleton Viewer (stick-figured agent), a Human GAN model for signing videos, and a 3D Avatar; and operates like the Google Translate service (Moryossef, 2023).

The last avatar (see Figure 5 f) is the avatar related to the RSL Adaptis project [2]. It has limited vocabulary on the website and a disproportionately large head, making some signs inaccurate (see Figure 15 in Appendix A, Adaptis performs DAD (in RSL) starting from the area of the intermediate nasal bones instead of the area of the forehead).

## 4. Questionnaire iterative development

Creating sign language interpreting systems emphasizes stakeholder review to determine usability before deployment by end-users. Numerous research studies have explored approaches for empirical evaluation of deaf individuals, including metrics and instruments for critical feedback (Berke et al., 2018). In user studies of sign language interpreting systems, questionnaires are commonly used to obtain subjective feedback (Huenerfauth et al., 2015; Berke et al., 2018).

In addition, the design of deaf-friendly survey elements is also critically important (Schnepp et al., 2011). Researchers are started acknowledging the importance of deaf-friendly questionnaire design and how accessibility, design, and delivery may affect respondents' critical judgments less than a decade ago. In Bosch-Baliarda et al. (2019), authors emphasized the necessity of ensuring full questionnaire accessibility for reliable research and legitimate outcomes.

Recent technology breakthroughs provide varied data collection techniques. Videos, photos, and picture books have helped deaf and hard of hearing research populations, but data gathering is challenging. Prior study has identified effective methods for employing signing avatars with deaf and hard-of-hearing individuals (Huenerfauth et al., 2015).

A key concern is creating deaf-friendly survey metrics to maximize the benefit of responses. The proposed questionnaire has gone through several user studies as an iterative development stage.

### 4.1. Online user study

For the online study, 18 respondents took part in the study (12 were native deaf signers and six were certified hearing interpreters). The ages ranged from 18 to 57 years (mean: 33 years), with a gender distribution of 4 males and 14 females. Education levels distribution: one person completed 9th grade, one 11th grade; four completed a bachelor's degree (1 deaf); the others completed a college degree. Both groups (deaf participants and interpreter) were evaluated separately to figure out if their opinions about signing avatars coincide or not.

It has been decided to compose 5 generic questions aimed at understanding deaf signers' opinion about the challenges they experienced, usability and perception in general, providing them 1-5 Likert scales with range from "Not at all" to "Yes" as options to reply: 1) Was it hard to understand separate signs without facial expressions?; 2) Do you think such an agent would be useful in the case of proficient interpreter absence?; 3) Do you find such a system interesting to use?; 4) Would such a system be useful for displaying ads?; 5) Would such a system be useful for announcing an alarm or emergency?

[2] https://adaptis.pro/en/

| Questions composed to obtain the general feedback |
|---|
| 1. Did you understand the meaning of the sentences right away? |
| 2. Do you think you understood correctly what the character is showing? |
| 3. Did you have to think out the meaning of the sentence because of the character's movement style? |
| 4. Do you consider such a system suitable for use? |
| 5. Do you find this type of interaction exciting? |
| 6. Would you trust such a character? |
| 7. Would you trust such a character to translate into SL? |
| 8. Will a similar character be useful for translation when there is no sign language interpreter? |
| 9. Was this communication effective? |
| 10. Would such a system be useful for advertising or entertainment content? |
| 11. Will such a system be useful for announcing important information at the PSC, Railway Station /Airport? |
| 12. Would such a system be useful for declaring an alert or emergency? |
| 13. Would you like to interact with such characters in the future? |

Table 1: Generic questions utilized for the first in-person user study.

| Section | Items | | |
|---|---|---|---|
| Anthropomorphism | Fake | - | Natural |
| | Machinelike | - | Humanlike |
| | Unconscious | - | Conscious |
| | Artificial | - | Lifelike |
| | Moving rigidly | - | Moving elegant |
| Animacy | Dead | - | Alive |
| | Stagnant | - | Lively |
| | Mechanical | - | Organic |
| | Artificial | - | Lifelike |
| | Inert | - | Interactive |
| | Apathetic | - | Responsive |
| Likeability | Dislike | - | Like |
| | Unfriendly | - | Friendly |
| | Unkind | - | Kind |
| | Unpleasant | - | Pleasant |
| | Awful | - | Nice |
| Perceived Intelligence | Incompetent | - | Competent |
| | Ignorant | - | Knowledgeable |
| | Irresponsible | - | Responsible |
| | Unintelligent | - | Intelligent |
| | Foolish | - | Sensible |

Figure 6: Four sections taken from Godspeed questionnaire

In conjunction with this, it was also decided to use sections of the Godspeed questionnaire (Bartneck et al., 2009). All these questions including Godspeed inquiries (see Figure 6) have been reformulated as questions and translated into K-RSL with slight elaborations and recorded as videos.

Unfortunately, such minimal elaborations revealed that participants were confused, and it has also become evident that several Godspeed items seem very similar in sign language: for example, "competent" and "intelligent" or "natural" and "organic" utilized the same signs with slightly different additional elaborations that ended up confusing and still seemed very similar for participants; therefore, the inclusion of elaborations used for translations became insufficient. So, further, more extensive elaborations are required to differentiate them. Participants also agreed that recognizing signs without facial expressions is extremely challenging and that these facial expressions are crucial not only for human-to-human sign communication, but also for comprehension in general.

An important finding of the online research is that each participant made several video calls to the separate interpreter, who was not involved as a participant, and helped with question translations into K-RSL before the user study and supported them during the user study to clarify inquiries throughout interactions with the signing avatars. It was allowed since it was online, while generally deaf signers come with interpreters. Thus, the presence of a human interpreter is essential for subjects' comfort and the assurance of being on the correct path. More information on the agents utilized, the questions, and the outputs obtained (including statistics) is described in Imashev et al. (2025b).

### 4.2. First in-person user study

Ten native deaf signers between the ages of 25 and 59 (mean age: 31.5; 4 male and 6 female participants) took part this time. Their educational background varied, and the majority of the participants had a college degree. The list of generic questions was improved and contained 13 of them (see Table 1). It is important to note that the selection of questions was not random and was based on a previous user study, where some subjects specifically indicated that they were not currently considering avatars for important tasks such as assisting with translations at the police or medical facilities, but rather for advertising or other more common purposes. All these questions were translated into K-RSL videos. Ten items from the Godspeed questionnaire were taken again, with elaborations to a greater extent since, for example, "Machinelike" might be understood as "Looks like a car", so this response option was formulated as "Looks like a robot". In order to reveal eye gaze motion patterns during the interaction with signing avatars, Tobii Pro 2 eyeglasses were used in this study.

This study revealed that: 1) the movements of the data-driven avatar exhibited a higher degree of physical credibility, which may elucidate why participants expressed a preference for its facial expressions over those of the manually animated counterpart, characterized by rigid articulations. Besides, participants demonstrated greater precision in their translations while engaging with the manually animated avatar (see Figure 10 A2), as opposed to the data-driven avatar, and exhibited the least fluctuation in mood throughout the interaction. The motions of this manually animated avatar were animated by a team of the research team under supervision of professional interpreter using iClone software; 2) the inclusion of tangible items for comparison, such as a genuine edible apple with a prop apple that is typically utilized in art schools or film production, should be used to provide a more profound comprehension than the abstract concepts of “artificial” and “realistic”, and others. More detailed information on the user study and statistics is available in Imashev et al. (2025b).

### 4.3. Second in-person user study

In total, for this user study we recruited 12 deaf signers aged 27–53: six male and six female, their mean age was 36.33. Four agents performed three different sentences in K-RSL (similar to the previous user study 12 sentences in total were composed: three distinct signing phrases in K-RSL for each agent, ensuring that the phrases were uniformly balanced with respect to the handshapes utilized and the complexity of the sentence structures and concepts articulated). After numerous trials of formulating proper elaborations for Godspeed items, it has been decided to propose a visual scale with not 1 to 5 or 1 to 10 range as response, but a thermometer-like 100-point scale that could become more suitable for participants to evaluate and for researchers to use especially in cases of small number of participants. Generic questions and sorting tasks were also utilized for the assessments. Several proposed scales are depicted in Figure 7, such as scales for Dead-Alive, Inert-Interactive, Ignorant-Knowledgeable.

In addition to images, tangible artifacts were also used as boundaries for the various ranges for better clarity (see Figure 8) and the Tobii Pro 2 kit to track eye trajectories, but this time in front of a big screen and a fixed distance.

Together with them a Funometer scale (Read et al., 2002) was also incorporated before and after interactions with each agent (see Figure 9).

More detailed information on the perceptions of agents by participants and statistical outputs available in Imashev et al. (2025a).

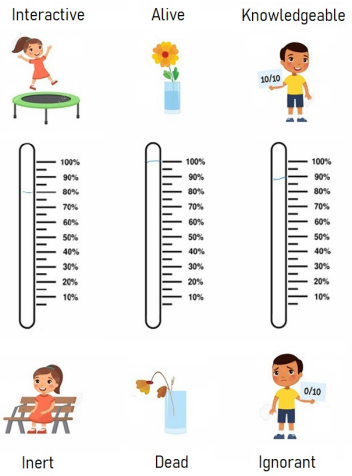



Figure 7: Godspeed questionnaire as visual thermometer scale provided with iPad tablet and stylus

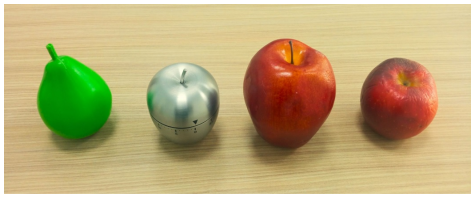


Figure 8: Tangible artifacts

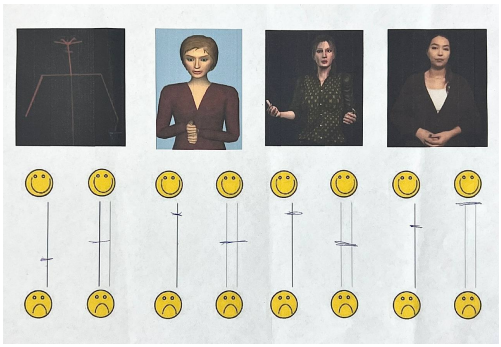


Figure 9: Funometer used for mood pre- and post-evaluations before and after interaction with signing agents (SignMT, SiGML, proposed and human)

## 5. EASIER evaluation tool

The EASIER project (Picron et al., 2024; Dimou et al., 2022a,b) underscores the importance of engaging deaf people in the creation and assessment of signing avatars to improve their clarity and cultivate confidence among users of synthetic signing technology. The project seeks to establish a collaborative environment in which sign language users can offer comments during the avatar development process. The fundamental element of this initiative is a questionnaire-based methodology aimed at collecting insights from various sign language groups,

notably focusing on users of Greek Sign Language (GSL), French Sign Language (LSF), German Sign Language (DGS), and Swiss German Sign Language (DSGS).

The technique comprises a systematic online questionnaire that evaluates essential performance factors, including the naturalness and comprehensibility of the signings performed by avatars. The questionnaire was based on the SUS questionnaire (Brooke, 2013) (with a 1 to 5 Likert scale from "Strongly disagree" to "Strongly agree") and has been designed to be user-friendly and accessible, allowing participants to offer feedback on different aspects and components of avatar performance. The preliminary pilot study (Dimou et al., 2022b) with GSL signers indicated a preference for the avatar "PAULA" over "FRANÇOISE" (Jennings et al., 2010), suggesting that PAULA's signing was regarded more natural and intelligible. The findings underscored the necessity for continuous user participation to further enhance avatars.

The results of the pilot user study indicate that although both L1 (native) and L2 (non-native) signers evaluated PAULA in a quite similar way, significant discrepancies were observed in their evaluations of FRANÇOISE. This suggests that L2 signers may have a more permissive perspective on average signing performance compared to L1 signers.

## 6. Results

The iterative outputs of the signing avatar evaluation methodology revealed that:

### 6.1. Online User Study

This subsection describes the results obtained in 4.1 conducted with avatar v1 (2.1).

All four agents executed sequences of sign language that exclusively incorporated open-palm signs, thereby facilitating a focused analysis of the effectiveness and reception of data-driven sign language representation. Specifically, Avatar 1 demonstrated the sign sequence "NOTHING NEW", Avatar 2 performed the "HELLO" sign twice, Avatar 3 communicated "IWILL STOP", and Avatar 4 conveyed "ILIKE FISH". Hearing interpreters acknowledged Agent 4 (see Figure 12 in Appendix A), the human interpreter, as the most adept in local sign language, while Agent 1 was seen as the least skilled in local sign language. Simultaneously, the deaf participants determined that Agent 4's sign language skill is marginally better than that of the other agents. This gap is speculated to arise from the fact that deaf users consider more than just hand configurations for communication (including non-manual markers); they may perceive that Agent 4 likely lacks expertise for particular signs she performed.

Table 2: Friedman test results over four agents, significant findings are in bold.

| Measurement | Friedman test output |
|---|---|
| **Fake - Natural** | $\chi^2(3) = 43.795; p = 0.000.$ |
| **Machinelike - Humanlike** | $\chi^2(3) = 39.281; p = 0.000.$ |
| Moving elegantly - rigidly | $\chi^2(3) = 6.614; p = 0.085.$ |
| **Stagnant - Lively** | $\chi^2(3) = 40.452; p = 0.000.$ |
| **Lifelike - Artificial** | $\chi^2(3) = 8.955; p = 0.03.$ |
| **Mechanical - Organic** | $\chi^2(3) = 42.022; p = 0.000.$ |
| Like-Dislike | $\chi^2(3) = 6.060; p = 0.109.$ |
| Competent - Incompetent | $\chi^2(3) = 6.944; p = 0.074.$ |
| Pleasant - Unpleasant | $\chi^2(3) = 3.358p = 0.340.$ |
| **Unintelligent - Intelligent** | $\chi^2(3) = 30.163; p = 0.000.$ |

Furthermore, a significant discovery from the online user study indicated that deaf natives had less optimism about avatar technology in contrast to hearing interpreters for live face-to-face translation, and they prefer the involvement of a human sign language agent-interpreter.

Additionally, Table 2 showed notable differences in six out of ten categories, which indicates that Agent 4 was perceived to be significantly more intelligent and realistic. These results revealed that significant advancements are necessary for our data-driven avatar to potentially compete with human agents in the future. This is how the iterative improvement started.

### 6.2. In-person User Study #1

This subsection describes the results obtained in 4.2 conducted with avatar v2 (2.2). For the first in-person user study, we compared a data-driven avatar (A1) with a manually animated avatar (A2) and a human agent (A3) (refer to Figure 10), which was comparable to the online study.




Figure 10: Signing agents used for this study from left to right: A1 - Data-driven avatar, A2 - manually animated avatar, and a novice human agent (A3).

In this instance, a novice human signer was involved and the decision is influenced by the lack of smoothness in the motions of both the data-driven and manually animated avatars, which appeared somewhat constrained. A comparison with a professional human signer may yield an unfair assessment, favoring the human signer significantly. They

performed 12 different signing phrases (see Table 3). These phrases were designed to attain balance in handshapes, the number of signs used in each sentence, and the cumulative complexity of the phrases.

Table 3: Sign sentences (as gloss sequences) performed by characters.

| Level | Avatar 1 (data-driven) | Avatar 2 (manually animated) | Human Agent |
|---|---|---|---|
| Easy1 | YOU HUNGER | YOU WORRY | YOU GLAD |
| Easy2 | YOU LIKE FISH | YOU BORN EARLY | YOU WANT WALK |
| Med. | YOU LIKE RAIN | WHERE YOU HOUSE | YOU HOUSE KEEPER |
| Hard | WHICH YOU | YOU HAVE | YOU DO |
| | NUMBER PHONE | DEAF RELATIVES | SEW FACTORY |

Cronbach's Alpha for the Human agent did not result in excellent internal consistency, as seen in Table 4, suggesting that individuals exhibited variability in their evaluations of the human agent.

Table 4: Godspeed Questionnaire mean and Cr.$\alpha$ valued (in brackets).

| | A1 | A2 | Human agent |
|---|---|---|---|
| Anthropom. | **0.75** (1.4) | **0.71**(1.5) | 0.6 (3.9) |
| Animacy | 0.59 (1.4) | **0.82** (1.5) | 0.53 (3.9) |
| Intelligence | **0.87** (1.9) | **0.77** (2.2) | 0.45 (3 ) |
| Likeability | **0.99** (3.85) | **0.89** (3.5) | **0.78** (2.9) |

Anthropomorphism and Animacy received a score of 3.9 out of 5, intelligence was scored at 3 out of 5, and likability was assessed at 2.9 out of 5. The phenomenon may arise because participants still needed further elaborations for enhanced explanations of closely related concepts, possibly interpreting knowledge of sign language as general intelligence or realism as questioning whether an individual is an actual interpreter rather than inquiring about being human.

The agents were shown on a laptop screen, and some participants recommended moving the screen's location to improve accessibility. Instances occurred when the screen size limited the participants' capacity to scrutinize the entire agent in depth, resulting in a focused concentration on the agent's center (see Figure 13 a and b in Appendix A) with little eye movement.

Presbyopia or age-related farsightedness was not expected. It caused a challenge twice when calibrating the glasses before the experiment began. This problem became apparent when the tracker's trajectory reflected the movements of agents' hands, despite being located in vacant areas of the screen. The participant's attention on the agents' hands is definitely apparent, but their focus on the agents' faces lacks comparable clarity (see Figure 13 c in Appendix A). Two instances of presbyopia were observed and subsequently excluded from the average heatmap composition; despite this, the questionnaire responses of the corresponding participants were included in the analysis.

### 6.3. In-person User Study #2

This subsection describes the results obtained in 4.3 conducted with avatar v3.1 (2.3). For the user study #3 (second in-person user study), the proposed signing avatar was compared with two existing solutions and the professional human CODA signer, the sentences performed are listed in Table 7 in Appendix A. One of the questionnaire sections was dedicated to asking participants how they understood the signings performance of agents. The numbers revealed that the participants understood the SiGML avatar slightly (non-significantly) (see Table 5): the results of the Dependent T-test indicated that there is a non-significant difference between comprehension of SiGML and aiSL-a: $t = -1.2, p = 0.268$.

Table 5: Evaluation on how participants understood agents.

| Agent | Mean (SD) | Shapiro-Wilk test |
|---|---|---|
| SignMT | 37.59 (15.61) | $W = 0.897, p = 0.243$ |
| SiGML | 89.70 (7.28) | $W = 0.9194, p = 0.4142$ |
| *aiSL-a* | 84.56 (9.02) | $W = 0.9439, p = 0.683$ |
| Human | 98.78 (3.67) | $W = 0.39, p < .001$ |

Despite the fact that the proposed avatar looked and performed more natural and less robotic (see Table 6).

Table 6: Mean values(SDs) for Sorting Task

| Measure | SignMT | SiGML | aiSL-a | Human |
|---|---|---|---|---|
| Anthrop. | 1,09(0.17) | 2,2(0.28) | 2,71(0.36) | 4(0.0) |
| Animacy | 1,22(0.3) | 2,11(0.4) | 2,78(0.32) | 3,83(0.3) |
| Likeability | 1,69(0.77) | 1,89(0.58) | 2,53(0.52) | 3,89(0.23) |
| Intelligence | 1,13(0.22) | 2,18(0.40) | 2,71(0.41) | 4(0.0) |

Based on heatmaps of interactions, more natural signings led deaf signers to focus on the interlocutor's head, meaning this effect may take place not only when deaf signers are communicating with human agents (which has been proven by Emmorey et al. (2009)), but with signing characters, too. In contrast, the less natural-looking Agent 1 had been watched mainly on hands (see Figure 14 in Appendix A).

Regarding the evaluation scales, both the Sorting Task and the Thermometer scale presented distinct advantages as assessment tools for signing agents. The Thermometer scale (0 - 100) revealed significant disparities between the human agent and other avatars in all categories, while the Sorting Task (which ended up conceptually similar to 1 - 5 Likert-scale) results offered a more readily apparent preference ranking among the signing agents (see Figure 11).

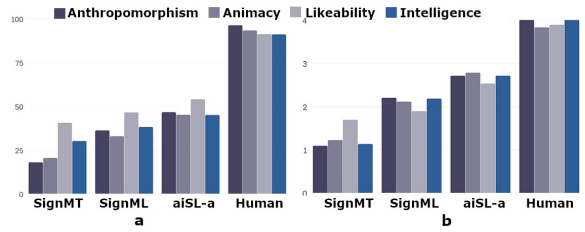



Figure 11: Visual comparison of Godspeed Thermometer(a) and Godspeed (b) Sorting Task means.

## 7. Limitations

It is highly preferable to conduct user studies with a larger number of people. Recruiting 30 or more participants (as stupulated by the Central Limit Theorem) to utilize the statistical methods allowed for normal distribution is still consistently problematic.

In our cases, native deaf female and younger signers often exhibited a greater willingness to participate in user studies, which influenced the imbalance. Each time we recruited different participants to avoid bias from previous interactions, it was quite challenging to involve participants due to the insular nature of the deaf community; establishing trust is also a long-lasting process. We hope that future familiarity with the research team and the participation of native deaf signers in the research will increase the interest and motivate more people to provide their feedback.

The results of three studies indicate the necessity of establishing a section titled “Perceived SL Expertise” in lieu of “Perceived Safety” from the Godspeed questionnaire, as the two questions pertaining to this subject in one particular case are insufficient to warrant a separate section of the questionnaire.

## 8. Conclusions and Future Work

Despite the fact that the proposed avatars already perform quite understandably and cover the demerits of other solutions, such as finger articulation, natural eye blink, mouthing, and smooth motions, and the aforementioned methodological explorations have already been appreciated in (Chemnad and Othman, 2025). The entire outcomes suggest the following for further improvement:

- A separate section for exterior attributes might be required. It is essential to consider including characteristics such as charisma and outfits, as shown in (Kipp et al., 2011b), to evaluate the likability of avatars.

- To formulate inquiries for the newly designated area titled Perceived SL Expertise.

- Concise phrases fail to convey a comprehensive understanding of signing avatars’ technical capabilities; hence, short tales, fairytales, or historical essays are more suitable for being performed by signing avatars.

- To provide additional artifacts or concepts in the central sections for all 100-point scales in the questionnaire.

- It is highly recommended to conduct follow-up research using more signing avatars and a larger pool of participants.

- The participation of a deaf interpreter as a member of the research team throughout the preparatory stages and the user studies itself will also be advantageous.

- A Text-To-Gloss LLM model trained to convert written text to proper word order gloss sequences, which will be comprehensible by native deaf signers, is needed for natural signing performance production by avatars.

Compared with the proposed questionnaire, EASIER online evaluation tool required about 20 minutes vs. 40 minutes in Online User Study, making it easier to follow and less oppressive for the participants. Moreover, despite the methodological similarities with the user studies performed, substantial discrepancies were overlooked in the conceptualization of these studies contrasted to EASIER, wherein participants were tasked with assessing fingerspelling, individually chosen lemmas, brief signing sequences incorporating the selected lemmas, and static visuals representing emotions, all of which offer a more nuanced understanding of the possible disadvantages of avatars to be assessed.

The implementation of 100-point scales in the proposed questionnaire allowed for a more nuanced and sensitive evaluation compared to the 1-5 Likert scale (see Figure 11). Although a 5-point scale is appropriate for a larger sample size with normally distributed data (Westland, 2022), some methodological adjustments are required for small groups (Gaworski and Daśko, 2026); a higher number of points may allow the capture of subtle differences in opinions and experiences that a simpler scale may overlook (Preston and Colman, 2000; Finstad, 2010; Diefenbach et al., 1993).

## 9. Acknowledgements

This iterative development of signing the avatar and questionnaire was supported during the following grants: Nazarbayev University under the Collaborative Research Project Grant 11022021CRP1502, A.S.; and by the Nazarbayev University Faculty Development Competitive Research Grant Program 2022–2024 “Kazakh-Russian Sign Language Processing: Data, Tools, and Interaction” - the award number is 11022021FD2902.

## 10. Bibliographical References


Nicoletta Adamo-Villani and Saikiran Anasingaraju. 2016. Toward the ideal signing avatar. *EAI Endorsed Transactions on e-Learning*, 3(11).

Nicoletta Adamo-Villani, Jason Lestina, and Saikiran Anasingaraju. 2016. Does character's visual style affect viewer's perception of signing avatars? In *E-Learning, E-Education, and Online Training: Second International Conference, eLEOT 2015, Novedrate, Italy, September 16-18, 2015, Revised Selected Papers 2*, pages 1–8. Springer.

Alice Audrezet, Bernadett Koles, Julie Guidry Moulard, Nisreen Ameen, and Brad McKenna. 2025. Virtual influencers: Definition and future research directions. *Journal of Business Research*, 200:115647.

Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1:71–81.

Larwan Berke, Sushant Kafle, and Matt Huenerfauth. 2018. Methods for evaluation of imperfect captioning tools by deaf or hard-of-hearing users at different reading literacy levels. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Marta Bosch-Baliarda, Olga Soler Vilageliu, and Pilar Orero. 2019. Toward a sign language-friendly questionnaire design. *The Journal of Deaf Studies and Deaf Education*, 24(4):333–345.

Yosra Bouzid, Mohamed Jemni, et al. 2016. The effect of avatar technology on sign writing vocabularies acquisition for deaf learners. In *2016 IEEE 16th international conference on advanced learning technologies (ICALT)*, pages 441–445. IEEE.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

John Brooke. 2013. Sus: a retrospective. *Journal of usability studies*, 8(2).

Khansa Chemnad and Achraf Othman. 2025. Perception and monitoring of sign language acquisition for avatar technologies: A rapid focused review (2020–2025). *Multimodal Technologies and Interaction*, 9(8):82.

Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. 2020. Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 20–40. Springer.

Mary Jo Davidson. 2006. Paula: A computer-based sign language tutor for hearing adults. In *Intelligent Tutoring Systems 2006 Workshop on Teaching with Robots, Agents, and Natural Language Processing*, pages 66–72.

Michael A Diefenbach, Neil D Weinstein, and Joseph O'reilly. 1993. Scales for assessing perceptions of health hazard susceptibility. *Health education research*, 8(2):181–192.

Athanasia-Lida Dimou, Vassilis Papavassiliou, Theodoros Goulas, Kyriaki Vasilaki, Anna Vacalopoulou, Stavroula-Evita Fotinea, and Eleni Efthimiou. 2022a. What about synthetic signing? a methodology for signer involvement in the development of avatar technology with generative capacity. *Frontiers in Communication*, 7:798644.

Athanasia-Lida Dimou, Vassilis Papavassiliou, John McDonald, Theodore Goulas, Kyriaki Vasilaki, Anna Vacalopoulou, Stavroula-Evita Fotinea, Eleni Efthimiou, and Rosalee Wolfe. 2022b. Signing avatar performance evaluation within easier project. In *Proceedings of the 7th international workshop on sign language translation and avatar technology: The junction of the visual and the textual: Challenges and perspectives*, pages 39–44.

Ramazan Duisenbek and Tamara Zhukabayeva. 2025. Real-time kazakh sign language alphabet recognition using svm and yolov8n. In *2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST)*, pages 1–6. IEEE.

Sarah Ebling and John Glauert. 2016. Building a swiss german sign language avatar with jasigning and evaluating it among the deaf community. *Universal Access in the Information Society*, 15(4):577–587.

Oussama El Ghoul and Achraf Othman. 2022. Virtual reality for educating sign language using signing avatar: The future of creative learning for deaf students. In *2022 IEEE Global Engineering*

*Education Conference (EDUCON)*, pages 1269–1274. IEEE.

Karen Emmorey, Robin Thompson, and Rachael Colvin. 2009. Eye gaze during comprehension of american sign language by native and beginning signers. *Journal of deaf studies and deaf education*, 14(2):237–243.

RM Farwell. 1976. Speechreading, a review of the research. *American Annals of the Deaf*, 121:19–30.

Kraig Finstad. 2010. Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of usability studies*, 5(3):104–110.

Cristina Gasch, Alireza Javanmardi, Ameer Khan, Azucena Garcia-Palacios, and Alain Pagani. 2025. Exploring avatar utilization in workplace and educational environments: A study on user acceptance, preferences, and technostress. *Applied Sciences*, 15(6):3290.

Marek Gaworski and Aleksandra Daśko. 2026. An approach to developing likert scale survey results based on the example of a research study involving a limited number of students. *Applied Sciences*, 16(5):2602.

Georgi Georgiev, Georgi Hristov, Plamen Zahariev, and Diyana Kinaneva. 2024. Integration of 3d scanning data and motion capture system for meta human animations. In *2024 5th International Conference on Communications, Information, Electronic and Energy Systems (CIEES)*, pages 1–6. IEEE.

Sylvie Gibet, Nicolas Courty, Kyle Duarte, and Thibaut Le Naour. 2011. The signcom system for data-driven animation of interactive virtual signers: Methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1):1–23.

Thomas Hanke. 2004. Hamnosys–representing sign language data in language resources and language processing contexts. In *sign-lang@ LREC 2004*, pages 1–6. European Language Resources Association (ELRA).

Matt Huenerfauth and Vicki Hanson. 2009. Sign language in the interface: access for deaf signers. *Universal Access Handbook. NJ: Erlbaum*, 38:14.

Matt Huenerfauth and Hernisa Kacorri. 2015. Best practices for conducting evaluations of sign language animation. *Journal on Technology and Persons with Disabilities*, 3:20–32.

Matt Huenerfauth, Pengfei Lu, and Hernisa Kacorri. 2015. Synthesizing and evaluating animations of american sign language verbs modeled from motion-capture data. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 22–28.

Alfarabi Imashev. 2017. Sign language static gestures recognition tool prototype. In *2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–4. IEEE.

Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. 2020. K-rsl: a corpus for linguistic understanding, visual evaluation, and recognition of sign languages. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Alfarabi Imashev, Nurziya Oralbayeva, Gulmira Baizhanova, and Anara Sandygulova. 2025a. Assessment of comparative evaluation techniques for signing agents: a study with deaf adults. *Journal on Multimodal User Interfaces*, 19(1):1–19.

Alfarabi Imashev, Nurziya Oralbayeva, Vadim Kimmelman, and Anara Sandygulova. 2022. A user-centered evaluation of the data-driven sign language avatar system: A pilot study. In *Proceedings of the 10th international conference on human-agent interaction*, pages 194–202.

Alfarabi Imashev, Nurziya Oralbayeva, and Anara Sandygulova. 2025b. An exploratory user study towards developing a unified, comprehensive assessment apparatus for deaf signers, specifically tailored for signing avatars evaluation: challenges, findings, and recommendations. *Multimedia Tools and Applications*, 84(26):30865–30902.

Vince Jennings, Ralph Elliott, Richard Kennaway, and John Glauert. 2010. Requirements for a signing avatar. In *sign-lang@ LREC 2010*, pages 133–136. European Language Resources Association (ELRA).

JR Kennaway, John RW Glauert, and Inge Zwitserlood. 2007. Providing signed content on the internet by synthesized animation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 14(3):15–es.

Michael Kipp, Alexis Heloir, and Quan Nguyen. 2011a. Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer.

Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011b. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 107–114.

Saule Kudubayeva, Dmitriy Ryumin, Marat Kalghanov, and Almatova Assem. 2016. Automated recognition system of statistic gestures via kinect sensor. In *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–4.

Anna Kuznetsova, Alfarabi Imashev, Medet Mukushev, Anara Sandygulova, and Vadim Kimmelman. 2021. Using computer vision to analyze non-manual marking of questions in krsl. In *Proceedings of the 1st international Workshop on automatic Translation for Signed and spoken languages (AT4SSL)*, pages 49–59.

Vincenzo Lombardo, Fabrizio Nunnari, and Rossana Damiano. 2010. A virtual interpreter for the italian sign language. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pages 201–207. Springer.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. 2017. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649.

MD McCallum. 2011. iclone 4.31 3d animation.

John McDonald. 2025. Considerations on generating facial nonmanual signals on signing avatars. *Universal Access in the Information Society*, 24(1):19–36.

John McDonald, Eleni Efthimiou, Stavroula-Evita Fotinea, and Rosalee Wolfe. 2024. Multilingual synthesis of depictions through structured descriptions of sign: An initial case study. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 244–253.

Donna A Morere. 2012. Measures of writing, math, and general academic knowledge. In *Assessing literacy in deaf individuals: Neurocognitive measurement and predictors*, pages 127–137. Springer.

Amit Moryossef. 2023. sign. mt: Real-time multilingual sign language translation application. *arXiv preprint arXiv:2310.05064*.

Lucie Naert, Caroline Larboulette, and Sylvie Gibet. 2020. A survey on the animation of signing avatars: From sign representation to utterance synthesis. *Computers & Graphics*, 92:76–98.

Lan Thao Nguyen, Florian Schicktanz, Aeneas Stankowski, and Eleftherios Avramidis. 2021. Automatic generation of a 3d sign language avatar on ar glasses given 2d videos of human signers. In *Proceedings of the 1st international workshop on automatic translation for signed and spoken languages (AT4SSL)*, pages 71–81.

Guanghan Ning, Jian Pei, and Heng Huang. 2020. Lighttrack: A generic framework for online top-down human pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1034–1035.

Hamid Ouhnni, Acim Btissam, Belhiah Meryam, Karim El Bouchti, Zaoui Seghroucheni Yassine, Souad Najoua Lagmiri, Benachir Rigalma, and Soumia Ziti. 2025. The evolution of virtual identity: a systematic review of avatar customization technologies and their behavioral effects in vr environments. *Frontiers in Virtual Reality*, 6:1496128.

Komal Parvez, Muzafar Khan, Javed Iqbal, Muhammad Tahir, Ahmed Alghamdi, Mohammed Alqarni, Amer Awad Alzaidi, and Nadeem Javaid. 2019. Measuring effectiveness of mobile application in learning basic mathematical concepts using sign language. *Sustainability*, 11(11):3064.

Frankie Picron, Davy Van Landuyt, Rehana Omardeen, Eleni Efthimiou, Rosalee Wolfe, Stavroula-Evita Fotinea, Theodore Goulas, Christian Tismer, Maria Kopf, and Thomas Hanke. 2024. The easier mobile application and avatar end-user evaluation methodology. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 276–281.

Carolyn C Preston and Andrew M Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, 104(1):1–15.

Lorna C Quandt, Athena Willis, Melody Schwenk, Kaitlyn Weeks, and Ruthie Ferster. 2022. Attitudes toward signing avatars vary depending on hearing status, age of signed language acquisition, and avatar type. *Frontiers in psychology*, 13:730917.

Janet C Read, Stuart MacFarlane, and Chris Casey. 2002. Endurability, engagement and expectations: Measuring children's fun. In *Interaction design and children*, volume 2, pages 1–23. Citeseer.

Floris Roelofsen, Lyke Esselink, Shani Mende-Gillings, Maartje De Meulder, Nienke Sijm, and Anika Smeijers. 2021. Online evaluation of text-to-sign translation by deaf end users: Some methodological recommendations (short paper). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 82–87.

Jerry Schnepp, Rosalee Wolfe, Brent Shiver, John McDonald, and Jorge Toro. 2011. Signquote: A remote testing facility for eliciting signed qualitative feedback.

Jerry C Schnepp, Rosalee J Wolfe, John C McDonald, and Jorge A Toro. 2012. Combining emotion and facial nonmanual signals in synthesized american sign language. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 249–250.

J Christopher Westland. 2022. Information loss and bias in likert survey responses. *PloS one*, 17(7):e0271949.

Rosalee Wolfe, John C McDonald, Thomas Hanke, Sarah Ebling, Davy Van Landuyt, Frankie Picron, Verena Krausneker, Eleni Efthimiou, Evita Fotinea, and Annelies Braffort. 2022. Sign language avatars: a question of representation. *Information*, 13(4):206.

Fu-Chia Yang, Christos Mousas, and Nicoletta Adamo. 2022. Holographic sign language avatar interpreter: A user interaction study in a mixed reality classroom. *Computer Animation and Virtual Worlds*, 33(3-4):e2082.

# A. Supplementary materials



Figure 12: Signing agents used for online user study: Agent 1 and Agent 2 are two options of the v1, Agent 3 is a manually programmed avatar created CUNY ASL program, signs that have meaning on KRSL were trimmed from a pre-recorded video, Agent 4 is an apprentice local sign language user

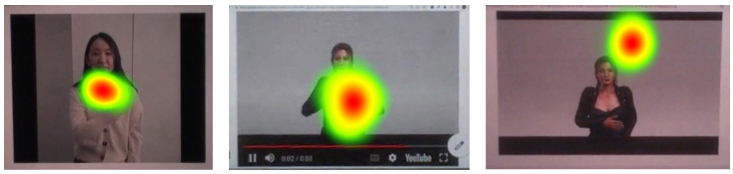



Figure 13: Examples of when image size is small enough to cover whole agent by eyes (a and b); Eye gaze shifted due to presbyopia (c).

Table 7: Sign sentences (as gloss sequences) performed by characters.

| Stick Figure Agent (SignMT) | SiGML avatar Anna | aiSL-a (proposed Data-Driven Avatar) | Human Agent |
|---|---|---|---|
| BIG AND GREY ELEPHANT | NEED LAW DOCUMENT/PAPER | I LIKE RAIN | YOU WATCH YOUTUBE |
| I AFRAID WOLF | DO SEWING STAFF | WHICH YOUR TELEPHONE NUMBER | WEATHER TODAY HUMID |
| BEAR EAT BERRIES | I DRINK COFFEE | I HOUSE KEEPER | WHICH YOUR FAVOURITE MOVIE |

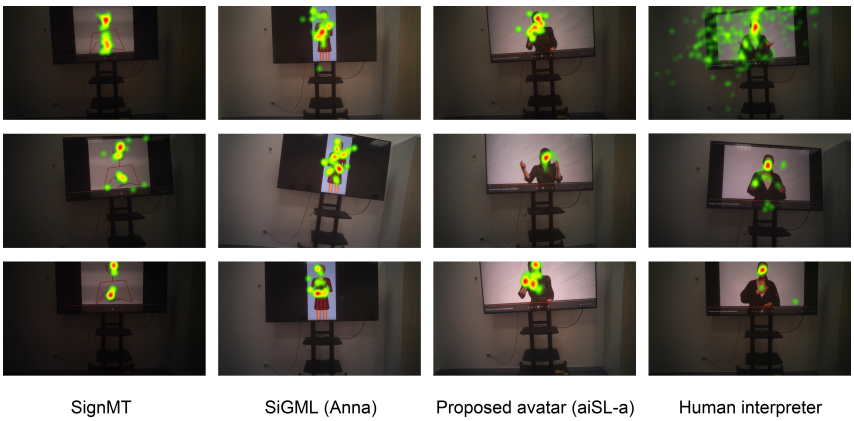



Figure 14: Averaged eye-tracking visualizations of participants' interaction with each agent.

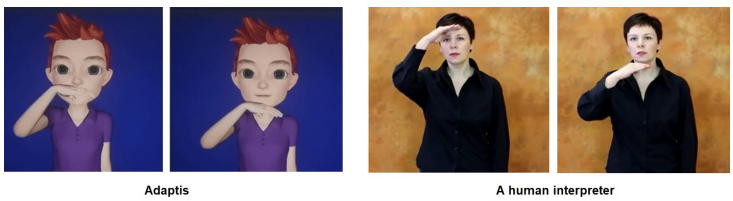



Figure 15: Comparison of the sign DAD of RSL is performed by Adaptis avatar and a human interpreter

# *Movement Coherence* in High Visual Load Environments: Implications for Attention in Mixed-Hearing Classes

**Mert Inan**[1], **Saki Imai**[1], **Anna Marshall**[2], **Tessa Karel**[2], **Malihe Alikhani**[1]
[1]Northeastern University [2]Open Up Pittsburgh
{inan.m, m.alikhani}@northeastern.edu

**Abstract**

Signed interpretation in movement based instruction creates high visual load environments in which spoken language, sign language, and physical demonstration compete for the same perceptual channel. We present a participatory multimodal observational study of mixed hearing movement and mindfulness classes in which Deaf, Hard of Hearing (DHH), and hearing participants practice together. Based on synchronized video recordings, instructor interviews, and participant surveys, we examine how alignment across demonstration, signed instruction, and bodily execution is achieved and restored in real time. Drawing on theories of grounding, repair, and sign language interaction, we conceptualize movement coherence as alignment across these parallel streams and describe how breakdowns trigger observable attention shifts and distributed repair across participants, interpreters, and instructors. Across sessions, we identify recurrent coordination strategies including peer checking, freeze and scan, interpreter repositioning, tactile cueing, and pacing adjustment. Our findings provide an empirically grounded account of grounding under attentional constraint in inclusive embodied settings, with implications for sign language interpretation, multimodal discourse, and the design of accessible movement instruction. This paper includes deidentified materials derived from recorded sessions, including selected keyframes, structured interactional annotations, and anonymized instructor and participant survey responses.

**Keywords:** movement coherence, Deaf community, participatory design, visual attention, sign language, multimodal communication, repair

## 1. Introduction

Participation in movement and mindfulness classes requires continuous inference about instructional intent across multiple modalities. What movement does the instructor intend? Has the interpreter accurately conveyed it? Are my neighbors doing what I think I should be doing? In mixed-hearing classes where Deaf, Hard-of-Hearing (DHH), and hearing participants practice together, understanding requires coordinates spoken instruction, signed interpretation and physical demonstration, through a single visual channel. These environments constitute *high visual load* settings, where multiple instruction streams compete for access to the same perceptual channel (Figure 1).

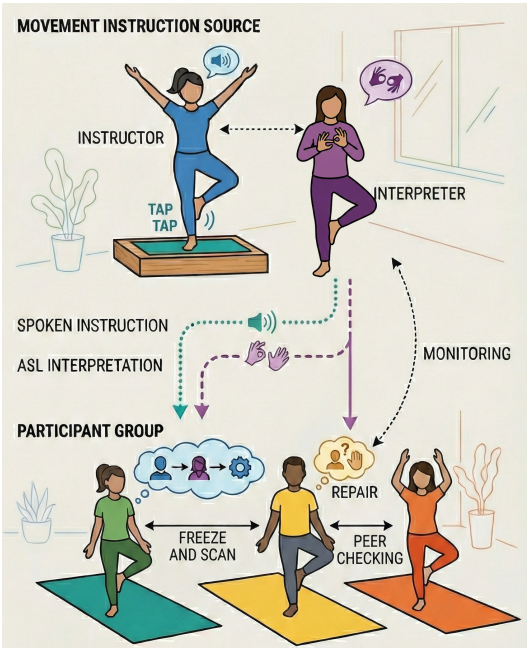



Figure 1: High visual load in mixed-hearing movement instruction. Spoken instruction, signed interpretation, and physical demonstration unfold in parallel, requiring participants to allocate visual attention across competing streams. We model coherence as the alignment between these streams and bodily execution; breakdowns produce observable attention shifts and repair behaviors that are captured in our annotation framework.

While such environments present challenges for accessibility and instruction, they also provide a unique opportunity to study how people maintain shared understanding when communication unfolds across parallel modalities. In these settings, understanding is not reflected in verbal agreement, but in whether participants' bodies align with the intended movement. When alignment breaks, participants visibly pause, scan the room, reposition themselves, or seek clarification. Interpreters adjust their position or correct movements, and instructors slow down or repeat demonstrations. Alignment is therefore continuously negotiated through observable repair behaviors.

We frame this coordination problem through



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 226–238
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Clark's theory of grounding and common ground in communication, which treats mutual understanding as a collaborative and distributed achievement rather than an individual process (Clark, 1996; Clark and Brennan, 1991). In Clark's account, interlocutors continuously coordinate to establish sufficient mutual understanding for the task at hand, initiating repair when alignment falters (Schegloff et al., 1977). In mixed-hearing movement instruction, grounding is not only linguistic but visibly embodied: alignment becomes observable in the coordination of bodies, gaze, and timing across participants, interpreters, and instructors.

Using this theoretical lens, we present a participatory observational study of nine mixed-hearing movement and mindfulness sessions. The study was conducted in collaboration with an inclusive movement and mindfulness education center whose accumulated pedagogical knowledge shaped our observational focus. Using synchronized video recordings, instructor interviews, and participant surveys, we analyze how participants, interpreters, and instructors coordinate attention and repair communication breakdowns in real time. This study is guided by three research questions. First, what coordination challenges arise in mixed-hearing movement instruction under high visual load? Second, how do participants, interpreters, and instructors maintain and repair movement coherence across parallel instruction streams? Third, how are these coordination and repair strategies experienced by participants and instructors?

Our findings show that high visual load environments create a recurring coherence–attention–repair cycle. When instruction streams diverge, participants shift attention, initiate repair through available modalities, and gradually restore alignment. We present a taxonomy of repair strategies distributed across actors and show convergent evidence from video annotation, instructor interviews, and participant surveys. We further identify a tension between comprehension and relaxation in mindfulness contexts, where sustained visual monitoring conflicts with inward attention, and we derive design implications for accessible movement instruction and multimodal communication systems.

This paper makes the following contributions:

1. We identify communication and coordination challenges in mixed-hearing movement instruction through a participatory study with instructors (§ 4).
2. We introduce movement coherence as alignment across demonstration, signed interpretation, and bodily execution, and operationalize it through a video annotation framework that tracks coherence breaks and repair sequences (§ 5).
3. We present a taxonomy of repair strategies initiated by participants, interpreters, and instructors, grounded in the theory of mind inferences each actor makes about others' states (§ 5.2).
4. We connect observational findings with participant and instructor surveys to understand how these repair strategies are perceived and experienced in practice (§ 5.4).

We release deidentified materials derived from recorded sessions, interactional annotations, and anonymized instructor and participant survey responses[1].

## 2. Related Work

Our account of movement coherence draws on three research traditions: grounding and common ground in communication, coherence relations in multimodal discourse, and visual attention in signed interaction. We synthesize these perspectives to characterize how participants construct coherent understanding across multiple information streams and how they adapt when coherence breaks.

**Grounding and Common Ground** Understanding in interaction is a collaborative achievement requiring coordination on both content and process (Clark, 1996). Clark and Brennan (1991) analyze grounding as the collective process by which participants reach mutual belief that they have understood one another to a criterion sufficient for current purposes; this account builds on earlier work establishing how speakers and addressees collaborate on contributions (Clark and Schaefer, 1989; Clark and Wilkes-Gibbs, 1986) and how they coordinate through evidence of continued attention (Clark et al., 1983; Schober and Clark, 1989). The effort required for grounding varies with communicative conditions including visibility, audibility, copresence, and shared knowledge (Clark and Brennan, 1991); Isaacs and Clark (1987) show that references between experts and novices require different grounding strategies than those between peers. Enfield (2006) extends this framework by examining the social consequences of common ground. Repair in conversation follows systematic patterns (Schegloff et al., 1977; Schegloff, 1987; Jefferson, 1987), with preference for self-initiated self-repair operating across diverse language communities (Dingemanse et al., 2015). In signed interaction, repair takes modality-specific forms (Manrique and Enfield, 2015; Manrique, 2016): interruption, gestural indication of confusion, and visual confirmation-seeking. Baker (1977) documented how turn-taking in American Sign Language relies on visual attention differently than in spoken discourse.

[1] https://github.com/sakimai/movement-coherence

**Coherence Relations in Multimodal Discourse** Discourse coherence theory (Hobbs, 1985, 1979) characterizes the inferences giving discourse units coherent joint interpretation through relations such as Elaboration, Explanation, and Parallel; Hobbs and Agar (1985) show how these relations enable coherence even in seemingly discontinuous discourse. Asher and Lascarides (2003) formalize these relations within Segmented Discourse Representation Theory, building on dynamic semantic approaches (Kamp and Reyle, 1993; Groenendijk and Stokhof, 1991). Grosz et al. (1995) model local coherence through relationships among focus of attention, choice of referring expression, and perceived coherence, building on earlier work on focus in dialogue (Grosz, 1977; Sidner, 1983). Centering theory posits that certain entities are more central than others, constraining reference choices (Gordon et al., 1993; Brennan, 1995). Walker et al. (1994) extends centering to Japanese, demonstrating cross-linguistic applicability. Stojnić et al. (2017) argue that pronouns are resolved to whatever entity is at the center of attention in coherent discourse; Stojnić et al. (2020) extend this to demonstratives, showing that pointing gestures update context by putting entities at the center of attention. Coherence relations extend to multimodal communication. Lascarides and Stone (2009) provide a formal semantic analysis showing how speech and gesture jointly contribute through structured relations, building on earlier work on nonverbal communication (Kendon, 2004; McNeill, 1992). Work on image-text coherence (Alikhani et al., 2020; Inan et al., 2021) demonstrates that different relations (Visible, Story, Subjective) affect comprehension and evaluation; Hiippala et al. (2021) extends this to diagrams. Inan and Alikhani (2024) show that coherence relations predict eye-gaze patterns during both comprehension and production, suggesting coherence shapes attentional processes during interpretation.

**Visual Attention in Signed Interaction** In sign language, coherence is realized through spatial and temporal organization. Inan et al. (2022) show that prosodic features such as intensification and signing space use carry meaning affecting comprehension; generation models failing to capture these features produce outputs signers find incoherent. Inan et al. (2025) address cross-linguistic coherence in sign-to-sign translation, showing alignment algorithms informed by linguistic structure improve quality. Research on sign language interpretation documents cognitive demands on interpreters (Napier, 2009), but less attention has been paid to demands on recipients. Cognitive load theory (Sweller, 1988; Chandler and Sweller, 1991) attributes learning difficulties to instructional designs exceeding working memory capacity, positioning load as an environmental feature. Alibali and Nathan (2014) show that teachers use gesture to reduce cognitive load, building on work linking gesture to spatial reasoning (Alibali et al., 2000; Goldin-Meadow, 2003). Mindfulness practices have documented benefits for DHH populations (Biegel et al., 2009; Black and Slavich, 2015; Werfel et al., 2025), yet these practices assume attentional conditions that high visual load environments disrupt.

**Participatory Research** Research with DHH communities raises methodological considerations (Singleton et al., 2014; Kusters et al., 2017). Participatory approaches position community members as partners (Bragg et al., 2019; Mack et al., 2020). Our study was conducted in partnership with an inclusive movement and mindfulness education center; community collaborators shaped research questions and interpretive frameworks.

## 3. Methodology

We conducted a participatory observational study with an inclusive movement and mindfulness education center offering wellness programming for DHH community members. The study combined observations of movement and mindfulness classes, semi-structured instructor interviews, and participant and instructor surveys. These sources allowed us to examine how participants construct understanding of instructor intentions and how communication breakdowns emerge and resolve.

**Participatory Partnership** We structured the collaboration following principles from Bragg et al. (2019) for research with Deaf communities. Organization staff participated in defining research questions, shaping observation protocols, and interpreting findings. A Deaf community liaison attended all observation sessions. We committed to returning findings in accessible formats and ensuring that any recommendations would serve community-identified needs.

**Observational Study** We conducted observations following methods for studying embodied interaction in naturalistic settings (Goodwin, 2000; Hutchins, 1995). Observations occurred during regularly scheduled classes over 10 weeks. Each class included eye yoga, gentle movements, and seated mindful breathwork. All classes were led by hearing instructors who provided spoken English instruction with simultaneous ASL interpretation by certified interpreters. Session duration was 50 minutes, with 4 to 10 participants per session. Two synchronized cameras captured complementary perspectives: a front-facing camera documented instructor and interpreter activity while an overview

camera captured participants. This dual-view approach enables analysis linking instruction production to participant reception (Mondada, 2006).

**Expert & Instructor Interviews** To complement observational data, we conducted semi-structured interviews with two instructors and an expert who have led movement and mindfulness programming. Interviews lasted approximately 50 minutes each and addressed teaching challenges, adaptations developed over time, and reflections on what supports or impedes participant engagement. We analyzed interview transcripts thematically, identifying patterns that informed our observational findings.

**Participant & Instructor Surveys** We additionally collected surveys from both participants and instructors. Participants completed surveys about attention strategies, fatigue, and preferred instructional modalities, while instructors completed post session surveys documenting observed confusion signals, instructional strategies used, and communication challenges encountered during each class. These surveys provided self report evidence that complemented the behavioral patterns observed in the video analysis.

**Cohort and Setting** Classes operated as mixed-hearing spaces open to Deaf, Hard-of-Hearing, and hearing participants. Our observed cohort comprised 4 participants (age range [18]-[60]; 4 women). Hearing status included Deaf (1), Hard-of-Hearing (1), and hearing (2). Among DHH participants, all used ASL as a primary or secondary communication mode. The mixed-hearing composition created a heterogeneous attention ecology: hearing participants attended primarily to the instructor while DHH participants attended primarily to the interpreter but also monitored the instructor's demonstrations. Different participants thus drew on different information sources for the same instructional content, facing different visual load profiles.

**Annotation Protocol** We developed the annotation scheme through iterative refinement with community partners and the research team. Each session was segmented into movement episodes corresponding to distinct poses or transitions. Within each episode, we coded four layers: (1) instruction components, marking the onset and offset of spoken description, signed interpretation, and physical demonstration; (2) attention events, including gaze shifts, interpreter repositioning, tactile cues, and eye closure instructions; (3) movement execution states, categorized as aligned, divergent, hesitant, or frozen relative to the demonstrated movement; and (4) repair sequences, marking the initiator (participant, interpreter, or instructor), the trigger (coherence break type), and the outcome (alignment restored or not). The repair categories reported in §5.2 emerged bottom-up from this coding process and were refined over multiple annotation passes.

We used ELAN (Wittenburg et al., 2006) for temporal annotation of the synchronized dual-camera recordings. The first author served as the primary annotator and coded all sessions. To assess reliability, a second annotator independently coded randomly selected movement segments (10 minutes of video) using the same scheme. We computed Cohen's Kappa (Fleiss and Cohen, 1973) for the inter-annotator agreement ($\kappa = 0.67$, *i.e.* substantial agreement). Disagreements were resolved through discussion, and the resulting consensus annotations were used to refine category boundaries for the remaining sessions. The frequency counts reported in Table 1 are based on the primary annotator's full coding of all sessions.

## 4. Challenges in Mixed-Hearing Movement Contexts

Instructor interviews show how the structural constraints of mixed-hearing environments create distinct cognitive and physiological difficulties for participants that make movement coherence fragile and continuously negotiated. These often place the goals of mindfulness in tension with the visual demands of instruction.

**The Paradox of Attention vs. Relaxation** A primary difficulty identified by instructors is the conflict between the visual vigilance required for comprehension and the somatic state required for mindfulness. Instructors noted that DHH participants often remained "activated" throughout the entire session. They continuously monitored the visual field for cues rather than "tapping in" to the relaxation experience. Hearing participants might close their eyes or soften their gaze to focus inward. In contrast, DHH participants were observed to physically "strain" to maintain visual contact with the instructor or interpreter. This constant state of alertness prevents the "drop-in" to the body that is central to yoga. The need to externally monitor effectively creates a barrier to internal proprioceptive awareness.

**The Validation Loop** This heightened activation manifests in what we term a "validation loop." Instructors reported that participants rarely moved on a single cue. Instead, they often required "two to three steps of extra validation" before committing to a pose. A participant might first observe the interpreter and then check the instructor. Finally, they might scan neighboring peers to confirm accuracy. This recursive checking process introduces latency and cognitive load. It disrupts the flow of

movement. Instructors attempted to mitigate this by moving around the room to remain visible. Yet the fundamental need for multi-source confirmation remained a persistent tax on participant attention.

**Heterogeneity and Pacing** Finally, the mixed-hearing cohort presents conflicting demands regarding pacing. Instructors noted a tension between participants who requested a faster flow and those requiring a gentler pace. In a signed environment, speed is a function of visibility. When the pace accelerates, the window for visual processing shrinks. Instructors reported that when the flow became too rapid, the "validation loop" broke down entirely. This caused participants to stop and signal confusion. This suggests that the speed of instruction is limited by the bandwidth of the visual communication channel.

These challenges suggest that coordination in mixed-hearing movement instruction is fundamentally constrained by visual attention and visual bandwidth. Participants must divide attention across multiple information sources while executing movements and engaging in mindfulness practices, making alignment across instruction streams fragile and continuously monitored. When this alignment fails, participants, interpreters, and instructors initiate visible repair actions to restore coordination.

## 5. Our Framework

The challenges reported by instructors describe the constraints of mixed-hearing movement instruction, where multiple instruction streams compete for visual attention and coordination must be maintained under high visual load. To understand how participants nonetheless coordinate their movements in this environment, we now turn to the movement recordings and apply two concepts to what we observed: *movement coherence* and *theory of mind*. We define both below, then apply them to specific cases from the recordings.

### 5.1. Embodied Meaning-Making

In the classes we observed, three streams of information unfold in parallel: the instructor's physical demonstration, the interpreter's signed rendering of the spoken instruction, and the participant's own bodily execution.

**Definition 1.** *We use the term **movement coherence** to refer to the alignment across these three streams: demonstration, signed interpretation, and participant execution. When the instructor extends their arm, the interpreter signs the corresponding description, and the participant extends their arm in the same way, coherence holds. When any of these streams diverge from the others, coherence breaks.*

**Algorithm 1** Movement Coherence Parsing

**Require:** Movement streams for instructor $I$, interpreter $S$, participant $P$; lag threshold $\tau$
**Ensure:** Coherence state sequence $L$
Extract movement features $M_I, M_S, M_P$
**for** each movement episode $t$ **do**
  Compute alignment and temporal lag between streams
  **if** streams are aligned and lag $\leq \tau$ **then**
    $L_t \leftarrow$ ALIGNED
  **else if** streams diverge or lag $> \tau$ **then**
    $L_t \leftarrow$ DIVERGENT
  **end if**
  **if** previous state was DIVERGENT and alignment is re-established **then**
    $L_t \leftarrow$ REPAIRING
  **end if**
**end for**
**return** $L$

To make this construct observable, we annotate coherence at the level of movement episodes using four states: aligned, divergent and repairing. A coherence break is marked when these streams fall out of alignment beyond a brief temporal lag, and a repairing is marked when alignment is re-established. This formulation enables us to track the temporal dynamics of coordination, quantify breakdowns, and relate them to attention shifts and repair strategies across actors. We describe these steps in Algorithm 1. Figure 2(a) shows coherence holding (yellow) across the three actors during aligned segments and breaking (red) when the streams fall out of sync.

Coherence across parallel information streams has been studied in other multimodal domains. Cumming et al. (2017) describe conventions that maintain spatial coherence across discontinuous film shots; Lascarides and Stone (2009) formalize coherence between gesture and speech; Alikhani et al. (2020) model coherence between images and their textual descriptions. We extend this line of work to a triadic embodied setting where demonstrations, signs, and bodily movements must cohere across three actors.

Coherence is also tied to attention. Stojnić et al. (2017) show that in linguistic discourse, coherence relations govern what is at the center of attention at any given point, and that attention determines how context-sensitive expressions are resolved. We observe the same structure in our data. When coherence holds, participants distribute attention across streams. When it breaks, attention narrows to a single stream that can resolve the mismatch. Figure 2(b) shows this cycle: coherence breaks trigger attention shifts, which trigger repair, which restores coherence.

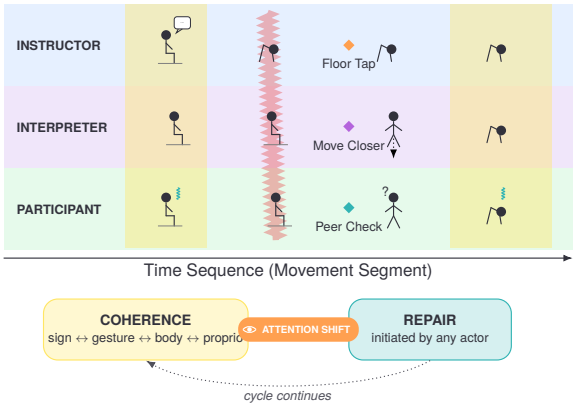



Figure 2: (a) Coherence (yellow) holds when signed instruction, demonstration, and bodily execution align across actors. Breaks (red) trigger attention shifts and repair actions (diamonds). (b) The coherence-repair cycle: attention shift detects breaks and initiates repair by any actor.

Coherence describes whether the three streams are aligned. But it does not explain how people figure out what to do when they come apart. For that, we need theory of mind (ToM): the capacity to infer what others intend and believe (Premack and Woodruff, 1978; Wimmer and Perner, 1983; Apperly, 2010; Schaafsma et al., 2015). Each actor in the room continuously models the others' intentions. The participant infers what movement the instructor wants them to do and whether the interpreter has accurately conveyed it. The interpreter monitors the participant's execution and infers their comprehension state. The instructor reads the group's movements to gauge whether the class is following. When coherence breaks, these inferences drive repair (Clark, 1996; Clark and Brennan, 1991; Schegloff et al., 1977). A participant who points to their own leg is signaling confusion; the interpreter who points back has inferred what the participant needs to know; an instructor who pauses has inferred that the group has fallen behind.

Clark's account of grounding provides a framework for understanding how participants collaboratively establish mutual understanding through visible coordination and repair. However, it remains largely interactional and does not explicitly model how individuals infer others' intentions when coordination fails. We therefore complement grounding with ToM, which captures the inferential processes by which participants interpret others' beliefs and intentions during breakdowns. In our setting, grounding is visible in the alignment of movements across actors, while theory of mind explains how participants decide what corrective action to take when this alignment is disrupted.

We observed this coherence-attention-repair sequence across multiple instances. However, interviews with instructors and experts suggest that these repair sequences are not occasional events but structural features of mixed-hearing movement instruction. Because multiple instruction streams compete for the same visual channel, maintaining coherence requires continuous monitoring, and coherence is frequently placed under pressure.

## 5.2. Repair Strategies

Through a bottom-up, data-driven approach we present a taxonomy of repair strategies (see Table 1). The categories emerged from iterative annotation of the video data and are not an exhaustive list. We organized them by who initiates the repair: participant, interpreter, or instructor.

Figure 3 illustrates a scenario where several of these strategies co-occur. During an eye yoga exercise, participants were instructed to focus on their thumb while shifting visual focus between the thumb and a point behind it. The exercise required sustained visual attention on a specific target while instruction about when to shift focus arrived through the same visual channel. Participants adapted by checking the interpreter between focus shifts, relying on peripheral vision, or following peer timing.

*Participant-initiated repairs.* When visual load exceeded capacity, participants employed adaptive strategies to restore alignment. Some are self-initiated repairs in the conversation-analytic sense (Schegloff et al., 1977); others are clarification-seeking actions directed at another actor (Clark, 1996). Both are triggered by coherence breaks and aim to restore alignment.

- **Peer-checking**: observing neighbors as a proxy for correct movement (Goodwin, 2000).
- **Freeze-and-scan**: pausing movement to sweep gaze across information sources (Kendon, 1967).
- **Strategic disengagement**: reducing monitoring during familiar sequences to conserve attentional resources (Hutchins, 1995).
- **Position-holding**: maintaining a previous pose during uncertainty (Clark, 1996).
- **Signing to the interpreter**: producing a gesture or sign directed at the interpreter to request clarification, such as pointing to a body part to ask "this one?" (Schegloff et al., 1977).
- **Communicating with peers or instructor**: verbally or gesturally addressing another actor for confirmation or request a repeat, showing uncertainty (Clark and Wilkes-Gibbs, 1986).

*Interpreter-initiated repairs.* Interpreters actively monitored participant behavior and adjusted their own positioning and output to restore alignment. These are other-initiated repairs (Schegloff et al., 1977) that operate across the visual and spatial modalities available in the signed environment.

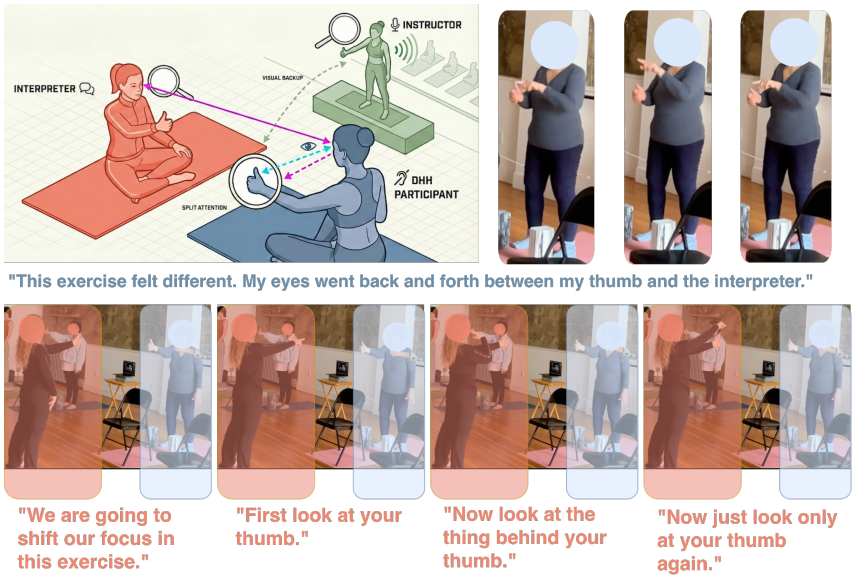



Figure 3: The eye yoga exercise illustrates visual load exceeding capacity at its most acute. The practice demands sustained visual focus on one's thumb while instruction about when to shift focus arrives through the same visual channel. Participants adapt by checking the interpreter between focus shifts, relying on peripheral vision, or following peer timing.

| Repair Strategy | Initiator | Trigger | Modality | N |
|---|---|---|---|---|
| Peer-checking | Participant | Uncertainty about correct movement | Visual (gaze to neighbors) | 7 |
| Freeze-and-scan | Participant | Maximum ambiguity at transition | Visual (gaze sweep) | 3 |
| Strategic disengagement | Participant | Familiar or low-stakes sequence | Visual (gaze softening) | 5 |
| Position-holding | Participant | Uncertain transition | Bodily (maintain prior pose) | 2 |
| Signing to interpreter | Participant | Need for explicit clarification | Gestural/signed (deictic, question) | 3 |
| Communicating with peers/instructor | Participant | Uncertainty, need for confirmation | Verbal/gestural (speech, sign, gesture) | 2 |
| Proximity adjustment | Interpreter | Participant posture limits view | Spatial (repositioning) | 11 |
| Persistent signing | Interpreter | Temporarily outside participant view | Linguistic (continued signing) | 11 |
| Error correction | Interpreter | Participant executes wrong movement | Gestural (pointing, demo) | 4 |
| Strategic non-intervention | Interpreter | Participant visually occluded | None (waiting) | 4 |
| Tactile cueing | Instructor | Participant cannot see visual cues | Tactile (floor tap) | 4 |
| Dual monitoring | Instructor | Ongoing instruction | Visual (checking interp. + participants) | 10 |
| Pacing adjustment | Instructor | Participant appears confused | Temporal (slowing, repeating) | 4 |

Table 1: Taxonomy of repair strategies observed in mixed-hearing movement classes, organized by initiator. Frequency counts (N) are from ELAN annotations across all sessions. Repair is distributed across all three actors, with each actor drawing on the modalities available to them.

- **Proximity adjustment**: moving into the participant's restricted visual field when their posture limits their view (Metzger, 1999).
- **Persistent signing**: continuing to sign even when temporarily outside the participant's view, maintaining the linguistic stream for when visual access resumes (Winston, 1994).
- **Error correction**: noticing a participant's execution error (e.g. wrong side in a mirrored movement) and producing a gestural correction (Schegloff et al., 1977).
- **Strategic non-intervention**: waiting for a participant to complete a visually occluded movement, recognizing that interruption may be more disruptive than partial information (Napier, 2002).

*Instructor-initiated repairs.* Instructors monitored both the interpreter and participants for signs of misalignment and adjusted the pace or modality of instruction accordingly.

- **Tactile cueing**: tapping on the floor to signal transitions through a non-visual channel, enabling communication regardless of visual orientation (Jain et al., 2015).
- **Dual monitoring**: checking both the interpreter's rendering and the participants' execution to verify that information is reaching the group (Clark, 1996).
- **Pacing adjustment**: slowing instruction or repeating a demonstration when participants appear confused, providing additional processing time (Schegloff et al., 1977).

We also observed *error propagation*: the interpreter accidentally performed a movement the instructor had not demonstrated, and the participant copied it. Coherence between participant and in-

terpreter held, but coherence with the instructor's actual instruction was broken. Repair mechanisms can introduce new coherence breaks when the participant cannot verify against multiple sources.

The repair strategies documented above all assume that visual access, even if limited, remains partially available. Participants can still glance at the interpreter, check a neighbor, or scan the instructor. The next section examines what happens when this assumption breaks down entirely.

### 5.3. Eyes-Closed: Visual Constraint at Maximum

Mindfulness components that instructed participants to close their eyes represented the most extreme form of visual constraint. With eyes closed, DHH participants lost access to their primary communication channel entirely. We observed 15 eyes-closed segments in which actors developed coordinated strategies to maintain coherence.

Before eyes-closed segments, participants caught the interpreter's eye and established which signs would indicate breathing in versus breathing out (*pre-occlusion coordination*). During forward folds and other positions that occluded the participant's view, the interpreter moved into the participant's restricted visual field (*interpreter proximity adjustment*). During fully occluded segments, the instructor tapped the floor to provide temporal structure (*tactile timing*). Participants varied in their response: some closed eyes fully, trusting the pre-established coordination; others kept eyes partially open, prioritizing comprehension monitoring over full participation in the mindfulness component.

### 5.4. Evidence from Instructor and Participant Surveys

While our observational analysis characterizes how coherence is maintained and repaired in real time, surveys from participants and instructors provide insight into how these dynamics are experienced and interpreted in practice. These reports reveal which aspects of the coherence–attention–repair cycle are perceptually salient to different actors and how they shape participation in the class.

**Instructor observed confusion signals algin with repair behaviors.** Instructor reports of confusion signals align closely with the repair strategies identified through video annotation. Instructors reported observing scattered gaze (4 sessions), peer copying (3 sessions), and uncertain posture or fidgeting (3 sessions) which are behavioral signals that correspond precisely to freeze-and-scan, peer-checking, and position-holding in Table 1 (see Figure 4, rows 1-3). This convergence strengthens the validity of the repair taxonomy and suggests that instructors recognize the same coherence breakdown signals that are visible in the recordings.

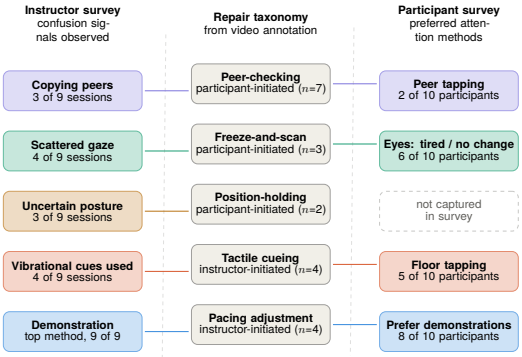



Figure 4: Convergence across instructor surveys, video annotated repair taxonomy, and participant surveys. Each row maps an instructor observed confusion signal to the corresponding repair strategy in Table 1 and, where captured, to the participant-endorsed attention method.

**Instructional strategies that reduce visual monitoring.** Surveys showed that both instructors and participants preferred instructional methods that reduced reliance on signed or verbal explanation, such as physical demonstration, floor tapping, and vibrational cues (Figure 4, rows 4-5). While demonstration is still visual, it differs from signed instruction in that participants can follow movement directly through embodied imitation rather than sustained linguistic attention. Floor tapping and vibrational cues further distribute information into non-visual channels. These strategies reduce the need for continuous monitoring of linguistic instruction.

**Fatigue reflects sustained coherence monitoring.** Post-class survey responses regarding eye fatigue further support the claim that participants must maintain sustained visual monitoring to preserve coherence. Participants reported varying levels of eye fatigue, suggesting heterogeneity in how much visual monitoring individuals maintain during class. This supports the interview findings that some participants prioritize comprehension monitoring while others prioritize relaxation, which reinforces the tension between attention and mindfulness identified earlier.

## 6. Discussion and Implications for Future Design

**Movement coherence extends grounding theory to embodied contexts.** Our observations and interviews extend grounding theory (Clark, 1996) to embodied, high visual load contexts. In

Clark's framework, successful communication requires continuous coordination between interlocutors to establish mutual understanding. Our data show how this coordination operates when linguistic and physical information streams compete for the same attentional channel, and when the goal of communication is bodily alignment rather than propositional agreement. Movement coherence provides a visible, continuous signal of grounding success or failure, and the repair sequences we documented show how interlocutors use these signals to detect and resolve breakdowns. Repair in this setting differs from spoken conversation, where it typically addresses a specific trouble source in the linguistic stream. In movement instruction, repair addresses misalignment between multiple parallel streams (demonstration, interpretation, execution), and any interlocutor can initiate it based on the signals available to them.

**Visual reliance creates a specific burden in mindfulness contexts.** Because DHH individuals communicate through sustained visual attention, their eyes are under constant cognitive and physical strain during instruction. Survey responses and instructor observations both indicate that participants experience visual fatigue and must continuously monitor multiple visual streams to maintain alignment. Mindfulness instruction asks participants to soften external monitoring and attend inward, while maintaining movement coherence requires sustained monitoring.

**Stillness and rest function as repair strategies.** Expert and instructor interviews suggest that moments of stillness function as a repair strategy for the nervous system. When visual and cognitive demands accumulate, participants need structured breaks to reset. Our expert recommended 3 to 10 minute breaks of stillness with eyes closed, focused on body awareness, after which participants perform differently. If comprehension monitoring and relaxation are incompatible within continuous instruction, then alternating between active instruction and structured rest may be more effective than attempting both simultaneously.

**Design implications.** Our findings suggest several implications for the design of inclusive movement instruction in high visual load environments:

- **Consistent structure reduces comprehension monitoring.** Repeated class structure allows participants to anticipate coherent movement and reduce visual monitoring, redirecting cognitive resources toward bodily experience.
- **Instruction should distribute information across sensory modalities and representations.** Our expert recommended activating tactile channels (hand-holding for connection, tapping chains for transitions, self-touch during breathwork), vibrational channels (gongs or vibrating instruments, smartwatch vibrations for timing cues), and proprioceptive engagement (grounding through seated positions, balancing poses). These methods reduce dependence on signed linguistic description and instead use embodied demonstration that participants can follow without sustained visual fixation on an interpreter.
- **Peer pairing distributes the work of maintaining coherence.** Pairs allow participants to observe and correct each other at close range, suggesting that dyadic models (e.g., an avatar guiding two paired participants) may be more effective than broadcast instruction.
- **Technology could support personalization and instructor awareness.** The behavioral markers we identified, such as freezing, scanning, peer-checking, and movement divergence, could be used by multimodal systems to detect coherence breakdowns in real time. Technologies such as AR overlays, avatars (Stoll et al., 2020), or haptic feedback devices could help distribute timing and instruction across visual and non-visual channels, allowing participants to maintain synchronization without continuous visual monitoring.

## 7. Conclusion

By observing movement coherence in mixed-hearing instruction, we provide an analysis framework grounded in everyday embodied communication. We consistently observe that when demonstration, interpretation, and participants' movements fall out of sync, someone in the room initiates a visible repair: participants pause and scan, interpreters reposition or correct, instructors slow down or repeat. Alignment is actively rebuilt in real time. We also observe a tension between comprehension and relaxation, where sustained visual monitoring limits the inward attention that mindfulness practices are designed to support. Expert and instructor interviews point toward concrete design principles for addressing these challenges: an asset-based orientation, multi-modal instruction, consistent class structure, peer pairing, and structured stillness. Throughout this work, participatory design has been essential to identifying challenges that matter to the DHH community and will remain essential to developing solutions that serve community needs. With the framework and data released through this study, we hope to spearhead research into multimodal communication specifically for *in situ* movement and mindfulness settings in mixed-hearing environments.

## 8. Limitations

Our cohort was drawn from a single organization, limiting generalizability across DHH communities, regional signing varieties, and class formats. Deaf-led classes without interpretation would likely show different dynamics. Our analysis privileges observable behavior; participants' subjective experience may differ from our interpretations. The participatory framework shaped research questions toward community priorities, potentially limiting attention to phenomena of primarily academic interest.

## 9. Ethical Considerations

This study was conducted under IRB approval (#25-08-35). All participants provided informed consent, and all data were deidentified prior to analysis and publication. We took care to minimize any potential risks to participants, and no harm resulted from the study.

## 10. Acknowledgements

We thank Open Up Pittsburgh and all the participants of the classes who volunteered for us to use their data. We thank Megan Flinn for their guidance and thoughtful advice on the study design, development of research questions, and feedback throughout the project.

## 11. Bibliographical References

Martha Alibali, Sotaro Kita, and Amanda Young. 2000. Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes - LANG COGNITIVE PROCESS*, 15:593–613.

Martha W Alibali and Mitchell J Nathan. 2014. Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. In *Video research in the learning sciences*, pages 349–365. Routledge.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Ian Apperly. 2010. *Mindreaders: The cognitive basis of "Theory of Mind"*. Psychology Press, Hove, UK.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

Charlotte Baker. 1977. Regulators and turn-taking in American Sign Language discourse. *On the other hand*, pages 215–236.

Gina M Biegel, Kirk Warren Brown, Shauna L Shapiro, and Christine M Schubert. 2009. Mindfulness-based stress reduction for the treatment of adolescent psychiatric outpatients: A randomized clinical trial. *Journal of Consulting and Clinical Psychology*, 77(5):855–866.

David S Black and George M Slavich. 2015. Mindfulness meditation and the immune system: A systematic review of randomized controlled trials. *Annals of the New York Academy of Sciences*, 1373(1):13–24.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

Susan E. Brennan. 1995. Centering attention in discourse. *Language and Cognitive Processes*, 10(2):137–167.

Paul Chandler and John Sweller. 1991. Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4):293–332.

Herbert H. Clark. 1996. *Using Language*. "Using" Linguistic Books. Cambridge University Press.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Herbert H Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22(2):245–258.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Samuel Cumming, Gabriel Greenberg, and Rory Kelly. 2017. Conventions of viewpoint coherence in film. *Philosophers' Imprint*, 17(1):1–29.

Mark Dingemanse, Seán G Roberts, Julija Baranova, et al. 2015. Universal principles in the repair of communication problems. *PloS one*, 10(9):e0136100.

Nicholas J Enfield. 2006. Social consequences of common ground. In *Roots of Human Sociality: Culture, Cognition and Interaction*, pages 399–430. Berg, Oxford.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Susan Goldin-Meadow. 2003. Hearing gesture: How our hands help us think.

Charles Goodwin. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522.

Peter C. Gordon, Barbara J. Grosz, and Laura A. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17(3):311–347.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–100.

Barbara J. Grosz. 1977. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'77, page 67–76, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A. Bateman. 2021. AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. *Lang. Resour. Eval.*, 55(3):661–688.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Jerry R Hobbs. 1985. On the coherence and structure of discourse. In Livia Polanyi, editor, *The Structure of Discourse*. Ablex, Norwood, NJ.

Jerry R Hobbs and Michael H Agar. 1985. The coherence of incoherent discourse. *Journal of Language and Social Psychology*, 4(3-4):213–232.

Edwin Hutchins. 1995. *Cognition in the Wild*. MIT Press, Cambridge, MA.

Mert Inan and Malihe Alikhani. 2024. Seeing eye-to-eye: Cross-modal coherence relations inform eye-gaze patterns during comprehension & production. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14494–14512, Torino, Italia. ELRA and ICCL.

Mert Inan, Piyush Sharma, Baber Khalid, Radu Soricut, Matthew Stone, and Malihe Alikhani. 2021. COSMic: A coherence-aware generation metric for image descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3419–3430, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mert Inan, Yang Zhong, and Malihe Alikhani. 2025. How to align multiple signed language corpora for better sign-to-sign translations? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4003–4016, Albuquerque, New Mexico. Association for Computational Linguistics.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling intensification for sign language generation: A computational approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.

Ellen A Isaacs and Herbert H Clark. 1987. References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1):26–37.

Dhruv Jain, Leah Findlater, Jamie Gilkeson, Benjamin Holland, Ramani Duraiswami, Dmitry Zotkin, Christian Vogler, and Jon E. Froehlich. 2015. Head-mounted display visualizations to support sound awareness for the deaf and hard of hearing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 241–250, New York, NY, USA. Association for Computing Machinery.

Gail Jefferson. 1987. On exposed and embedded correction in conversation. In *Talk and Social Organisation*, pages 86–100. Multilingual Matters.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

Annelies Kusters, Maartje De Meulder, and Dai O'Brien. 2017. *Innovations in deaf studies: The role of deaf scholars*. Oxford University Press.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.

Kelly Mack, Danielle Bragg, Meredith Ringel Morris, Maarten W. Bos, Isabelle Albi, and Andrés Monroy-Hernández. 2020. Social app accessibility for deaf signers. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Elizabeth Manrique. 2016. Other-initiated repair in Argentine Sign Language. *Open Linguistics*, 2(1).

Elizabeth Manrique and Nicholas J Enfield. 2015. Suspending the next turn as a form of repair initiation: Evidence from Argentine Sign Language. *Frontiers in Psychology*, 6:1326.

David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Melanie Metzger. 1999. *Sign Language Interpreting: Deconstructing the Myth of Neutrality*. Gallaudet University Press, Washington, DC.

Lorenza Mondada. 2006. Video recording as the reflexive preservation and configuration of phenomenal features for analysis. In *Video analysis: methodology and methods : qualitative audiovisual data analysis in sociology*, pages 51–68. Lang.

Jemina Napier. 2002. *Sign Language Interpreting: Linguistic Coping Strategies*. Douglas McLean, Coleford, UK.

Jemina Napier. 2009. *Exploring linguistic and cultural identity: My personal experience*.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Sara M. Schaafsma, Donald W. Pfaff, Robert P. Spunt, and Ralph Adolphs. 2015. Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, 19(2):65–72.

Emanuel A Schegloff. 1987. Recycled turn beginnings: A precise repair mechanism in conversation's turn-taking organisation. *Talk and Social Organisation*, pages 70–85.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Michael F Schober and Herbert H Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.

Candace L Sidner. 1983. Focusing in the comprehension of definite anaphora. *Computational Models of Discourse*, pages 267–330.

Jenny L. Singleton, Gabrielle Jones, and Shilpa Hanumantha. 2014. Toward ethical research practice with deaf participants. *Journal of Empirical Research on Human Research Ethics*, 9(3):59–66. PMID: 25746786.

Una Stojnić, Matthew Stone, and Ernest Lepore. 2017. Discourse and logical form: pronouns, attention and coherence. *Linguistics and Philosophy*, 40(5):519–547.

Una Stojnić, Matthew Stone, and Ernest Lepore. 2020. Pointing things out: In defense of attention and coherence. *Linguistics and Philosophy*, 43(2):139–148.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks. In *International Journal of Computer Vision*, volume 128, pages 891–908. Springer.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Marilyn A Walker, Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232.

Krystal L. Werfel, Jessica Mattingly, and Emily Lund. 2025. Yoga as a fatigue intervention for children who are deaf and hard of hearing. *Perspectives of the ASHA Special Interest Groups*, 10(4):1180–1189.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Elizabeth A Winston. 1994. An interpreted education: Inclusion or exclusion? *Gallaudet University Communication Forum*, 3:55–62.

## 12. Language Resource References

Wittenburg, Peter and Brugman, Hennie and Russel, Albert and Klassmann, Alex and Sloetjes, Han. 2006. *ELAN: A professional framework for multimodality research*. The Language Archive, Max Planck Institute for Psycholinguistics.

# A Comparative Analysis of Traditional and Contemporary Visual Features for Computational Annotation of Irish Sign Language

**Sarmad Khan**[*†], **Simon D. McLoughlin**[†], **Irene Murtagh**[*†]
[*]ADAPT: AI-Driven Digital Content Technology, [†]Technological University Dublin
Dublin, Ireland
sarmad.khan@adaptcentre.ie, simon.d.mcloughlin@tudublin.ie, irene.murtagh@adaptcentre.ie

**Abstract**

Automatic annotation of sign language data is critical for advancing linguistic research and developing sign language technologies, yet it remains a major bottleneck due to the inherently motion-based and multi-modal nature of signing. Irish Sign Language, like many sign languages, presents challenges for computational annotation and sign language processing due to limited annotated corpora and the inherent difficulty of reliably annotating movement, trajectories, and coarticulation across manual and non-manual articulators. This paper presents an automated computational framework for gloss-level annotation support in Irish Sign Language, designed to assist scalable corpus annotation by learning motion-related cues directly from sign language videos. Using ELAN-aligned segments from the Signs of Ireland Corpus, we compare contemporary self-supervised visual representations with traditional pose-based features derived from explicit skeletal tracking, evaluating three feature configurations: DINOv2, MediaPipe, and multi-modal fusion. Our results show that self-supervised visual embeddings achieve the highest average accuracy 86.12 %, outperforming both multi-modal fusion 84.28 % and pose-based representations 76.74 %. This indicates that recent visual models can implicitly encode linguistically relevant motion information, including articulator movement and transitional dynamics, reducing the need for explicit landmark extraction in practical annotation pipelines. Overall, this work provides empirical guidance and a deployable computational framework to support computational annotation and enrichment of sign language corpora.

**Keywords:** Irish Sign Language, Automated Computational Annotation, Self-Supervised Learning, Sign Language Corpora

## 1. Introduction

The automatic annotation of Sign Language has the potential to advance corpus-based linguistic research, support the development of sustainable language resources, and facilitate sign language technologies that address communication barriers (Bragg et al., 2019). Sign languages are expressed through continuous movement: lexical and grammatical meaning is realised through coordinated movement trajectories of the hands, body, and face, unfolding continuously over time. Capturing this motion in a form that is both linguistically meaningful and computationally tractable remains a central challenge for the creation and analysis of sign language resources.

While substantial progress has been made for sign languages such as American Sign Language (ASL) and British Sign Language (BSL), Irish Sign Language (ISL) remains under-resourced with respect to annotated and data-driven processing resources (Khan et al., 2025). ISL is officially recognised as a native language of Ireland under the Irish Sign Language Act 2017. It serves as the first or preferred language for approximately 5,000 Deaf people in Ireland, with an estimated 40,000 people using ISL more broadly (Irish Deaf Society, 2024). Limited annotated resources, together with ISL's distinct grammar, gender-based linguistic variation, and multi-modal articulation, constrain progress in computational modelling of the language (Murtagh, 2021). Due to the visual-gestural nature of Sign Language, expert manual annotation is labour-intensive and difficult to scale, particularly when identifying gloss boundaries within continuous, coarticulated motion. Recent advances in artificial intelligence offer new support for sign language annotation by learning visual and temporal representations directly from video (Lugaresi et al., 2019; Caron et al., 2021). Pose-based methods explicitly model articulator movement via skeletal landmarks, whereas contemporary self-supervised vision models learn visual representations without manual feature design. For sign languages such as ISL, it remains unclear which visual feature extraction approaches best support scalable gloss-level annotation. From an annotation perspective, the central challenge is not only achieving high recognition accuracy, but enabling reliable computational annotation processes that reduce annotator workload across continuous, coarticulated signing.

This paper presents an automated computational annotation support pipeline for ISL that compares traditional pose-based representations with contemporary self-supervised visual features, as well as their multimodal fusion. We focus on evaluating how different representations encode motion across gloss-level sign segments and how this impacts their suitability for annotation-oriented workflows. Using ELAN-aligned data from an ISL cor-



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 239–247
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

pus, we conduct a comprehensive ablation study across 24 model configurations (8 classifiers $\times$ 3 feature representations) to assess representation robustness, efficiency, and scalability.

The main contributions of this paper are twofold:

- A comparative benchmark of traditional pose-based features (MediaPipe), contemporary self-supervised visual features (DINOv2), and their multimodal fusion for ISL gloss-level annotation support, highlighting the role of implicit motion representation.

- An empirical insight relevant to sign language corpora: self-supervised visual embeddings outperform multimodal fusion in both accuracy and computational efficiency, suggesting that implicit motion cues may suffice for scalable annotation support without explicit skeletal modelling.

The remainder of this paper is structured as follows: Section 2 reviews related work on sign language corpora and motion-based representations, Section 3 describes the dataset and preprocessing pipeline, Section 4 details the methodology, Section 5 presents experimental results, Section 6 discusses implications for sign language annotation and resource development, and Section 7 concludes the paper.

## 2. Related Work

Research on computational processing of sign languages has expanded rapidly in recent years, driven by advances in computer vision and machine learning aimed at improving accessibility and supporting linguistic analysis. Early work in sign language recognition primarily relied on convolutional neural networks and handcrafted motion descriptors to model visual gestures. Studies on Irish Sign Language (ISL) demonstrated that motion-based representations, such as Motion History Images and CNNs, can effectively capture hand shape and movement for isolated recognition tasks (Khan et al., 2024, 2025). While effective for constrained settings such as fingerspelling, these approaches are limited in their ability to represent the complex spatio-temporal dynamics required for gloss-level annotation in continuous signing.

Subsequent research introduced pose estimation to explicitly model articulator movement using skeletal landmarks. Frameworks based on MediaPipe and similar toolkits enabled the extraction of hand, body, and facial keypoints, which were combined with CNNs, recurrent networks, or attention mechanisms to improve recognition performance (Shukla and Gupta, 2025; Siju and Selvam, 2024). These approaches demonstrated that explicit modelling of articulator trajectories can support sign recognition, particularly in real-time or resource-constrained environments. However, pose-based pipelines introduce additional complexity and remain sensitive to occlusion, viewpoint variation, and tracking errors, which are especially problematic in large-scale corpus annotation. More recent work has shifted toward transformer-based architectures and self-supervised learning, which learn visual representations directly from raw video without manual feature specification. Vision transformers and masked-feature learning have been shown to capture rich spatial relationships and short-range temporal context through attention mechanisms, enabling more flexible modelling of sign language motion (Ravi Kiran et al., 2025; Caron et al., 2021). These representations have proven effective for both isolated and continuous sign language tasks and reduce reliance on explicit skeletal tracking. Parallel to advances in recognition models, several studies have highlighted the challenges of manual sign language annotation. Identifying sign boundaries within continuous motion, dealing with coarticulation, and integrating manual and non-manual features make annotation slow, subjective, and difficult to scale (Tian et al., 2024). Annotation is extremely time-intensive, with detailed linguistic annotation often requiring hours per minute of video, limiting the size and diversity of available corpora. Automated and semi-automated annotation tools have therefore become an important research focus, particularly for low-resource sign languages. Similarly, Low et al. (Low et al., 2025) propose a segment-aware visual tokenisation framework for gloss-free sign language translation, achieving computational efficiency through semantically informed temporal reduction. This work demonstrates that segment-level visual representations can reduce sequence length by 50% while maintaining translation quality.

Despite these developments, relatively little work has systematically examined which visual representations are most suitable for scalable, annotation-oriented processing of low-resource sign languages such as Irish Sign Language. Much of the existing research has focused on resource-rich sign languages, particularly American Sign Language (ASL) and British Sign Language (BSL), where large annotated datasets enable the use of complex models and extensive training regimes. In contrast, sign languages like ISL face additional constraints related to limited data availability, uneven distribution of gloss categories, and restricted annotation capacity. Many existing approaches focus either on pose-based or appearance-based features in isolation, and naive multi-modal fusion strategies remain underexplored in annotation contexts, especially for low-resource settings. This work addresses this gap by evaluating self-supervised visual representations, pose-based features, and their fusion within a unified computational annotation pipeline,

with an emphasis on motion representation, computational efficiency, and practical deployment for sign language corpus development.

## 3. Dataset and Preprocessing

### 3.1. Corpus Description

This study uses the **Signs of Ireland Corpus**, a linguistically annotated collection of Irish Sign Language (ISL) recordings developed at Trinity College Dublin (Leeson, 2004). The data comprises approximately 10 hours of naturalistic narrative signing collected in 2004 as part of a corpus development project involving members of the Irish Deaf community and researchers at Trinity College Dublin (Leeson, 2004). The corpus includes 40 Deaf native or early learners of ISL (24 female, 16 male), spanning four age groups (18–30, 30–45, 45–60, 65+) and five regions across Ireland (Dublin, Galway, Wexford, Waterford, Cork). Recordings consist primarily of narrative stories and retellings, including a subset of the Frog Story dataset, with time-aligned gloss annotations provided in ELAN XML format.

### 3.2. ELAN Parsing and Gloss Alignment

Each video in the Signs of Ireland Corpus is accompanied by an ELAN XML file containing time-aligned annotations for every gloss instance (Wittenburg et al., 2006). These annotations define precise temporal boundaries for the start and end of each sign, enabling frame-level segmentation and feature extraction. Formally, each gloss annotation is represented as:

$$A_i = \{ID_i, t_{\text{start}}, t_{\text{end}}, g_i\} \tag{1}$$

where $ID_i$ is the unique annotation identifier, $t_{\text{start}}, t_{\text{end}}$ is the temporal boundaries of the gloss (in milliseconds) and $g_i$ lexical gloss label.

To ensure consistency between the video data and its annotations, each gloss was automatically matched to its corresponding video file using the `MEDIA_URL` field embedded in the ELAN XML metadata. This process established a one-to-one mapping between gloss annotations and their respective video segments. After automated alignment and validation, a total of 3,865 gloss instances were successfully mapped from ELAN annotations to video clips, forming the basis for subsequent frame extraction and feature computation.

### 3.3. Gloss Filtering and Curation

To ensure high-quality input for training and evaluation, a four-stage filtering pipeline was applied to the raw ELAN gloss annotations:

1. **Frequency Thresholding**: Only glosses with at least 20 distinct occurrences were retained to ensure sufficient representation and statistical reliability per class.

2. **Lexical Normalisation**: All gloss labels were converted to `UPPERCASE` and cleaned by removing non-linguistic characters (e.g., `!`, `?`, numeric suffixes). The final curated the vocabulary of 55 glosses consists entirely of established lexical forms (e.g., `MOTHER`, `WORK`, `DAY`), as compound and reduplication variants did not meet the minimum frequency threshold for inclusion.

3. **Linguistic Exclusion**: Non-lexical, ambiguous, or structurally inconsistent items were excluded to improve dataset coherence, as follows:

   - *Classifiers* (e.g., `CL-THING`, `CL-PERSON`): excluded as they represent productive, gradient forms rather than established lexical entries.
   - *Indexing signs* (e.g., `INDEX1`, `INDEX2`): removed as they are referential and context-dependent rather than lexically stable.
   - *Name signs and anonymised placeholders* (e.g., `_JOHN`, `*XYZ`): excluded as they are person-specific and not generalisable.
   - *Emblems and expressive gestures* (e.g., `THUMBS-UP`, `GESTURE`): excluded as they lack consistent phonological form across signers.

4. **Manual Review**: The final subset was manually reviewed to ensure:

   - Semantic and visual consistency across all occurrences.
   - Preference for established lexical glosses (standardised, dictionary-like entries) over productive glosses (context-specific, creative, or infrequent signs) (Sutton-Spence, 2007).

As a result, the dataset was reduced from an initial pool of 3,865 annotated gloss instances to a curated vocabulary of 55 distinct ISL gloss labels. While the full corpus contained annotations across hundreds of glosses, filtering by frequency ($\geq 20$) and linguistic consistency retained approximately 1,950 gloss instances mapped to these 55 target glosses. This curation prioritises lexical consistency and sufficient per-class representation, which is critical for stable learning in less-researched sign language settings. Alternative preprocessing strategies would likely influence model behaviour. For example, retaining lower-frequency glosses would increase vocabulary coverage but introduce greater class imbalance and variability, potentially reducing

classification reliability. Similarly, less strict filtering may introduce inconsistencies in gloss labels, affecting both training stability and evaluation outcomes. The final dataset, therefore reflects a deliberate trade-off between linguistic coverage and statistical robustness.

## 3.4. Dataset Statistics

All gloss instances were uniformly sampled to a fixed number of frames to ensure consistent representation across samples. For each gloss instance with duration $[t_{\text{start}}, t_{\text{end}}]$, we extract 20 frames using uniform temporal spacing:

$$t_i = t_{\text{start}} + i \cdot \frac{t_{\text{end}} - t_{\text{start}}}{19}, \quad i \in \{0, 1, \ldots, 19\} \quad (2)$$

This ensures consistent temporal coverage regardless of the original gloss duration, with frames indexed from $i = 0$ (start) to $i = 19$ (end). Table 1 summarises the key properties of the curated ISL dataset used in this study, including vocabulary size, sampling strategy, and annotation format.

Table 1: Summary of Curated Dataset Statistics

| Metric | Description |
|---|---|
| Gloss Vocabulary | 55 distinct ISL gloss labels |
| Gloss Instances | ∼1,950 annotated gloss segments |
| Frame Resolution | $448 \times 448$ |
| Sampling Strategy | 20 uniformly spaced frames per gloss instance |
| Data Split | 80% training / 20% testing (stratified by gloss) |
| Annotation Format | ELAN XML with gloss-level temporal segmentation |

## 3.5. Feature Extraction

To construct semantically rich representations for Irish Sign Language gloss classification, we extract two complementary feature modalities from each video frame. These feature representations are designed to support gloss-level computational annotation by capturing motion cues that distinguish lexical signs within continuous signing.

### 3.5.1. DINOv2 Visual Features

DINOv2 is a self-supervised vision model that learns transferable features via self-distillation (Oquab et al., 2023). For each input frame $I$, the image is resized to $448 \times 448$ pixels and normalised to the range $[-1, 1]$ using standard scaling. The processed frame is passed through the DINOv2 ViT-S/14 encoder, yielding two levels of representation:

- **CLS Token Vector** ($\mathbf{f}_{\text{cls}} \in R^{384}$): A global descriptor obtained from the special classification token at the transformer's output. It captures coarse-level context such as signer pose, location, and signing space.
- **Patch Tokens** ($\mathbf{P} \in R^{256 \times 384}$): Representations corresponding to $16 \times 16$ spatial patches of the input frame, enabling localised visual analysis.

To remove background noise and emphasise signer-specific regions, we apply **PCA-based foreground selection** over the patch tokens:

$$M_{\text{fg}} = \left\{ p_i \in P \;\middle|\; \frac{p_i - \min(p)}{\max(p) - \min(p)} < 0.4 \right\} \quad (3)$$

The selected foreground tokens are averaged to produce a local structure embedding:

$$\mathbf{f}_{\text{patch}} = \frac{1}{|M_{\text{fg}}|} \sum_{p \in M_{\text{fg}}} p \quad (4)$$

Finally, the visual feature vector is constructed by concatenating the global and local descriptors:

$$\mathbf{f}_{\text{visual}} = [\mathbf{f}_{\text{cls}}; \mathbf{f}_{\text{patch}}] \in R^{768} \quad (5)$$

### 3.5.2. MediaPipe Skeletal Features

To complement appearance features, we extract skeletal motion and pose information using MediaPipe Holistic (Grishchenko and Bazarevsky, 2020). For each frame, the following landmark sets are obtained:

- **33 Body Pose Landmarks:** Each with $(x, y, z, v)$ coordinates, giving $33 \times 4 = 132$ features.
- **21 Hand Landmarks per Hand:** Both hands are represented using $(x, y, z)$ for each keypoint, totaling $2 \times 21 \times 3 = 126$ features.

The full skeletal vector is then formed as:

$$\mathbf{f}_{\text{pose}} = [\text{pose}; \text{hands}] \in R^{258} \quad (6)$$

This pose vector captures articulator positions, joint geometry, and inter-limb relationships relevant to distinguishing ISL glosses.

## 3.6. Multi-modal Feature Fusion

To integrate complementary information from both modalities, we concatenate the DINOv2-based visual embeddings with the MediaPipe-based skeletal features:

$$\mathbf{f}_{\text{multi}} = [\mathbf{f}_{\text{visual}}; \mathbf{f}_{\text{pose}}] \in R^{1026} \quad (7)$$

This combined representation encodes both high-level semantic content (via DINOv2) and

fine-grained spatial articulations (via pose landmarks), enriching the feature space for classification. Table 2 provides a detailed summary of all extracted features, their dimensionalities, and semantic meanings:

Table 2: Extracted Feature Representations

| Feature | Dim. | Description |
|---|---|---|
| DINOv2 CLS | 384 | Global visual embedding |
| DINOv2 Patch Mean | 384 | PCA-selected foreground patches |
| MediaPipe Body | 132 | 33 body landmarks $(x, y, z, v)$ |
| MediaPipe Hands | 126 | 21 landmarks $\times$ 2 hands $(x, y, z)$ |
| **Visual (DINOv2)** | **768** | CLS + patch features |
| **Pose (MediaPipe)** | **258** | Body + hand landmarks |
| **Multi-modal** | **1026** | Visual + pose features |

# 4. Methodology

## 4.1. Experimental Design

We conduct a comparative ablation study to systematically identify the most effective feature representation for Irish Sign Language (ISL) gloss classification. The experimental framework evaluates three distinct feature configurations MediaPipe (traditional skeletal tracking), DINOv2 (contemporary vision features), and their multi-modal fusion across eight diverse classification algorithms spanning gradient boosting, kernel methods, neural networks, and probabilistic models.

This design enables rigorous comparison of traditional versus contemporary approaches while empirically testing whether multi-modal fusion yields complementary benefits over single-modality representations. Figure 1 illustrates the comprehensive methodological framework employed for ISL gloss classification, detailing each stage from data preprocessing to model evaluation.

## 4.2. Feature Configurations

We evaluate three feature representations that embody different paradigms for SL understanding (detailed extraction methodology provided in Section 3.5):

### 4.2.1. MediaPipe (Traditional Skeletal Features)

Explicit articulator tracking via Google MediaPipe Holistic, extracting 258-dimensional skeletal keypoint features comprising 33 pose landmarks (x, y, z, visibility) and 21 hand landmarks per hand (x, y, z). This traditional approach leverages geometric relationships between hand configurations, spatial positions, and body pose to represent signing gestures through explicit articulator trajectories.

### 4.2.2. DINOv2 (Contemporary Vision Features)

Holistic visual representation via Meta AI's DINOv2-ViT-S/14 self-supervised vision transformer, yielding 768-dimensional feature vectors. We extract the 384-dimensional CLS token capturing global image semantics and mean-pool the top 40% most salient patch features (384 dimensions), selected via PCA-based saliency analysis to isolate signing-relevant regions. This contemporary approach captures holistic visual patterns including motion blur, spatial context, hand-object interactions, and appearance cues without requiring explicit keypoint detection.

### 4.2.3. Multi-modal Fusion

Simple early fusion concatenates DINOv2 visual features (768D) with MediaPipe skeletal features (258D), yielding a (1026D) multi-modal representation. This configuration tests whether combining global visual patterns with explicit articulator geometry provides complementary discriminative information for robust ISL gloss annotation.

## 4.3. Classifier Diversity

We evaluate three feature representations across eight standard classifiers: XGBoost, CatBoost, Random Forest, Support Vector Machines, Logistic Regression, Multi-Layer Perceptron, K-Nearest Neighbors, and Naive Bayes spanning ensemble, kernel-based, neural, probabilistic, and instance-based learning paradigms to ensure a robust and model-agnostic comparison.

## 4.4. Training Protocol and Hyperparameter Optimisation

All classifiers are trained using an 80:20 stratified train–test split. Hyperparameters follow standard best practices with empirical tuning and early stopping to control overfitting. Table 3 summarises the training configurations adopted for each classifier.

## 4.5. Evaluation Metrics

We evaluate classification performance using a standard set of metrics, capturing both overall accuracy and class-level behaviour. These include accuracy, macro- and weighted F1-scores to account for class imbalance, as well as Top-$K$ accuracy to reflect annotation-oriented use cases where multiple candidate glosses may be presented. Model performance is evaluated against expert-annotated ELAN gloss labels from the Signs of Ireland Corpus, which serve as the ground truth. Predictions

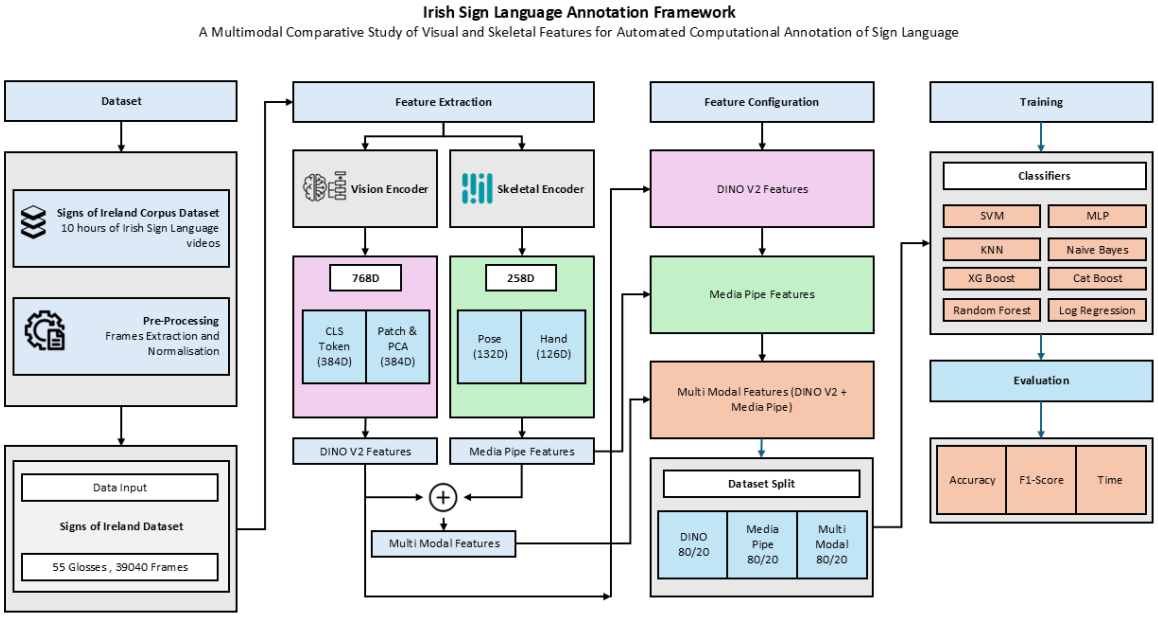



Figure 1: End-to-end computational pipeline for Irish Sign Language gloss annotation, illustrating dataset preparation, visual and skeletal feature extraction, multi-modal fusion, and model evaluation.

Table 3: Key Hyperparameter Configurations

| Classifier | Configuration |
|---|---|
| XGBoost | 1000 trees, depth 6, lr 0.15, early stopping |
| CatBoost | 1000 iters, depth 6, lr 0.1, early stopping |
| Random Forest | 300 trees, max depth 20, Gini split |
| SVM | RBF kernel, $C$=1.0 |
| Logistic Regression | L2 regularisation, $C$=1.0, 1000 iters |
| MLP | Hidden layers (200,100), ReLU, reg $= 0.001$ |
| KNN | $k$=7, Euclidean distance, weighted voting |
| Naive Bayes | Smoothing $= 10^{-8}$ |

Table 4: Evaluation Metrics

| Metric | Description |
|---|---|
| Accuracy | Overall proportion of correct predictions |
| Top-K Accuracy | True label appears among the top K predicted labels |
| Macro F1 Score | Average F1 score across all classes (equal weighting) |
| Weighted F1 Score | F1 score adjusted for class imbalance |
| Log Loss | Measures prediction confidence and penalises errors |
| Per-Class F1 | F1 score reported individually for each class |

Table 5: Implementation Environment and Tools

| Component | Tools and Libraries |
|---|---|
| Programming | Python 3.10+ |
| Visual Features | PyTorch 2.0+, DINOv2 (via `torch.hub`) |
| Skeletal Features | MediaPipe Holistic v0.10.0 |
| Hardware | GPU-accelerated (Google Colab Pro, CUDA-enabled) |

are compared directly with these annotations using standard classification metrics.Table 4 summarises the metrics used.

### 4.6. Experimental Setup

All experiments were conducted in Python 3.10+ using modern machine learning and vision libraries. Table 5 outlines the key software and hardware setup. All experiments used fixed random seeds (`random_state=42`) for data splitting, initialisation, and validation. The complete experimental workflow is implemented, with structured JSON metadata stored for full reproducibility and traceability.

## 5. Results

### 5.1. Overall Performance

Across all classifiers, DINOv2 features achieve the highest average accuracy, outperforming both multi-modal fusion and MediaPipe-based representations. The strongest individual result is obtained

using XGBoost with DINOv2 features. Table 6 provides average accuracy across all classifiers.

Table 6: Average classification accuracy across all eight classifiers.

| Feature Configuration | Avg. Accuracy (%) |
|---|---|
| DINOv2 | **86.12** |
| MediaPipe | 76.74 |
| Multi-modal | 84.28 |

## 5.2. Cross-Model Comparison

Several key observations emerge from this comprehensive comparison:

- **DINOv2 dominance:** DINOv2-based features achieve the highest accuracy for the majority of classifiers, with particularly strong performance for gradient boosting methods, where XGBoost reaches a peak accuracy of 95.91%.
- **Performance Variation**: Naive Bayes performs poorly across all configurations (13.97–25.32%), as its feature independence assumption is violated by the correlated DINOv2 embeddings and kinematically dependent MediaPipe landmarks, rendering it unsuitable for sign language feature representations.
- **Limited benefit of fusion:** Multi-modal fusion does not consistently outperform DINOv2 representations and, in several cases, results in reduced accuracy, suggesting redundancy or interference between visual and pose features.
- **Classifier sensitivity:** MediaPipe features perform competitively for some ensemble models but degrade substantially for linear and probabilistic classifiers, indicating limited suitability of explicit skeletal representations for certain learning paradigms.
- **Model robustness:** Contemporary self-supervised visual embeddings generalise more reliably across diverse classifier families than pose-based or fused representations, supporting their use in annotation-oriented workflows.

## 5.3. Computational Efficiency Analysis

An important consideration for practical deployment is the trade-off between accuracy and computational cost. Table 7 compares average accuracy and training time across all eight classifiers for each feature configuration.

- **DINOv2:** Provides the best balance between accuracy and computational cost across classifiers, making it well suited for scalable annotation workflows.

Table 7: Model-wise efficiency comparison: accuracy vs. training time

| Model | DINOv2 | | MediaPipe | | Multi-modal | |
|---|---|---|---|---|---|---|
| | Acc. | Time | Acc. | Time | Acc. | Time |
| XGBoost | 95.91 | 272 | 95.12 | 163 | 95.61 | 3,881 |
| CatBoost | 95.84 | 299 | 95.17 | 134 | 94.74 | 4,646 |
| Random Forest | 95.43 | 13 | 94.83 | 7 | 85.07 | 20 |
| MLP | 95.26 | 217 | 91.09 | 117 | 94.74 | 15 |
| KNN | 95.35 | 0.1 | 86.74 | 0.03 | 92.57 | 42 |
| Log. Reg | 94.45 | 50 | 60.05 | 22 | 94.56 | 4 |
| SVM | 93.99 | 96 | 76.96 | 47 | 91.61 | 7 |
| Naive Bayes | 22.69 | 0.5 | 13.97 | 0.1 | 25.32 | 0.5 |
| **Avg.** | 86.12 | 119 | 76.74 | 61 | 84.28 | 1,077 |

- **Multi-modal:** Incurs substantially higher training cost without corresponding accuracy gains, indicating limited efficiency for practical deployment.
- **MediaPipe:** Trains fastest but yields lower accuracy, favouring speed-oriented scenarios over annotation quality.

## 5.4. Per-Class Analysis

To evaluate gloss-wise classification performance, we analyse per-class F1-scores from the best-performing model (XGBoost with DINOv2 features). Table 8 shows the 10 highest scoring and 10 lowest scoring glosses out of 55 total classes.

Table 8: Higher and lower scoring ISL glosses

| Higher Scoring Glosses | F1 | Lower Scoring Glosses | F1 |
|---|---|---|---|
| DRINK | 0.99 | DAY | 0.94 |
| SMALL | 0.99 | TIME | 0.94 |
| MOTORBIKE | 0.99 | NOT | 0.94 |
| PLAY | 0.99 | LOOK | 0.94 |
| THREE | 0.99 | FUNNY | 0.94 |
| DEAF | 0.99 | FATHER | 0.93 |
| BOY | 0.98 | WHERE | 0.93 |
| SAY | 0.98 | MY | 0.93 |
| SISTER | 0.98 | FINISH | 0.91 |
| TWO | 0.98 | SAME | 0.90 |

Overall, the model demonstrates strong and consistent performance across glosses, with 58.2% of classes achieving F1-scores of 97% or higher. Only a small proportion of glosses (10.9%) fall below 95% F1, indicating robustness even for less frequent or potentially ambiguous signs.

## 5.5. Confusion Matrix Analysis

Figure 2 presents the complete $55\times55$ confusion matrix for the best model, exhibiting strong diagonal concentration with 95.91 % correct classifications. The matrix demonstrates error distribution with predominantly small or zero off-diagonal elements, indicating that misclassifications are rare and not systematically biased.

# 6. Discussion and Conclusion

This work addresses automated computational annotation of sign language corpora, where the main

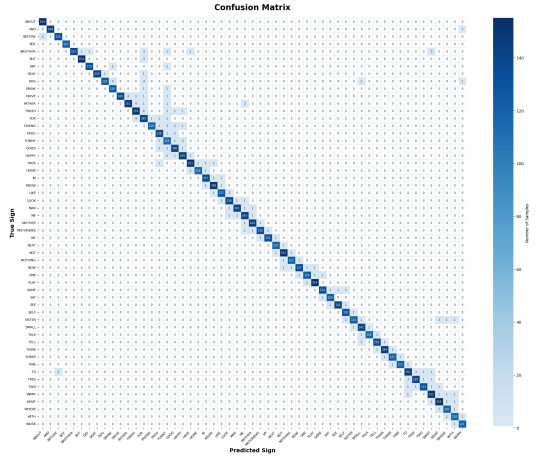



Figure 2: Confusion matrix for XGBoost with DINOv2 features

bottleneck is the cost and consistency of gloss-level labelling in continuous, motion-rich signing. For Irish Sign Language (ISL), this challenge is intensified by limited annotated data and the difficulty of identifying gloss boundaries under coarticulation. Our results show that self-supervised visual embeddings provide a robust baseline for annotation assistance on high-frequency glosses, consistently outperforming pose-based and multi-modal representations across classifiers. From a corpus development perspective, DINOv2 offers the most favourable accuracy–efficiency trade-off, achieving strong performance while avoiding the added complexity and computational cost of pose estimation and feature fusion. The limited benefit of naive multi-modal fusion further suggests that explicit skeletal landmarks may not be essential when contemporary visual models already capture linguistically relevant motion cues. These findings motivate practical recommendations for ISL annotation workflows: prioritising self-supervised visual representations, treating pose estimations as optional rather than essential, and employing efficient classifiers such as gradient boosting. In summary, this paper presents an automated pipeline for ISL gloss-level annotation and offers empirical insight into motion-aware visual representations that balance accuracy and computational efficiency. The approach supports faster and more consistent gloss suggestions for sign language corpus development. The set of 55 glosses reflects a focus on high-frequency, consistent signs to ensure reliable learning, which serves as a controlled benchmark. Future work will apply explainable artificial intelligence methods to better understand model decisions and refine the annotation framework.

## Acknowledgements

This research was supported by the ADAPT Centre for AI-Driven Digital Content Technology at Technological University Dublin. The ADAPT Centre is funded by Research Ireland through the Research Centres Programme. The authors also thank the contributors and annotators involved in the creation of the Signs of Ireland Corpus.

## Bibliographical References

Danielle Bragg, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic — simultaneous face, hand and pose prediction, on device. Google Research Blog.

Irish Deaf Society. 2024. Irish sign language. https://www.irishdeafsociety.ie/irish-sign-language/.

Hafiz Muhammad Sarmad Khan, Simon D. Mcloughlin, and Irene Murtagh. 2025. Comparative evaluation and utilization of convolutional neural network architectures for irish sign language recognition. *Journal of Artificial Intelligence*, 16(1).

Sarmad Khan, Irene Murtagh, and Simon D. McLoughlin. 2024. Investigating motion history images and convolutional neural networks for isolated irish sign language fingerspelling recognition. In *Proceedings of the LREC-COLING 2024 11th Workshop on Sign Languages*.

Lorraine Leeson. 2004. Signs of ireland corpus: a collection of irish sign language video data from 40 signers of ireland.

Jian He Low, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Sage: Segment-aware gloss-free encoding for token-efficient sign language translation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5011–5020.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Irene Murtagh. 2021. The nature of verbs in sign languages: A role and reference grammar account of irish sign language verbs. *TEANGA,*

*the Journal of the Irish Association for Applied Linguistics*, 11:67–99.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Gali Ravi Kiran, Varada Srinadh, V Ankitha, and S Surekha. 2025. Sign language translator using transformer model. *International Journal of Environmental Sciences*, 35(1):84–92.

Manish Shukla and Harsh Gupta. 2025. Advancing sign language interpretation with transfer learning and multimodal features. *Research Square Preprint*.

Ijas Siju and Prabu Selvam. 2024. A novel approach for lightweight sign language recognition leveraging google mediapipe and deep neural net. In *Proceedings of the 2024 International Conference on Signal Processing and Integrated Networks*, pages 746–751. IEEE.

Rachel Sutton-Spence. 2007. Signs in BSL – established or productive? Technical report, University of Bristol, Graduate School of Education.

Yingli Tian, Jianbo Su, Lan Ni, and Yuchun Fang. 2024. Editorial: Bridging the gap: Ai and sign language recognition– a path toward inclusive communication. *Frontiers in Computer Science*, 6:1203805.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research.

# HeSLEx: A novel online questionnaire for heritage sign language research

**Evgeniia Khristoforova, Roman Poryadin**
University of Cologne, University of Cologne
evgeniia.khristoforova@uni-koeln.de, roman.poryadin@gmx.de

**Abstract**

In this paper, we present *Heritage Sign Language Experience (HeSLEx)*, a novel online sociolinguistic questionnaire adapted from the heritage spoken language survey HeLEx (Tomić et al., 2023) to provide a standardized community profile prior to data collection. Heritage sign languages are minority sign languages used by Deaf signers in migration contexts and thus offer a unique window on bilingualism in the visual modality. HeSLEx is designed to be visual-first: most content is delivered as videos featuring signing in Russian Sign Language (RSL) by community member in a JavaScript/`jsPsych` interface. To accommodate heterogeneous RSL comprehension, each video includes optional Russian and German text hidden behind a "Show text" button. HeSLEx adds sign-specific modules, including participant and parental hearing status; modality-appropriate proficiency ratings (signing/comprehension for RSL and German Sign Language; reading/writing for Russian and German); educational histories and language(s) of instruction; interactional contexts central to Deaf life (including Deaf clubs); and Deaf-centered identity and language-attitude measures. Many items use slider scales to yield continuous predictors. The tool is designed to be adaptable to other sign language pairs in the framework of heritage language research and beyond.

**Keywords:** heritage sign languages, sociolinguistic survey, RSL, DGS, jsPsych

## 1. Introduction: community focus

Monolingual speakers are often treated as the gold standard in language research (Genesee, 2022). Yet neither monolingualism nor the auditory modality is intrinsic to the human language faculty. Heritage sign languages thus provide a particularly informative testing ground to explore an interface between the visual modality and bilingualism.

Heritage language can be loosely defined as a minority language, often (but not exclusively) used in migration contexts: families use a minority language (in migrant contexts, the language used in their home countries) at home and in the diaspora, while also using a dominant societal language in official contexts (at school, at work, etc.) or simply outside of the home setting (Montrul and Polinsky, 2021). Children in such families – either born in the country of residence or migrating before puberty – often acquire and use the minority (heritage) language primarily at home. Due to limited input and restricted contexts of use, the language of heritage speakers is often described as divergent from that of monolingual speakers.

It goes without saying that sign languages can also be heritage languages: deaf families may migrate to a new country and form a diaspora, while acquiring and using the local sign language alongside their heritage sign language. The pressure to use the societal sign language may be even greater for deaf signers than for speakers, as it may determine access to interpretation services in the country of residence.

To our knowledge, there have so far been no comprehensive studies of heritage sign languages (but see Pichler et al. (2018) for a discussion of children of hearing adults as heritage signers). The present project focuses on one such heritage signing community, described in Section 1.1. At the same time, it is highly likely that other heritage signing communities exist beyond the one examined in this study.[1] Heritage Deaf communities in migration contexts present a unique opportunity to explore language at the intersection of bilingualism, migration, and the visual modality.

In exploring the language of heritage signers, whether through experimental or corpus-based methods, a necessary first step is a holistic description of their sociolinguistic profiles. This is especially important for heritage language users, as studies on heritage spoken languages have repeatedly shown that this population exhibits substantial variability in linguistic behavior, including variable acceptability judgments across participants and variable preferences for different grammatical patterns. This variability has been linked to the complexity and diversity of the social contexts in which heritage speakers are embedded (Nagy and Gadanidis, 2021; Özsoy and Blum, 2023; Rinke and Flores, 2025). In the case of heritage signers specifically, although dedicated research is still lack-

[1] At least one additional community of heritage Russian Sign Language signers resides in Israel, formed within the same migration wave from the USSR (Svetlana Dachkovsky, p.c.). Potentially, a community of heritage German Sign Language signers may also be found in Israel due to the similar historical context.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 248–255
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ing, variability may be even higher for the following reasons:

- Unlike most investigated heritage languages, the great majority of sign languages, has no standardized school curriculum and, in many cases, no widely recognized linguistic standard or dedicated institution to safeguard it.[2] This is known to contribute to substantial variation even among monolingual language users (seeDeumert (2001) and Ameka (2016), among others, for discussion on spoken languages) and may thus be enhanced in the heritage setting.
- Related to the point above, unlike most investigated heritage spoken languages, sign languages in general lack a writing system that would be widely used and suited to the needs of the language. Instead, signers use the writing system of the spoken language of surrounding. This fuels variation in already monolingual contexts (Sebba, 2007) and, arguably, even more so in heritage contexts, since (unlike hearing heritage speakers) heritage signers do not have access to a vast library of texts in their heritage language (with the partial exception of videos on social media), which further limits their input thus leading to divergent attainment and increase language viability.
- Deaf communities are always minority communities in their own countries. Consequently, a diaspora using a given sign language in another country is typically even smaller. This is not necessarily the case for all heritage communities; for instance, the community of heritage Turkish speakers in Germany is estimated in the hundreds of thousands, if not millions (Özsoy and Blum, 2023). While equivalent estimates for heritage signing communities are still lacking (see Section 1.1 for an approximation), there is no doubt that they constitute only a fraction of the size of most investigated heritage-speaking communities. Smaller communities may thus contribute to increased idiolectal variation (Raviv et al., 2019; Thompson et al., 2020).
- Almost all signers, even those who use a single sign language, are bilingual in the sense that many can read and write in the surrounding spoken language. This may itself contribute to variation, but the situation for heritage signers may be even more complex (Fricke et al., 2019; Benmamoun et al., 2013; Polinsky and Scontras, 2020). Depending on age of migration, these signers may or may not receive schooling in the country of origin and/or in the country of current residence, which affects which written languages they acquire and use. It is therefore possible that some signers are bilingual across both visual and written modalities. How such a complex situation influences linguistic outcomes is at present unclear, but it is plausible that a high degree of inter-signer variability should be expected.

The multitude of the factors listed above suggests that, to interpret heritage data, researchers should first document the sociolinguistic dimensions outlined above, and others discussed throughout the paper.

While heritage spoken language research (itself a relatively recent subfield) has developed survey instruments for community profiling, heritage sign language research lacks a standardized procedure for this stage. The present paper proposes one such instrument: the *Heritage Sign Language Experience* questionnaire (HeSLEx). HeSLEx builds on an existing heritage spoken-language survey – *Heritage Language Experience* (HeLEx) – but is adapted to the specific design requirements of research in sign language communities. The current version of the survey was developed in the framework of our project on deaf heritage signers of Russian Sign Language (RSL) living in Germany who are also fluent in German Sign Language (*Deutsche Gebärdensprache*, DGS). The present version is tailored to the RSL–DGS community described below, but it is designed to be easily adaptable to and comparable across other sign language dyads (i.e., pairs of sign languages, e.g., RSL-DGS in the present project).

## 1.1. Target community

The project investigates deaf RSL signers born into deaf families who migrated from the USSR/Russia to Germany. Migration largely took place in the 1980s. Many families settled in North Rhine–Westphalia and formed a local community. Although its size is not formally documented, community members estimate roughly 30–50 RSL signers in the region. We focus on two generations. The first generation (G1) acquired RSL before migration, learned DGS after arrival, and now uses DGS daily, while RSL is mainly restricted to family and community settings. The second generation (G2) consists of ***heritage signers***: individuals born in Germany to RSL-signing families or who migrated

[2] According to Fridman and Özsoy (2024), half of all heritage research is dedicated to heritage speakers of Spanish. Other frequent research targets include heritage Korean, Chinese, Russian, and Turkish. These are all highly codified languages with standard curricula and high availability of resources. Many of these languages are studied in English-speaking countries and Germany, where they are used in the contexts of large diasporas.

before puberty. Most of them would grow up using RSL at home but acquired DGS rather early on in local deaf schools.

## 2. Heritage Language Experience survey

HeLEx is a state-of-the-art comprehensive survey covering linguistic experience and related life-history factors (Tomić et al., 2023). It includes (a) self-assessed proficiency in heritage, societal, and other languages in speaking, listening, reading, and writing (Figure 1); (b) frequency of language use across social contexts (home, extended family, work/school, leisure, community); (c) self-reported code-switching and code-mixing in those contexts; and (d) self-identification and language attitudes. Together, these measures support descriptive profiling (including, e.g., *language entropy*, an estimate of exposure to different languages across social/communicative contexts (Gullifer and Titone, 2020)) and allow subsequent linguistic work to be related inferentially to sociolinguistic predictors, as has been done in prior heritage spoken-language studies (Montrul and Polinsky, 2021).

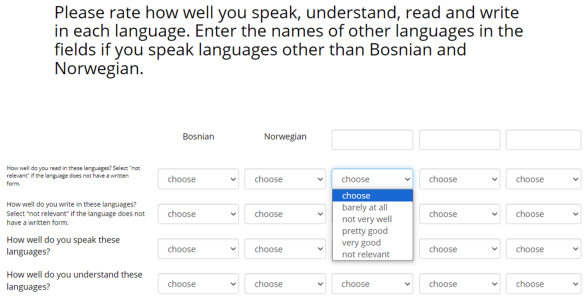



Figure 1: An example of the HeLEx questionnaire for heritage speakers of Bosnian in Norway (Tomić et al., 2023); translated into English for illustrative purposes.

However, HeLEx cannot be used in signing communities without substantial adaptation: it is not designed for deaf participants and does not capture Deaf sociocultural factors relevant to acquisition and use. We therefore developed He**S**LEx to better capture the specifics of Deaf migrant communities.

## 3. Heritage Sign Language Experience

Starting with the dimensions of inquiry already included in HeLEx (see above), in HeSLEx we implement the following changes.

### 3.1. Visual-first design

In HeLEx, all questions and instructions are presented in written form as also the case in Figure 1. In state-of-the-art sign language research, however, it is customary to minimize written text for multiple reasons. First, exposure to written text may bias signers toward using spoken-language-like structures (Morford et al., 2011). Although this may not straightforwardly affect responses to a sociolinguistic survey, it may become an issue if the survey is used immediately before linguistic data collection. Second, from an ethical standpoint, signers should be able to receive information in their native language. This is especially important in sign language research, which relies on participants' linguistic knowledge and therefore should respect their linguistic rights – including the right to have information presented visually and in sign language.[3]

In accordance with the considerations above, written content is replaced by videos containing signed versions of the respective texts. To enable this technically, we developed HeSLEx in `jsPsych`, a JavaScript framework designed for creating behavioral experiments that run in a web browser (de Leeuw, 2015). Unlike the Gorilla platform and similar platforms such as Qualtrics – which are user-friendly but slightly less flexible – `jsPsych` allows detailed customization and thus supports a highly visual, accessible layout for deaf users while preserving the original HeLEx template.[4] More specifically, we use the `jsPsychSurveyHtmlForm` plugin, which supports a customized HTML layout, precise positioning of videos and response interfaces (e.g., scales, text windows, dropdown menus), and additional tools described in the following subsections.

### 3.2. Community-matched video delivery + optional text

Instructions, questions, and consent materials are recorded by a Deaf RSL signer from the target community (a G2 signer fluent in both RSL and DGS). However, heritage signers may vary in their comprehension of the home language (RSL in our context), especially with respect to technical terminology in the consent form and instructions. In order to provide participants with an optimal ex-

[3] In addition, from the standpoint of recruitment and data quality, participant comfort and ease (improved by using the native language as the primary *modus operandi*) may positively affect both responses and willingness to continue collaborating with the researcher.

[4] A further advantage of `jsPsych` is that we can develop a dedicated HeSLEx plugin so researchers can insert their own video materials via a template without detailed JavaScript coding; developing this plugin is a key future goal.

perience, each video is accompanied by Russian and German text. Both texts are collapsed under *Pokazat russkiy tekst/ Deutschen Text anzeigen* (*Show Russian/German text*) buttons so that participants who can complete the survey in sign are not forced to be exposed to the written language. The setup is illustrated in Figure 2.

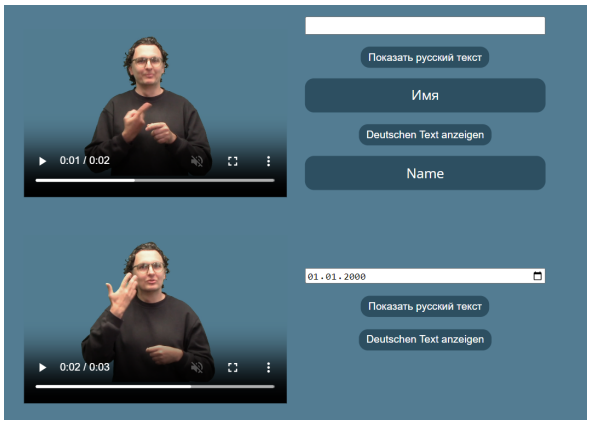



Figure 2: An example of the HeSLEx questionnaire for heritage signers of RSL in Germany. The featured questions concern the name and date of birth of the participants. For illustrative purposes, the Russian and German texts are expanded for the first question; the second question shows how participants usually see the questions – i.e., with the texts collapsed.

In a future version of the survey, it will also be possible to choose between RSL and DGS versions of the questionnaire (this choice will be recorded as a separate variable), with the possibility of changing the selection during completion.

### 3.3. Family hearing status and language background

Although the original HeLEx includes questions on the language background of parents/caregivers, it predictably does not include questions about their hearing status. At the same time, parental hearing status (recorded separately for the mother and father, or two caregivers) and language repertoire are known to potentially impact sign language acquisition (e.g., age of onset, properties of the input such as native vs. L2 input, and regional variants) (Mayberry and Kluender, 2018). This parameter is especially important in the context of heritage signers, as the family is often the primary language source, while contact with the larger community may be limited. Participant identity (hearing, deaf, hard of hearing, CODA) is also recorded as explained below. See also Crasborn and Hanke (2003) for a more extended list of adaptations proposed for metadata collection for sign language corpora.

### 3.4. Education and other institutions as key sociolinguistic predictors

Educational context – specifically deaf schools and their types – is a salient factor in sign language acquisition. The particular deaf school may determine the sign language variant preferred by an individual, including dialectal patterns (Quinn, 2010; Stamp et al., 2014; Eichmann and Rosenstock, 2014), while the type of school – i.e., oralist, bilingual, or sign-language-medium – may greatly influence sign language proficiency (Meristo et al., 2007; Hall et al., 2017).

In migration contexts, the situation may be even more complex, since heritage signers born in the country of origin may attend multiple deaf schools that differ in language policy and in the signed and spoken languages used.

To do justice to these parameters, participants are asked to select a language for different levels of education: RSL, DGS, Russian, and German. Free-choice text windows are provided in case none of these languages are applicable. The education levels include primary school, secondary school, vocational school/colleges, and university. Selecting multiple languages is possible, allowing identification of bilingual education. The question for the first two education levels –i.e., primary and secondary school –is illustrated in Figure 3.

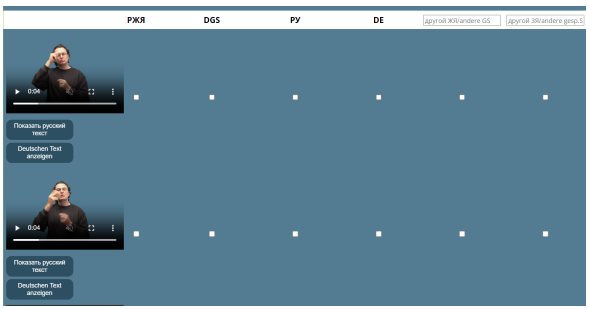



Figure 3: An example of the HeSLEx questionnaire for heritage signers of RSL in Germany. The featured question concerns the language used across education levels.

Another institution that plays a particular role in Deaf communities (and especially in migrant Deaf communities) is Deaf meeting centers/clubs. Most major cities have at least one such meeting center, which serves as an epicenter of social life for Deaf people. Migrant Deaf communities also have their own Deaf clubs or regular meetings, and these are essentially the main ways for heritage signers to communicate in their heritage language outside of the family context. To account for this, we add Deaf clubs/meeting centers as a distinct interactional context alongside home/family/work/leisure/official affairs already specified in HeLEx. In addition, so-

cial media is added as another context, since heritage signers (especially younger generations) may consume signed content in various sign languages and International Sign on social media platforms.

## 3.5. Deaf-centered identity and attitudes

In addition to language profiling, the original HeLEx puts emphasis on the identity of the heritage speaker, their attitudes toward the heritage and societal languages, and their perception of societal attitudes toward their heritage language use. Identity (e.g., as Russian or German) and language attitudes may influence daily language use and the desire to keep the language active or pass it on to children.

In sign language research, Deaf identity is an additional important variable, as it may similarly influence the use of sign language. We therefore probe signers' identity as deaf, hard of hearing, CODA, or hearing. Note that we ask about *identity*, not objective hearing status, since these do not always overlap, yet both matter for language use.

Similarly, the perceived societal attitude toward the use of a heritage sign language conflates attitudes toward minority language use and toward sign language use, adding another level of complexity. One may hypothesize that DGS signers may be more sympathetic to heritage RSL signers than hearing speakers of German are toward heritage Russian speakers, due to shared experiences related to deafness. At the same time, heritage RSL signers may receive less governmental support for services in heritage RSL than heritage speakers of Russian receive for services in Russian. The survey is therefore adapted to the intersectionality between migration and deafness, including attitudinal questions addressing both.

## 3.6. Minor adjustments

### 3.6.1. Modality-appropriate proficiency measures

Instead of inquiring about self-reported speaking and listening proficiency in heritage and societal languages, we add parallel questions about signing and comprehension. This is implemented both for self-assessment and for questions concerning frequency of language use. It is important to consider these aspects separately, as comprehension in heritage signers may be significantly more advanced than production. We do, however, retain the questions about reading and writing in the spoken languages used in the country of origin and in the country of residence (e.g., Russian and German). It would be interesting to examine how exposure to these spoken/written languages interacts with the use of heritage sign language and with language identity and attitudes. Additionally, heritage signers (as well as heritage speakers) often do not read or write in the minority language at all (or do so very little) (Benmamoun et al., 2013). It is therefore of interest to examine how limited exposure to the spoken/written heritage language relates to signing practices in heritage sign language.

### 3.6.2. Response format optimization

On a more technical note, although not directly related to the sign language context, some changes have been made to the response format. Instead of many multiple-choice items in HeLEx (see Figure 1), we use slider scales in HeSLEx (Figure 4). This has two advantages. First, scalar results are continuous, thus being more informative for descriptive summaries and better suited to inferential models (e.g., linear mixed-effects regression) (Schütze and Sprouse, 2013). Second, scales are generally more visual and usually require less text, thus being more suitable for a sign language survey.




Figure 4: An example of the HeSLEx questionnaire for heritage signers of RSL in Germany. The featured question concerns self-reported proficiency in comprehension and production in RSL and DGS; the response is represented as a scale from "Null" to "High level"; the slider can be placed anywhere on the scale, including between the ticks.

## 3.7. Planned development

As mentioned previously, it is desirable in sign language research to limit exposure to written text. Accordingly, typed responses are also undesirable. For items that currently require typed responses (e.g., occupation or name), we aim to add an option for participants to provide video responses stored in the dataset (e.g., as base64-encoded media). In the current version of `jsPsych`, this is not yet implemented. There is a dedicated *jsPsychExtensionRecordVideo* extension in addition to the *jsPsychInitializeCamera* plugin, which together allow recording video responses. However, using these plugins does not allow HTML-based customization; consequently, the layout structures, scales, and multiple answers per window would be lost. Second, video

recording does not allow previewing the recorded video or even viewing the webcam output during recording, making the recording essentially blind. This format is suboptimal. Our goal is therefore to create a customized extension allowing video recording with a webcam view and a preview, enabling typed answers to be replaced.

A further long-term goal is to create a dedicated plugin based on `jsPsych` that can be easily reused for other sign-language dyads in heritage sign language research. Currently, all videos, translation texts, and response tools are included directly in the code of each *trial* (one full window of questions submitted by clicking the "Continue" button). A plugin would allow researchers to simply provide video files and translation texts, and to select the desired response format. In other words, a plugin would make HeSLEx reusable for further bilingual sign language research on any sign language dyads within heritage language contexts (including research on CODAs as heritage signers) and beyond.

## 4. Pilot usage and subsequent edits

HeSLEx was piloted with three deaf participants, one on site and two online. In addition to completing the questionnaire, pilot participants were asked to provide informal feedback, in written or signed form, on their experience with the questionnaire from the user's perspective.

Pilot participants reported that HeSLEx was visually legible and well aligned with the needs of Deaf bilingual communities. The on-site participant and one of the online participants completed the questionnaire on a laptop using the Chrome browser. These participants reported no problems with the layout, but noted that the text on the scale ticks could be larger or bolded; this change has been implemented in the current version of the survey. The second online participant completed the survey on a smartphone. This participant reported that the layout on some pages (e.g., the informed consent page) became distorted on a smartphone, with page elements appearing "crumbled together". The participant noted, however, that this did not significantly hinder use of the questionnaire. The necessary adjustments will therefore be implemented in the next version of the survey.

The two online participants, as well as the reviewers, noted that a button allowing participants to return to the previous page would be useful. The `jsPsychSurveyHtmlForm` plugin primarily used in the experiment does not include a built-in option for such a button. In the current version of `jsPsych` (v8.2.3), the only plugin that supports returning to the previous page is the dedicated instruction plugin `jsPsychInstructions`. However, this plugin cannot serve as the basis for the experiment, since it does not allow customized HTML code or other forms of user input. To address this issue, we implemented backward navigation by replacing the original fixed sequence of separate `jsPsychSurveyHtmlForm` trials with a controller, `surveyPages`, that displays questionnaire pages from an indexed list. Whenever the page changes, the script saves the current responses and restores them when a participant returns to an earlier page by pressing the "Back" ("Nazad / Zurück") button. In other words, the button moves to the previous page in the list and re-renders it with the previously entered responses, allowing participants to review and, if necessary, revise their answers. The corresponding JavaScript code is available via the OSF link below.

With respect to the languages used in the questionnaire, all participants found the videos containing the questions in RSL clear and reported that, for them, consulting the written text was unnecessary. One participant mentioned double-checking the German and Russian texts in the informed consent section of the survey, which arguably contains the most technical signs in the questionnaire. We would like to reiterate, however, that greater variability in RSL comprehension is to be expected in a larger participant group.[5]

Importantly, as anticipated, participants found the scales intuitive to use and used them in a graded way, placing the slider not only on the predetermined ticks but also between them.

On the technical side, an important outcome of the pilot was the confirmation that data from questionnaires completed on different types of devices and browsers could be successfully uploaded from

[5] We are grateful to the anonymous reviewer who drew our attention to the fact that, in the previous version of the survey, many drop-down menu responses were labelled only in Russian. Although implementing signed drop-down menu options is, unfortunately, technically suboptimal, in the current version of the questionnaire we have added German equivalents for all drop-down menu options. Additionally, the reviewer rightly pointed out that having the first option in the drop-down list pre-selected by default may create ambiguity as to whether this option was actively chosen by the participant or whether the participant skipped the question. To address this issue, we added a default "..." option at the top of each list; the appearance of this option in the responses is treated as indicating that the participant skipped the question. Although no straightforward equivalent solution is available for the scales, we are currently working on recording not only the slider position, but also whether the participant interacted with the slider at all. This is not a complete solution, since a participant may judge the midpoint of the scale to be the most appropriate response and therefore leave the slider untouched. Nevertheless, this approach provides a somewhat better basis for disambiguating the survey output.

the server. This was indeed the case.[6]

An offline demo version of HeSLEx (in RSL, Russian, and German) in HTML and `.js` formats can be downloaded from the following link: https://osf.io/jgdxv/.

## 5. Acknowledgments

The present project received financial support from the Rubicon research programme, project no. 019.243SG.004, funded by the Dutch Research Council (NWO).

We would like to thank Dr. Anastasia Bauer (University of Cologne) for her support and guidance. Special thanks go to Dr. Onur Özsoy (University of Cologne), whose advice during the early stages of the project was invaluable. We are also particularly grateful to the participants in the ViCom Hub Cologne project meeting at the University of Cologne on 21 January 2026 (organized by Dr. Anastasia Bauer, Dr. Cornelia Loos, and Dr. Lara Pearson), who provided valuable feedback in the final stages of the survey's development. Finally, we are deeply indebted to the German RSL Deaf community, whose experiences inspired the present research and without whom it would not have been possible.

## 6. Bibliographical References

Felix K. Ameka. 2016. The uselessness of the useful: Language standardisation and variation in multilingual contexts. In *Prescription and tradition in language: Establishing standards across time and space*, volume 165, pages 59–71.

Elabbas Benmamoun, Silvina Montrul, and Maria Polinsky. 2013. Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical Linguistics*, 39(3-4):129–181.

Onno Crasborn and Thomas Hanke. 2003. Metadata for sign language corpora. Background document for an ECHO workshop, May 8–9, 2003, Radboud University Nijmegen.

Joshua R. de Leeuw. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1):1–12.

Ana Deumert. 2001. Language variation and standardization at the cape (1880–1922): a contribution to afrikaans sociohistorical linguistics. *Journal of Germanic Linguistics*, 13(4):301–352.

Hanna Eichmann and Rachel Rosenstock. 2014. Regional variation in german sign language: The role of schools (re-)visited. *Sign Language Studies*, 14(2):175–202.

Melinda Fricke, Megan Zirnstein, Christian Navarro-Torres, and Judith F. Kroll. 2019. Bilingualism reveals fundamental variation in language processing. *Bilingualism: Language and Cognition*, 22(1):200–207.

Clara Fridman and Onur Özsoy. 2024. Sentiments towards heritage languages and their speakers. *Heritage Language Journal*, 21(1):1–33.

Fred Genesee. 2022. The monolingual bias: A critical analysis. *Journal of Immersion and Content-Based Language Education*, 10(2):153–181.

Jason W. Gullifer and Debra Titone. 2020. Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition*, 23:283–294.

Wyatte C. Hall, Leonard L. Levin, and Melissa L. Anderson. 2017. Language deprivation syndrome: a possible neurodevelopmental disorder with sociocultural origins. *Social Psychiatry and Psychiatric Epidemiology*, 52(6):761–776. Epub 2017-02-16. PMID: 28204923. PMCID: PMC5469702.

Rachel I. Mayberry and Robert Kluender. 2018. Rethinking the critical period for language: New insights into an old question from American Sign Language. *Bilingualism: Language and Cognition*, 21(5):938–944.

Marek Meristo, Kerstin W. Falkman, Erland Hjelmquist, Mariantonia Tedoldi, Luca Surian, and Michael Siegal. 2007. Language access and theory of mind reasoning: evidence from deaf children in bilingual and oralist environments. *Developmental Psychology*, 43(5):1156–1169.

Silvina Montrul and Maria Polinsky, editors. 2021. *The Cambridge Handbook of Heritage Languages and Linguistics*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, Cambridge, UK.

Jill P. Morford, Erin Wilkinson, Agnes Villwock, Pilar Piñar, and Judith F. Kroll. 2011. When deaf signers read english: Do written words activate

[6] If a questionnaire is exited before completion on a computer, the data entered up to that point remain recorded. This is not the case when the questionnaire is completed on a phone, where all data are lost if the browser window containing the unfinished questionnaire is closed. In the context of an online questionnaire, this is only a minor issue, since a partial dataset, even if recorded, is unlikely to be included in the analysis.

their sign translations? *Cognition*, 118(2):286–292. Epub 2010-12-08. PMID: 21145047. PMCID: PMC3034361.

Naomi Nagy and Timothy Gadanidis. 2021. Heritage language variation and change – how complex is it? *Heritage Language Journal*, 18(2).

Deborah C Pichler, Diane Lillo-Martin, and Jeffrey L Palmer. 2018. A short introduction to heritage signers. *Sign Language Studies*, 18(3):309–327.

Maria Polinsky and Gregory Scontras. 2020. Understanding heritage languages. *Bilingualism: Language and Cognition*, 23(1):4–20.

Gary Quinn. 2010. Schoolization: An account of the origins of regional variation in british sign language. *Sign Language Studies*, 10(4):476–501.

Limor Raviv, Antje Meyer, and Shiri Lev-Ari. 2019. Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, 286(1907):20191262.

Esther Rinke and Cristina Flores. 2025. Systematic and predictable variation in heritage grammars: The role of complexity, diachronic change, and linguistic ambiguity in the input. In Roberta D'Alessandro, Michael T. Putnam, and Silvia Terenghi, editors, *Heritage Languages and Syntactic Theory*. Oxford University Press, Oxford.

Carson T. Schütze and Jon Sprouse. 2013. Judgment data. In R. J. Podesva and D. Sharma, editors, *Research methods in linguistics*, pages 27–50. Cambridge University Press.

Mark Sebba. 2007. Between language and dialect: orthography in unstandardised and standardising vernaculars. In *Spelling and Society: The Culture and Politics of Orthography around the World*, pages 102–131. Cambridge University Press, Cambridge.

Rose Stamp, Adam Schembri, Jordan Fenlon, Ramas Rentelis, Bencie Woll, and Kearsy Cormier. 2014. Lexical variation and change in british sign language. *PLOS ONE*, 9(4):e94053.

Bill Thompson, Limor Raviv, and Simon Kirby. 2020. Complexity can be maintained in small populations: A model of lexical variability in emerging sign languages. In Andrea Ravignani, Chiara Barbieri, Molly Flaherty, Yannick Jadoul, Ezra Lattenkamp, Hannah Little, Mauricio Martins, Katie Mudd, and Tessa Verhoef, editors, *The Evolution of Language: Proceedings of the 13th International Conference (Evolang13)*, pages 440–442. The Evolution of Language Conferences, Nijmegen.

Aleksandra Tomić, Yulia Rodina, Fatih Bayram, and Cécile De Cat. 2023. Documenting heritage language experience using questionnaires. *Frontiers in Psychology*, 14:1131374.

Onur Özsoy and Frederic Blum. 2023. Exploring individual variation in turkish heritage speakers' complex linguistic productions: Evidence from discourse markers. *Applied Psycholinguistics*, 44(4):534–564.

# Comparison of Low Bitrate Quantizers for Encoding Swedish Sign Language

**Anna Klezovich**[1], **Johanna Mesch**[1,2], **Gustav Eje Henter**[1], **Jonas Beskow**[1]
[1]Speech, Music and Hearing, KTH Royal Institute of Technology, [2]Stockholm University
Stockholm, Sweden
{annkle, ghe, beskow}@kth.se, johanna.mesch@su.se

**Abstract**

This paper investigates the bitrate–distortion trade-off of different discrete representations for Swedish Sign Language (STS) using the STS Mocap v1 motion capture dataset. We compare the K-Means algorithm with the Residual Vector Quantized Variational Autoencoder (RQ-VAE) to determine how efficiently each method preserves salient motion information at low bitrates. The results show that RQ-VAE consistently achieves lower reconstruction error than K-Means at matching bitrates, particularly for body motion, and better preserves the signing space volume. We further demonstrate that quantized representations can serve as conditioning for a flow-matching generative model, producing plausible but still imperfect sign sequences at low bitrates. These findings highlight the advantages of vector quantized models for efficient sign language motion encoding.

**Keywords:** compression quality trade-off, input bitrate, Swedish Sign Language, motion capture data, 3D sign language generation

## 1. Introduction

Having a strong representation learning model as a step before a generative model can significantly improve the quality of the generative model output. This has been shown by many recent works in the domains of sign language (Cruz and Bejarano, 2024, Xie et al., 2024, Malmberg et al., 2024), co-speech gesture (Guichoux et al., 2025), dance (Alexanderson et al., 2023, Siyao et al., 2023), and human motion generation (Jiang et al., 2023, Zhang et al., 2024), as well as in spoken speech and audio generation papers (SoundStream by Zeghidour et al. (2021), WavLM by Chen et al. (2022)). One possible explanation for this is that those learned representations are able to capture small discrete patterns relevant for each domain, for example, dance moves for dance generation or in the case of speech and sign language — phonological information. Representation learning models should ideally be trained on large amounts of data, which is why finding an optimal compression quality trade-off can potentially offer faster and more efficient training that preserves all the relevant information.

For sign language translation, promising attempts have been made to apply large language models for sign language translation (Gong et al., 2024), which requires sign language to be represented with an information rate (i.e. bitrate) of a similar order of magnitude as that of the text. This motivates our investigation in bitrate vs quality for quantized sign language representations.

Quantization implies mapping the data from a continuous space to a smaller discrete space, while preserving meaningful information. For example, Vector Quantized Variational Autoencoder (VQ-VAE) type models (van den Oord et al., 2017, (Zeghidour et al., 2021)) comprise an encoder, a codebook that learns quantized representations, and a decoder. A VQVAE model is jointly trained to condense and quantize the data into the codebook and then reconstruct it back into the original space with the decoder, optimized through a reconstruction loss. These quantized representations from the codebook are then used as inputs to generative models.

Determining the optimal size of the codebook and the lowest bitrate at which the most salient information is preserved (Malmberg et al., 2024), especially when changing between motion domains, remains a challenging task.

In this paper, we investigate two types of representation in terms of their bitrate-distortion tradeoff in the sign language domain based on the Swedish Sign Language (STS) motion capture dataset, STS Mocap v1 (Klezovich et al., 2025). We compare K-Means representations with Residual VQVAE (RQ-VAE) codebook representations (Zeghidour et al., 2021) in order to find out whether a more advanced and complex RQ-VAE outperforms K-means at the same bitrates or not.

## 2. Background

Multiple papers have already investigated different types of representation learning methods for the domain of sign language generation and understanding. For example, Cruz and Bejarano (2024) investigated the use of the RVQ-VAE quantizer for generative interpolation of sign language poses in



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 256–261
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2D. Xie et al. (2024) proposed a VQ-VAE model called Pose-VQVAE as a step in translation tasks from sign language gloss sequences to sign language 2D poses. The authors fed the learnt sign language representations from a codebook as an input into a diffusion-type model. Another representation learning approach for sign languages is the SHuBERT model introduced in (Gueuwou et al., 2025). This is a self-supervised framework for sign language representation that adapts masked token prediction to multi-stream visual input — jointly modeling hand, face, and body pose streams via cluster prediction following a HuBERT paper idea from the audio synthesis domain (Hsu et al., 2021). In (Malmberg et al., 2024), the authors applied a representation learning model from the generic human motion domain to 2D sign language data, namely (Jiang et al., 2023), which is also a VQVAE-type model.

Beyond sign language, similar representation learning strategies have proven effective in other human motion domains. (Alexanderson et al., 2023) applied diffusion models to audio-driven motion synthesis for co-speech gesture, dance, and locomotion. Guichoux et al. (2025) also used RQ-VAE as a motion tokenizer, but in a unified co-speech gesture and audio synthesis domain.

The RQ-VAE model (Zeghidour et al., 2021) that we use in this study builds on the idea introduced in the original VQVAE model paper by van den Oord et al. (2017). RQ-VAE is designed for audio processing codecs, but has also proven effective in other domains. The difference between RQ-VAE and VQVAE is that the former has multiple subsequent quantizers, with each of them quantizing the residual error of the previous quantizers. As a result, this method performs joint compression and enhances the audio at a lower bitrates than other similar methods.

K-Means clustering (Bradley and Fayyad, 1998) offers a baseline for investigating bitrate-distortion trade-offs, as its bitrate is determined directly by the number of clusters. Even though the K-Means algorithm is a classical machine learning based quantization method that is entirely different from VQ-VAE, the CVQ-VAE Zheng and Vedaldi (2023) paper showed that the classic VQ-VAE can even benefit from merging with K-Means. The CVQ-VAE model uses online K-Means-style anchor updates to remedy a codebook collapse problem and reinitialize dead codes.

# 3. Methods

## 3.1. Motion Capture Dataset

The STS Mocap v1 dataset (Klezovich et al., 2025) used in this study comprises of approximately 4 hours of high quality motion capture data (body, hands, and face) for Swedish Sign Language (STS). The data consist of Swedish sentence translations from *8sidor* (MTM and Hillblom, 2025), a newspaper written in simplified Swedish (around 3.5 hours), and signs and sentences featuring these signs from STS dictionary (Svenskt teckenspråkslexikon, 2025) (around 30 minutes). All of the data was recorded with a single CODA (aka. Child of Deaf adult(s)) signer in consultation with a second, deaf L1 signer.

The mocap data was transformed into both 3D coordinates and joint angles (as exponential maps) with the help of the *pymo* module (Alemi, 2019). For K-means clustering and RQ-VAE quantization experiments, we used joint angles, while 3D coordinates were used in evaluation to get metrics in centimeters. We did not use 3D positions in the training/fitting stage and used 3D angles instead, because 3D positions do not convey a head rotation.

The data were split into three body parts: body, right hand, and left hand (face data was not considered for the purposes of this study). The hands and the body have different motion dynamics in sign language, the hands have more joints, move faster, and produce more complex poses. We wanted to model them separately to take into account that one of the methods could be better for the hands or for the body than the other.

## 3.2. RQ-VAE and K-Means Implementation

RQ-VAE in this paper is based on Python library *vector-quantize-pytorch* (Wang, 2021) implementation of residual vector quantizer. The encoder and decoder have a very simple structure: each consists of three linear layers with ReLU activations between them.

K-Means algorithm is implemented with *scikit-learn* Python library.

## 3.3. Bitrates Calculation

To calculate bits per second for K-Means algorithm, we employed the formula

$$bitrate = log_2(k) * fps$$

where $k$ is the number of K-means clusters, and $fps$ stands for frames per second.

In RQ-VAE the number of quantizers signify the number of stages of quantization and each subsequent quantizer encodes a residual error of the previous one, adding more codes and hence more bits of information per frame. So the formula for bitrate for RQ-VAE case requires multiplication by the

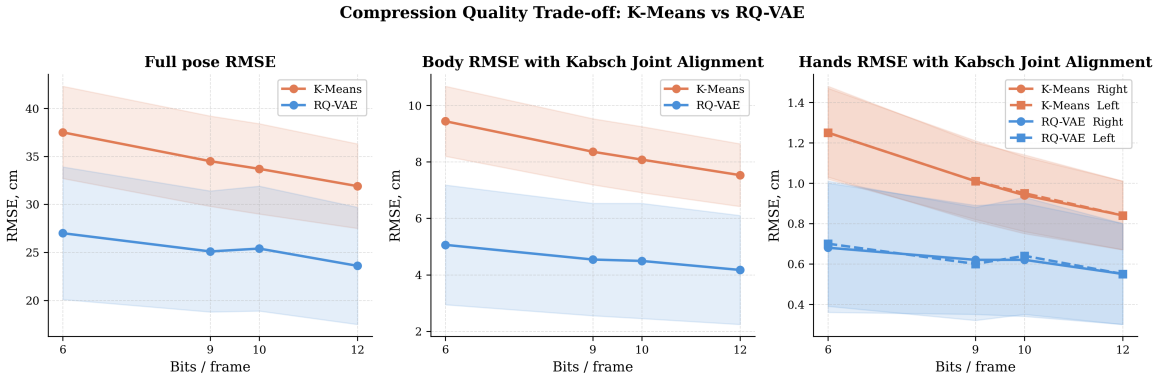



Figure 1: Bitrate–RMSE for K-Means and RQ-VAE.

| Algorithm | K | | Bits/frame | Bitrate |
|---|---|---|---|---|
| K-Means | 64 | | 6 | 180 |
| K-Means | 512 | | 9 | 270 |
| K-Means | 1 024 | | 10 | 300 |
| K-Means | 4 096 | | 12 | 360 |
| | **D** | **Q** | | |
| RQ-VAE | 8 | 2 | 6 | 180 |
| RQ-VAE | 8 | 3 | 9 | 270 |
| RQ-VAE | 32 | 2 | 10 | 300 |
| RQ-VAE | 8 | 4 | 12 | 360 |

Table 1: Bits per frame in each experiment. All experiments have the same frame rate of 30 frames per second.

number of quantizers $Q$ ($D$ stands for the codebook size):

$$bitrate = log_2(D) * Q * fps$$

Table 1 lists the experiments that we carried out with their hyperparameters. The lower values of codebook size $D$ are chosen so that it is possible to match RQ-VAE experiments with K-Means in terms of bitrate. For example, bitrate of 16 would requires 65 536 clusters already, while the number of frames in the whole train set is 626 664, which is only 9.6 times higher than the number of clusters in this case. Selecting that many clusters brings this method closer to a temporal subsampling rather than clusterisation.

### 3.4. Evaluation Metrics

As for the reconstruction quality, we calculate the root mean square error (RMSE) on 3D positions of the original pose vs. the reconstructed pose. The RMSE values for full poses are calculated this way on the raw 3D positions and essentially show the averaged distances between joints in centimeters.

In order to have a closer look at different body parts, since the quantization was performed on them separately, we first align translation and rotation of frames for each body part with the Kabsch algorithm (Kabsch, 1976). This is conceptually similar to the Procrustes alignment, but does not do the scaling, so that the measures stay in real units, in our case centimeters. It is important to do the alignment first, because otherwise the RMSE error would say more about the global positions of body parts than the differences between the hand shapes or arms positions in the body pose. After alignment is done, we calculate the RMSE in the same way for the aligned body poses and the aligned hand poses.

Last but not least, we use an expressivity-related metric. For that we calculate the convex hull of the wrist joints point cloud. The intuition is the bigger the volume of this space, the more expressive the hands are. Convex hull volume was calculated with the help of *scipy* Python library. We first extract wrists 3D coordinates both from reconstructed sequences and original. Then convex hull volume is calculated over a point cloud of these coordinates for each animation sequence. For better interpretability, we present these values as a percentage: $(V_{reconstructed} \div V_{original}) * 100$. This value would answer how much of the original signing space the hands are taking up in the reconstructions.

## 4. Results

The comparison of K-Means with RQ-VAE in Figure 1 shows that RQ-VAE outperforms K-Means in most scenarios in terms of compression quality trade-off. For the full pose RMSE calculated on raw 3D coordinates, RQ-VAE performed slightly better than K-Means for all bitrates, but more so for lower bitrates. If we look at body RMSEs calculated with Kabsch joint alignment, RQ-VAE showed consistently lower RMSEs with no overlap of standard errors, on average 1.8 times lower than for K-means at the same bitrates. The picture is more interest-

| Bits/frame | K-Means: Convex Hull Volume | RQ-VAE: Convex Hull Volume |
|---|---|---|
| 6 | 55.3% | 68.2% |
| 9 | 76.6% | 91.3% |
| 10 | 81.6% | 86.8% |
| 12 | 87.9% | 109.3% |

Table 2: Percentage of used space out of total space of motion. For reference 0.81 $m^3$ is a ground truth average convex hull volume of the wrists.

ing for the hands RMSEs (also done with Kabsch joint alignment) — the results overlapped more at higher bitrates, while RQ-VAe was still on average better than K-Means. Given that we used very low bitrates in our experiments, and according to our visual inspection of the reconstructions, the most likely explanation is that both methods fall equally short in reconstructing hand poses. The error for the hands was in general smaller than for the body for both K-Means and RQ-VAE at all bitrates, but this is because the hands perform smaller motions than the body.

We rendered[1] two sequences into a video for a visual comparison with the original sequence. This video compares side-by-side sequences reconstructed for 10 bits per frame, namely one sequence created with K-Means 1 024 clusters, and another one created with RQ-VAE with 32 codebook size and 2 quantizers. In a still frame from this visualization in figure 2, aside from obvious errors due to the generally low bitrate, the signing space is visually smaller in both of the reconstructions compared to the original. To compare the signing space size across our methods, we calculated the convex hull volume of the hands point clouds based on 3D coordinates of the wrists and the percentages of space volume used by the hands in the reconstruction compared to the original space volume. The results are given in Table 2. They show that the percentage of the convex hull volume used increases as the bitrate increases. For K-Means, this process was relatively linear, while for RQ-VAE, it was not, and it might also correlate with the choice of codebook size and other hyperparameters. The Table also shows that the percentage of used convex hull volume in RQ-VAE reconstructions was higher than with K-Means at the same bitrates.

To demonstrate that these quantized sequences do in fact produce reasonable generated animations after passing them as input to diffusion-type models, we trained a flow-matching model. To be more precise, we employed an OT-CFM model introduced by (Lipman et al., 2022) and adaped in Mehta et al. (2024)'s Matcha-TTSG model. As the input, we took angular (exponential maps) cluster centroid for each frame from K-means with $k = 1\,024$. It showed promising results with a smooth motion[2]. Even though the synthesized motion resembled the input sequence from a test set,

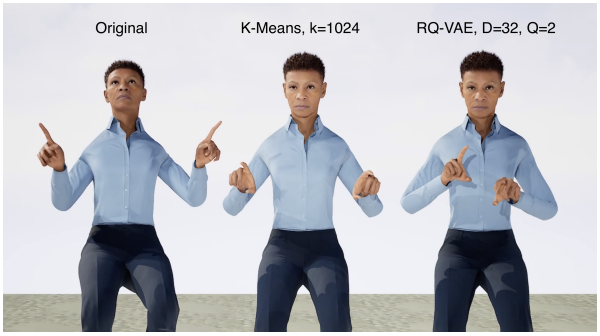



Figure 2: Still frame out of a video render. Left to right: original mocap sequence; reconstruction of that sequence with K-means, $k = 1\,024$; reconstruction of that sequence with RQ-VAE, $D = 32$, $Q = 2$

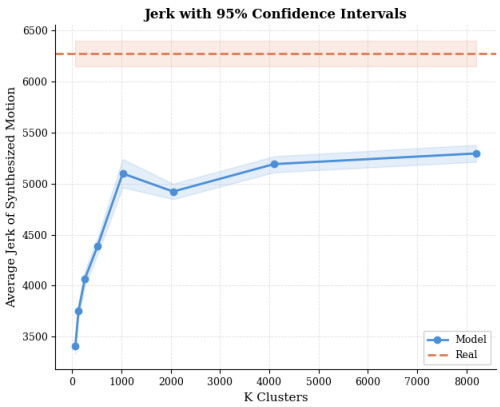



Figure 3: Jerk, CI Vs. number of K-Means clusters, trained and clustered on 2D positions. For comparison the jerk on the original dataset is 6 275.

[1] Link to reconstructions example render. The data is rendered out in Unreal Engine 5.4.4 with the help of Metahuman avatars (Epic Games, 2025). Left to right: original mocap sequence; reconstruction of that sequence with K-means, $k = 1\,024$; reconstruction of that sequence with RQ-VAE, $D = 32$, $Q = 2$

[2] Link - Original mocap sequence Vs. sequence synthesized with the model trained on K-means, $k = 1\,024$ angular cluster centroids.

it had too many errors in reconstructing the handshapes and overall motion to be semantically and phonologically correct. For a better result, it is important to have a stronger representation learning model both in terms of complexity and bitrate.

Since we were testing a very low number of clusters and codebook sizes to match the bitrates for the main compression–quality trade-off experiment, we also examined whether the metrics on the model synthesis Vs. original test data grow linearly or not at these low bitrates. Figure 3 shows that K-Means started producing results closer to the real data distribution in terms of the average jerk only at 1 024 clusters (the "knee" on the graph). The average jerk over the whole dataset in 2D was 6 275, CI [6 150, 6 401]. OT-CFM models conditioned on values lower than 1 024 did not capture enough information and their synthesized sequences were closer to the mean pose than to the data. Given that K-Means plateaus after 1 024 clusters in a 2D conditioning setting and also does not produce entirely accurate poses in either 2D conditioned or exponential maps conditioned settings, it is possible that this method reached its limitations when applied to this complex data. However, VQ-VAE type models, such as RQ-VAE of sufficient size is a promising future research direction.

## 5. Conclusions

In this paper, we compared two quantization methods — the K-Means algorithm and the RQ-VAE qunatizer — in terms of how well they encode Swedish Sign Language motion capture sentences at different bitrates. We compared them at four bitrates between 6 and 12 bits per frame. The results showed that RQ-VAE quantization outperforms the K-Means type quantization at matching bitrates in terms of RMSE in 3D joint space; it did so more strongly for the body than for the hands joints. When comparing the wrist motion expressivity of reconstructions Vs. the original wrist motions, we found that RQ-VAE used a higher percentage of signing space than K-Means at matching bitrates. For K-Means, the percentage of used signing space is with the growing bitrate, while the same is not true for RQ-VAE. This means other factors must be affecting this metric, such as the codebook size of other hyperparametres.

We also demonstrated that these sign language representations can be used for the sign language generation downstream task. To do so, we trained a flow-matching generative model, Matcha-TTSG (Mehta et al., 2024), on K-Means cluster centroids in exponential maps 3D space, at $k = 1\,024$. This model synthesized reasonable sign language sequences that resembled the original sequences. However, it still made errors in hand shape and body pose, because the bitrates used were too low for this task.

## 6. Discussion and Limitations

Even though the sign language generation task possibly requires a stronger quantizer at a higher bitrate, the lack of expressivity highlighted in this study is not an issue for diffusion and flow-matching models as long as the quantized motion does not have erroneous poses in it. Diffusion and flow-matching models can be used together with the classifier-free guidance method (Lipman et al., 2022), which helps to extrapolate the synthesis towards a stronger conditioning using unconditional model outputs.

In the future perspective, it would also be interesting to finetune an LLM to predict these discrete representation from text and use them for machine translation downstream task, from text to 3D signing. This text-to-sign experiment could help evaluate the quality of representations from a perspective of linguistic usefulness. In a current experiment setup it is possible to compare bitrate only against reconstruction quality of the 3D motion, which does not necessarily mean that these representations preserve semantically and phonologically relevant information.

## 7. References


Omid Alemi. 2019. PyMO: Motion capture library. Accessed: 2025-10-14.

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Trans. Graph.*, 42(4).

Paul S. Bradley and Usama M. Fayyad. 1998. Refining Initial Points for K-Means Clustering. In *International Conference on Machine Learning*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Fidel Omar Tito Cruz and Gissella Bejarano. 2024. Generative Interpolation of Sign Language Poses using RVQ-VAE. In *Latinx in AI @ NeurIPS 2024*.

Epic Games. 2025. Metahuman creator overview. Accessed: 2025-06-05.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. LLMs are Good Sign Language Translators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18362–18372, Los Alamitos, CA, USA. IEEE Computer Society.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H Liu. 2025. SHuBERT: Self-Supervised Sign Language Representation Learning via Multi-Stream Cluster Prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810.

Téo Guichoux, Théodor Lemerle, Shivam Mehta, Jonas Beskow, Gustav Eje Henter, Laure Soulier, Catherine Pelachaud, and Nicolas Obin. 2025. Gelina: Unified Speech and Gesture Synthesis via Interleaved Token Prediction. *arXiv:2510.12834*. Working paper or preprint.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. MotionGPT: Human Motion as a Foreign Language. *Advances in Neural Information Processing Systems*, 36:20067–20079.

Wolfgang Kabsch. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923.

Anna Klezovich, Johanna Mesch, and Jonas Beskow. 2025. Motion Capture Driven Avatars for Swedish Sign language. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow Matching for Generative Modeling. *arXiv:2210.02747*.

Fredrik Malmberg, Anna Klezovich, Johanna Mesch, and Jonas Beskow. 2024. Exploring Latent Sign Language Representations with Isolated Signs, Sentences and In-the-Wild Data. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 219–224, Torino, Italia. ELRA and ICCL.

Shivam Mehta, Ruibo Tu, Simon Alexanderson, Jonas Beskow, Eva Szekely, and Gustav Henter. 2024. Unified Speech and Gesture Synthesis Using Flow Matching. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024)*, pages 8220–8224.

Myndigheten MTM and Marie Hillblom. 2025. 8 Sidor: Nyheter på lätt svenska. Accessed: 2025-10-14.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2023. Bailando ++: 3D Dance GPT with Choreographic Memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–15.

Svenskt teckenspråkslexikon. 2025. Swedish Sign Language Dictionary online. Accessed: 2025-10-14.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. *Advances in Neural Information Processing Systems*, 30.

Phil Wang. 2021. Vector Quantize PyTorch. `https://github.com/lucidrains/vector-quantize-pytorch`. Accessed: 2026-02-20.

Pan Xie, Qipeng Zhang, Peng Taiying, Hao Tang, Yao Du, and Zexian Li. 2024. G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6234–6242.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2024. MotionGPT: Finetuned LLMs Are General-Purpose Motion Generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7368–7376.

Chuanxia Zheng and Andrea Vedaldi. 2023. Online Clustered Codebook. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22798–22807.

# Exploring Aspects of Spontaneous Signing in the DGS Corpus

**Maria Kopf, Reiner Konrad, Gabriele Langer, Marc Schulder, Lutz König**
Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany
{maria.kopf, reiner.konrad, gabriele.langer, marc.schulder, lutz.koenig}@uni-hamburg.de

**Abstract**

Most use of sign language is spontaneous, unplanned, embedded in a one-to-one situation and transient. General sign language corpora aim at such naturalistic data. Thus it can be expected that they include phenomena of spontaneous language similar to the ones described for spontaneous speech in vocal languages: that is, (dis)fluencies such as pauses, hesitations, errors, false starts and repairs as well as discourse markers. In this paper we explore which of the known phenomena of spontaneous language from previous research on vocal and sign languages could be identified in the DGS Corpus using the annotations at hand. We describe our search strategies, consider additional annotation tiers for spontaneous language, and provide examples for the phenomena identified.

**Keywords:** spontaneous signing, disfluency, annotation, DGS Corpus, German Sign Language

## 1. Introduction

In our ongoing research on German Sign Language (*Deutsche Gebärdensprache*, DGS) with the *DGS Corpus* (Prillwitz et al., 2008), we come across passages that appear to be semantically incorrect. For example, a current price is given in Deutschmarks instead of euros[1]. Other passages at first glance look like annotation errors, for instance the manual sign EARRING1A^ (used for meanings like 'woman' or 'girl') being accompanied by the mouthing "junge" (boy).[2] Upon closer inspection, we can determine these to be either errors or slips of the hands, a typical phenomenon in spontaneous signing (Keller et al., 2003). This begs the question: what other phenomena of spontaneous signing can we find in our data, and how do we deal with them in our annotations? And conversely, how can we use our existing annotations to detect different phenomena of spontaneous signing?

Most sign language use is spontaneous face-to-face interaction embedded in the present communication situation, transient and largely unplanned. Other times, signing is informed by careful planning, review, recording, correcting and rerecording, produced for permanent use – a conceptually "written" form of signing (Koch and Oesterreicher, 1985).

Most linguistic sign language corpora, while recorded, aim at a more naturalistic form of signing for their data, closer to spontaneous signing than to edited signed text. Data collection formats are preferred in which sets of signers spontaneously sign with each other on topics of interest or in "staged communicative events" (Himmelmann, 1998, pp. 185–186). Thus, the data to be dealt with in these sign language corpora can be seen as language-in-interaction and be expected to contain similar characteristics and challenges as spontaneous speech events of vocal language[3], studied under the terms *disfluencies, turn taking* or *repairs*.

With this article we present a first attempt to identify phenomena of spontaneous signing that occur in the *DGS Corpus*. To do these phenomena justice, we propose an additional annotation scheme to supplement the basic annotation of the corpus. We do not offer a comprehensive typology in the way Shriberg (1994) does for English, as we only looked for phenomena in stretches of monologue-like signing where one signer holds the floor (turn).

After an overview of related work and introducing the *DGS Corpus* (section 2) we discuss how annotation schemes could be expanded or supplemented to permit more in-depth analyses of spontaneous signing (section 3). We then illustrate a number of phenomena with examples from the corpus and discuss how they (and the search for them) differs from their counterparts in vocal language (section 4). We conclude with a discussion of our findings and outlook to future steps (section 5).

## 2. Background

### 2.1. Spontaneous Language

Spontaneous language use has been studied from different fields with different frameworks and goals. The study of phenomena of spontaneous language

[1] https://ling.meinedgs.de/release4/html/2025500_en.html#t00154443

[2] https://ling.meinedgs.de/release4/html/1212402_en.html#t00033228

[3] We use the term vocal language for audio-phonetic languages (Sallandre and Garcia, 2013), commonly also referred to as *spoken languages*.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 262–274
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

has a long history with psycholinguistics under the term *disfluency*, focussing on language production. Conversation analysis examines it from the point of discourse structure, calling it *talk-in-interaction*. Computational linguistics analyses it for the purpose of automatic speech processing and language acquisition studies it for the sake of improving fluency in L2 speakers. See Kosmala (2022) for an overview and intriguing combination of the different frameworks. One of the most fundamental works in the study of disfluency is Shriberg (1994), which presents a comprehensive classification of disfluency types for American English. In 1999 the *Disfluency in Spontaneous Speech (DiSS)* workshop series was established from which a broad canon of literature has grown. While multimodal accounts are quite established, studies on sign languages have not yet been part of the workshop series.

In studies approaching spontaneous language from the discourse analysis framework, sign languages are used to compare phenomena cross-linguistically and cross-modally (cf. Floyd et al., 2016). In the field of sign language linguistics the interest on discourse management and interactional phenomena is growing as well. Corpus-based studies comparing vocal and sign languages (here French Belgian Sign Language (*Langue des Signes de Belgique Francophone*, LSFB) are compiled in Degand et al. (2019). For a brief overview of additional studies see Lutzenberger et al. (2024). Most literature, though, focusses on individual phenomena. The palm-up gesture as discourse marker has received significant attention (Lepeut and Shaw, 2025; McKee and Wallingford, 2011; Notarrigo and Meurant, 2014). Gabarró-López (2020b) gives an overview of studies on discourse markers in sign languages, covering American Sign Language (ASL), Venezuelan Sign Language (*Lengua de Señas Venezolana*, LSV), Spanish Sign Language (*Lengua de Señas Española*, LSE) and LSFB. Gabarró-López (2020a) adapted a vocal languages annotation protocol for discourse markers for use with LSFB. Notarrigo and Meurant (2019) study repetitions in spontaneous signing, also for LSFB.

For DGS we found a small number of studies dedicated to spontaneous signing. Hansen (2012) discusses several aspects of discourse studies such as coherence, cohesion and context. Volk (2016) reports on discourse functions of the palm-up gesture in DGS, as does Kaul (2024) who analyses data from the *Public DGS Corpus*. Mpadanes (2020) studies repairs in DGS, also based on *DGS Corpus* data. Based on annotations of filmed monologues, Oviedo et al. (2018) studies filled and unfilled pauses in the context of fluency and perceived disfluency. Keller et al. (2003) studies articulation errors and their repairs, based on a collection of signed stories chosen to evoke slips of the hands.

## 2.2. Sign Language Annotation

Besides documenting and archiving, one of the first steps in building sign language corpora is basic annotation, in which each manual sign is identified by labelling it with an ID-gloss (Johnston, 2010) or directly linking token to type entries in a lexical database (Hanke and Storz, 2008). Further annotations depend on the goals and the resources of each project and are documented in annotation guidelines. See Hodge and Crasborn (2022) regarding good practices in annotation.

Phenomena of spontaneous language which are annotated in several corpora are repetitions, false starts and gestures (Kopf et al., 2022). For repetitions the distinction between grammatical and non-grammatical forms is decisive for annotation. In grammatical forms the repetition is used to mark aspect or plural; in non-grammatical forms the sign is repeated for other reasons, e. g. to intensify or gain time for language planning. For the focus of this article only non-grammatical forms are of interest. False starts are marked with a suffix in the Auslan Corpus and STS Korpus (Johnston, 2024; Mesch and Wallin, 2021). In the Auslan Corpus a false start is identified by a repair sequence in the direct or near context. The STS Korpus additionally annotates pauses due to hesitations and lengthenings in the form of a hold. The way in which gestures are annotated differs in how exactly gestures are defined. One gesture that nearly all corpora annotate with a special ID-gloss is the palm-up gesture (Kopf et al., 2022).

An aspect that occurs when annotating spontaneous sign language is the use of the articulators, i. e. the hands, for non-linguistic purposes such as scratching one's nose (Crasborn et al., 2020; Konrad et al., 2022). For the purpose of this article we exclude such uses.

Another challenge when annotating spontaneous signing lies in the multimodality of language. As Braem and Volterra (2023, p. 32) state: "the facial, gestural, and other prosodic behaviors produced simultaneously with either speech or signing had traditionally usually been relegated to a kind of limbo with other phenomena that were considered 'paralinguistic' or 'nonlinguistic'". But in the study of spontaneous language these "nonlinguistic" behaviours are crucial. The difference between repair sequences and rhetorical stylistic devices in signed statements of the form "*A, no, B*" can often only be recognised by subtle differences in facial expressions and head movement. During pauses in the manual production, whole sections of the uttered statement can be produced on only the face. Research on disfluency in speaking and signing does recognize this and adjusts methods and theories (again, see Braem and Volterra, 2023). Still, the disregard of non-manual activity is evident today in

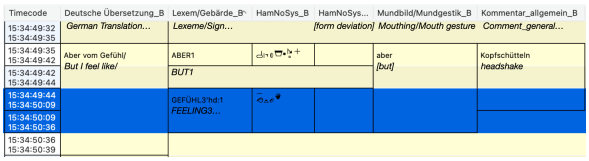



Figure 1: iLex transcript with tiers for translation, token, mouthing/mouth gesture, and comment.

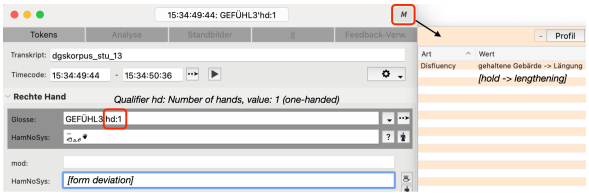



Figure 2: iLex token-tag window with metadata.

sign language annotation as the main focus in corpus annotation continues to be on manual signs, in part due to resource restrictions. While mouthings and mouth gestures are often annotated, other non-manuals are not part of a basic annotation.

### 2.3. DGS Corpus

The *DGS Corpus* (Prillwitz et al., 2008) is an annotated reference corpus of DGS. It contains monologues and dyadic discourses that were captured in the context of various collection tasks (Nishio et al., 2010). The recordings from its first data collection campaign (2010–2012), which the analyses in this article are based on, cover 560 hours of language production by 330 participants from all parts of Germany. Over 50 hours of the *DGS Corpus* have been released publicly as the *Public DGS Corpus* in a number of different formats (Konrad et al., 2024a), which includes the fully annotated research dataset *MY DGS – annotated* (Konrad et al., 2024b).

The corpus contains annotations and translations. Translations into German are done by sign language interpreters; broken-off utterances are marked by a slash. Using the iLex annotation tool (Hanke and Storz, 2008), annotators time-align the translations sentence by sentence, then segment and lemmatise tokens (token-type matching). Mouthing and mouth gesture are also annotated in this first annotation pass, including a note on the deviation from the type's citation form. Gestures and non-manual signals like head nods and shakes are annotated with special gesture types when occurring on their own or less formally in a comment tier when they co-occur with a manual sign.

In the lemma revision process, tokens with form deviation are reviewed and either the deviation is specified using the Hamburg Notation System (HamNoSys) (Hanke, 2004) or tokens are grouped into form features called 'qualifiers' (see Loos and Konrad, 2022). Additional information can be specified via metadata entries.

In addition to its basic annotation, the *DGS Corpus* labels so-called non-tokens for lexicographic work (Langer et al., 2016). Non-tokens are sign uses that should not count as evidence for a sign type or its use by a particular signer. In addition to metalinguistic references and contrastive untokens, non-tokens include various cases of slips of the hands, false starts, aborted signs, malformed or deviated sign forms, gestures, and trying out or searching for signs. Therefore, offering an overlap between non-tokens and phenomena of spontaneous signing.

Figure 1 shows a snapshot of an iLex transcript[4] with timestamps and the tiers "German Translation", "Token", "Type citation form" (HamNoSys), "Token deviation" (HamNoSys), "Mouthing/Mouth gesture", and "Comment". The signer interrupts their signing (the translation ends with a slash) and the second sign is held with a duration of nearly one second (42 frames). Figure 2 shows the token-tag window of the second sign with the qualified type gloss and the metadatum *Disfluency*.

## 3. Additional Annotations

In order to spot tokens or translation tags of interest for this study we implemented the metadata category *Disfluency* (cf. section 2.3). For annotating the corresponding sequences we introduced two tiers in the transcripts: *Disfluency_types* and *Disfluency_parts*. Tags in *Disfluency_types* can span over more than one token, tags in *Disfluency_parts* are used to label tokens within repetitions and repairs. Parts of a repetition are the sign's first instantiation and its repetition. Usually one sign is repeated, but in example 1[5] a combination of two signs meaning "I participated in all the cities" is part of a non-adjacent repetition (see link in footnote to view the example). Figure 3 shows the annotation of this in iLex.

Parts of a repair are *reparandum*, *interregnum*, and *reparans* as introduced by Shriberg (1994) (cf. Kosmala, 2022, p. 51; Piu and Bove, 2007, p. 399), *interregnum* being optional. In example 2[6] the signer recounts Diana, Princess of Wales, caring greatly about children (phrased in German and DGS as *"having a heart for children"*). He first signs HEART2B, PERSON1 and FOR1 accompanied

[4] Public transcript matching fig. 1: https://ling.meinedgs.de/release4/html/1182343_en.html#t00050035

[5] https://ling.meinedgs.de/release4/html/1414312_en.html#t00073006

[6] https://ling.meinedgs.de/release4/html/1413485_en.html#t00035311

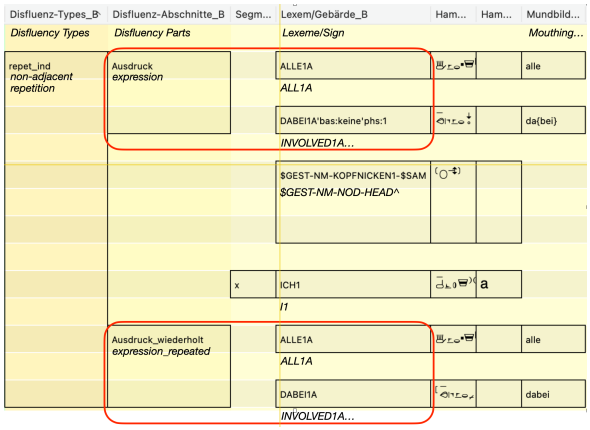



Figure 3: Disfluency annotation for a repetition.

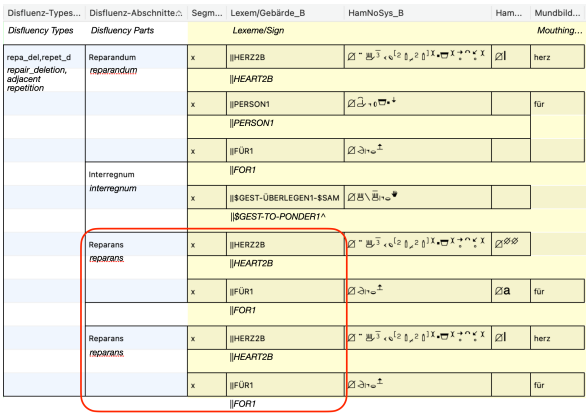



Figure 4: Disfluency tiers for repair and repetition.

by the mouthing “herz für” (heart for), then hesitates while gesturing, and repeats the sequence HEART2B FOR1 without PERSON1 two times. Its disfluency annotation is shown in fig. 4.

As Crible (2018, p. 78) shows for vocal languages, one sequence can display several disfluency types. In our current work we did not opt for macro-labels of sequences but allowed the selection of more than one type in the *Disfluency_types* tag, as for example 2, where we tagged a repair with deletion (see section 4.1) and an adjacent repetition (see section 4.2).

We agree with Crible (2017) that “[i]t is not [certain] whether these annotation systems could be applied to languages other than their original”. Nevertheless, we hope to offer ideas and perspectives which can be tested for other sign languages.

An added benefit of this work is the ongoing quality control it provides for corpora, as annotation artefacts or missing annotations are detected while looking for disfluencies.

## 4. Selected Phenomena

We follow Shriberg in her definition of (dis)fluencies as any “stretch of linguistic material [which] must be deleted to arrive at the sequence the speaker ‘intended’” (Shriberg, 1994, p. 1, as quoted in Crible, 2017). We do not only concentrate on material that interrupts the fluency of the sign stream but also include material that helps retaining it (“fluencemes” in the terminology of Götz (2013)). Comparing (dis)fluency typologies from Shriberg (2001), Kosmala (2022), Notarrigo and Meurant (2019), and Blache et al. (2017), we decided to concentrate on the following phenomena: *Pauses, Repetitions, Elisions, Lengthening, Repairs,* and *Articulation Errors* (sections 4.1 to 4.6). Based on Crible (2018) we added *Discourse Markers* as a further phenomenon (section 4.7).

We now describe how we searched for these phenomena of spontaneous DGS using several strategies: the database query capabilities of iLex, existing annotation, previous work with the corpus data and frequency counts. Queries resulted in lists of candidates that were manually searched for cases of the phenomena. Each phenomenon is introduced and we describe the subphenomena we encountered, illustrated by examples.

### 4.1. Repairs

A repair is a sequence where the signer detects a problem, halts the production and then continues. The halting can occur after the sign is completed, creating a pause, or during sign execution, resulting in a broken off sign or a sign with a reduced movement or missing repetition. The continuing part can be a correction of the prior utterance (in the form of a deletion, insertion or substitution), but it also can be a complete new phrase, as is the case for false starts. Following Shriberg (2001) we focus on five kinds of repairs: repairs of articulation errors, repairs including a deletion, insertions, substitutions, and false starts.

To detect repair sequences we first used several queries detecting break-offs in the translation tier (indicated by a slash) at the beginning, middle or end of a translation tag, also in combination with tokens marked as aborted in the gloss tier with a HamNoSys-symbol.

Not all shortened sign forms are aborted signs and marked as such with the symbol for break-offs. Shortened sign executions with missing repetitions are marked through a qualifier indicating the number of repetitions or absence of repeated motion (phases). We used queries to find all cases where a sign with a repeated motion in its citation form is signed without repetition (*phase:1*) or without movement at all (*phase:0*) and signed again with repetition within the next five signs.

Another query to find self-repair candidates identified repetition of a sign after a break-off by looking

at translation tags with a break-off slash and checking whether the last sign of that tag re-occurred within the first four signs of the next translation tag. Cases with significant pauses between translation tags were filtered out, as this pause implies a pause in the production, followed by a resumption or change of topic.

A deletion was illustrated in example 2 (section 3), where the signer deletes PERSON1 in the repair sequence. A false start can be seen in example 3[7] where the signer interrupts his signing by signing NO1A and then restarts with a different phrasing.

In example 4[8] an incorrect sign is substituted, the signer having made a mistake in naming a district of Hamburg—not Blankenese, but Sülldorf—and corrects this with a shake of the head while signing OR4B before showing the manual alphabet sign for S.

In addition to substitutions of incorrect signs we found repairs with substitutions of lexical variants. In example 5[9] the signer talks about a person interviewing a woman. She starts with the onset of the sign WOMAN1A, which involves tracing the shape of the breast. Then she aborts the execution of this sign by blending smoothly to a full execution of the lexical variant WOMAN4A. The signer uses WOMAN1A fourteen times in the corpus, and WOMAN4A only three times. It is not clear why in example 5 she changes to the less frequently used variant. Possible explanations are the attempt to replace WOMAN1A by another lexical variant because of an ongoing discussion about the political correctness of the sign WOMAN1A or a commonly observed pattern of signing multiple lexical variants in a row for concepts with rich regional variation, possibly as a strategy to ensure understanding across dialectal boundaries.

Example 6[10] (also seen in fig. 5) shows the signer starting to sign ENGLAND4, stop and change to ENGLAND3, both times breaking off the execution, then fully produce ENGLAND4. In the entire conversation segment of about 16 minutes, the signer uses ENGLAND3 six times and ENGLAND4 seventeen times. Shortly before, his conversation partner used ENGLAND3, so this is most likely an attempt by the signer to adapt his signing to his interlocutor. (Later on in the conversation the signer uses ENGLAND3 once, then switches back to ENGLAND4 and even later he fully signs both signs in a row.)



| | | | |
|---|---|---|---|
| ... | **ENGLAND4*** | **ENGLAND3*** **ENGLAND4** | |
| | $INDEX1 | $INDEX1* $INDEX1* | ... |
| ... | **e{ngland}** | **england** | |
| | It's from **England**. | | |

Figure 5: Repair with repetitions and break-offs in example 6[10]. Shown are the three signs meaning 'England', the first two broken off. Below are glosses for dominant and non-dominant hand, mouthing, and translation. For "e{ngland}", only "e" was mouthed.

## 4.2. Repetitions

Repetitions are cases where basic meaningful elements appear more than once. Notarrigo and Meurant (2019) give a comprehensive analysis and typological description of repetitions based on corpus data of spontaneous signing in LSFB. They distinguish three form groups of repetition: adjacent, non-adjacent (with one or more elements in between), and framing (non-adjacent with the function to highlight a central element). They report that in sign language linguistics literature adjacent repetition is identified as a disfluency marker indicating hesitation or repair. The authors looked for repetitions of manual signs in the stretch of one signer's turn, excluding interactional factors. Although nearly half of the repetitions had a pragmatic function, there were only few cases where repetition indicates hesitation or repair.

In this work, we limit our analysis to utterances of single signers, excluding aspects of immediate interaction with the addressee. However, as is discussed for example 6, wider interactional contexts could still play a role. To find adjacent repetitions we used a query looking for two neighbouring tokens with the same gloss name irrespective of numbers and letters for variants or extensions for qualifiers.

A case of adjacent repetition can be seen in example 7[11], where the signer recounts her school days. TEACHER2 is signed twice. There is no pragmatic or grammatical reason for this repetition.

A non-adjacent repetition can be seen in example 8[12]. The signer talks about the British royal family moving into a castle. The sign TO-MOVE1

[7] https://ling.meinedgs.de/release4/html/1210997_en.html#t00065404

[8] https://ling.meinedgs.de/release4/html/1179224_en.html#t00164519

[9] https://ling.meinedgs.de/release4/html/1248862_en.html#t00043921

[10] https://ling.meinedgs.de/release4/html/1248090_en.html#t00050405

[11] https://ling.meinedgs.de/release4/html/1418889_en.html#t00021205

[12] https://ling.meinedgs.de/release4/html/1248505_en.html#t00032409

is signed two times with the sign TO-LIVE4 in between. There is no manual or non-manual marking which would indicate a repair or emphasis.

Another pattern, which may be called lexical doubling, is seen quite often in the corpus: Two consecutive lexical variants for one concept, often sharing one mouthing, where each part is also a stand alone sign for the concept. Examples of this pattern are FAMILY1 FAMILY3 as in example 9[13] and BEAUTIFUL1A BEAUTIFUL3 as in example 10[14]. This kind of redundancy might simply be a strategy to ensure understanding when several lexical or even regional variants exist for the same concept, as already discussed in section 4.1.

In other cases, the second sign might have a visual or illustrating function due to its iconic motivation, e. g. providing information on the visual or spatial properties of the object, setting it up in space or providing a hold (ground) to be referred to subsequently. In example 11[15] the signer describes the equipment of a bus. First signing BUS2 (iconic motivation: using a steering wheel), followed by BUS1A (tracing the shape of the vehicle with two C-handshapes). The latter variant can be located anywhere in the signing space, in contrast to the fixed place of articulation (PoA) of the first. The two variants are followed by an index and the signs CURTAINS2 and CURTAINS3, which are executed at the same PoA as BUS1A, indicating the curtains were in the bus. Here the visual space is used to draw the whole picture of the bus with its curtains.

A case with lexical variants and a modified form can be seen in example 12[16]. The signer is retelling a picture story where a cage is picked up. He uses three strategies to sign the concept 'cage'. A one-handed modification of the sign CAGE4 in front of the face, as if he were inside the cage looking through the bars, followed by an unmodified CAGE4, and then the lexical variant CAGE2.

### 4.3. Elisions

Elisions are cases where an element of a word or the whole word is not realised (Blache et al., 2017; Crible, 2018). As morphemes are often executed simultaneously in sign languages, elisions are quite hard to detect. Looking at the phonological parameters of a sign, a sign executed in front of the body instead of the head could be seen as an elision of the PoA, but there still is a PoA, it just changed to a less marked PoA (discussed as "lowering" in the literature). The same is true for handshapes and orientation, one always has a handshape and orientation when signing, but it might deviate from the citation form. However, a sign with a path movement in its canonical form can be executed without movement. So to detect elisions we concentrated on movement using the already existing annotation with the movement qualifier *phase*, described in section 4.1, e. g. LOUD1C in example 13[17] in which the signer describes the learning environment at her boarding school.

The deletion of a whole word is easier to detect, as it is part of a repair sequence (cf. section 4.1). To find cases where the movement is not completely deleted but only reduced we searched for tokens being considerably shorter than the average duration of their type. To detect outliers, we calculated the z-score[18] of each sign's duration and inspected tokens with a score of $z < -1$.

With this detection strategy we found many cases with a shortened movement. This might in part be an artefact of annotation, which allows changing the playback speed or even looking at individual frames. That way, annotators may detect handshapes, often between two clearly produced signs, and annotate them as tokens with deviations, which are unlikely to be visible to the naked eye. Still, they might be seen as a reduced sign originating in a semi-realised planning event, as in the case of example 14[19] where the signer talks about playing squash. She signs WORKOUT1 with a shortened movement, followed by TO-PLAY2 without any markers for a repair.

Both strategies will detect cases that should not be considered aspects of spontaneous signing, but are deliberate markings to indicate an interruption in a process, what Liddell and Johnson (1989) call "unrealized-inceptive aspect". One such case is example 15[20], where the signer describes the passage of time with the sign PROCEEDING1^, which is aborted to show the interruption when she got pregnant, and resumed afterwards. Therefore, candidates for elision have to be double checked for their status as true elisions.

[13] https://ling.meinedgs.de/release4/html/1431428_en.html#t00011118

[14] https://ling.meinedgs.de/release4/html/1210997_en.html#t00061437

[15] https://ling.meinedgs.de/release4/html/1429709-16344227-16451821_en.html#t00015207

[16] https://ling.meine-dgs.de/release4/html/1413240_en.html#t00000402

[17] https://ling.meine-dgs.de/release4/html/1212402_en.html#t00064531; note that in public data, qualified types are indicated by an asterisk, while the gloss is LAUT1C'phs:0 in the internal corpus.

[18] Z-score, or standard score, indicates the number of standard deviations that a value is above or beyond the population's mean value. It is defined as $z = \frac{x-\mu}{\sigma}$, x being the value, $\mu$ the mean, and $\sigma$ the standard deviation.

[19] https://ling.meinedgs.de/release4/html/1212402_en.html#t00085710

[20] https://ling.meinedgs.de/release4/html/1418889_en.html#t00055607

### 4.4. Lengthening

Lengthenings or prolongations are executions of words with longer durations of phonemes (or syllables) than usual. Other phenomena can also be prolonged, e. g. filled pauses or discourse markers (Crible et al., 2024). Oviedo et al. (2018) describes lengthening as one form of filled pauses.

To search for lengthened signs, we again used the z-score of sign durations, as we did for shortenings (section 4.3), this time searching for longer than usual instances. Long outliers can be more extreme than short ones, so we searched for signs with $z > 5$ for our initial investigation. We found lengthenings of two forms: either the sign is repeated multiple times[21] or the hand is held in the first or last PoA of the sign.

In example 16[22] the signer describes different vocations that are part of the education in a school in Leipzig. He repeats $LIST1 several times while trying to remember the vocations. He looks at his interlocutor to start with, but after the sign LEIPZIG1A his gaze wanders to the top left corner of the room while he is thinking. He continues signing different vocations, his gaze still to the top left corner. Arriving at number four of his list he again lengthens the sign with a reduplication followed by a filled pause ($GEST^ in the current public release, internally already revised to $GEST-TO-PONDER1^). Finally he remembers the fourth vocation and his gaze is back on the interlocutor again.

In example 17[23] the signer seems to have difficulties recalling where the last sign language festival he attended took place. After repeating the signs for 'sign language festival', he holds the sign A-MOMENT-AGO1A. During the whole sequence of repeating and holding, his head is bowed, he looks down, and the tip of his tongue is visible in the corner of his mouth. Once he recalls the location he nods and continues to sign IN1 BERLIN1A.

In contrast to the described unintended forms of prolongations, lengthening can also be used intentionally to indicate an intensifier. Example 18[24] shows the signer prolonging LONG-TEMPORAL4A as well as their mouthing of "lang" (long) to indicate that the referenced time span was particularly long.

### 4.5. Pauses

Pauses can be described as interruptions of information production. Gabarró-López and Meurant (2016, p. 85) define them as "periods of no signing at all" and include stops with crossed hands and relaxed hands in the neutral signing space or along the body.

Most authors differentiate between unfilled or silent pauses and filled pauses. Unfilled pauses are often defined by a timespan of silence, i. e. the interruption of the sound signal in vocal languages or missing or "neutral" (Oviedo et al., 2018) activity of the hands and face in sign languages, e. g. Oviedo et al. (2018) specify a minimum of 160 milliseconds for DGS. Filled pauses are stretches filled with "conventional and neutral phonetic form[s], (e. g. 'euh' in French)" (Crible, 2018, p. 72). For DGS, Oviedo et al. (2018) describes four kinds or filled pauses: a) crossed hands accompanied with non-manual activity, b) a longer transitioning between two signs (due to a break in or slow execution of the movement), c) non-morphological prolonging of sign segments (e. g. a prolonged hold at the beginning of a sign), and d) gestures without meaning.

In the *DGS Corpus* silent pauses are not annotated and the time length of tokens are determined by the rules of narrow transcription, i. e. transition movements are not part of the token. This is important to keep in mind when calculating time between tokens. For detecting stretches without signing, other than between utterances, we used a query that checks whether there are long periods within a translation tag without any sign token and without inputs from an interlocutor. To find the most clear cut cases we started out with pauses of a full second. In future passes, this period can be reduced to match the approach of Oviedo et al. (2018).

A part of filled pauses is covered by existing annotations and therefore already annotated. Relevant types include $GEST-OFF1^[25] (the palm-up gesture), $GEST-TO-FOLD-ONES-HANDS[26] (used when the signer folds their hands within a turn), $GEST-NO-IDEA1^[27] (a gesture meaning 'no idea'), and five variants with the gloss name $GEST-TO-PONDER (gestures used while or for thinking, three of which occur in the public corpus and can thus be found in its type index.[28])

We found both kinds of pauses, filled and unfilled. Within filled pauses we detected the four kinds described by Oviedo et al. (2018): crossed hands with non-manual elements, long transitions, lengthening (see also section 4.4), and gestures.

[21] Multiple repetitions without breaks are annotated as one token in the *DGS Corpus*.

[22] https://ling.meinedgs.de/release4/html/1414563_en.html#t00105700

[23] https://ling.meinedgs.de/release4/html/1414312_en.html#t00090244

[24] https://ling.meinedgs.de/release4/html/1178133_en.html#t00000845

[25] https://ling.meinedgs.de/release4/types/type16636_en.html

[26] https://ling.meinedgs.de/release4/types/type72381_en.html

[27] https://ling.meinedgs.de/release4/types/type60946_en.html

[28] https://ling.meinedgs.de/release4/ling/types_en.html

In example 19[29] the signer recounts a football tournament. As he tries to remember which team he played against, he folds his hands, looks up and furrows his brow.

When working with corpus data it is apparent that not all non-activity of the hands are pauses in a strict sense. Sometimes manual signing is halted while the face, head or torso is used in a CA to add information to the utterance or the head and facial expression are used to comment on the signing, as in example 20[30], and thus should not be interpreted as a pause at all, so just checking for activity of the hands is not enough. Even cases with hardly any manual and non-manual activity must be viewed in context. While they can be indicators of active thinking as in example 21[31] (pause between ATHLETICS1 and TO-SWAP3A) they might also be re-enactments of thinking or hesitating in a constructed action (CA), although our observations indicate that in most such cases they are explicitly marked.

### 4.6. Articulation Errors

Articulation errors, or slips of the hands, are cases of incorrect signing in an otherwise correct production. For vocal languages it is estimated that on average within 1,000 words or 10 minutes there is one slip of the tongue (Fromkin, 1980). For sign languages, Keller et al. (2003) show that slips of the hands follow the same patterns found in vocal languages regarding the types of errors (e. g. anticipations, perseverations, substitutions, blends, fusion, etc.), monitoring behaviour (repairs) and affected units (from phonological parameters up to full signs and even phrases). The only difference between them and slips of the tongue lies in their distribution: complete signs are affected more than complete spoken words, whereas morphemes and phrases are less affected in sign languages than in vocal languages.

Articulation errors are already a subcategory of repairs (see section 4.1), but as only half of all slips of the hands are repaired by the signers themself (Keller et al., 2003) there is a need for a distinct category, annotating articulation errors that occur outside of repair sequences. To find articulation errors, we relied on pre-existing work that marked non-tokens during lexicographic analysis (see section 2.3). Developing detection strategies like those described in previous sections proved challenging, as there are many valid reasons for deviations from the citation form that would have to be filtered out. And clearly distinguishing articulation errors from phenomena such as assimilations, signing style or lax signing also proves to be extremely difficult. Newkirk et al. (1980) suggest that slips of the hand change a parameter of the sign completely while juncture assimilations substitute only part of the parameter, e. g. the handshape is changed only towards the end of the sign. The distinction between different linguistic phenomena should be refined for future work on articulation errors.

A phonological articulation error can be seen in example 22[32] where the signer talks about her final exam. While signing the two signs GRADUATION-OR-COMPLETION2 and TO-EXAMINE1 meaning 'final exam', a metathesis of the handshapes happens. The wrong execution of TO-EXAMINE1 is aborted and the sign is repeated in it's correct form. The two examples from section 1 represent a lexical substitution (Deutschmark instead of euros) and a lexical anticipation (girls anticipated instead of boys). Both examples also show a mismatch between the manual sign and the mouthing.

### 4.7. Discourse Markers

There exist numerous approaches for defining the category of discourse markers (DM), however, they still are, in the words of Crible (2018, p. 34) a "slippery linguistic category". For our purposes we define them as signs that are used on a metadiscursive level to structure discourse.

In previous research, some signs and gestures, especially palm-up, have been discussed with regard to discourse organisation. Palm-up is attested for many sign languages with a variety of functions. Volk (2016), Debreslioska et al. (2025), and Kaul (2024) show that palm-up gesture has, among others, discourse regulating functions in DGS, which is also attested for several other sign languages (e.g. McKee and Wallingford, 2011; Lepeut and Shaw, 2025). Some authors also discuss individual signs that function as DM, such as HEY and WELL in ASL by Hoza (2011), FINE in ASL by Metzger and Bahan (2001), different kinds of buoys in LSFB by Gabarró-López (2019), as well as SAME and palm-up in LSFB by Gabarró-López (2020b), or ENCAMBIO (INSTEAD-OF) and ES-DECIR (THAT-IS-TO-SAY) in LSE by Villameriel (2010). Other studies try to detect all discourse markers in a given piece of signing, such as Pérez (2006) for LSV, Gabarró-López (2025) for LSFB and Villameriel (2008) for LSE.

Furthermore, as Crible (2018) showed in her bottom-up research on French and English, there are language-specific as well as universal patterns

[29] https://ling.meinedgs.de/release4/html/1209006_en.html#t00024826

[30] https://ling.meine-dgs.de/release4/html/1419797_de.html#t00091547

[31] https://ling.meinedgs.de/release4/html/1414123_en.html#t00015311

[32] https://ling.meinedgs.de/release4/html/1212402_en.html#t00123445

of DMs, so not all described DMs might apply to DGS. To our knowledge there is no comprehensive study of DMs in DGS, however, some DGS signs have been mentioned serving the function of DMs, e. g. palm-up (Debreslioska et al., 2025), BEISPIEL (EXAMPLE) and BEDEUTUNG (MEANING) (Hansen, 2012), and DONE (Jahn et al., 2021).

Crible (2018, p. 82) notes "a highly pervasive and prominent use of DMs in spoken language" and reports a relative frequency of DMs around 55 occurrences per thousand words. Thus, in a corpus of spontaneous signing, signs that function as DMs can be expected to be among the most frequent elements. Consequently, we scanned the 100 most frequent signs of the *DGS Corpus* for possible candidates. Of those 100, we identified over twenty that we suspect of having a discourse structuring function. Among them are signs that previously had been identified as DM (e. g. palm-up, DONE), signs that have a comparable semantic origin as the DM signs mentioned for other sign languages (THEN1A[33], EQUAL1A[34], various kinds of buoys) and several signs that suggested themselves during lexicographic analysis or annotation as having DM functions, e. g. MEANING1, illustrated by example 23[35].

This can only be a first glimpse into an area not thoroughly investigated so far and further analysis on discourse markers in DGS is highly desired. The possible candidate signs identified could be a good starting point for such an analysis to detect, verify, and describe their functions as DMs.

## 5. Discussion and Future Work

Errors and slips of the hands in signing are hard to identify. In order to spot a mistake one has to rule out that the signing is intentional and correct. DGS includes a large amount of lexical variation, a great amount of liberty of how to execute a specific sign, and flexibility in syntactic structures. Therefore, unless a mistake is clearly marked by a self-correction as in example 4, one always has to consider the possibility that occurrences such as repetitions have an intentional linguistic or communicative purpose. In contrast to vocal languages repetitions of lexical variants (lexical doubling) as well as pairing an iconically motivated sign with a conventional sign seem to be sign language specific communicative strategies that might not only be found in spontaneous but also in planned signing.

As the examples given in section 4 show, phenomena of spontaneous signing go far beyond repairs of errors. Basic annotation that focusses on manuals can only be used to a limited extent to detect them. For more comprehensive searches, detailed annotation of non-manuals (body posture, head movement, facial expression, eye gaze, etc.) is also required. Building on this, a dedicated annotation of phenomena as we discuss in section 3 may be undertaken. Scaling this sufficiently for frequency-based analyses as, for example, Crible (2018) did for two vocal languages, is, unfortunately, outside the scope of the DGS-Korpus project. However, future projects will hopefully use the *DGS Corpus* to explore such matters.

Aside from manual annotations, another potential source of information to guide researchers may be kinematic data gained from automatic pose recognition models (Schulder and Hanke, 2020). While the semantic interpretation of signing with computer vision methods, especially at the level of nuance relevant to linguistic analysis, remains an open challenge, pose information may already be helpful, e. g. for spotting unfilled pauses.

Spontaneous language also offers the potential to revisit the assumptions and paradigms of current annotation schemata. One such undertaking in the related field of gesture studies is the *MultiModal MultiDimensional* (M3D) labelling system (Rohrer et al., 2025), which exhibits a very broad view on annotating linguistic phenomena.

Finally, while the separate presentation of phenomena gives the impression that they can be clearly distinguished and delineated, this is not necessarily the case. We found many instances in which several phenomena occurred in parallel or were intertwined. Crible (2018) and Shriberg (2001) use the term *complex* for these cases. Accounting for such phenomenon complexes in annotation schemes will require further deliberation.

## 6. Acknowledgements

This document has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

[33] https://ling.meinedgs.de/release4/types/type14891_en.html

[34] https://ling.meinedgs.de/release4/types/type13890_en.html

[35] https://ling.meinedgs.de/release4/html/1246100_en.html#t00053300

## 7. Bibliographical References


Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prévot, and Stéphane Rauzy. 2017. The corpus of interactional data: A large multimodal annotated resource. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1323–1356. Springer Netherlands, Dordrecht.

Penny Boyes Braem and Virginia Volterra. 2023. Through the signed language glass: Changing and converging views in spoken and signed language research. In Terry Janzen and Barbara Shaffer, editors, *Signed Language and Gesture Research in Cognitive Linguistics*, pages 23–48. De Gruyter Mouton.

Onno A. Crasborn, Inge Zwitserlood, Els van der Kooij, and Richard Bank. 2020. Annotation conventions for the Corpus NGT. Annotation convention, Centre for Language Studies & Department of Linguistics, Radboud University, Nijmegen, Netherlands.

Ludivine Crible. 2017. Discourse markers and (dis)fluency in English and French: Variation and combination in the DisFrEn corpus. *International Journal of Corpus Linguistics*, 22(2):242–269.

Ludivine Crible. 2018. *Discourse Markers and (Dis)fluency: Forms and functions across languages and registers*, volume 286 of *Pragmatics & Beyond New Series*. John Benjamins Publishing Company, Amsterdam.

Ludivine Crible, Ivana Didirková, Christelle Dodane, and Loulou Kosmala. 2024. Towards an inclusive system for the annotation of (dis)fluency in typical and atypical speech. *Clinical Linguistics & Phonetics*, 38(4):381–398.

Sandra Debreslioska, Anna Kuder, and Pamela Perniss. 2025. The palm-up open hand gesture across language modalities: A comparison of German and DGS (German Sign Language). *Journal of Pragmatics*, 249:99–119.

Elisabeth Degand, Laurence Meurant, Gaëtanelle Gilquin, and Anne Catherine Simon, editors. 2019. *Fluency and Disfluency across Languages and Language Varieties*. Number 7 in Corpora and Language in Use. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium.

Simeon Floyd, Elizabeth Manrique, Giovanni Rossi, and Francisco Torreira. 2016. Timing of visual bodily behavior in repair sequences: Evidence from three languages. *Discourse Processes*, 53(3):175–204.

Victoria Fromkin. 1980. Introduction. In *Errors in linguistic performance: slips of the tongue, ear, pen, and hand*, pages 1–12. Academic Press, New York, NY.

Sílvia Gabarró-López. 2019. Describing buoys from the perspective of discourse markers: A cross-genre study of French Belgian Sign Language (LSFB). *Sign Language & Linguistics*, 22(2):210–240.

Sílvia Gabarró-López. 2020a. Are discourse markers related to age and educational background? A comparative account between two sign languages. *Journal of Pragmatics*, 156:68–82.

Sílvia Gabarró-López. 2020b. Discourse markers, where are you? Investigating the relationship between their functions and their position in French Belgian Sign Language conversations. *Sign Language Studies*, 20(2):231–263.

Sílvia Gabarró-López. 2025. Discourse markers in French Belgian Sign Language (LSFB) dialogues and their translation into French: A corpus-based study. *Languages*, 10(9):243.

Sílvia Gabarró-López and Laurence Meurant. 2016. Slicing your SL data into basic discourse units (BDUs). Adapting the BDU model (syntax + prosody) to signed discourse. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 81–88, Portorož, Slovenia. European Language Resources Association (ELRA).

Sandra Götz. 2013. *Fluency in Native and Nonnative English Speech*, volume 53 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company, Amsterdam.

Thomas Hanke. 2004. HamNoSys – representing sign language data in language resources and language processing contexts. In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 1–6, Lisbon, Portugal. European Language Resources Association (ELRA).

Thomas Hanke and Jakob Storz. 2008. iLex – a database tool for integrating sign language corpus linguistics and sign language lexicography. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 64–67, Marrakech, Morocco. European Language Resources Association (ELRA).

Martje Hansen. 2012. Textlinguistik: Gebärdensprache im Kontext. In Hanna Eichmann, Martje Hansen, and Jens Heßmann, editors, *Handbuch deutsche Gebärdensprache: sprachwissenschaftliche und anwendungsbezogene Perspektiven*, number Band 50 in Internationale Arbeiten zur Gebärdensprache und Kommunikation Gehörloser, pages 199–224. Signum, Seedorf.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics: An Interdisciplinary Journal of the Language Sciences*, 36(1):161–196.

Gabrielle Hodge and Onno Crasborn. 2022. Good practices in annotation. In Trevor Johnston, Julie A. Hochgesang, and Jordan Fenlon, editors, *Signed Language Corpora*, pages 46–89. Gallaudet University Press, Washington, D. C.

Jack Hoza. 2011. The discourse and politeness functions of HEY and WELL in American Sign Language. In Cynthia B. Roy, editor, *Discourse in signed languages*, number 17 in Sociolinguistics in deaf communities series, pages 69–95. Gallaudet University Press, Washington, D.C.

Elena Jahn, Gabriele Langer, and Cornelia Loos. 2021. Are we DONE yet? a corpus-based analysis of DONE as a clause-boundary diagnostic in DGS. Poster presented at Formal and Experimental Advances in Sign Language Theory (FEAST 2021).

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Trevor Johnston. 2024. Auslan Corpus annotation guidelines. v2. Technical report, Monash University & Macquarie University.

Kim Josephine Kaul. 2024. *Palm-up Geste. Ein Vergleich zwischen Laut- und Gebärdensprachen*. Term paper, Georg-August-Universität, Göttingen, Germany.

Jörg Keller, Annette Hohenberger, and Helen Leuninger. 2003. Sign language production: slips of the hand and their repairs in German Sign Language (DGS). In *Cross-linguistic perspectives in sign language research. Selected papers from TISLR 2000*, volume 41 of *Internationale Arbeiten zur Gebärdensprache und Kommunikation Gehörloser*, pages 307–333, Amsterdam, Netherlands. Signum Verlag.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36(1):15–43.

Reiner Konrad, Thomas Hanke, Amy Isard, Marc Schulder, Lutz König, Julian Bleicken, and Oliver Böse. 2024a. Corpus à la carte – improving access to the Public DGS Corpus. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 390–399, Turin, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2024b. MY DGS – annotated. Public Corpus of German Sign Language, 4th release. Dataset, University of Hamburg.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. Public DGS Corpus: Annotation conventions. Project Note AP03-2018-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.

Maria Kopf, Marc Schulder, Thomas Hanke, and Sam Bigeard. 2022. Specification for the harmonization of sign language annotations. EASIER project deliverable D6.2, Universität Hamburg.

Loulou Kosmala. 2022. Rethinking (dis)fluency within the scope of interactional linguistics and gesture studies. *Studia Universitatis Babeș-Bolyai Philosophia*, 67(2):49–66.

Gabriele Langer, Thomas Hanke, Reiner Konrad, and Susanne König. 2016. "Non-tokens": When tokens should not count as evidence of sign use. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 137–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Alysson Lepeut and Emily Shaw. 2025. Interactive functions of palm-up: Cross-linguistic and cross-modal insights from ASL, American English, LSFB and Belgian French. *Languages*, 10(9):239.

Scott K. Liddell and Robert E. Johnson. 1989. American Sign Language: The phonological base. *Sign Language Studies*, 64(1):195–277.

Cornelia Loos and Reiner Konrad. 2022. Detailed annotation and qualifiers. Project Note AP03-2021-01, DGS-Korpus project, IDGS, Universität Hamburg, Hamburg, Germany.

Hannah Lutzenberger, Lierin De Wael, Rehana Omardeen, and Mark Dingemanse. 2024. Interactional infrastructure across modalities: A comparison of repair initiators and continuers in British Sign Language and British English. *Sign Language Studies*, 24(3):548–581.

Rachel McKee and Sophia Wallingford. 2011. 'So, well, whatever': Discourse functions of *palm-up* in New Zealand Sign Language. *Sign Language & Linguistics*, 14(2):213–247.

Johanna Mesch and Lars Wallin. 2021. Annoteringskonventioner för teckenspråkstexter. version 8. Technical report, Avdelningen för teckenspråk, Institutionen för lingvistik, Stockholms universitet, Stockholm.

Melanie Metzger and Ben Bahan. 2001. Discourse analysis. In Ceil Lucas, editor, *The Sociolinguistics of Sign Languages*, pages 112–144. Cambridge University Press, Cambridge.

Elena Georgia Mpadanes. 2020. Approaching repairs in German Sign Language conversations: an analysis of data from the DGS corpus. Master's thesis, University of Hamburg, Hamburg.

Don Newkirk, Edward S. Klima, Carlene Canady Pedersen, and Ursula Bellugi. 1980. Linguistic evidence from slips of the hand. In *Errors in linguistic performance: slips of the tongue, ear, pen, and hand (12th International Congress of Linguists, 1977; Working Group on Speech Errors)*, pages 165–198, Vienna, Austria. Academic Press.

Rie Nishio, Sung-Eun Hong, Susanne König, Reiner Konrad, Gabriele Langer, Thomas Hanke, and Christian Rathmann. 2010. Elicitation methods in the DGS (German Sign Language) Corpus project. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 178–185, Valletta, Malta. European Language Resources Association (ELRA).

Ingrid Notarrigo and Laurence Meurant. 2014. Nonmanuals as markers of (dis)fluency. In *Proceedings of the LREC2014 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, pages 135–142, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ingrid Notarrigo and Laurence Meurant. 2019. Conversations spontanées en langue des signes de Belgique francophone (LSFB) : fonctions et usages de la répétition. *Lidil*, 60.

Alejandro Oviedo, Thomas Kaul, Katharina Urbann, Reiner Griebel, and Leonid Klinner. 2018. Zur bedeutung von Pausen in der Deutschen Gebärdensprache – eine linguistische Annäherung. *DAS ZEICHEN*, 108:116–125.

Yolanda Pérez. 2006. Marcadores manuales en el discurso narrativo en la lengua de señas venezolana. *Letras*, 48(72):349–363.

Marie Piu and Rémi Bove. 2007. Annotation des disfluences dans les corpus oraux. In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. REncontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues (Posters)*, pages 397–406, Toulouse, France. ATALA.

Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and Arvid Schwarz. 2008. DGS Corpus project – development of a corpus based electronic dictionary German Sign Language / German. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 159–164, Marrakech, Morocco. European Language Resources Association (ELRA).

Patrick Louis Rohrer, Ulya Tütüncübasi, Júlia Florit-Pons, Ingrid Vilà-Giménez, Núria Esteve-Gibert, Ada Ren-Mitchell, Stefanie Shattuck-Hufnagel, and Pilar Prieto. 2025. Multidimensional labeling of gesture in communication: the M3D proposal. *Corpus Pragmatics*, 9(4):411–433.

Marie-Anne Sallandre and Brigitte Garcia. 2013. Epistemological issues in the semiological model for the annotation of sign languages. In Laurence Meurant, Aurélie Sinte, Mieke Van Herreweghe, and Myriam Vermeerbergen, editors, *Sign Language Research, Uses and Practices: Crossing Views on Theoretical and Applied Sign Language Linguistics*, pages 159–178. De Gruyter Mouton.

Marc Schulder and Thomas Hanke. 2020. OpenPose in the Public DGS Corpus. Project Note AP06-2019-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD dissertation, University of California, Berkeley.

Elizabeth Shriberg. 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

Saúl García Villameriel. 2008. La interpretación de los marcadores del discurso de la lengua de signos española. In *Actas del III Congreso de la Federación Española de Intérpretes de Lengua de Signos y Guías-Intérpretes (FILSE)*.

Saúl García Villameriel. 2010. EN-CAMBIO y ES-DECIR: origen de dos marcadores discursivos de la lengua de signos en el español oral. In *Actas del IX congreso de lingüística general*, Valladolid, Spain.

Elisabeth Volk. 2016. Discourse functions of palm-up in German Sign Language (DGS). Poster presented at the 5th Meeting of the Formal and Experimental Advances in Sign Language Theory Colloquium (FEAST 2016), Venice, Italy.

# Two-Handed Signs and Handedness: Phonological Implications for Sign Language Structure

**Lisa Lepp, Mirella De Sisto, Dimitar Shterionov**
{L.B.Lepp, M.DeSisto, D.Shterionov}@tilburguniversity.edu
Tilburg University
Warandelaan 1, 5037AB Tilburg

**Abstract**

Handedness —the use of one versus two hands in sign production— has traditionally been discussed in relation to dominance and symmetry conditions, yet it remains underrepresented in formal phonological models of sign languages. This paper argues that handedness constitutes a core phonological parameter that directly influences the structure and interaction of movement, handshape, location, and orientation. Building on hierarchical and dependency-based approaches, we propose an adapted phonological dependency model that explicitly integrates handedness in the representation of manual articulators. In one-handed signs, features are specified for a single active hand. In two-handed signs, feature distribution is constrained by symmetry and dominance conditions, which regulate whether the hands must share features or may differ in a structurally restricted way. This structural encoding accounts for variation phenomena such as weak add, weak prop, and weak drop as constrained adjustments within the phonological system. From a technical perspective, this refinement suggests more formal restrictiveness and empirical discriminability within the feature geometries, reduced representational ambiguity, and improved empirical testability across theoretical, corpus-based, and computational implementations, strengthening the interface between phonological theory and sign language technology.

**Keywords:** handedness, one- vs. two-handed signs, sign language phonology, sign language technology

## 1. Introduction

One of the most unique properties of sign languages (SLs) is the availability of two active manual articulators (Brentari, 1998; Crasborn and van der Kooij, 2013; Sandler, 2017; Van der Kooij, 2002). Lexical signs may be produced with one or two hands, and across SLs, both options are roughly equally represented in the lexicon[1] (Crasborn, 2011; McKee et al., 2024; Sandler, 2017). Despite this, the role of handedness in the phonological structure has not been included. Handedness itself is often treated as a derived aspect of lexical sign production. That is, handedness is not considered as an individual feature but rather as a side effect of the main hand.

In this paper, we take handedness as a starting point rather than as a secondary feature. From this perspective, handedness is not only about the number of involved articulators, but also about how phonological dependencies are distributed across them. With that respect, we investigate to what extent the choice between one-handed or two-handed signs formalizes the internal structure of signs. We treat handedness as the primary factor constraining the realization of other parameters. These parameters function as fundamental units of phonological structure, comparable to sublexical units in spoken languages. Thus, the contribution of this paper is to propose an adapted SL phonological hierarchy model where handedness is not a secondary property, but a structurally represented parameter.

The question whether a structural position for both manual articulators is needed in the phonological analysis of lexical signs is raised in earlier work of Van der Kooij (2002), as the involvement of one or two hands is potentially distinctive. Using variation provides evidence about the distribution and limits of phonological categories.

To analyze handedness as a phonological parameter and assess its place within established or new frameworks, we develop a hierarchical view of SL phonology. We argue that many well-known generalizations—such as dominance, symmetry, and patterns of weak-hand participation (Battison, 1978; Brentari, 1998; Crasborn and Van der Kooij, 2023; McKee et al., 2024) —are more naturally captured if handedness is represented explicitly within the phonological hierarchy of manual articulators. This addition does not require abandon-

[1] As Van der Kooij (2002) analyzed for the SignPhon database with 3084 signs, 54% of the signs were one-handed versus 46% of two-handed.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 275–286
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ing existing models; rather, it involves a small but consequential shift in perspective that makes the structural role of the second hand visible within the representation itself. That is, handedness gives a useful perspective for re-examining how phonological dependencies are organized in SLs.

This representational shift also has implications beyond theory, particularly for how structure is reflected in descriptive and corpus practices. If handedness is hierarchically encoded, annotation systems should capture not only whether one or two hands are involved, but also how phonological features are distributed across them. Not all corpora provide equal information for both hands, and some do not even annotate the weak hand (Sisto et al., 2022). We propose instead that each hand should be encoded as a distinct articulator with separate specifications for handshape, orientation, location, and movement. This would allow corpora to represent structured feature distribution across hands rather than reducing handedness to a binary distinction.

This paper is structured as follows: we start with a discussion of the evolution of phonological hierarchy models in SL linguistic theories in Section 2. Then, in Section 3, we describe the widely adopted manual phonological parameters, i.e., movement, handshape, orientation, and location. Section 4 is dedicated to motivating our choice to include the parameter *handedness* as a sixth parameter in future SL linguistics research (see Figure 4). We end with an explanation of the new adapted phonology dependence model in Section 5.

## 2. Phonological hierarchies in sign language phonology

Early work on SL phonology established the now-familiar view that lexical signs can be decomposed into a small number of discrete manual parameters. In the seminal model proposed by Stokoe (1980), handshape, location, and movement are introduced as abstract phonological units, laying the foundation for later parameter-based analyses. Subsequent work refined and expanded this inventory, adding orientation and location, while maintaining the assumption that phonological form is structured and systematic rather than holistic (Brentari, 1998; Sandler, 2014; Van der Kooij, 2002).

While early parameter models primarily treated these units as co-equal components, later approaches increasingly emphasized hierarchical organization and dependency relations among phonological features (Brentari, 1998; Crasborn, 2001; Crasborn and van der Kooij, 2013; van der Hulst and van der Kooij, 2020). In this view, not all features in language systems are equally important. Some have a more prominent role of some features in comparison with others. In a hierarchical structure, there is a head feature and a dependent feature (Anderson and Ewen, 1987), leading to the proposal and establishment of hierarchical representations. Hierarchical representations are motivated by empirical patterns, such as markedness asymmetries, as well as by the need to model both simultaneous and sequential structure within a single representational framework. Many studies (Battison, 1974, 1978; Crasborn, 2001; Fenlon et al., 2017; Van der Kooij, 2002; McKee et al., 2024; Padden and Perlmutter, 1987; Sandler, 2014, 2017) among others, showed how the phonological manual structure of a lexical item is built up using this hierarchical description. Within these models, parameters such as handshape and movement are often treated as structurally prominent, as they trigger strong constraints on the realization of other features (Fenlon et al., 2017; van der Kooij and Crasborn, 2016; van der Kooij and van der Hulst, 2005). Orientation and location, in turn, are typically analyzed as dependent or relational parameters whose interpretation is conditioned by higher-level articulatory structure (Sandler, 2013; Fenlon et al., 2017)[2]. Despite the increasing formalization of hierarchical structure, handedness has not yet been systematically[3] incorporated into these representation structures. As handedness determines the organization of articulatory dominance, symmetry, and internal manual specification,the number of hands directly constrains the distribution and realization of other phonological parameters. If hierarchical models aim to capture asymmetries, dependency relations, and structural prominence, then arguably handedness constitutes a candidate for structural encoding rather than peripheral annotation.

[2]Non-manual parameters (NMP) such as eyebrow movement, cheek tension, tongue position, and body movement are not included in the scope of this article.

[3]By "systematically," we refer to the consistent and architecture-wide encoding of hierarchical relations between articulators. While some models do not incorporate such representations at all, others assume them in principle but implement them only in selected analyses, rather than as a general representational commitment.

## 3. Phonological parameters

The phonological structure of SLs can be described in terms of a limited set of manual parameters that jointly define lexical contrast, i.e., a small difference in one of the manual parameters can distinguish one sign from another[4]. Minimal pairs illustrate this contrast: they are pairs of signs that are identical in all parameters, except for one. The *Nederlandse Gebarentaal* (NGT), i.e., the sign language of the Netherlands, signs for DONDER (THUNDER) in Figure 1a, and BOOM (TREE) in Figure 1b illustrate this concept[5].



(a) An example of the one-handed sign THUNDER in NGT.



(b) An example of the two-handed sign TREE in NGT.

Figure 1: An example of a minimal pair in *NGT* illustrating contrast in handedness.

We first provide a brief description of the core parameters[6] (Anderson and Ewen, 1987; Battison, 1974, 1978; Crasborn, 2011; Crasborn et al., 2000; Crasborn and van der Kooij, 2013; Fenlon et al., 2017; Sandler, 1986, 2013, 2014; Stokoe, 1980; van der Kooij and Crasborn, 2016), and then discuss the new proposal.

[4]We acknowledge that Brentari (1998); Sandler (1996); Sandler and Lillo-Martin (2006) refer to parameters as classes and class nodes, but we have decided to follow the more general definition of parameters.

[5]Figures 1, 2, 3, and 5 are reproduced from the website of the Dutch Sign Language Centre (https://www.gebarencentrum.nl/)

[6]As noted by a reviewer, this paper does not discuss individual phonological features in detail; rather, it focuses on the classes of parameters and their specifications. The phonological features are mentioned implicitly throughout this article.

### 3.1. Movement

Movement is a defining property of SL-syllables[7]. In SLs, this organization can be expressed by the movement of the hand, or by the movement of only the fingers (Van der Kooij and Crasborn, 2008), and is phonologically obligatory (Brentari, 1998; Crasborn et al., 2000; Van der Kooij, 2002; Liddell and Johnson, 1989; Sandler, 2013, 2014). A distinction is made between *path movement* and *local movement*. Path movement (PM) involves a transition between two explicitly specified positions. Local movements are further categorized in hand-internal movement and orientation change (Van der Kooij, 2002; Sandler, 2014), with hand-internal movements referring specifically to the movement of the fingers and the wrist. Local movements therefore refer to lexical changes in finger position or changes in the orientation of the hand (Van der Kooij, 2002; van der Kooij and Crasborn, 2016; Sandler, 2013). A single syllable may contain at most one instance of each movement type, allowing simultaneous but constrained combinations (Crasborn and van der Kooij, 2013; van der Hulst and van der Kooij, 2020; van der Hulst, 1993; van der Kooij and Crasborn, 2016).

Distinguishing path and local movement is crucial for modeling the temporal structure of signs. By pairing two values of a setting (for example, [high] vs. [low], [contralateral] vs. [ipsilateral], or [proximal] vs. [distal]), the movement path can be precisely defined within a given location domain, without positing movement as an independent phonological primitive (van der Hulst and van der Kooij, 2020). In corpus annotation, for example, using ELAN, movement is typically encoded either on a single tier or through a set of defined labels, making explicit whether a sign involves path movement, local movement, or both. This distinction clarifies which movement features are obligatory or optional and how they interact with other parameters.

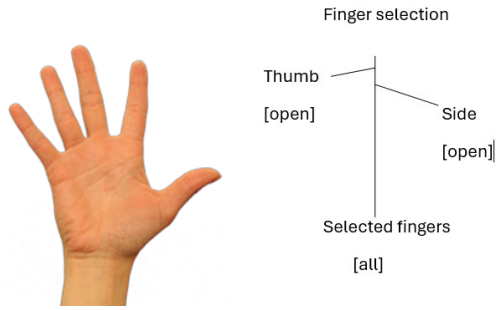
### 3.2. Handshape

Handshape is described as the form of the hands during the articulation of a sign, determined by the selection of fingers and their internal configuration (Brentari, 2011; Crasborn et al., 2000; Van der Kooij, 2002; Sandler, 2014). Finger se-

[7]Following Brentari (1998), a syllable in sign language is a prosodic unit organized around a movement nucleus, optionally preceded and/or followed by holds.

lection[8] distinguishes between foregrounded and back grounded fingers (Van der Kooij, 2002), while finger configuration specifies properties such as curve, width, aperture (Van der Kooij, 2002)[9]. The curve is the specification of the flexion of the fingers. Aperture specifies the open-close relationship between the thumb and the selected fingers, and width refers to the degree of spreading or opposition between them.

Simple handshapes, also called unmarked handshapes, involve minimal specification—for example, selecting all fingers in an open configuration, as shown in 2a, and in the phonological dependency hierarchy in Figure 2b.



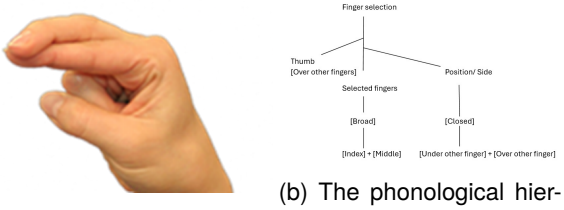

(a) Example of [all] selected fingers, thus a simple handshape.

(b) The phonological hierarchy model based on the 5-handshape in Figure 2a.

In contrast, complex handshapes, or marked handshapes[10], require multiple dependent features, as shown in Figure 3a and the related phonological hierarchy model in Figure 3b.

### 3.3. Orientation

Orientation encodes the facing of the hand palm with respect to space, the body, or another articulator. Previous research has approached the encoding of orientation from three complementary perspectives. The first approach, referred to as handshape-linked-orientation, defines the orientation via the direction of the palm of the handshape (Brentari, 2011; Crasborn and van der Kooij, 2013; Fenlon et al., 2017; Van der Kooij, 2002). In contrast, the movement-linked-orientation encodes which part of the hand faces the final location (Van der Kooij, 2002) and remains constant throughout the articulation of the sign. van der Hulst and van der Kooij (2020) suggests four types for the movement-linked-orientation: (i) Body – end contact; (ii) Weak hand – end contact; (iii) Space – path movement, and (iv) Body – continuous contact.




(a) Example of the R-handshape, thus a complex handshape.

(b) The phonological hierarchy model for the complex R-handshape in Figure 3a.

A third approach to orientation adopts a contact-based perspective. Fenlon et al. (2017) defines orientation as the relationship between the active hand and the place of articulation, specifying which side of the hand—most notably the radial (thumb) side—is involved. Orientation is often reduced to features such as [palm up] or [palm down], and defined relative to body, space, or another articulator.

### 3.4. Location

Location is the area of articulation (e.g., head, torso, arms, neutral space)[11]. It can be further divided into sub-locations that are contrastive (Van der Kooij, 2002). An empirical argument in favor of the distinctive locations is the *generalized double contact constraint*. This constraint indicates that if a sign involves two distinct contacts with the body, those contact points are restricted in their location (Van der Kooij, 2002). Evidence from NGT shows that this constraint is not limited to double contact movements, but also applies to continuous contact movements, which likewise remain within a single distinctive location (Van der Kooij, 2002). These constraints are part of a broader set of phonological rules, such as the symmetry and

[8] In Crasborn et al. (2000); Van der Kooij (2002); Sandler (2014), handshapes with either [one] or [all] features are discussed. However, in cases with [two], [three], or [four] selected fingers, the elements associated with [one] or [all] can combine and form dependency relations.

[9] Brentari (2011) describes fully open, bent (closed), flat-open, flat-closed, curved open, curved closed, and fully closed as seven options for joint configuration in ASL-handshapes.

[10] Markedness correlates with complexity, articulatory difficulty, frequency, order of acquisition, and other factors in the formalization of a handshape (van der Hulst and van der Kooij, 2020; Lepp, 2025; Sandler, 2014).

[11] Van der Kooij (2002) also calls Space and weak hand as areas.

dominance conditions, which will be discussed in the next section. Furthermore, signs may share the same phonological location while differing in the specific part of the hand that establishes contact, indicating that phonological location cannot be reduced to a single point in space (Mauk and Tyrone, 2012).

Location is distinguished from setting: location refers to the major articulatory area, while setting specifies finer-grained positional or movement-related distinctions within that area (e.g., high/low, ipsilateral/contralateral, proximal/distal, contact, and others) (Crasborn et al., 2000; van der Hulst and van der Kooij, 2020; Sandler, 1986) Some of the linked features are [radial], [back], [tips], [palmar], and [ulnar] (van der Kooij and Crasborn, 2016).

## 4. Handedness

As noted above, a defining characteristic of SLs is the availability of two manual articulators (van der Kooij and Crasborn, 2016; Crasborn, 2011; Sandler, 2017) which leads to the notion of handedness. Van der Kooij (2002) hypothesized that handedness may be lexically and phonologically specified and that one-handed and two-handed signs may exhibit distinct structural features. Following her idea, in this section, we introduce *handedness* as a new parameter in the existing model of phonological features in SLs. Building on these observations, we formally integrate handedness into the phonological hierarchy.

### 4.1. Handedness in existing models

A long-standing question in sign language phonology concerns how the two manual articulators should be represented in the phonological structure. Van der Kooij (2002) explicitly raised the issue of whether handedness should receive a structural position in the hierarchy, noting that earlier hierarchical models (Brentari, 1990; van der Hulst, 1996; Sandler, 1986) risk over-generalization if both hands are fully specified at the lexical level. Subsequent work, including (Padden and Perlmutter, 1987) and more recently (Crasborn and Van der Kooij, 2023), further examined the role of the non-dominant hand in discourse and questioned whether two independent structural positions are always required, or whether a single underlying representation might suffice in some cases. These authors do not treat handedness as a fully formalized phonological parameter. Rather, they identify a representational tension: some two-handed configurations appear lexically stored, while others arise through morphological or post-lexical processes. To address this, Crasborn and Van der Kooij (2023) propose a three-level architecture consisting of (i) a lexical level, (ii) a surface (spelled-out) level, and (iii) a post-lexical level. According to their view, the lexical representation alone cannot account for phenomena that span multiple lexical items or depend on signer-specific dominance[12]. An intermediate surface level is therefore required to determine which hand is dominant and to specify the distribution of features such as finger selection, orientation, and position of the non-dominant hand. Only at the post-lexical level are the left and right hands fully specified separately, allowing the model to capture processes in which one hand, or specific features, persists longer than the other.

This discussion directly motivates our proposal. The need for multiple representational levels and for mechanisms that determine dominance and feature distribution across articulators indicates that handedness is not a marginal or purely phonetic property. Previous accounts treat handedness as an architectural issue rather than a formal parameter; we integrate it directly into the phonological hierarchy. In doing so, we maintain the need for multiple representational levels, avoid lexical overgeneralization, and capture the distribution of features across articulators within the core phonological system.

### 4.2. Handedness and asymmetry

A functional asymmetry is observed between the dominant (strong) hand and the non-dominant (weak) hand (Crasborn, 2011; Padden and Perlmutter, 1987). The dominant hand typically serves as the active articulator, while the non-dominant hand functions as a passive articulator or a place of articulation (Brentari, 2011; Crasborn and Van der Kooij, 2023; Van der Kooij, 2002; McKee et al., 2024).

These asymmetrical signs are subject to two well-established phonological constraints: the symmetry constraint and the dominance constraint (Battison, 1978; van der Kooij and Crasborn, 2016; Sandler, 2013). The symmetry constraint requires

[12]Signers may be right- or left-hand dominant; thus, "dominant" is relative to the individual signer (McKee et al., 2024).

that if both hands are active, they share handshape, movement, and location. The dominance constraint applies when one hand is passive: if the non-dominant hand serves as a place of articulation, it must either have an unmarked handshape or share the handshape of the dominant hand (Van der Kooij, 2002). These constraints suggest that handedness concerns not only the number of articulators, but also the distribution of phonological dependencies across them. Such restrictions are difficult to explain if handedness is treated only as a surface realization choice; instead, Crasborn and Van der Kooij (2023) points to a structurally constrained role for the second hand within the phonological system.

### 4.3. Handedness configurations

Because handedness varies in constrained ways, it must be represented as a phonological parameter. This interpretation is further supported by the recent work of McKee et al. (2024), who show that handedness is systematically manipulated in phonetic and discourse-driven contexts[13]. As they show for New Zealand SL, signs may undergo several specifications, such as:

1. Weak add: A weak hand is added to a one-handed sign.
2. Weak prop: The non-dominant hand partially or fully copies the handshape and movement of the dominant hand, consistent with the dominance constraint (Frishberg, 1975)
3. Weak drop: The weak hand of a lexically two-handed sign is omitted.
4. Doubling: A one-handed sign becomes a symmetrical two-handed sign.
5. Echo articulation and mirroring: The weak hand participates incompletely or at a reduced level (Sandler, 1993; McKee et al., 2024).

These manipulations of the non-dominant hand respect phonological constraints such as the dominance constraint and the symmetry constraint, showing that handedness interacts systematically with the phonological system. This supports the claim that handedness is a manipulable phonological parameter rather than a fixed lexical property.

[13]The focus of their research analyzed specific communicative functions that are socially conventionalized and reflected in grammar and lexicon over time (McKee et al., 2024).

## 5. Adapted phonological dependency model

The previous sections outlined the manual phonological parameters and their interaction with handedness. While previous work addressed only specific aspects—for instance, Sandler (2014) focused on hand configuration, and Van der Kooij (2002) did not incorporate the newer parameter handedness—a comprehensive account remained lacking. Building on these hierarchical approaches (Van der Kooij, 2002; Sandler, 2014, 2017), the present section formalizes the interactions among manual articulators within a phonological dependency model, incorporating hierarchical organization and head-dependent relations (Battison, 1974, 1978; Fenlon et al., 2017; Crasborn, 2001; McKee et al., 2024; Padden and Perlmutter, 1987). Morgan (2022) discusses the theoretical distinction between dominant and non-dominant hand features (often expressed as [1H] and [2H]) within phonological analyses of Kenyan Sign Language. However, this work does not present a fully implemented dependency model that incorporates these features. The present paper develops exactly such a formalization.

### 5.1. Model description

The present model extends earlier work in two main ways. First, it introduces handedness as an explicit structural node within the hierarchy–shown in green letters in Figure 4. Second, it formalizes the asymmetric distribution of phonological specifications across the two hands–shown via the dashed lines. In contrast to previous accounts[14], handedness is not treated as an implicit or peripheral factor, but is integrated directly under the manual articulator node. This constitutes a key innovation of the model. Because the hands are asymmetrical, the non-dominant hand may carry independent specifications for movement, handshape, orientation, and location. Although the model does not explicitly encode temporal structure, it is not intended to represent the sequential organization of signs, which is instead assumed to emerge at the level of articulatory realization.

For the research community, this provides a concrete framework for analyzing cross-linguistic differences in one- and two-handed signs, un-

[14]The extent to which the weak hand is phonologically independent has been widely debated in Brentari (1993).

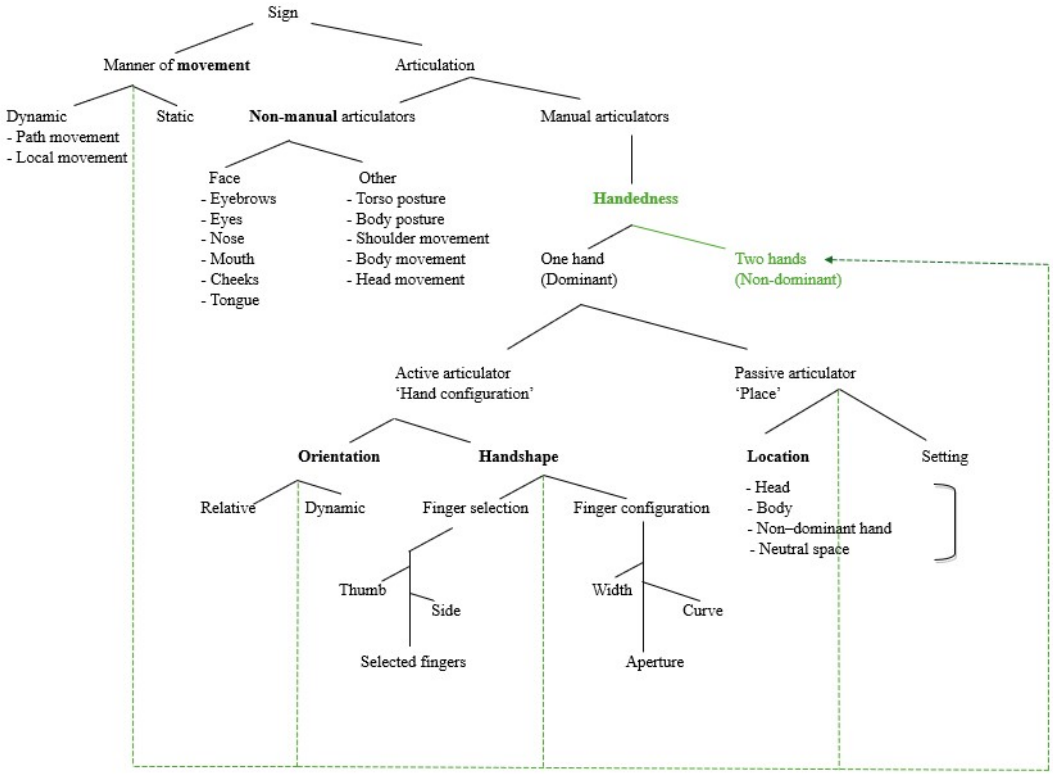



Figure 4: Adapted phonological hierarchy dependency model

derstanding constraint interactions, and exploring the interface between phonology and articulation in SLs. It also offers a practical tool for formal modeling, corpus annotation, and computational applications where handedness must be represented explicitly. While the Global Sign Bank (https://signbank.cls.ru.nl/) already distinguishes between one- and two-handed signs -and the handshapes-, the proposed model refines this distinction further by encoding the dependency relations between the hands. In corpus annotation terms, for example in ELAN (2025), an additional tier could be introduced to encode whether the non-dominant hand carries independent specifications, and, if so, which phonological parameters are specified independently.

## 5.2. Model interpretation

The adapted phonological hierarchy dependency model builds on earlier hierarchical accounts of sign language phonology, in which manual and non-manual components are organized across tiers. In the present model, this structure is retained but extended with a bidirectional interpretation. Specifically, the model can be interpreted in two directions: (1) bottom-up, or (2) top-down. Each of these directions corresponds to two modes of processing: the first reflects the production phase, whereas the second represents the comprehension phase. From the production perspective, during the production of a lexical item, the movement patterns and articulators, together with their associated subcomponents, serve to formalize the sign's linguistic content. Viewed from the comprehension perspective, sign recognition begins with the perception of finger selection and handshape configuration, which provide the earliest and most salient cues for lexical access (Brentari, 1998; Emmorey and Corina, 1990; Mauk and Tyrone, 2012). In contrast to previous accounts, however, the present model explicitly formalizes how these early cues are mapped onto the hierarchical structure. These cues inform the reconstruction of the hand configuration, which in turn supports the interpretation of the manual-articulator tier (with non-manuals as its sister tier), the movement tier as a parallel branch, and, ultimately, the sign as a whole.

To illustrate how these parameters are captured within a unified single phonological hierarchy dependency model, we analyze the one-handed sign THUNDER (Figure 1a), shown in Appendix A, and

the two-handed sign TREE (Figure 1b), shown in Appendix B. For each sign, two representations are provided: an original hierarchy (Figure 5 and 7), and a revised hierarchy (Figure 6 and 8)[15].

THUNDER (Figures 5 and 6) is characterized by a zigzag, downwards (path) movement. It is produced with a single articulator in neutral space, starting from a high position and moving downwards (indicated by '–>'). The handshape consists of a closed palm oriented toward neutral space, with the thumb bent across the fingers and one selected finger fully extended. As the second hand does not give any information in this sign, the notation in the new hierarchy contains as features 'X', e.g., no information available.

TREE (Figures 7 and 8), on the other hand, has two active articulators. This sign is produced with symmetrical parameters, with the same underlying representations, and thus both hands have the same orientation, handshape, location, and orientation. Both hands produce a zigzag, downward movement.

A comparison of THUNDER and TREE highlights two key differences between the original and revised hierarchies. First, while the original model does not systematically encode the distribution of features across articulators, the revised model explicitly distinguishes between active and non-active articulators. This is reflected in the extra specification of the non-active hand in THUNDER versus the full specification of both hands in TREE. Second, the revised hierarchy captures the parallelism in symmetrical two-handed signs by representing shared feature structures across articulators, rather than duplicating independent specifications.

Taken together, these differences demonstrate that the revised model more transparently encodes handedness and articulator activity within a single hierarchical structure, while maintaining a consistent representation of shared phonological parameters such as movement. This supports the claim that handedness is not an external annotation, but an integral part of the phonological representation.

[15]One of the reviewers requested two fully worked examples of signs represented under both an existing model and the adapted model as we suggested to make the structural role of handedness more concrete. This was actually a very good suggestion, which we thus adopted.

## 6. Conclusion

This paper proposes a new hierarchical model of SL phonology. We have argued that handedness should be represented as a structurally encoded phonological parameter within a proposed hierarchical model of sign language structure. By integrating handedness directly under the manual articulator node, the model makes explicit how phonological features are distributed across articulators and how variation shows where the boundaries within the system are. Our next step is to validate this approach with automated tools and with linguists.

Encoding these distinctions in annotation also brings technical benefits. Hierarchical articulator structure enables more precise corpus searches, improves dataset interoperability, and supports computational modeling. Distinguishing lexical weak-hand features from gradient or discourse-driven variation helps automated systems separate stable contrasts from performance effects, strengthening reproducibility and integration with technological applications.

## Bibliographical References


John M. Anderson and Colin J. Ewen. 1987. *Principles of Dependency Phonology*. Cambridge University Press.

Robbin Battison. 1974. Phonological deletion in American Sign Language. *Sign Language Studies*, 5:5–19.

Robbin Battison. 1978. *Lexical Borrowing in American Sign Language*. Linstok Press, Silver Spring, MD, USA. Discusses lexical borrowing and phonological constraints in ASL.

Diane Brentari. 1990. *Theoretical Foundations of American Sign Language Phonology*. Ph.D. thesis, University of Chicago, Chicago.

Diane Brentari. 1993. *Current Issues in ASL Phonology*. Academic Press, San Diego.

Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. Cambridge University Press, Cambridge, UK.

Diane Brentari. 2011. Handshape in Sign Language Phonology. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice,

editors, *The Blackwell Companion to Phonology*, pages 195–222. Wiley-Blackwell, Malden, MA.

Onno Crasborn. 2011. The other hand in sign language phonology. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice, editors, *The Blackwell Companion to Phonology*, pages 223–240. Wiley-Blackwell, Malden, MA.

Onno Crasborn, Harry van der Hulst, and Els van de Kooij. 2000. Phonetic and phonological distinctions in sign languages. Intersign Introduction Chapter (draft).

Onno Crasborn and Els van der Kooij. 2013. The phonology of focus in Sign Language of the Netherlands. *Journal of Linguistics*, 49(3):515–565. Published online 11 April 2013.

Onno Crasborn and Els Van der Kooij. 2023. *The Emergence of the Second Hand in Sign Language Phonology: From Underlying to Surface Representations*, pages 319–344. Cambridge University Press, Cambridge.

Onno A. Crasborn. 2001. *Phonetic Implementation of Phonological Categories in Sign Language of the Netherlands*. Phd thesis, Universiteit Leiden, Leiden, The Netherlands. Promotoren: Colin Ewen & Harry van der Hulst.

ELAN. 2025. Computer software version 7.0. Technical report, Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands.

Karen Emmorey and David Corina. 1990. Lexical recognition in sign language: Effects of phonetic structure and morphology. *Perceptual and Motor Skills*, 71(3 Suppl.):1227–1252.

jordan Fenlon, Diane Brentari, and Kearsy Cormier. 2017. *The Phonology of Sign Languages*.

Nancy Frishberg. 1975. Arbitrariness and iconicity: Historical change in American Sign Language. *Language*, 51(3):696–719.

Lisa Lepp. 2025. *Lexical Acquisition of Sign Language of the Netherlands by Deaf Refugee Children: A Study of Phonological Complexity*, pages 315–344.

Scott K. Liddell and Robert E. Johnson. 1989. American Sign Language: The phonological base. *Sign Language Studies*, 64:195–278.

E. Mauk and Martha E. Tyrone. 2012. Location in ASL: Insights from phonetic variation. *Sign Language & Linguistics*, 15(1):128–146.

Rachel McKee, Mireille Vale, George Major, Sara Pivac Alexander, and Miriam Meyerhoff. 2024. Two hands are powerful. Handedness variation and genre in New Zealand Sign Language. *Language & Communication*, 100.

Hope E. Morgan. 2022. *A Phonological Grammar of Kenyan Sign Language*. De Gruyter Mouton, Berlin/Boston.

Carol Padden and David M. Perlmutter. 1987. American Sign Language and the architecture of phonological theory. *Natural Language & Linguistic Theory*, 5:335–375.

Wendy Sandler. 1986. The spreading hand autosegment of American Sign Language. *Sign Language Studies*, (50):1–28.

Wendy Sandler. 1993. Hand in hand: The roles of the nondominant hand in sign language phonology. *The Linguistic Review*, 10(4):337–390.

Wendy Sandler. 1996. The prosodic model of sign language phonology. *Sign Language & Linguistics*, 1(1):1–36.

Wendy Sandler. 2013. The phonological organization of sign languages. *Linguistics and Language Compass*, 6(3):162–182.

Wendy Sandler. 2014. The emergence of phonetic and phonological features in sign language. *Nordlyd*, 41(2):183–212.

Wendy Sandler. 2017. The challenge of sign language phonology. *Annual Review of Linguistics*, 3:43–63.

Wendy Sandler and Diane Lillo-Martin. 2006. *Sign Language and Linguistic Universals*. Cambridge University Press.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shteriono, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.

William C. Stokoe. 1980. Sign language structure. *Annual Review of Anthropology*, 9:365–390.

Harry van der Hulst. 1993. Units in the analysis of signs. *Phonology*, 10(2):209–241.

Harry van der Hulst. 1996. On the other hand. *Lingua*, 98:121–143.

Harry van der Hulst and Els van der Kooij. 2020. Sign language phonology – theoretical perspectives. In Josep Quer, Roland Pfau, and Annika Herrmann, editors, *The Routledge Handbook of Theoretical and Experimental Sign Language Research*, pages 1–25. Routledge, London / New York.

Els Van der Kooij. 2002. *Phonological Categories in Sign Language of the Netherlands: The Role of Phonetic Implementation and Iconicity*. Ph.D. thesis, Leiden University / LOT Netherlands Graduate School of Linguistics, Utrecht, The Netherlands. PhD thesis on phonological representation and the role of iconicity in NGT sign phonology.

Els Van der Kooij and Onno Crasborn. 2008. Syllables and the word-prosodic system in Sign Language of the Netherlands. *Lingua*, 118(9):1307–1327.

Els van der Kooij and Onno Crasborn. 2016. Phonology.

Els van der Kooij and Harry van der Hulst. 2005. On the internal and external organization of sign language segments: Some modality-specific properties. In Marc Van Oostendorp and Jeroen van de Weijer, editors, *The Internal Organization of Phonological Segments*, volume 89 of *Studies in Generative Grammar*, pages 153–180. Mouton de Gruyter, Berlin / New York.

## A. The phonological hierarchy dependency models for THUNDER

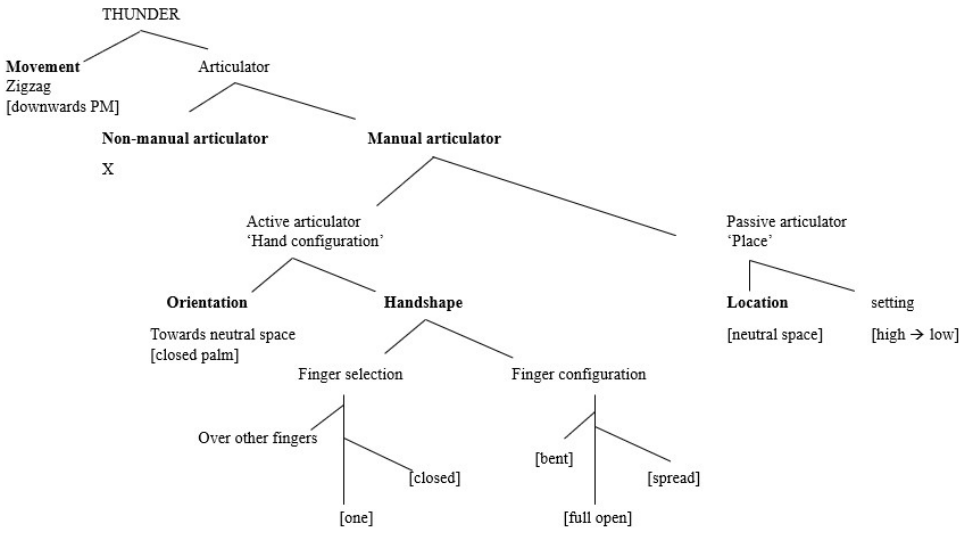



Figure 5: The old phonological hierarchy dependency model for the one-handed sign THUNDER

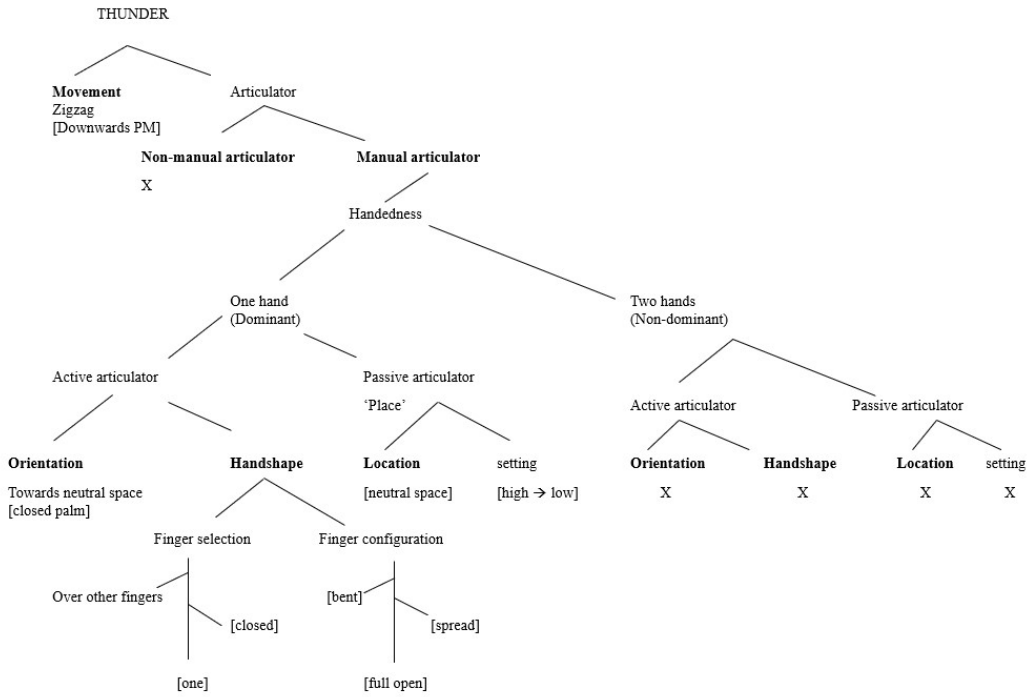



Figure 6: The new phonological hierarchy dependency model for the one-handed sign THUNDER

## B. The phonological hierarchy dependency models for TREE

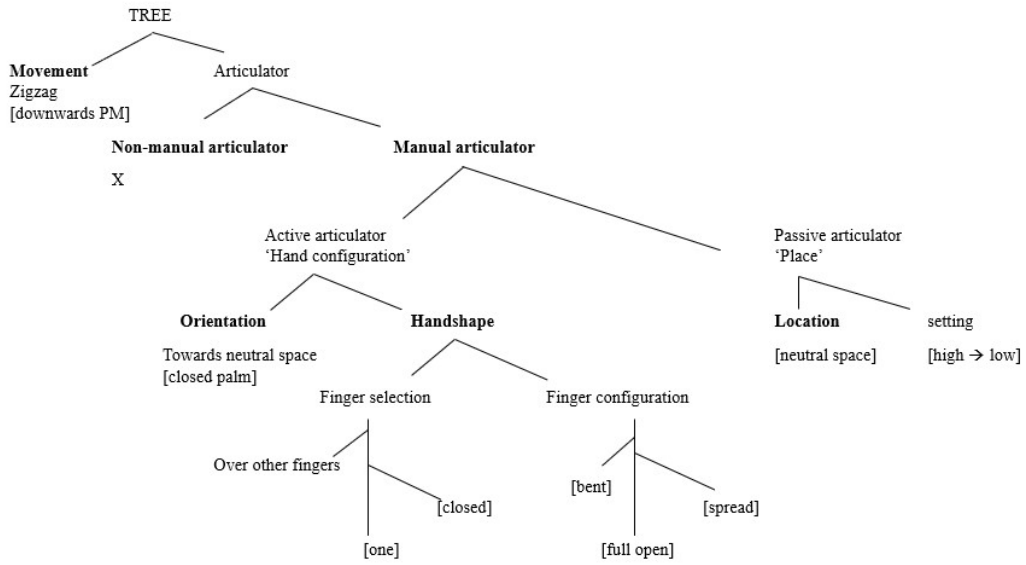



Figure 7: The old phonological hierarchy model for the two-handed sign TREE

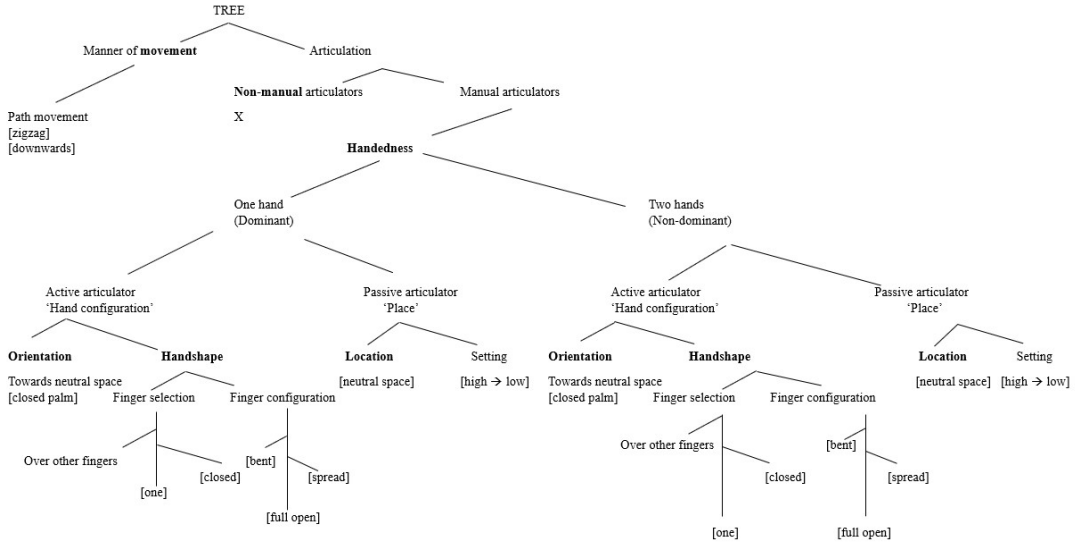



Figure 8: The new phonological hierarchy model for the two-handed sign TREE

# HNS2CF: A Mapping Tool from HamNoSys to SL CatForm

**Lisa Loy, Hope E. Morgan**
Universität Hamburg, Universität Hamburg
lisa.loy@uni-hamburg.de, hope.morgan@uni-hamburg.de

**Abstract**

Over the past six decades, a variety of systems have been developed for representing sign language forms, from Stokoe Notation (Stokoe, 1960) to SignWriting (Sutton, 1999) and lexical database schemas. Each was designed with specific goals and applications, leading to a fragmented landscape of representations. To enable greater interoperability and data sharing among sign language users and researchers, we propose a robust approach to translating between notation systems. As a first step in this direction, we introduce a formal mapping framework between HamNoSys and the SL CatForm coding schema, describe its implementation, and present empirical evidence of its performance. An extensive evaluation of mapping mismatches revealed improvements to the mapping logic needed to further advance the HNS2CF mapping tool. However, the initial version of the system already achieves an overall accuracy of 76.7% and an in-depth analysis reveals that many apparent mismatches stem from annotator disagreement rather than mapping errors, indicating that the tool's actual accuracy is even higher. These results demonstrate the feasibility and promise of establishing mapping mechanisms across sign representation systems.

**Keywords:** phonological notation, HamNoSys, coding translation

## 1. Introduction

Many different ways of representing sign forms have been designed over the past 60+ years, from Stokoe Notation (Stokoe, 1960) to SignWriting (Sutton, 1999), to fields in lexical databases and much in between. Each of these systems was developed with certain goals and uses in mind. In order to promote flexibility for language users and researchers alike to fit a notation system to any given purpose, we argue there should be a robust effort to translate between systems so that data is not isolated. Indeed, this was the goal of the SignTyp conference and project over 15 years ago (Channon and van der Hulst, 2008), which sought to create a master dataset that creates a mapping between various notation systems, especially those used in research. Unfortunately, the momentum from that conference was lost and its goal was not followed up. In this paper, we pick up this thread again and introduce the first version of HNS2CF, an automated tool, mapping from the Hamburg Notation System (HamNoSys; Hanke, 2004) to the new SL CatForm coding schema (Morgan, 2026) to move towards interoperability in sign language corpora and resources.

This paper has three parts and main contributions: first, our HNS2CF mapping tool, including the formal mapping rules/relations and an implementation thereof; second, the Python code used to build the mapping tool[1]; and third, an empirical evaluation of the system's accuracy on German Sign Language (Deutsche Gebärdensprache, DGS) data.

[1]The code repository is vailable at https://codeberg.org/lisaloy/HNS2CF.git

## 2. Background

### 2.1. Hamburg Notation System

The Hamburg Notation System (HamNoSys) is a phonetic notation system designed to transcribe signs in a language-independent way. The notation system was originally developed in the 1980s and has undergone several improvements and extensions (see Prillwitz et al., 1987 for version 1 and Prillwitz et al., 1989 for version 2). In this paper, we use HamNoSys version 3 (Hanke, 2004). This notation system encodes signs as linear strings of iconic symbols that represent the manual part of a sign (handshape, orientation, location, movement) and optionally non-manual features (facial expression, mouth, gaze, etc.). The system was developed to be used in academic research rather than as a writing system for everyday use by Deaf communities. The HamNoSys alphabet encompasses roughly 200–210 characters (see Figure 1 for some examples). These symbols can be categorised into the following main groups: **handshape** (hand configuration, plus modifiers for finger selection, bending, and shape details), **orientation** (direction of palm and finger facing), **location** (where the sign is articulated), **movement** (path movements, in-place movements, repetitions, size, and manner), and **symmetry** markers (indicating how the hands relate to each other in two-handed signs). HamNoSys follows a specific syntax and is inherently combinatorial, mean-



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 287–296
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ing individual parameters of a sign are usually not encoded in only one character but rather a combination of multiple characters and diacritics. A basic handshape, for instance, is annotated using a main handshape character, combined with a character for finger specifications and relations and a character for palm orientation, with the ability to specify additional properties (e.g., bending, thumb position, etc.). An example of a HamNoSys notation for the sign PAIN3 is provided in Figure 1[2].

HamNoSys has been used extensively in areas such as: corpus annotation (e.g., Konrad et al., 2020), language documentation (e.g., Schmaling, 2000), sign language recognition (e.g., Koller et al., 2016), and sign animation/avatar technology (e.g., Elliott et al., 2004).

### 2.2. SL CatForm Coding Schema

SL CatForm (Morgan, 2026) is a more recent phonological coding schema, designed to represent minimal units of categorical form in sign languages for research purposes, particularly for computational analysis of large datasets. It derives from database fields for coding a range of phonological characteristics. SL CatForm builds on the descriptive analysis of one specific sign language, Kenyan Sign Language (KSL; as documented in Morgan, 2022). Version 1.0 includes only values from KSL (Morgan, 2026), but can be modified to include additional language-specific values. In fact, the same basic values have been used to encode phonological data of Israeli Sign Language in ISL-LEX (Morgan et al., 2022), and only a few new values were required for handshapes and path shapes. Thus, the basic categories (variables) are likely applicable in other sign languages.

The schema contains 42 variables representing aspects of sign form, with each variable consisting of values in single or double alphanumeric characters (see Figure 1 for an illustration). These variables are placed in a fixed order in the SL CatForm string for use in computational scripts; and they are grouped into six categories: **articulator** (variables *#0-6; overall sign type, handedness, and balance*), **handshape** (variables *#7-19*; whole handshape as well as individual features), **orientation** (variables *#20-21*; absolute palm and finger direction), **location** (variables *#22-25*; where the sign is articulated and lateral symmetry), **core movement** (variables *#26-34*; overall movement types and axes/shape of path movements), and **manner of movement** (variables *#35-42*; how movements are distributed in syllables; e.g., by repetition or a change in dominance). An example of a sign coded using the SL CatForm coding schema is given in Figure 1.

Compared to HamNoSys, the variables in the schema are fixed and independent, therefore the system does not employ rules for sequential ordering; i.e., a syntax. Yet, each variable requires a value, unlike HamNoSys which has the option of leaving out irrelevant or unspecified information. Also unlike HamNoSys, SL CatForm is not designed for ease of human readability since forms are represented by the fewest possible characters. It was also created to represent phonological types that are explicitly categorical — e.g., they occur in minimal pairs — rather than gradient phonetic phenomena. Relatedly, SL CatForm was developed in alignment with a particular model of sign language structure, described in Morgan (2022).[3] It also encodes information about syllable structure, which is not overtly encoded into HamNoSys.

SL CatForm has only recently been made publicly available, so it is not cited in previous literature; however a current study on phonological distance uses the SL CatForm string structure (Morgan et al., 2026).

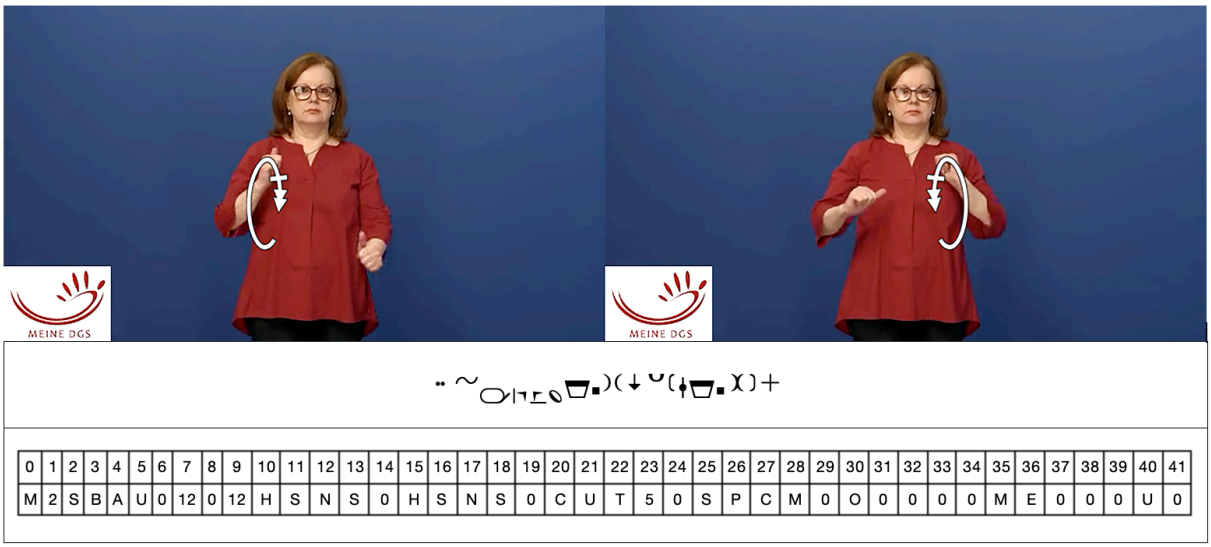
## 3. Mapping Specifications

The goal of the HNS2CF mapping system is to convert HamNoSys strings (source domain) to SL CatForm strings (target domain). To achieve this, we started by working from the documentation in the current version of HamNoSys (Hanke, 2004) and the documentation for SL CatForm (Morgan, 2026), looking for overlaps in the information encoded in each system. For some phonological information, the mapping was straightforward; for example, path shape, the number of hands, or the finger/palm orientation. For these phonological aspects, the mapping system simply matches certain characters in the HamNoSys string (e.g., movement characters for path shape, symmetry markers for number of hands, or finger extension characters for absolute finger orientation) to the corresponding value in the SL CatForm system.

Other phonological information is only explicitly encoded in one system or the other. For instance, SL CatForm encodes lexical mouthing in variable *#6*, whereas there is no information on lexical mouthing encoded in HamNoSys. Due to its combinatorial nature, HamNoSys, in turn, offers a wider variety for possible handshapes, resulting in handshapes which do not (yet) exist in SL CatForm.

[2]All images used as examples of signs in this paper were taken from the Public DGS Corpus release 3 (Konrad et al., 2020).

[3]This model is based on the Dependency Model (van der Kooij, 2002), but incorporates observations about syllable structure from the Prosodic Model (Brentari, 1989).



| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M | 2 | S | B | A | U | 0 | 12 | 0 | 12 | H | S | N | S | 0 | H | S | N | S | 0 | C | U | T | 5 | 0 | S | P | C | M | 0 | O | 0 | 0 | 0 | 0 | M | E | 0 | 0 | 0 | U | 0 |

Figure 1: The DGS sign PAIN3 (top) with the corresponding HamNoSys annotation (middle) and coding in SL CatForm (bottom)

Finally, there is some phonological information that is only implicitly encoded in one system or the other. For example, for path direction, HamNoSys uses its many arrow-based characters to unambiguously represent specific path directions, including diagonals. In contrast, SL CatForm uses a combination of three variables (*#28–30*) to encode path axis and directionality, but with less precision, and diagonals that move along three axes at once are not fully described. On the other hand, SL CatForm explicitly encodes the number of syllables of a sign, whereas this information is only implicitly encoded in the HamNoSys string. For such cases, the mapping required more complex logic; this is discussed in Section 3.1

After a comprehensive comparison of the two notation/coding systems and their inventories, we decided that two variables in the SL CatForm coding schema had to be excluded because the information needed to code these variables is not in HamNoSys: *#6 lexical mouthing* and *#42 secondary movement*. Specifically, the latter refers to how changes in the joints or palm orientation are distributed across syllables; e.g., the difference between finger flexion timed to the start and end of a path movement versus flexion repeating multiple times on a path. However, for full transparency, we keep these variables in their original position in the SL CatForm string, using the placeholder ”xx”, and consider them in the evaluation.

### 3.1. Mapping Relations

In order to transform these observations about how information is encoded in each system into a translational mapping in Python code, a classification of mappings was created. Table 1 shows a summary of four overall types of mapping relations that were used to extract information for each SL CatForm variable from the HamNoSys string. The most common mapping relation (25/40 variables in Table 1) from HamNoSys to SL CatForm was `n:1`, meaning multiple characters in the HamNoSys string map to a single value within a variable in the SL CatForm string. Eighteen variables have the most straightforward `1:1` mapping; i.e., direct matches from one HamNoSys character to one value in SL CatForm. Additionally, we developed two more mapping relations, `combination` and `pattern`. The `combination` mapping relation describes instances where the matching value for a variable could not be extracted from a single character in the HamNoSys string but from a fixed combination of individual HamNoSys characters. Similarly, the mapping relation `pattern` looks for combinations of HamNoSys characters; however, these are not fixed characters but can be any character from a set of characters. Once a `pattern` is found, this triggers one of two scenarios: a) set a specific value for a specific variable, or b) set a slicing point and split the HamNoSys string into two (or more) parts, then look only in one of the parts for matches. We present one example per mapping relation for the sign PAIN3 (presented in Figure 1) in Table 2. The mapping relation `pattern → set slicing point` was not found and activated in the PAIN3 example, so Figure 2 shows how this mapping relation would work for a two-handed sign with two different handshapes.

### 3.2. Technical Specifications

Our HNS2CF mapping tool was written in Python and consists of six functions, one for

| SL CatForm category | SL CatForm variable | 1:1 | n:1 | combination | pattern | |
|---|---|---|---|---|---|---|
| | | | | | set value | set slicing point |
| ARTICULATOR | 0 | | ✓ | | ✓ | |
| | 1 | | ✓ | | ✓ | |
| | 2 | | | | ✓ | ✓ |
| | 3 | | ✓ | | | |
| | 4 | ✓ | | ✓ | | |
| | 5 | ✓ | | | | |
| | 6 | | | | | |
| HANDSHAPE | 7 | | | | ✓ | ✓ |
| | 8 | | | | ✓ | ✓ |
| | 9 | | ✓ | | ✓ | ✓ |
| | 10 | | ✓ | ✓ | ✓ | ✓ |
| | 11 | ✓ | ✓ | | | ✓ |
| | 12 | ✓ | ✓ | | ✓ | |
| | 13 | ✓ | | | | ✓ |
| | 14 | | ✓ | | | ✓ |
| | 15 | | ✓ | ✓ | ✓ | ✓ |
| | 16 | ✓ | ✓ | | | ✓ |
| | 17 | ✓ | ✓ | | ✓ | |
| | 18 | ✓ | | | | ✓ |
| | 19 | | ✓ | | | ✓ |
| ORIEN-TATION | 20 | ✓ | ✓ | ✓ | ✓ | |
| | 21 | ✓ | ✓ | | ✓ | |
| LOCATION | 22 | | ✓ | | ✓ | |
| | 23 | ✓ | ✓ | ✓ | | |
| | 24 | ✓ | ✓ | ✓ | | ✓ |
| | 25 | ✓ | | ✓ | ✓ | |
| CORE MOVEMENT | 26 | ✓ | ✓ | | | |
| | 27 | ✓ | ✓ | ✓ | ✓ | |
| | 28 | | ✓ | | ✓ | |
| | 29 | | ✓ | | | |
| | 30 | ✓ | ✓ | ✓ | ✓ | |
| | 31 | | | | ✓ | ✓ |
| | 32 | | | ✓ | ✓ | ✓ |
| | 33 | | | | ✓ | |
| | 34 | | | ✓ | ✓ | |
| MANNER OF MOVEMENT | 35 | ✓ | ✓ | | ✓ | |
| | 36 | | ✓ | | | |
| | 37 | | | | ✓ | ✓ |
| | 38 | ✓ | | | | |
| | 39 | | | | | ✓ |
| | 40 | | ✓ | | | |
| | 41 | | | | | |

Table 1: Summary table of HNS2CF mapping relations used per variable (#0-#41) in the SL CatForm string

the mapping of each of the six categories in the SL CatForm coding schema (see section 2.2.). Each function contains a set of rules and conditionalities employing the four mapping relations described above. There are seven lists in our code: `handshapechars` (containing all handshape characters), `locationchars` (containing all location characters), `movementchars` (containing all movement characters), `orientationchars` (containing all orientation characters), `palmchars` (containing only palm orientation characters), `fingerchars` (containing only finger orientation characters), and `symmetrychars` (containing all symmetry characters). These lists are accessed at various points throughout the code, especially during `n:1` mappings and to set up `patterns`. The HamNoSys strings were examined for combinations by indexing through a for-loop. To map the HamNoSys to the *whole handshape* variables in the SL CatForm string (variables *#7*, *#8*, and *#9*), we created a dictionary containing tuples of HamNoSys character combinations and their corresponding values in SL CatForm. This dictionary is accessed in the code to iterate through the handshape combinations and find matches in the HamNoSys annotation string. Similarly, to map the HamNoSys locations to the *major area* variable (*#22*) in the SL CatForm string, we created a dictionary containing HamNoSys location characters and their corresponding values for *major location* in SL CatForm. This dictionary is accessed to perform two operations: a) to find the corresponding value in the SL CatForm string for locations in the HamNoSys strings, and b) to set the value for variable *#22* to '2' only if there are two locations specified in the HamNoSys, which **do not** fall within the same major area. Finally, patterns were stored as Python variables, triggering TRUE if a pattern is found in the HamNoSys string or FALSE if a pattern is not found. Based on this boolean, a new set of mapping conditionalities was either activated or ignored.

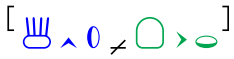
| mapping relation | information encoded in HamNoSys | target variable + value in SL CatForm |
|---|---|---|
| 1:1 | ~ hands move alternatingly | #4 hands move simult./altern. 'A' (alternating) |
| n:1 | ☤ or ☓ hands brush body or touch body | #28 path axis 1 'M' (midsagittal) |
| combination | ⬭ + ✓ fist handshape with thumb out | #7 whole handshape H1 '12' (A-thumb) |
| pattern → set value | no ↣ followed by any location character location does not change | #24 2nd location '0' (no 2nd location) |

Table 2: Examples for each mapping relation for the sign PAIN3 (see Figure 1)

## 3.3. Methodological Choices

Besides excluding the two variables (*#6 lexical mouthing* and *#42 secondary movement*) but keeping the placeholder 'xx' for readability, it was necessary to make certain methodological choices to account for ambiguities emerging during the creation of the HNS2CF mapping tool. For instance,





| 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 0 | 1 | A | S | S | S | 0 | A | S | U | U | 0 |

| | |
|---|---|
| information encoded in HamNoSys | two-handed sign with two different handshapes |
| pattern → set slicing point | any orientation character followed by (plus in HamNoSys) followed by any handshape character |
| target variable in SL CatForm | #7 + #10-14 for H1 and #9 + #15-19 for H2 |

Figure 2: Example for the mapping relation `pattern → set slicing point` for a two-handed sign (TREE2) with two different handshapes. Only the relevant handshape information from HamNoSys (top) and SL CatForm (bottom) are presented

the content of HamNoSys strings can vary from only the most basic information for a valid string to as much detail as the researcher/annotator can discern and chooses to annotate. For this reason, some information was encoded in HamNoSys that was not encoded in the SL CatForm coding schema. At the same time, there were also instances in which the HamNoSys strings for some signs lacked information; most often, this was for phonologically under-specified features. For example, the initial movement to a touch is typically unmarked in HamNoSys but encoded in SL CatForm. Another example is handshapes in which the thumb was not explicitly coded in HamNoSys due to its precise position being underspecified, such as in the V-handshape. However, thumb position is obligatorily encoded for all handshapes in SL CatForm. For the V-handshape, for instance, the thumb is marked as restraining/crossed. In these cases, we treated the thumb position as a match because it is a permissible thumb position for this handshape. Also, dispersed signs (i.e., including 'two touch' signs, like the sign DEAF1A[4], among others), are not uniformly encoded in HamNoSys. Therefore, we created a forced choice between dispersed signs and two-location signs: if there are two locations within the same major area (e.g., *forehead* and *chin* are both on the head), these are treated as dispersed signs within the same phonological location. When signs have two locations in different major areas (e.g., on the head and torso), these are encoded as two locations.

## 4. Data

There are four datasets in the current study: 1) the HamNoSys dataset, 2) a test dataset for evaluation, 3) an existing database of KSL signs coded in SL CatForm, and 4) the output dataset of DGS transformed into SL CatForm. First, the **HamNoSys dataset** is a list of types[5] in the Public DGS Corpus release 3 (Konrad et al., 2020) with a total of 14,476 types and their corresponding HamNoSys strings. This dataset was used for spot checks and to test intermediary steps in the creation of the HNS2CF tool, as well as being the source domain for the HNS2CF mapping. The total runtime for the HNS2CF mapping system on this dataset of 14,476 types was 1:05 minutes.

Second, to evaluate the accuracy of our HNS2CF mapping system, we created our **test dataset** by randomly selecting 100 sign types from the HamNoSys dataset. These were then manually coded according to the SL CatForm coding schema. This manual coding only used reference videos of the sign types and was therefore completely independent of the HamNoSys notations. During the analysis, however, we found two outliers, where the sign was coded as one-handed in the HamNoSys string but the corresponding video showed a signer producing the sign with two hands, or vice versa. The two outliers were removed from the test dataset and the final evaluation was performed on the 98 remaining signs in our test dataset. The total runtime for the HNS2CF mapping system on this test dataset of 98 types was 0.4 seconds.

The third dataset is **1,880 KSL signs coded in the SL CatForm coding schema**, contained in the KSL Lexical Database (Morgan, 2022). This dataset was used as a reference dataset at two points. First, it was used to address ambiguities arising from the theoretical mapping rules (section 3). Second, it was used in the evaluation step (section 5) to verify how certain phonological phenomena were coded in the past. This unpublished

[4] https://doi.org/10.25592/dgs.corpus-3.0-type-14040

[5] A type is equivalent to a citation form in a dictionary, distinguishable from individual tokens in a corpus.

dataset is part of a different project and the authors were granted access for the purpose of verifying mapping decisions.

The fourth dataset is the dataset of **DGS signs transformed into SL CatForm**. Currently, this dataset is not publicly available, pending further improvements to the mapping.

## 5. Evaluation Study

In order to find out how accurate the HNS2CF mapping tool worked, we created an evaluation study to compare the automated SL CatForm output with the hand-coded SL CatForm in the test dataset of 98 DGS signs. The results are divided into five categories: a) **matches** (i.e. the manual coding and the HNS2CF output were the same), b) **mismatches** (i.e. the manual annotator and the HNS2CF system assigned different values for a variable), c) **no match found** (i.e. the HNS2CF mapping tool did not find a match for this variable and kept the placeholder 'xx'), d) **no manual coding** (i.e. the manual coder was unsure and left the variable blank), and e) **mapping not possible** (i.e. variables *#6* and *#42*, which were not encoded in the HamNoSys strings).

### 5.1. General Results

Out of the 4,116 compared values (98 types * 42 variables), we found 76.7% (3,155 values; green) were matches, 16.9% (699 values) were mismatches 0.6% (24 values; red) were instances where the mapping tool could not find a suitable match; 1,1% (42 values; yellow) were empty manual annotation values because the manual coder was unsure of the correct value in the SL CatForm coding schema. These yellow values were due to the manual coder not being fluent in DGS; to avoid misinterpretations of the categorical form, it was left blank. Lastly, 4.8% (196 values; 98 types * 2 variables; grey) were not encoded in HamNoSys. Figure 3[6] shows an overview of the 98 signs in the test dataset and the result of the mapping for each of the 42 variables in the SL CatForm coding schema. As the first version of the HNS2CF mapping, these overall results were better than expected.

We further investigated if there was a difference in mapping performance by phonological parameter. When looking at the distribution according to the six categories in the SL CatForm coding schema (Table 3), one can see the mapping system performed least accurately for **orientation** with 52.1% of all values (91 values) being mismatches, whereas the system performed best for **articulator** with only 3.9% of all values (27 values) being mismatches. We attribute the poor performance for orientation to two factors. First, absolute orientation may be particularly sensitive to subjective differences between coders. Second, there is a notable difference in specificity between the two schemas; SL CatForm has many fewer categories than HamNoSys for orientation, so that edge cases may end up in a different category in the mapping.

It is interesting to note that instances where the mapping system could not find a suitable match (0.6%; 24 values) fall within only the **handshape** category (22 values) and **core movement** category (2 values). The empty manual annotation values (1.1%; 42 values), however, are spread across **orientation** (3 values), **location** (6 values), **core movement** (19 values), and **manner of movement** (14 values).

We also investigated which signs performed particularly well or poorly. The five signs with the fewest mismatches were EBB1[7] with 0 mismatches out of 42 possible matches, FOR2[8] with 1 mismatch, as well as END2[9], TO-PRAY1B[10], and WE1A[11] with 2 mismatches each. EBB1, the only sign with 100% accuracy, is a very simple two-handed sign with flat hands, symmetrical in neutral space, palms down, fingers outward, that moves straight downward once. The five signs with the most mismatches were PAIN3[12] and FEELING6[13] with 14 mismatches each, KINDERGARTEN1A[14] and CLOTHES1C[15] with 15 mismatches each, and PERFORMANCE3[16] with 17 mismatches. PAIN3 is shown in Figure 1. It is also two-handed and symmetrical, with two A-thumb handshapes, but with alternating arced or circular movement that have brushing contact on the upper torso. Therefore, the amount of phonological complexity may play a role in the mapping success.

### 5.2. Individual Variable Results

Furthermore, we were interested in how individual variables within each phonological parameter fared. These results are important because they pinpoint where improvements can be made in the mapping.

[6] This figure is intended to give an overview of the error/match distribution across the test dataset rather than a detailed analysis for each sign.

[7] https://doi.org/10.25592/dgs.corpus-3.0-type-14858
[8] https://doi.org/10.25592/dgs.corpus-3.0-type-13533
[9] https://doi.org/10.25592/dgs.corpus-3.0-type-15796
[10] https://doi.org/10.25592/dgs.corpus-3.0-type-14352
[11] https://doi.org/10.25592/dgs.corpus-3.0-type-14641
[12] https://doi.org/10.25592/dgs.corpus-3.0-type-13869
[13] https://doi.org/10.25592/dgs.corpus-3.0-type-13954
[14] https://doi.org/10.25592/dgs.corpus-3.0-type-15111
[15] https://doi.org/10.25592/dgs.corpus-3.0-type-18667
[16] https://doi.org/10.25592/dgs.corpus-3.0-type-17699

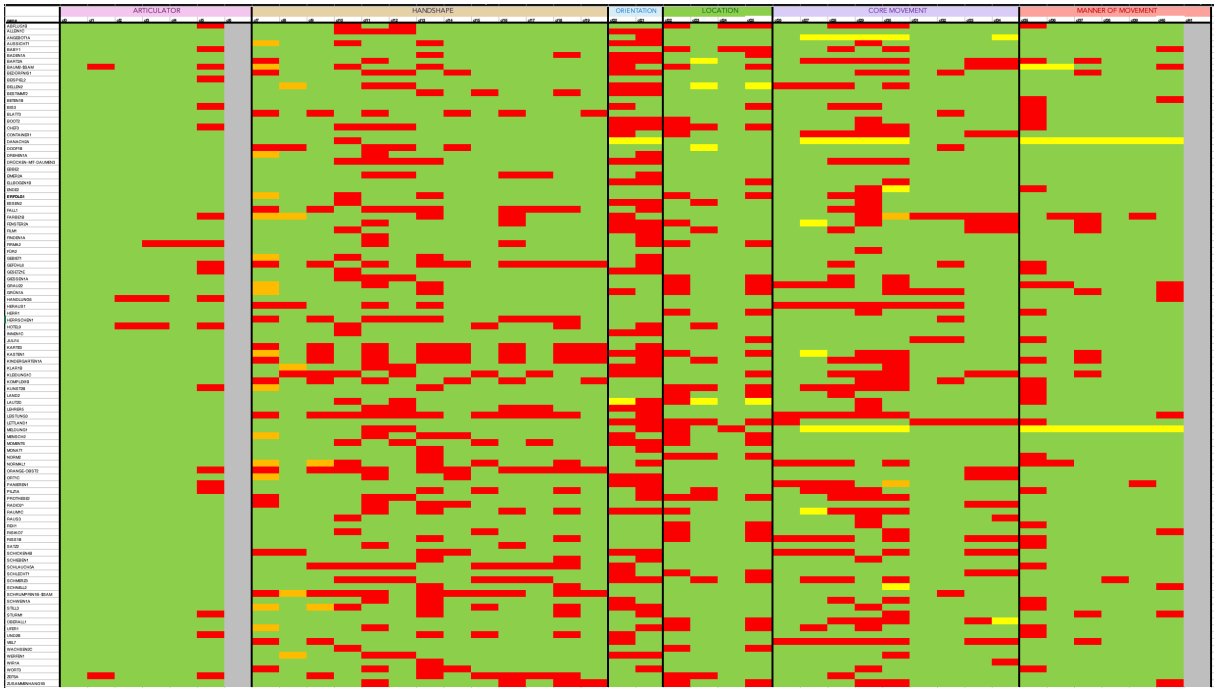



Figure 3: Colour-coded visualisation of the full evaluation results (matches = green, mismatches = red, no match found = orange, no manual coding = yellow, impossible mappings = grey). They x-axis represents the 42 SL CatForm variables, the y-axis represents the 98 signs in our test dataset

| | matches | mismatches | no match found | no manual coding | not possible |
|---|---|---|---|---|---|
| **articulator** | 561 (81.8%) | 27 (3.9%) | 0 (0%) | 0 (0%) | 98 (14.3%) |
| **handshape** | 987 (77.5%) | 265 (20.8%) | 22 (1.7%) | 0 (0%) | 0 (0%) |
| **orientation** | 102 (52.1%) | 91 (46.4%) | 0 (0%) | 3 (1.5%) | 0 (0%) |
| **location** | 306 (78.1%) | 80 (20.4%) | 0 (0%) | 6 (1.5%) | 0 (0%) |
| **core movement** | 684 (77.6%) | 177 (20.1%) | 2 (0.2%) | 19 (2.2%) | 0 (0%) |
| **manner of movement** | 515 (75.1%) | 59 (8.6%) | 0 (0%) | 14 (2.1%) | 98 (14.3%) |
| **total** | 3155 (76.7%) | 699 (16.9%) | 24 (0.6%) | 42 (1.1%) | 196 (4.8%) |

Table 3: Match distribution according to the six main categories in the SL CatForm coding schema

We found that the HNS2CF mapping system performs the most accurately for variables *#0* (*overall sign type*; 100%), *#4* (*hands moving simult./alt.*; 98.9%), *#1* (*number of hands*; 97.9%), *#38* (*switch dominance*; 96.9%), *#24* (*second location*; 96.9%), *#2* (*same shape*; 96.2%), and *#39* (*switch orientation*; 95.9%).

In contrast, the mapping system performs least accurately for these five variables: *#21* (*direction fingers facing*; 46.9%), #20 (*direction palms facing*; 57.1%), *#11* (*H1 joint config.*; 57,1%), *#29* (*path axis 2*; 58.2%), and *#30* (*path directionality*; 59.2%). We analyse each of these in more detail below.

For **variable *#21, direction fingers facing***, most mismatches occurred when the manual coder assigned the value 'Y' (dynamic), meaning that the finger orientation changed during the sign, but the HNS2CF mapping system assigned a different value. However, we did not observe a predictable tendency or pattern; i.e. the mapping system did not confuse 'Y' for a single other value but assigned various other values instead of 'Y'. This is due to the complex logic needed to extract information about finger orientation change in HamNoSys. Our HNS2CF mapping logic did find multiple instances of finger orientation change — even 9 instances the manual coder did not recognise as

such — however, the system missed 8 instances of finger orientation change the manual coder did recognise and coded. Overall, within the 51 mismatches for *#21*, there were 31 mismatches where the mapping system actually worked correctly but the HamNoSys did not match the manual coder; these can be taken as successes. The other 19 mismatches were due to errors in the code for the HNS2CF tool, and can be addressed in future revisions of the tool.

For **variable *#20, direction palms facing***, most mismatches were caused by the mapping system assigning the value 'G' (diagonal) when the manual coder assigned a different value. However, there was no systematic patterning with a different value chosen by the manual coder. We found our HNS2CF logic still needs improvement in extracting the correct information from the HamNoSys due to the complexities of how palm orientation is encoded in HamNoSys (dependent also on the finger orientation). However, the 40 mismatches for variable *#20* are mostly not due to coding errors (only 12 mismatches) but due to differences in what the HamNoSys encodes and the values our manual coder assigned (25 mismatches). They should therefore be considered more of an annotator disagreement – that is, two annotators perceived certain aspects of a sign differently, leading to different annotations, which are not necessarily correct nor incorrect – than a mapping error.

For **variable *#11, H1 joint config.***, we observed that the HNS2CF mapping system tended to assign the value 'U' (curved) when the manual coder assigned the value 'H' (hooked). This was the case in 13 mismatches out of 42 (30.9%), but is not an error in the way the HNS2CF tool was coded but is either due to perception differences of what is considered curved or hooked, or it may also be slight phonetic differences between DGS and KSL handshape categories. There were, however, also 21 mismatches caused by an error in the HNS2CF logic: the mapping system incorrectly assigned the value 'C' (closed) when the manual coder assigned other values, such as the value 'S' (straight; 8 mismatches), the value 'U' (8 mismatches), or the value 'B' (bent; 5 mismatches). This coding error will be fixed in an updated version of the HNS2CF mapping system.

For **variable *#29, path axis 2***, most mismatches were caused by the mapping system assigning the value 'M' (midsagittal) when the manual coder assigned the value '0' (no second axis; 11 mismatches out of 37). Overall, the mapping system was more likely to confuse '0' values for other values (20 mismatches; 54.1%), meaning our HNS2CF mapping tool tends to code a second axis when, according to the manual coder, there is no second axis. Because this information is not directly encoded in the HamNoSys string, the mismatches are most likely caused by the mapping logic not being refined enough yet.

Finally, for **variable *#30, path directionality***, a tendency of the mapping system was to assign a different value when the manual coder assigned the value 'A' (away from body; 6 mismatches). However, there was no systematic patterning with a different value assigned by the mapping system. In general, we found the logic in our HNS2CF tool needs further refinement for diagonals and back-and-forth movements, as these caused 7 out of the 33 total mismatches within variable *#30*. However, it is important to note that variable *#30* was the variable with the highest number of missing manual annotations (5 values; 5.1%), suggesting that path directionality might be more complex to encode and various coders perceive path directionality differently. Indeed, out of the 33 mismatches for this variable, 13 were due to differences in what the HamNoSys encodes and the values our manual coder assigned rather than due to issues with the mapping tool and its implementation.

## 6. Discussion

This first version of the HNS2CF mapping tool already performs fairly well in the evaluation study, with 76.7% of variable matches in 98 random DGS signs. Further, as we unpacked the reasons for the non-matched values — i.e., everything that is not green in Figure 3 and Table 3 — we find that many mismatches do not point to problems with the HNS2CF mapping itself, but have other sources, such as the manual coder not feeling confident in coding the phonological form from one video alone (yellow cells in Figure 3 and Table 3), or the manual coder making mismatches for finger orientation that were later shown to be mapped correctly from the HamNoSys.

We also identified several ways to improve the mapping. One of the most straightforward to implement is to add language-specific values in certain phonological inventories in SL CatForm. The evaluation study had two handshape values that were not in SL CatForm. Handshape, location, and path shape inventories are known to differ across languages, and values like these can easily be expanded with new SL CatForm values. Other targets for improving the actual mapping tool were identified in the individual variable results in the previous section; e.g., refining the mapping logic for *path axis 2*.

We also found some indication that the degree of sign complexity may lead to more mismatches. We reached this conclusion by looking at the five signs with the highest and lowest accuracies. Those with the highest accuracy all have

the following characteristics in common: they are signs in neutral space that have only path movement (no handshape or orientation movement; i.e., no secondary movement) and are marked as symmetrical for lateral symmetry. In other words, they were very simple signs. In contrast, the signs with the lowest accuracy were all two-handed and contained mismatches in orientation and in path axis 1; and many had secondary movements (handshape and/or orientation changes in the sign).

## 7. Conclusion & Future Directions

In this paper we have introduced the HNS2CF mapping system and tool. We presented the formal mapping relations, have outlined the system's implementation, and have shown empirical evidence of the system's quality. Despite it being only version 1 of this mapping system, we can already report an overall accuracy of 76.7%. Furthermore, the tool's total runtime of only 1:05 minutes on the DGS dataset with 14,476 types shows that the tool works efficiently, even with large datasets. However, a detailed analysis of mismatches in the evaluation study identified several places where the mapping logic should be revisited. Yet, the analysis also showed that the actual accuracy of the HNS2CF mapping tool is even higher than 76.7% as we have found multiple mismatches which were not due to errors with the tool but correctly mapped what was encoded in the HamNoSys to values in the SL CatForm coding schema. These mismatches therefore signal annotator disagreement rather than mapping errors.

We will continue to improve the current version of the HNS2CF mapping system, adding additional values to the SL CatForm coding schema where needed and adjusting the schema further to better fit DGS. Additionally, the efficiency of our mapping tool could potentially be further optimised by building a HNS2CF compiler. In line with this, we aim to build a reversed system, mapping from SL CatForm to HamNoSys, which could also serve as a validation tool and offer valuable insights into the kind and amount of information lost when mapping between those two systems. With some further adjustments, the mapping system has the potential to be extended not only to include a wider variety of sign languages and dialects, but its basic scaffolding could also be extended to other notation systems, for example, SignWriting. Further, the mapping system could be integrated into annotation pipelines to create even richer resources and larger datasets annotated in parallel in multiple notation systems and coding schemata.

## 8. Limitations

A test dataset of 100 signs is only a snapshot and may not be sufficiently representative of the DGS data; yet, annotating signs manually is time-consuming. After all, there are 4,100 separate data points within even this small dataset. Another limitation we encountered in the manual coding is grounded in the data itself. The SL CatForm coding schema is based on phonological categories. Without an existing phonological analysis of DGS describing its categorical forms, it was difficult to force some articulatory types into SL CatForm values. Finally, the HamNoSys data was annotated by multiple annotators whereas the SL CatForm data was annotated by a single annotator. In order to be able to fully account for annotator disagreement, this imbalance should be resolved and analysed as its own variable in future studies.

## 9. Acknowledgements

We would like to thank the DGS corpus team at the University of Hamburg for their input and continued support.

Research in this paper was funded by the European Union (ERC, *SemaSign*, No.101117395). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## 10. Bibliographical References

Diane Brentari. 1989. *A prosodic model of sign language phonology*. MIT Press, Cambridge, MA.

Rachel Channon and Harry van der Hulst. 2008. The Phonetics and Phonology of Sign Languages First SignTyp Conference June 26-28, 2008. [Accessed 03-04-2026].

Ralph Elliott, John Glauert, Vince Jennings, and Richard Kennaway. 2004. An overview of the SiGML notation and SiGMLSigning software system. In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 98–104, Lisbon, Portugal. European Language Resources Association (ELRA).

Thomas Hanke. 2004. HamNoSys – representing sign language data in language resources and language processing contexts. In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 1–6, Lisbon, Portugal. European Language Resources Association (ELRA).

Oscar Koller, Hermann Ney, and Richard Bowden. 2016. Automatic alignment of HamNoSys subunits for continuous sign language recognition. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 121–128, Portorož, Slovenia. European Language Resources Association (ELRA).

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release.

Hope Morgan, Amy Isard, and Ahn Dang. 2026. Improving phonological distance measures for signs: the CatFormCompare tool. In *Proceedings of the LREC2026 12th Workshop on the Representation and Processing of Sign Languages: Language in Motion*, Palma, Mallorca (Spain). ELRA Language Resources Association (ELRA).

Hope E. Morgan. 2022. *A Phonological Grammar of Kenyan Sign Language*. De Gruyter Mouton, Berlin, Boston.

Hope E. Morgan. 2026. SL CatForm coding schema v.1: A field-based phonological coding for sign languages (Version 1.0). Technical report, Universität Hamburg.

Hope E. Morgan, Wendy Sandler, Rose Stamp, and Rama Novogrodsky. 2022. ISL-LEX v.1: An online lexical resource of Israeli Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 148–153, Marseille, France. European Language Resources Association (ELRA).

Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, Jan Henning, et al. 1987. *HamNoSys. Hamburg Notation System for Sign Language. An Introduction*. Zentrum für deutsche Gebärdensprache, Hamburg, Germany.

Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, Jan Henning, et al. 1989. *HamNoSys. Version 2.0. Hamburg Notation System for Sign Languages. An Introductory Guide*. Signum, Hamburg, Germany.

Constanze Schmaling. 2000. *Maganar Hannu, language of the hands: A descriptive analysis of Hausa Sign Language*. Hamburg: Signum.

William C. Stokoe. 1960. *Sign language structure: An outline of the visual communications systems of the American Deaf*. Silver Springs, MD: Linstock Press.

Valerie Sutton. 1999. SignWriting: On the occasion of its 25th anniversary November 1999. *Sign Language & Linguistics 2.2*.

Els van der Kooij. 2002. *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity*. LOT Publications, Leiden, Netherlands.

# Emotion Recognition in German Sign Language with Facial Action Units

**Cristina Luna-Jiménez*, Lennart Eing*, Sergio Esteban Romero†, Tanja Schneeberger‡, Patrick Gebhard‡, Fabrizio Nunnari‡, Elisabeth André***

*University of Augsburg
{cristina.luna.jimenez, lennart.eing, elisabeth.andre}@uni-a.de
†Universidad Politécnica de Madrid
{sergio.estebanro}@upm.es
‡German Research Center for Artificial Intelligence (DFKI)
{tanja.schneeberger, patrick.gebhard, fabrizio.nunnari}@dfki.de

**Abstract**

Emotion Recognition research in Sign Languages is still in its infancy. Still today, there exists a lack of knowledge about appropriate annotation guidelines and the impact that facial expressions, body postures and head positions have in recognizing emotions while signing, considering that sign language encompasses manual and non-manual cues with linguistic purposes. In this article, we present an acquisition protocol to record acted emotions in German Sign Language under four scenarios (High-Valence and High-Arousal, High-Valence and Low Arousal, Low-Valence and High-Arousal, and Low-Valence and Low-Arousal). The goal is to provide a reference dataset to explore the use of machine learning techniques for an automated classification of emotions in sign language utterances. As a baseline reference, we trained static models with features extracted from the facial muscle activations. The best model achieved an accuracy of 68.84% and a F1 of 67.96% with a random forest trained on the statistics extracted from Action Units. These results highlight the importance of facial expression in sign language, not only for carrying linguistic information but also for transmitting emotions. Results also indicate challenges in detecting emotions in the High-Valence and Low Arousal scenario, which suggests future investigation lines to explore.

**Keywords:** Emotion Recognition, Sign Language, Machine Learning

## 1. Introduction

Emotions play an important role in our communication and decision-making. In spoken conversations, emotional content can be conveyed from our facial expressions, or posture, among others (Luna-Jiménez et al., 2022; Zhang et al., 2023). However, Sign Languages are characterized by the use of manual and non-manual channels to transmit, simultaneously, linguistic and affective content. For example, Pendzich et al. (Pendzich et al., 2022) observed that conditionals in German Sign Language (DGS) can be accompanied by brow raise, chin raiser or upper lid raise expressions. Similarly, mouthing can disambiguate semantics of certain signs in DGS (e.g., the same hand movements depicting a square can refer to 'square', 'page' or 'letter' depending on the mouthing (Konrad et al., 2022)). This richness of facial expressions for conveying linguistic meanings represents a challenge for existing automatic emotion recognizers when the person is signing since they were not trained to differentiate that certain facial expressions are associated to linguistic meanings.

Although previous research exists in Affective Computing for Facial Emotion Recognition (FER) in spoken languages (Livingstone and Russo, 2018; Vlasenko et al., 2007), the studies in Sign Languages is rather limited. Therefore, the main contributions of this article are summarized as:

- A proposal of an acquisition framework to elicit and record acted emotions in Sign Language, following previous research in Affective Computing and adapting it to DGS. This framework was evaluated by performing a corpus acquisition with professional deaf interpreters under controlled conditions, signing different glosses in DGS.

- An analysis of the acquired dataset by proposing a baseline with classical Machine Learning models on Facial Emotion Recognition in Sign Language. The top performance was obtained employing Action Units with a Random Forest, achieving an average accuracy of 68.84% and F1 of 67.96%, following a subject-wise leave-one-out strategy. These results highlight that non-manual features, as the facial expressions, not only carry semantic and linguistic information in DGS but also valuable keys for performing emotion recognition.

To our knowledge, this is the first study that explores how to record an acted dataset for performing emotion recognition in German Sign Language (DGS) and evaluate the contribution of facial features for predicting emotions. This study establishes a first exploration in a controlled scenario to, in the future, extend to other affective computing paradigms (e.g. study of emotions in the wild) and



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 297–305
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

acquisition methodologies given the infancy state of the field of emotion recognition in Sign Language.

The remainder of the paper is organized as follows: Section 2 describes the related work, summarizing previous research in emotion recognition in sign language. Section 3 introduces the corpus acquisition process, including the scenario and recording procedure. Then, Section 4 describes the corpus evaluation with machine learning algorithms, indicating the extracted features and the trained models. Afterwards, Section 5 presents the obtained results and the confusion matrices of the predictions per subject. Finally, Section 6 summarizes the main conclusions and future research directions.

## 2. Related Works

### 2.1. Emotions in Sign Language

In sign language linguistics, much work investigated how non-manual features serve both affective and grammatical functions.

For example, in a study on Irish sign language (ISL), Kulshreshtha (2024) found that participants used the doubled-wh signs in the context of anger, surprise, and excitement.

Kirst (2024) reports on an experiment aimed at studying the "layering", i.e., "when multiple phonological and paralinguistic elements co-occur". In this experiment, they observed that signers employ three meta-strategies to cope with the communication of emotions on the top of a message: 1) separation, by sequencing the emotion expression before or after the grammatical information; 2) addition, by introducing, for example, rapid blinking and jaw drops; 3) competition, by increasing muscle activation when, for example, both a question modifier and expression of anger require brows furrowing.

Recently, Schwarzenberg et al. (2024) also investigated the interaction between lexemes and non-manuals, showing inconsistency in the use of facial expressions accompanying a sign for an emotional concept (e.g., sadness).

All of the above studies suggest how the use of facial expressions for emotion communication in sign languages differs from spoken communication.

In addition, in sign languages emotion communications is not limited to hands. Hietanen et al. (2004) conducted a meticulous study on the perception of emotions from hands movement in Finnish sign language. The study was focusing on only two emotions (happy, anger) plus neutral. The emotions were judged by non-signers. The results show indeed a significant capability to distinguish between neutral and anger. If emotions from hands can be perceived by non-signers, surely it is part of the language, and it suggests that it is worth investigating in deeper detail the role of body movement for emotion communication.

### 2.2. Automatic Emotion Recognition

From a computer science viewpoint, research in emotion recognition has advanced in several lines pursuing to study emotions from different perspectives depending on the naturalness and genuineness of the elicited and expressed emotions.

In the first group, datasets on acted emotions are characterized by clean expressions of emotions, in which actors are hired to display them with high arousal (Livingstone and Russo, 2018; Vlasenko et al., 2007). The second group contains elicited emotions, in which stimulus such as images, sounds or conversations are employed to induce a specific emotion in the participants (McKeown et al., 2012), which after being exposed to the stimulus display a more natural emotional reaction. Finally, the third group of datasets consists of 'spontaneous' emotions captured for example from news or TV shows (Grimm et al., 2008), in which natural reactions to events are recorded from the facial and body expressions of the participants. Although this last set is the most natural or genuine, it suffers often from an imbalance in the available classes.

However, although these datasets exist normally for spoken languages, the same pairs are starting to appear and being investigated only recently for sign language. For example, the eJSL dataset (Funakoshi and Zhu, 2025) enters in the category of 'acted' datasets in which two signers represented emotions in 78 pre-selected utterances, resulting in 1,092 videos, recorded in Japanese Sign Language (JSL). This dataset was recorded by a deaf native professional actor, mimicking seven emotions plus neutral. An example of a 'spontaneous' dataset in DGS is FePh (Alaghband et al., 2020), in which emotions for 3,359 image samples from the RWTH-PHOENIX-Weather 2014 (Koller et al., 2015) were annotated based on facial expressions according to the seven 'universal emotions' proposed by P. Ekman (Ekman, 1999) plus a 'None' label in case the emotion was not in the group of seven. Finally, the EmoSign dataset (Chua et al., 2025), in American Sign Language (ASL), is a subset of 200 videos of an average duration of 4.8 seconds annotated by three deaf signer deaf interpreters. This dataset would be considered as part of the 'spontaneous' emotion dataset since videos were extracted from a larger corpus, the ASLLRP (American Sign Language Linguistic Research Project) (Neidle et al., 2022) which was recorded with real interactions of deaf participants and annotated posteriorly. The subset contains labels of sentiments on 7-point Likert scale, and 10 emotions. Additionally, it includes descriptions of manual and non-manual emotional

indicators such as facial expressions, head, mouth and body movements. In addition to the dataset, they provide baselines for the study of emotions in ASL with Multimodal Large Language Models.

Due to the limit number of available datasets with annotations of emotions, researchers interested on performing automatic emotions recognition are starting to employ them. For example, Nunnari et al. (2025) used a subset of 2,547 frames from FePh to recognize emotions in Sign Language videos with a light CNN architecture to perform emotion recognition. Nonetheless, there exists still open-challenges to address, such as the disentanglement of emotional and grammatical functions of facial expressions (Kimmelman et al., 2020; Kirst, 2024), which requires of further development of datasets and annotation schemes. Additionally, the understanding of whether previous approaches and features developed for facial and body emotion recognition (Luna-Jiménez et al., 2022; Martinez-Martin and Fernández-Caballero, 2025) could be also effective for predicting emotions while signing is still a open research question.

## 3. Corpus Acquisition

In order to create a scenario for collecting and eliciting emotions, we followed the valence-arousal approach introduced by Posner et al. (2005). In this theory, 'valence' describes how pleasant or unpleasant a situation is. For example, situations with negative valence are perceived as unpleasant or bad, which tends to trigger feelings that we tend to avoid. In contrast, situations with positive valence are perceived as pleasant or good. The second axis of the space is 'arousal', which refers to the level of activation or excitement in a given situation. High arousal means that someone is internally very activated or excited, while low arousal means that someone is internally not very activated (e.g. calm, relaxed, or even sleepy).

Specifically, in our scenario, four sub-regions of the valence-arousal space were considered: High-Valence and High-Arousal (HV-HA), High-Valence and Low Arousal (HV-LA), Low-Valence and High-Arousal (LV-HA), and Low-Valence and Low-Arousal (LV-LA). We gave additional examples and colors to these scenarios to make them easier to understand and to explain. For example, the HV-HA was assigned to the color 'Yellow', the HV-LA was assigned to the color 'Green', the LV-HA with the color 'Red' and LV-LA with the color 'Blue'. Figure 1 shows the valence-arousal diagram designed for the explanation of the four scenarios.

### 3.1. Material preparation

In order to record samples of signs with different emotional states, the first step was to explain the experiments and the goal of the acquisition. With this aim, we created a serie of slides to follow during the acquisition protocol in order to guide to the signer in the glosse to perform and the emotions to be expressed in each of the specific scenarios. The protocol consisted on the following steps:

- First, the signer was welcomed to the experiment and introduced to the task. The task consisted on recording at least ten repetitions of ten pre-selected glosses in four emotional scenarios. For eliciting the emotions of each scenarios, we asked the participant to put themselves in each of the situations and scenarios. After that, the recordings start.
- After the introduction to the experiment, the second step was to introduce briefly the concept of arousal and valence. The introduction was done with short descriptions as well as with graphic elements consisting on the Self-Assessment Manikin (SAM) scale and images extracted from the OASIS corpus (Kurdi et al., 2017) that demonstrated to elicit the emotional states explained in the description, as shown in Figure 1. For example, the picture of the kids on the left corner elicits high valence feelings; whereas the picture of the garbage on the right corner-down elicits emotions with negative valence.

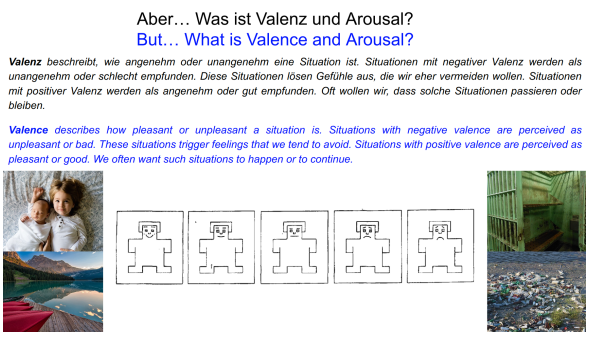



Figure 1: Description of Valence, originally only the German version is provided, the English version is added for the publication.

- After the description of valence and arousal, the next step was to clarify the scenarios and the axis system. For that part, we employed the designed valence-arousal diagram incorporating colors, the SAM images, and additional descriptions, as it can be seen in Figure 2
- Once the experiments were explained, the description of the scenarios starts, as well as the acquisition process. Each of the scenarios

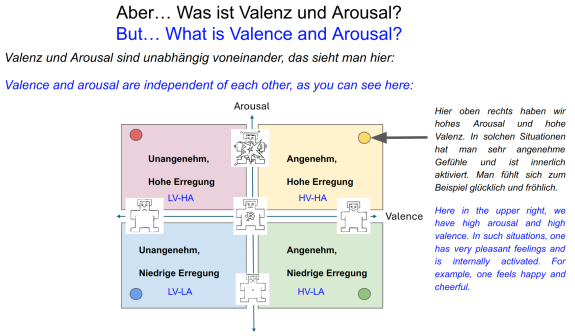



Figure 2: Graph with the Valence-Arousal diagram for explaining each scenario, together with an example, in this case for the yellow scenario of High Valence and High Arousal (HV-HA). The English translation (text in blue) is added for the publication but it was not in the original version.

follow the same structure, first, the valence-arousal graph is shown, indicating in which of the four scenarios we are, then, the scenario is described with a short text and a vignette associated with it (See Figure 3). After that, the recording process starts having a first video with a demo sign with neutral emotion, and followed by the picture of the scenario between sign transitions, in order to let the participant remember the scenario. The same process is repeated for the 4 scenarios.

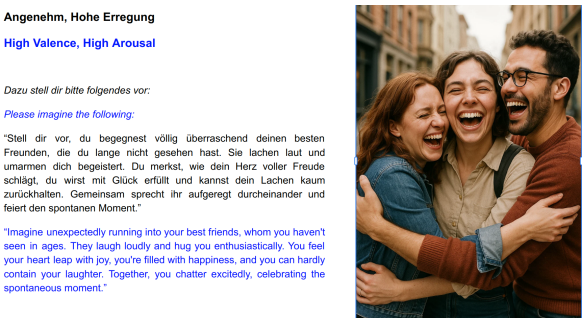



Figure 3: Stimuli for the High Valence and High Arousal scenario (HV-HA). The English translation appears in blue.

- Finally, the acquisition process concludes by thanking the participants for their time.

### 3.2. Participants and Recordings

For the recordings, we pre-selected a vocabulary of 10 signs in DGS to evaluate the data collection approach proposed and study whether emotions could be distinguished and recognized in this initial set of glosses to verify the validity of the approach. These signs correspond to the meanings of: JA (Yes), NEIN (No), KEINE-AHNUNG (No idea), PERSON (Person), LEUTE (People), AUTO (Car), FEUER (Fire), MEDIZIN (Medicine), ANDERS (Different), MEHR (More).

In order to record high-quality signs, the recording process was performed in the studios of the professional deaf interpreters. In total, three deaf interpreters participated in the recordings, two men and one woman. Only frontal cameras were used during the recording process and a green-screen in the background. All the recordings are in Full HD resolution (1920x1080) at 50 fps. As the recordings of each sign were captured in a continuous way, an annotation step was required. This annotation step was performed by a DGS-proficient professional with linguistic background that was in charge of selecting when signs start and end in the videos and annotate them with the ELAN tool.

After the annotations, videos were split automatically, resulting in 1,540 short-videos of one gloss signed with one of the specific emotions of each scenario. Videos have different durations in length ranging from 0.3 to 3.58 seconds ($\mu = 1.70; \sigma = 0.42$). Figure 4 shows the distribution of samples per subject and emotions in the collected dataset.

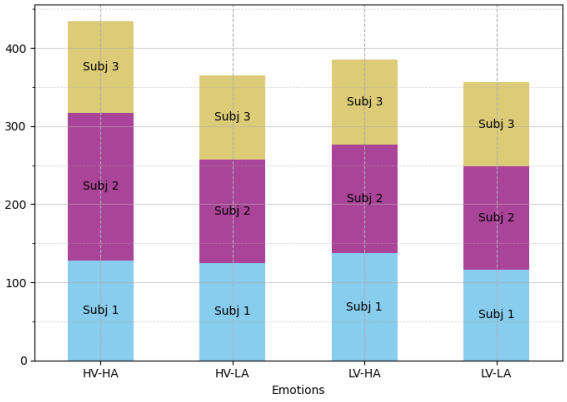



Figure 4: Distribution of the number of files per emotion and subjects. X-axis shows the four scenarios, and the y-axis the number of videos. Finally, colors represent the subjects.

## 4. Corpus Evaluation on Emotion Recognition

In order to test whether the acquisition methodology and approach was able to capture different emotions across signs in DGS, we trained emotion recognizers on the acquired corpus. As only three subjects were recorded, a subject-wise three-fold Leave-One-Out approach was followed, training in each fold with two subjects and testing the model in the third subject.

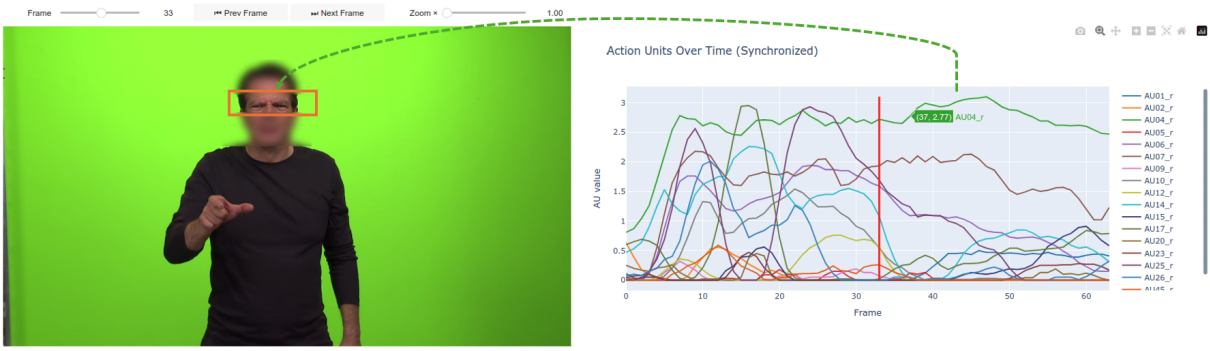



Figure 5: On the right, it can be observed the frame 33 of one of the videos of the 'Angry' scenario of one of the deaf interpreters signing 'PERSON'. On the left, the intensity of the Action Units per frame. Notice that in the AU 04 ('Brow Lowerer') is active with high intensity across the whole video. According to previous studies in facial emotion recognition (Ekman et al., 2002), the AU 04 is related to emotions such as 'Angry' or 'Disgust'.

### 4.1. Facial Features Extraction

Although pose and manual cues are key in sign language, non-manual features such as facial expressions or mouthing could carry linguistic and emotional information. For this reason, we extracted facial features to evaluate emotion recognition in Sign Language.

To recognize emotions from facial expressions, we employed the Action Units (AUs). AUs constitute the fundamental elements of the Facial Action Coding System (FACS), which categorizes human facial movements based on their observable muscle configurations and intensities. Each AU represents a specific facial muscle activation and is typically linked to changes in facial expression that reflect variations in an individual's psychological or emotional state, and traditionally they achieved considerable good results for recognizing emotions in spoken languages from facial expressions (Luna-Jiménez et al., 2022).

Among the available tools for extracting Action Units from video frames, we selected the OpenFace toolkit (Baltrusaitis et al., 2018). This framework has been extensively employed in facial emotion recognition for spoken languages and provides both the intensity of 17 AUs (AU1,2,4,5,6,7,9,10,12,14,15,17,20,23,25,26,45) and the presence of 18 AUs (same as intensity plus AU28). AU presence is represented as a binary value, equal to one when the AU is detected; whereas AU intensity is a continuous measure ranging from 0 to 5. These outputs are generated by two separate models that share a common pre-processing pipeline. First, facial regions are aligned to derive geometric and appearance-based features; then, these features are fed into two independent SVM classifiers, which estimate the AU presence and intensity for each frame, respectively.

Figure 5 represents the continuous Action Units generated by OpenFace for one of the frames of the videos of the 'Low Valence and High Arousal' scenario, in which the interpreter is signing 'person'. As it can be observed, the AU4 ('Brow Lowerer') is active during the whole video, which usually correlates with emotions that appear in the Low Valence and High Arousal quadrant such as anger, annoying or frustration (Kohler et al., 2004).

After extracting the 35 Action Units of each video frame, we calculated statistical aggregations from them in order to adapt the temporal dimension to static models. To accomplish this, the mean, standard deviation, maximum, minimum, skew, kurtosis and median statistics were calculated for each of the features across their temporal dimension, resulting in a total of 245 features. This approach, employed as our baseline, offers two primary advantages: first, its conceptual simplicity; and second, the robustness conferred by the averaging effect. To illustrate this, consider a video in which several frames deviate from a prototypical emotional expression, e.g., due to the subject closing their eyes. By employing statistics across the temporal dimension, the influence of such anomalous frames on the overall emotion recognition system is mitigated, assuming that the remaining frames predominantly reflect the target emotion.

### 4.2. Models Training

After adapting the features, we evaluated a range of statistical learning models implemented in the scikit-learn library. Specifically, we considered support vector machines (SVMs) with radial basis function (RBF) and linear kernels, logistic regression, ridge classifiers, decision trees, random forests, and multilayer perceptrons (MLPs). For the SVMs, Logistic

Regression (LR), and Ridge Classifiers (RC), the regularization parameter was varied over the values 1, 0.1, 0.01, 0.001, 0.0001. For the Random Forest (RF) models, we explored different numbers of estimators, namely 10, 20, 30, 40, 50, 100, 200, 300. Similarly, for the Decision Trees (DT), we evaluated different values for the minimum number of samples required to split an internal node using the same set of values (10, 20, 30, 40, 50, 100, 200, 300). For the MLP models, we explored the same set of values for the number of hidden units in the hidden layer (10, 20, 30, 40, 50, 100, 200, 300). All remaining hyperparameter were set to their default values as provided by the scikit-learn library.

## 5. Results on Facial Emotion Recognition

As can be observed in Table 1, the 245 features extracted from facial expressions seem to be promising features for performing facial expression recognition in Sign Language too. The random forest with 300 estimators, that received at its input the statistics computed from the regression values of the Action Units achieved an accuracy of 68.84% and a F1 of 67.96%, surpassing the Zero-Rule approach by +43.76 percentage points in accuracy and +57.89 in F1. This model is closely followed by the linear-SVM, the Logistic Regressor and the MLP.

| Model | Hyperparam. | ACC | F1 |
|---|---|---|---|
| ZeroRule | - | 25.08 | 10.07 |
| RBF-SVM | C=1 | 50.56 | 49.80 |
| DT | Min. Split = 300 | 54.65 | 52.28 |
| RC | C=0.01 | 61.97 | 58.93 |
| MLP | Neurons = 20 | 65.72 | 64.43 |
| LR | C=0.1 | 66.89 | 65.06 |
| Linear-SVM | C=0.001 | 67.32 | 65.38 |
| **RF** | **Estimators=300** | **68.84** | **67.96** |

Table 1: Top models configurations - emotion recognition.

In order to understand variations across different subjects on their facial expressions, we evaluated the confusion matrices in the test sets of each of the folds of the subject-wise LOO strategy. Figure 6 shows these confusion matrices in which colors represents each of the subjects. The firs observation suggests that, as in spoken languages, different individuals rely differently in facial expressions to transmit emotions. For illustration, if we analyze subject two (in pink) and three (in orange), we can observe that facial expressions are much more relevant for recognizing emotions in them than for the subject one (in blue), reaching F1 values over 70% versus the 57% of accuracy obtained for the first subject.

Regarding the complexity on recognizing emotions across individuals, it seems that the emotion of the high valence, low arousal (HV-LA) scenario, resembling a calm feeling, is the worst detected by the models with an accuracy of 17.60% (22/125) for the first subject, 40.15% (53/132) for the second subject, and 40.75% (44/108) for the third subject. On the other hand, the high valence, high arousal (HV-HA) scenario, resembling emotions like happy, was the easiest to detect by the models, correctly recognizing the 84.56% (367/434) percent of the samples of the happy class. From the perspective of the data acquisition scenario, these results may indicate that the participants found more complex recording in the calm scenario maybe because the behavioral cues associated to this emotion could be not as familiar as for example those related to emotions such as happy or angry. Indeed, the calm emotion is not that frequently annotated, as commented in (Livingstone and Russo, 2018) and it is not part of the considered theoretical 'universal emotions' described by P. Ekman (Ekman, 1999). However, it is still one of the most important quadrants of the 'circumplex model of affect' (Posner et al., 2005), for this reason, it was considered in the study.

## 6. Conclusions

As in many languages (spoken and signed), facial expressions carry rich information that is key for fully understanding the transmitted message and the intentions of the communicator. These facial expressions complement the transmitted message and reduce disambiguation. However, although the field of Affective Computing have advanced in spoken languages, in sign languages the exploration of how emotions are expressed while signing has barely started.

In this article, we proposed a method to acquire a dataset of acted emotions, following previous datasets in the literature released for spoken languages, such as RAVDESS (Livingstone and Russo, 2018), in which professional actors are asked to perform emotions 'as if they were feeling them'; in the same way, in this dataset professional deaf interpreters were asked to put themselves in four different emotional scenarios of low arousal and low valence, low arousal and high valence, high arousal and low valence, and high arousal and high valence and perform 10 different signs.

In order to evaluate the suggested approach, we applied traditional methods in facial emotion recognition to extract Action Units across the videos. With this simple approach, the models were able to achieve an average F1 of 67.96% for the face modality, indicating a promising investigation line to perform emotion recognition in sign language,

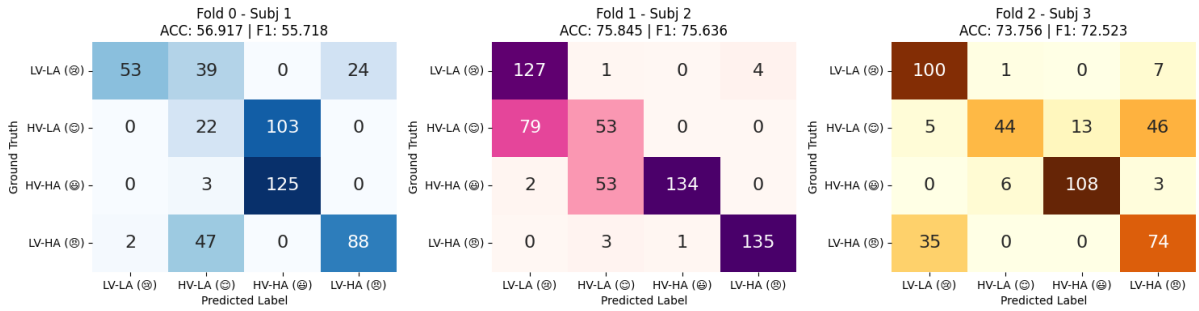



Figure 6: Confusion Matrices of the RF model for facial emotion recognition per subject for the test set of each fold of the LOO. Colors indicate the results of each subject.

and reinforcing the validity of the data acquisition.

As future lines, we would evaluate whether the expressed emotions are perceived in a similar way for deaf and non-deaf individuals, to study perceptual differences when you know the language versus when you do not. Additionally, we will extend the analysis to other non-manual channels (e.g. body posture) to understand the impact of other features and modalities in emotion recognition in Sign Language. Finally, one of the future steps would be to evaluate longer sentences and interactions, as well as annotating them in terms of the transmitted emotions.

## 7. Limitations

While the study provides preliminary results about the relevance of facial expressions, including more features and subjects in the training and evaluation of the models could benefit to the overall understanding of expressed emotions in Sign Language. However, recruiting deaf participants is challenging, specially considering that recordings can not always being made anonymous given the richness of DGS in terms of the information transmitted by manual and non-manual cues. Additionally, we acknowledge that extracted features are dependent on the libraries, so they could be biases and more accurate for some individuals than for others. Finally, as other datasets in Affective Computing, the first version of our acquisition was based on an acted scenario, in which the deaf interpreters needed to put themselves in hypothetical scenarios to express emotions, hence, models' performance may vary in the real-world, given the nuances associated to emotions. For this reason, these results should be understood as a first step and work in progress in the understanding of expressiveness of emotions in Sign Language.

## 8. Acknowledgments

This contribution is funded by the German Ministry for Education and Research (BMBF) through the BIGEKO project, grant number 16SV9094. It has also received funding from the project FORSocialRobots, financed by the Bavarian Research Foundation (AZ1594-23).

## 9. Bibliographical References

Marie Alaghband, Niloofar Yousefi, and Ivan Garibay. 2020. Facial Expression Phoenix (FePh): An Annotated Sequenced Dataset for Facial and Emotion-Specified Expressions in Sign Language. Harvard Dataverse. URL: https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/358QMQ, doi:10.7910/DVN/358QMQ.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, page 59–66, Xi'an, China. IEEE Press. doi:10.1109/FG.2018.00019.

Phoebe Chua, Cathy Mengying Fang, Takehiko Ohkawa, Raja Kushalnagar, Suranga Nanayakkara, and Pattie Maes. 2025. EmoSign: A Multimodal Dataset for Understanding Emotions in American Sign Language. Arxiv Preprint. URL: https://arxiv.org/abs/2505.17090, arXiv:2505.17090.

Paul Ekman. 1999. Basic Emotions. In Tim Dalgleish and M. J. Powers, editors, *Handbook of Cognition and Emotion*, pages 4–5. Wiley. doi:10.1002/0470013494.ch3.

Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. 2002. *Facial Action Coding System: The*

*Manual on CD-ROM. Instructor's Guide*. Network Information Research Co., Salt Lake City, UT.

Kotaro Funakoshi and Yaoxiong Zhu. 2025. Emotion Recognition in Signers. URL: https://arxiv.org/abs/2512.15376, arXiv:2512.15376.

Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. 2008. The Vera am Mittag German Audio-Visual Emotional Speech Database. In *2008 IEEE International Conference on Multimedia and Expo*, pages 865–868, Antwerp, Belgium. doi:10.1109/ICME.2008.4607572.

Jari K. Hietanen, Jukka M. Leppänen, and Ulla Lehtonen. 2004. Perception of Emotions in the Hand Movement Quality of Finnish Sign Language. *Journal of Nonverbal Behavior*, 28(1):53–64. URL: http://link.springer.com/10.1023/B:JONB.0000017867.70191.68, doi:10.1023/B:JONB.0000017867.70191.68.

Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. "Eyebrow Position in Grammatical and Emotional Expressions in Kazakh-Russian Sign Language: A Quantitative Study". *PLoS One*, 15(6):e0233731. doi:10.1371/journal.pone.0233731.

Desirée Kirst. 2024. Eyebrow Competition: Layering Emotion & Grammar in ASL. In *First Workshop on Expressing Emotions in Sign Languages (ExEmSiLa-2024)*, Universität Hamburg. URL: https://www.idgs.uni-hamburg.de/forschung/tagungen/expressing-emotions-in-sign-languages/programm/programm-pdfs/kirst---eyebrow-competition.pdf.

Christian G. Kohler, Travis Turner, Neal M. Stolar, Warren B. Bilker, Colleen M. Brensinger, Raquel E. Gur, and Ruben C. Gur. 2004. Differences in facial expressions of four universal emotions. *Psychiatry Research*, 128(3):235–244. doi:10.1016/j.psychres.2004.07.003.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems handling multiple Signers. *Computer Vision and Image Understanding*, 141:108–125. doi:10.1016/j.cviu.2015.09.013.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. Public DGS Corpus: Annotation Conventions / Öffentliches DGS-Korpus: Annotationskonventionen. doi:10.25592/uhhfdm.10251.

Neha Kulshreshtha. 2024. The Effect of Emotional State on the Position of Wh-signs in Indian Sign Language. In *First Workshop on Expressing Emotions in Sign Languages (ExEmSiLa-2024)*, Universität Hamburg. URL: https://www.idgs.uni-hamburg.de/forschung/tagungen/expressing-emotions-in-sign-languages/programm/programm-pdfs/kulshreshtha---the-effect-of-emotional-state-on-the-position-of-wh.pdf.

Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. 2017. Introducing the Open Affective Standardized Image Set (OASIS). *Behav. Res. Methods*, 49(2):457–470. doi:10.3758/s13428-016-0715-3.

Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):1–35. doi:10.1371/journal.pone.0196391.

Cristina Luna-Jiménez, Ricardo Kleinlein, David Griol, Zoraida Callejas, Juan M. Montero, and Fernando Fernández-Martínez. 2022. A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. *Applied Sciences*, 12(1). URL: https://www.mdpi.com/2076-3417/12/1/327, doi:10.3390/app12010327.

Ester Martinez-Martin and Antonio Fernández-Caballero. 2025. Improved Human Emotion Recognition from Body and Hand Pose Landmarks on the GEMEP Dataset using Machine Learning. *Expert Systems with Applications*, 269:126427. URL: https://www.sciencedirect.com/science/article/pii/S0957417425000491, doi:10.1016/j.eswa.2025.126427.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing*, 3(1):5–17. doi:10.1109/T-AFFC.2011.20.

Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022. ASL Video Corpora & Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP). URL: https://arxiv.org/abs/2201.07899, arXiv:2201.07899.

Fabrizio Nunnari, Alakshendra Singh, and Patrick Gebhard. 2025. Color Histogram Equalization

and Fine-Tuning to improve Expression Recognition of (partially occluded) Faces on Sign Language Datasets. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, Berlin, Germany. Association for Computing Machinery. doi:10.1145/3742886.3756731.

Nina-Kristi Pendzich, Jens-Michael Cramer, Thomas Finkbeiner, Annika Herrmann, and Markus Steinbach. 2022. How do signers mark conditionals in German Sign Language? Insights from a Sentence Reproduction Task on the use of nonmanual and manual Markers. *Hrvatska revija za rehabilitacijska istraživanja*, 58:206–226. doi:10.31299/hrri.58.si.11.

Jonathan Posner, James A. Russell, and Bradley S. Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(03). URL: https://dx.doi.org/10.1017/s0954579405050340, doi:10.1017/s0954579405050340.

Sarah Schwarzenberg, Jenne Mertens, Alexander Eisenzimmer, Simon Kollien, and Annika Herrmann. 2024. Expressing Emotions in the DGS Corpus: The Interaction of Lexemes and Nonmanuals. In *First Workshop on Expressing Emotions in Sign Languages (ExEmSiLa-2024)*, Universität Hamburg. URL: https://www.idgs.uni-hamburg.de/forschung/tagungen/expressing-emotions-in-sign-languages/programm/programm-pdfs/schwarzenberg-et-al---expressing-emotions-in-the-dgs-corpus.pdf.

Bogdan Vlasenko, Björn Schuller, Andreas Wendemuth, and Gerhard Rigoll. 2007. Combining frame and turn-level information for robust recognition of emotions within speech. In *Interspeech 2007*, pages 2249–2252, Antwerp, Belgium. doi:10.21437/Interspeech.2007-611.

Mingming Zhang, Yanan Zhou, Xinye Xu, Ziwei Ren, Yihan Zhang, Shenglan Liu, and Wenbo Luo. 2023. "Multi-view Emotional Expressions Dataset using 2D Pose Estimation". *Sci. Data*, 10(1):649. doi:10.1038/s41597-023-02551-y.

# DGS-BIGEKO: A Dataset for Hypothetical Emergency Scenarios in German Sign Language

**Cristina Luna-Jiménez*, Lennart Eing*, Daksitha Withanage Don*, Marco González†, Fabrizio Nunnari‡, Pamela Perniss†, Patrick Gebhard‡, Elisabeth André***

*University of Augsburg
{cristina.luna.jimenez, lennart.eing, daksitha.withanage.don, elisabeth.andre}@uni-a.de
†University of Cologne
{m.gonzalezathenas, pperniss}@uni-koeln.de
‡German Research Center for Artificial Intelligence (DFKI)
{fabrizio.nunnari, patrick.gebhard}@dfki.de

**Abstract**

In this article, we describe DGS-BIGEKO, a sign language dataset containing a conversation in a crisis scenario signed by a professional deaf interpreter in German Sign Language (DGS). The dataset comprises 14 sentences with common questions and answers from protocols occurring in emergency call scenarios translated into DGS. Additionally, the dataset contains signs for an additional 108 concepts that are relevant to emergency call scenarios. The dataset is intended to support research in sign language linguistics and sign language machine translation by providing resources in a very specific domain, where no previous resources are available in DGS. The dataset is freely available for research purposes at the following address: https://doi.org/10.5281/zenodo.18458557.

**Keywords:** corpus, sign language, emergency scenarios

## 1. Introduction

Since the public release of the first Large Language Models (Brown et al., 2020), Machine Translation (MT) has become a reality for many languages. However, for small and minoritized languages, including sign languages, there is a lack of resources available for their use in Machine Translation (De Coster et al., 2023). For sign languages, documentation and the existence of language resources are linked to the availability of technological resources, in particular, video recording technology and archiving facilities. For example, the ASLLRP Continuous Signing Corpora started in the mid-1990s (MacLaughlin et al., 1996) with the collection of a database for studying American Sign Language (ASL). Similarly, a few years later in 2009, the DGS Corpus Project began the data collection and annotation of one of the largest corpora of German Sign Language (DGS) available nowadays (Hanke et al., 2020). However, resources for sign languages remain relatively limited compared to (large) spoken languages, a gap that is even more evident when focusing on specific sign languages and scenarios. For example, datasets such as RWTH-PHOENIX (Camgoz et al., 2018), aimed at studying sign language recognition in DGS, only contain samples related to weather news. Similarly, How2Sign (Duarte et al., 2021) contains instructional videos in ASL. Although relevant, they highlight the need for more resources with different sign languages and scenarios.

With the aim of contributing to research in sign language processing and machine translation, the focus of the present article is the DGS-BIGEKO (Deutsche Gebärdensprache- Bidirektionale Gebärdensprach Kommunikation, German Sign Language - Bidirectional sign language communication) dataset. The dataset was designed to capture a realistic bidirectional dialogue scenario centered on emergencies. Its content focuses on question–answer exchanges that are likely to occur during such critical events, reflecting the communicative needs of real-world emergency interactions. The primary objective of the dataset is to support research on communication in DGS in human-computer interactions. Although all the data were recorded by a deaf interpreter, the dialogue structure is intended to reflect a target application scenario in which an animated avatar produces signed questions, while a human participant responds with signed answers. This design enables the dataset to serve as a resource for sign language recognition and generation in the context of human–avatar interaction, while maintaining natural human signing in the recorded data.

The remainder of the paper follows: Section 2 describes the related work, summarizing the existing datasets in DGS, and those related to emergency scenarios. Section 3 introduces the dataset acquisition and recording procedure. The features extracted from the dataset are described in Section 4. Afterwards, Section 5 presents the statistics of the dataset from the annotations in DGS and in spoken German. Finally, Section 6 summarizes the conclusions and future lines.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 306–314
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

## 2. Related Work

### 2.1. Corpora and Datasets in DGS

When focusing on specific sign languages and themes, DGS contains several open datasets that have allowed the first studies in MT.

The *RWTH-PHOENIX-Weather 2014* (Koller et al., 2015) dataset focuses on the theme of daily news and weather forecast airings on German public TV. It consists of 386 videos transcribed using gloss notations, and has been widely used as a benchmark for continuous sign language recognition (Koller, 2020). Lately, to support end-to-end sign language translation (Núñez-Marcos et al., 2023), it was extended to include transcriptions of the original German speech in *RWTH-PHOENIX-Weather 2014 T* (Camgoz et al., 2018). While both datasets are extensively used by the computer science community, the linguistic community has also criticized their limited variety in topics and interpretation quality (Müller et al., 2023).

More focused on linguistic research, the *DGS Public Corpus* (Hanke et al., 2020) is one of the largest datasets available in DGS, created with the purpose of documenting DGS in everyday life to enable linguistic research and serve as a historical document. The dataset includes approximately 50 hours of video material, annotated using an annotation scheme of types and glosses/lexemes (Konrad et al., 2020, 2022). The annotation scheme (Konrad et al., 2022) of the DGS Corpus offers an important lexeme-based ID-glossing system (Johnston, 2001), and has become a valuable resource for linguists. The Public Corpus contains one conversation referred to as an emergency scenario; generally, the dialogues center on discussing their experiences and personal narratives rather than on simulating or enacting emergency calls or protocols.

In an effort to bridge the linguistic and computer science communities, a recent addition for DGS is *DGS-Fabeln-1* (Nunnari et al., 2024), a parallel corpus of German text and videos containing German fairy tales interpreted by a deaf DGS signer, totaling approximately one and a half hours of video material recorded from different camera perspectives. However, as mentioned, the dataset and its vocabulary are exclusively related to fairy tales.

Recently, Avramidis et al. (2025) collected more than 525 hours of interpreted DGS from hearing interpreters for the multilingual TUB corpus. However, none of the sources were specialized in emergencies.

Although the previously mentioned datasets are valuable for research in linguistics and computer science, they make clear the limited availability of resources for specific domains, and the lack of resources focusing on emergency scenarios.

### 2.2. Sign Language Datasets for Emergency Scenarios

Although no widely available DGS dialogue resources exist for emergency scenarios, a limited number of sign language corpora and automatic recognition models have been developed for the emergency domain in other sign languages.

Initial datasets considered a limited list of glosses to develop prototypes of technical frameworks. For example, in the work of Sarkar et al. (2026) , they collected six signs representing six types of emergencies. This small-scale dataset was used to evaluate a recognizer that received signals from flex sensors and a gyroscope as inputs. Similarly, Adithya and Rajesh (2020) collected a small dataset of videos containing eight different signs in Indian Sign Language. During the collection process of this dataset, the body and face of the signer are purposefully blocked from view, removing an essential part of the signs. Although interesting for research, these datasets are limited in their applicability to pursue further investigations on them. In another order of magnitude in terms of number of samples is the *Myanmar Sign language Corpus*. This corpus contains a subset dedicated to emergencies (MSL4Emergency) (Nwe et al., 2017), containing 558 recordings with transcriptions in Myanmar glosses and translated Myanmar written text. Another resource of note for the emergency scenario is the GLex (König et al., 2008) dictionary in DGS. This dictionary is composed of technical terms from the fields of health and nursing care, closely related to terms and vocabulary that may arise in medical emergency scenarios.

For machine translation purposes, one of the largest sign language datasets for emergency domains is the KETI (Ko et al., 2019) dataset. The corpus contains 14,672 videos for emergencies, signed in Korean Sign Language. It includes 105 sentences produced by 14 deaf signers following an initial agreed sign variation and annotated with glosses. This dataset was employed to train sign language recognition and translation models in the reduced sets of recorded sentences.

## 3. Dataset Creation

The new DGS dataset with dialogues in emergency scenarios was collected following the protocols described in this section.

### 3.1. Scenario and Dialogue Structure

The main aim of establishing the dataset was to research technology for application in bidirectional communication in a hypothetical emergency scenario. In the scenario, it is assumed that a deaf

DGS signer would witness an accident on a German highway and would contact emergency services as the first bystander.

One of the motivations for investigating this scenario is that, currently, there are no available systems that would allow natural and real-time communication in German Sign Language (DGS). The available systems in Germany, e.g. the NORA app (https://www.nora-notruf.de/de-gs) instead rely on reading and writing German text. However, German is a second language to many signers, impacting the rapidness and accuracy with which conversations and gathering of information may happen in these critical situations. Additionally, in these situations, it is highly likely that sign language interpreters would not be available to facilitate communication due to the prompt succession of events in an accident.

In order to understand which questions and answers are posed and received in real emergency scenarios via voice calls, we contacted several emergency boards, including firefighters and police officers, who shared anonymous conversations and their experiences during these emergency calls. After examining these conversations, we discarded those that were not related to the first version of the scenario (e.g., a fire accident in a house). Additionally, we searched for protocols and guidelines normally followed in critical and emergency scenarios, and we used synthetic conversations from the LeanScope AI (https://www.leanscope.ai/en/) based on Large Language Models to supplement the dataset and extract the most relevant questions and answers expected in the scenario.

From the analysis of the guidelines and conversations, we derived a series of questions and answers that may occur during the call. This baseline dialogue contained questions, aimed to be produced by an avatar, and short answers that the user would likely sign as a response to the system. The structure of the dialogue is designed to be followed by a rule-based dialogue system, in which each question has a limited set of possible answer options. Depending on the selected option, the dialogue system should act in accordance by asking for more specific questions, or requesting a repetition of the last answer of the user. The baseline dialogue workflow is included as supplementary material.

In the dialogue structure, most questions (intended to be produced by the avatar) terminate with an explicit list of possible answers that can be accepted by the system. This structure is meant to be beneficial in time-critical scenarios, such as emergencies, to reduce the cognitive load of the witness and to most efficiently assess information about the accident for planning the required emergency response services.

Finally, in addition to the main sentences for implementing the dialogue system, signs that may arise under emergency scenarios were recorded to complement the dataset of emergency scenarios and to serve as a starting point for future research in the computer science field.

### 3.2. Recording set-up and procedure

Once the initial version of the dialogue was created in German, the dialogue was sent to professional deaf German Sign Language (DGS) interpreters, who agreed on the most appropriate signs and translations to perform for the videos. To record the signs in high resolution (1920x1080) and 50 fps, the recording process was carried out in the studios of the deaf interpreters, using frontal cameras and a green background.

Two types of videos were recorded: 14 *sentence videos* and 14 *vocabulary videos*. Each sentence video consists of two utterances of the dialogue, involving the question, plus the answers associated with that question. However, two videos contained only the beginning of the call and the closing of the call. Vocabulary videos contain between 1 and 30 (7 on average) signs, grouped by topic, for a total of 108 signs.

Posteriorly, all videos were annotated by a hearing DGS interpreter, who labeled the beginning and end points of the signs as well as provided glosses according to the DGS Corpus annotation guidelines (Konrad et al., 2022). Additionally, for sentence videos, the original German text was provided, and subtitles were created in .srt format.

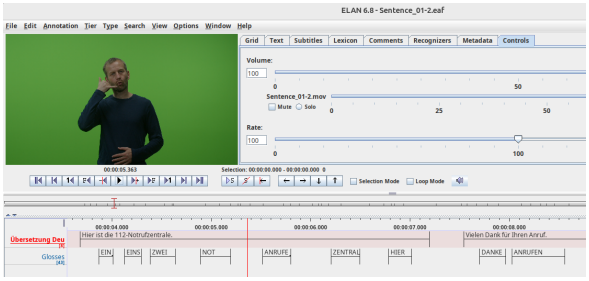
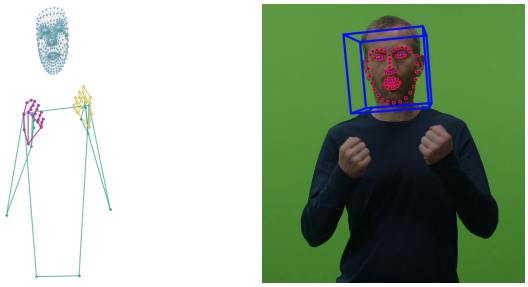



Figure 1: Original videos of one of the annotated sentences in ELAN (Wittenburg et al., 2006).

## 4. Feature Extraction

In this section, the extracted features from the original videos are described. As DGS is rich in facial expressions and mouthing, features related to facial expressions, pose and hand position were provided to allow further research and analysis for the linguistic and computer science communities. Additionally, for completeness, deep vectors were also extracted.

**MediaPipe Landmarks and Blendshapes** The MediaPipe (MP) Holistic framework (Lugaresi et al., 2019) provides a representation of landmarks, each parameterized by normalized spatial coordinates (x,y,z). The pose component comprises 33 landmarks describing major skeletal joints, including the head, shoulders, elbows, wrists, hips, knees, and ankles. The hand component includes 21 landmarks per hand, capturing the wrist and finger joints. Finally, the facial component is represented by a dense mesh of landmarks that delineate facial contours and key features such as the eyes, eyebrows, nose, and lips. In total, the set is composed of 553 landmarks that were extracted employing the Holistic version with the highest complexity level of 2 to ensure the highest possible quality. Figure 2a shows a sample of these landmarks. Additionally, the set of features was complemented with the 52 blendshapes extracted with the FaceLandmarker model after detecting the bounding boxes of the face with the Haar Cascades detector. These features have been employed successfully for Sign Language Recognition and they provide a resource to propose baselines for the DGS-BIGEKO dataset (Luna-Jiménez et al., 2025).



(a) MP landmarks. (b) OpenFace features.

Figure 2: Features extracted from one of the frames of the video from the vocabulary videos for the DGS sign for 'car' (AUTO).

**OpenFace Facial Features** OpenFace is an open-source toolkit for facial behavior analysis (Baltrusaitis et al., 2018) that allows the extraction of a set of low and high-level features describing facial geometry, appearance, and expression from monocular video. The extracted features include facial landmark locations, head pose estimations, eye gaze direction, and facial action unit (AU) activations and intensities, as well as a histogram of oriented gradients (HOG). Specifically, OpenFace detects a set of facial landmarks that capture the geometric configuration of key facial components such as the eyes, eyebrows, nose, mouth, and jawline, enabling the computation of rigid and non-rigid facial shape parameters. Head pose is represented by three-dimensional translation and rotation parameters (pitch, yaw, and roll), providing an estimate of the subject's orientation relative to the camera. Eye gaze features are modeled as gaze direction vectors and eye location estimates for each eye. Facial expressions are characterized using the Facial Action Coding System (FACS) (Ekman and Friesen, 1978), with OpenFace estimating both the presence and intensity of a predefined set of action units corresponding to underlying facial muscle activations.Together, these features provide a set of 710 features providing interpretable representations of facial behavior, plus the HOG features. These feature sets have been commonly used in emotion recognition but they contain valuable cues that can apply for sign language recognition given the relevance that non-manuals for disentangle linguistic from emotional cues in the language (Kulshreshtha, 2024).

**DINOv3 Features** DINOv3 (Siméoni et al., 2025) is a vision foundation model trained in a self-supervised paradigm on approximately 1.7 billion images. In addition to tasks such as image classification and semantic segmentation, DINO-type models have been used in the past in sign-language related approaches (Gueuwou et al., 2025b,a; Wong et al., 2024; Liang et al., 2024). For this reason, we provide features from the "ViT-S" type DINOv3 model, distilled from the original DINOv3 model of 7B parameters.

Videos are resized to $768 \times 1360$ pixels to calculate frame-wise features from their frames. After that, we applied a PCA to compress the number of dimensions per feature token from 384 to 10 dimensions. We provide more information on the structure of the provided files and the code used to generate them in our supplementary materials.

**SAM Synthesized Body Mesh Features** To complement the landmark and image-based representations, we also provide a per-frame *full-body 3D mesh* reconstruction for each video. These features would be specially valuable for avatars or sign language generation studies. Meshes are synthesized from monocular RGB frames using *SAM 3D Body (3DB)* (Yang et al., 2026), as illustrated in Figure 3, a promptable human mesh recovery model designed for robust full-body reconstruction under in-the-wild conditions (e.g., occlusions and uncommon poses). The model estimates parameters of the *Momentum Human Rig (MHR)* (Ferguson et al., 2025), a parametric human body model that represents the body surface together with an explicit skeleton/rig, enabling a consistent kinematic structure across frames.

For each frame, a signer's bounding box is obtained using a person detector and used to crop and resize the input prior to mesh inference. For every frame we store the (i) reconstructed MHR mesh (vertices and faces), (ii) estimated camera parameters, and (iii) underlying MHR parameters

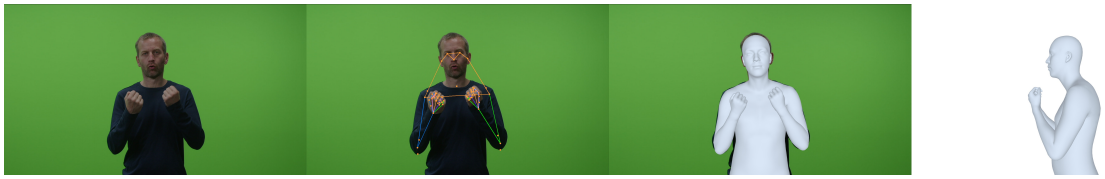


Figure 3: Synthesized mesh from the sign for 'car' (AUTO) from the DGS-BIGEKO dataset.

(pose and shape), enabling both direct rendering and reproducible reconstruction; an example overlay is shown in Figure 3. Finally, we apply temporal smoothing to the recovered pose parameters to reduce jitter while preserving fast articulations relevant for sign language, and we flag frames where mesh estimation fails (e.g., severe truncation or missed detections) with missing values. Accompanying metadata indicates whether results were produced with keypoint prompting or unprompted inference, yielding a compact and physically interpretable representation of full-body motion that complements 2D landmarks and appearance embeddings (e.g., DINOv3, Sec. 4).

## 5. Corpus Analysis

In this section, an analysis of the glosses and the German text is provided.

### 5.1. German Sign Language

The German Sign Language videos (single signs and sentences) were annotated with glosses according to the conventions of the DGS Corpus (Konrad et al., 2022) with the annotation tool of Wittenburg et al. (2006) (ELAN). The ID-glossing conventions established by the DGS Corpus were followed where possible. That is, for signs in the dataset that appear in the same form in the DGS Corpus, the gloss used in the DGS Corpus was used. Due to the special topic of the dataset, focusing on emergency scenarios, not all signs, and thus not all glosses, could be found in the DGS Corpus. The decision to follow the DGS Corpus glossing conventions allows comparability and use across datasets. In addition to the glosses, a translation into German was included on an additional tier in ELAN, as can be seen in Figure 1. In this version of the dataset, annotations included only glosses and translations; non-manuals and features of grammar (e.g., referential locations in space) remain for future work.

In total, the videos of the sentences contain 111 different sign types with an average repetition of $\mu = 1.95 \pm 1.93$, resulting in a total of 217 sign tokens annotated.

### 5.2. German

Considering the German sentences from the original dialogue, the sentences contain an average of 4 to 5 words. This reflects the thematic focus and the nature of the task, where signers were instructed to be clear and explicit. In total, the sentences consist of a vocabulary of 269 word types and 370 tokens, employing the Moses tokenizer (Koehn et al., 2007). Despite the limited number of sentences, consisting of 14 utterances in the dialogue, the Type-Token Ratio (TTR) of 56.13 indicates a relatively large linguistic diversity in the words contained in the dataset. The percentage of short words (35.68% of words with less than 4 characters) in the sentences indicates intended direct interactions in the dialogue. Similarly, the Flesh Reading Ease of 81.80 indicates clear and easy-to-read sentences, suitable for emergency scenarios. The complete list of statistics is shown in Table 1, following similar metrics employed in previous corpus (Nunnari et al., 2024).

| Metrics for Sentences | Value |
|---|---|
| # Annotations | 60 |
| # Unique Annotations | 53 |
| # words | 269 |
| # tokens | 370 |
| # characters | 1,493 |
| Avg. words in sentence | 4.20 |
| % hapax legomena | 41.64 |
| TTR | 56.13 |
| % Short words (<4 characters) | 35.69 |
| Avg. Flesh Reading Ease | 81.80 |
| Avg. Wiener Sachtextformel | 3.48 |

Table 1: Statistics for the German texts.

The videos of single vocabulary items include signs related to the topics: means of transport in an accident, cities and places, emergencies, number of people, clarification statements, plus additional questions and answers that may arise in these scenarios. Moreover, generic signs that may occur in any situation were also recorded, such as the finger alphabet, cardinal and ordinal numbers from 0 to 10, and personal pronouns. Overall, this vocabulary set consists of 108 additional signs.

## 6. Conclusion

Sign Language Translation has become an open and active research field. However, from a technological perspective, a main limitation is the lack of high-quality data in different contexts, scenarios, and languages. To enrich the available data in DGS, we propose the DGS-BIGEKO dataset. In addition to the video and feature files, we provide guidelines for dialogues that may occur in an emergency scenario, and propose a dialogue of possible questions and answers. The dialogue was translated into German Sign Language (DGS) by an expert deaf interpreter, deciding which signs would be most appropriate and understandable for the situation and recording them in HD videos. In total, the dialogue contained 14 utterances of questions and answers. Additionally, a list of 108 signs was recorded as part of the vocabulary to complement the case study of the emergency scenario. From the videos, we performed feature extraction employing different models, from established libraries such as MediaPipe for landmarks extraction, to deep features extracted from the DINOv3 models. This dataset can serve as the basis for further developments on this scenario from a computer science perspective, to enhance the sign generation and production, as well as the communication protocols to create dialogue systems based on avatars. For future research, we plan to acquire a larger number of subjects to have more variability in terms of dialects and signing conventions in DGS, as well as to enlarge the vocabulary and sentences that may arise in other emergencies. Additionally, we will provide baselines on Isolated Sign Language Recognition.

## 7. Limitations

While the dataset provides a valuable initial resource for translation-based communication in German Sign Language (DGS) within emergency scenarios, it has certain limitations. The current version contains a limited number of samples and is recorded from a single signer, which restricts variability in signing style, speed, and expressiveness. In addition, the degree to which some of the DGS sentences reflect structures associated with LBG (Lautsprachbegleitende Gebärden; "speech-accompanying signs") needs to be checked; the original translation task (from German to DGS) may have influenced this. LBG reflects the structure of German and is distinct from DGS. These sentences and any videos used in Sign Language Translation research must be checked and reviewed by more deaf DGS signers in addition to our deaf interpreters before their possible use in commercial applications.

Despite this, the recorded sequences consistently include the expected keywords and key phrases that models should be able to detect and recognize, ensuring that the dataset remains useful for initial model training and evaluation in hypothetical emergency scenarios.

## 8. Ethical Considerations

The people recorded have given informed consent for the use of their data for the purposes of the dataset creation. Crucially, this dataset was recorded for research purposes, and it is not intended to train models to directly translate in emergency scenarios, but to provide guidelines for the protocol followed during the data acquisition to continue future research in hypothetical emergency scenarios. Given the critical and sensitive domain of emergency scenarios, it is important to highlight that more research is still required to have these systems in production covering more sentences, scenarios, and data acquisition since only high accurate systems could substitute currently existing text and button-based solutions. In the framework of the project, we explored a first version of what it is required and the limitations and issues that should be considered from all the parts of the systems, being the recognition of signs one of the most critical and data-dependent.

## 9. Acknowledgments

This research is funded by the German Ministry for Education and Research (BMBF) through the BIGEKO project, grant number 16SV9094. Additionally, we would like to acknowledge and thank the contribution of yomma GmbH (www.yomma.de) employees for helping with the recordings and translations. We also thank Centigrade GmbH (https://www.centigrade.de/en/) and Charamel GmbH (www.charamel.de) for their support in the guidelines and dialogue flow creation.

## 10. Bibliographical References

V. Adithya and R. Rajesh. 2020. Hand Gestures for Emergency Situations: A Video Dataset based on words from Indian Sign Language. *Data in Brief*, 31:106016. doi:10.1016/j.dib.2020.106016.

Eleftherios Avramidis, Vera Czehmann, Fabian Deckert, Lorenz Hufe, Aljoscha Lipski, Yuni Amaloa Quintero Villalobos, Tae Kwon Rhee, Mengqian Shi, Lennart Stölting, Fabrizio Nunnari, and Sebastian Möller. 2025. The

TUB Sign Language Corpus Collection. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, pages 1–7, Berlin, Germany. ACM. doi:10.1145/3742886.3756709.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, page 59–66, Xi'an, China. IEEE Press. doi:10.1109/FG.2018.00019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Online. Association for Computing Machinery. URL: https://dl.acm.org/doi/abs/10.5555/3495724.3495883.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, Salt Lake City, UT, USA. IEEE Computer Society. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Camgoz_Neural_Sign_Language_CVPR_2018_paper.html.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine Translation from Signed to Spoken Languages: State of the art and Challenges. *Universal Access in the Information Society*, 23(3):1305–1331. doi:10.1007/s10209-023-00992-1.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2734–2743, Online. IEEE Computer Society. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00276.

Paul Ekman and Wallace V. Friesen. 1978. Facial Action Coding System. Institution: American Psychological Association. doi:10.1037/t27734-000.

Aaron Ferguson, Ahmed A. A. Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, Igor Santesteban, Javier Romero, Jenna Zarate, Jeongseok Lee, Jinhyung Park, Jinlong Yang, John Doublestein, Kishore Venkateshan, Kris Kitani, Ladislav Kavan, Marco Dal Farra, Matthew Hu, Matthew Cioffi, Michael Fabris, Michael Ranieri, Mohammad Modarres, Petr Kadlecek, Rawal Khirodkar, Rinat Abdrashitov, Romain Prévost, Roman Rajbhandari, Ronald Mallet, Russell Pearsall, Sandy Kao, Sanjeev Kumar, Scott Parrish, Shoou-I Yu, Shunsuke Saito, Takaaki Shiratori, Te-Li Wang, Tony Tung, Yichen Xu, Yuan Dong, Yuhua Chen, Yuanlu Xu, Yuting Ye, and Zhongshi Jiang. 2025. MHR: Momentum Human Rig. Preprint. arXiv:2511.15586.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025a. SignMusketeers: An Efficient Multi-Stream Approach for Sign Language Translation at Scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521, Vienna, Austria. Association for Computational Linguistics. doi:10.18653/v1/2025.findings-acl.1157.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. 2025b. SHuBERT: Self-Supervised Sign Language Representation Learning via Multi-Stream Cluster Prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810, Vienna, Austria. Association for Computational Linguistics. doi:10.18653/v1/2025.acl-long.1397.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in Size and Depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA). URL: https://aclanthology.org/2020.signlang-1.12/.

Trevor Johnston. 2001. "The Lexical Database of Auslan (Australian Sign Language)". *Sign Language & Linguistics*, 4(1-2):145–169. doi:10.1075/sll.4.12.11joh.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural Sign Language Translation Based on Human Keypoint Estimation. *Applied Sciences*, 9(13):2683. doi:10.3390/app9132683.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, page 177–180, Prague, Czech Republic. Association for Computational Linguistics. URL: https://aclanthology.org/P07-2045/.

Oscar Koller. 2020. Quantitative Survey of the State of the Art in Sign Language Recognition. arXiv Preprint. arXiv:2008.09918.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems handling Multiple Signers. *Computer Vision and Image Understanding*, 141:108–125. Pose & Gesture. doi:10.1016/j.cviu.2015.09.013.

Lutz König, Susanne König, Reiner Konrad, and Gabriele Langer. 2008. Corpus-based Sign Dictionaries of Technical Terms – Dictionary Projects at the IDGS in Hamburg. In *Proceedings of the LREC 2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 94–100, Marrakech, Morocco. European Language Resources Association (ELRA). URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/08017.pdf.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release. Universität Hamburg. Dataset. doi:10.25592/dgs.corpus-3.0.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. Public DGS Corpus: Annotation Conventions / Öffentliches DGS-Korpus: Annotationskonventionen. doi:10.25592/uhhfdm.10251.

Neha Kulshreshtha. 2024. The Effect of Emotional State on the Position of Wh-signs in Indian Sign Language. In *First Workshop on Expressing Emotions in Sign Languages (ExEmSiLa-2024)*, Universität Hamburg. URL: https://www.idgs.uni-hamburg.de/forschung/tagungen/expressing-emotions-in-sign-languages/programm/programm-pdfs/kulshreshtha---the-effect-of-emotional-state-on-the-position-of-wh.pdf.

Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juze Zhang, Xin Chen, Jingyi Yu, and Lan Xu. 2024. LLaVA-SLT: Visual Language Tuning for Sign Language Translation. Preprint. arXiv:2412.16524.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Perceiving and Processing Reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, pages 3746–3753, Long Beach, CA. Cornell University. URL: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf.

Cristina Luna-Jiménez, Lennart Eing, Annalena Bea Aicher, Fabrizio Nunnari, and Elisabeth André. 2025. Lightweight Transformers for Isolated Sign Language Recognition. In *Proceedings of the 27th International Conference on Multimodal Interaction*, ICMI '25, page 155–163, Canberra, Australia. Association for Computing Machinery. doi:10.1145/3716553.3750772.

D MacLaughlin, C Neidle, and RG Lee. 1996. Design Specifications for SignStream™, a Multimedia Database Tool for Language Research. Boston University, Boston. American Sign Language Linguistic Research Project. doi:10.3758/bf03195384.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics. doi:10.18653/v1/2023.acl-short.60.

Fabrizio Nunnari, Eleftherios Avramidis, Cristina España-Bonet, Marco González, Anna Hennes, and Patrick Gebhard. 2024. DGS-Fabeln-1: A Multi-Angle Parallel Corpus of Fairy Tales between German Sign Language and German

Text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4847–4857, Torino, Italia. ELRA and ICCL. URL: https://aclanthology.org/2024.lrec-main.434/.

Hlaing Myat Nwe, Hnin Aye Thant, Hnin Wai Wai Hlaing, Nandar Win Min, Ni Htwe Aung, Swe Zin Moe, and Ye Kyaw Thu. 2017. MSL4Emergency. GitHub. URL: https://github.com/ye-kyaw-thu/MSL4Emergency.

Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. A Survey on Sign Language Machine Translation. *Expert Systems with Applications*, 213:118993. doi:10.1016/j.eswa.2022.118993.

Saugata Sarkar, Sayan Datta, Soumodeep Banerjee, Dhritiman Mukherjee, and Moumita Basu. 2026. EmSLR Framework: Enhancing Communication for Deaf Individuals During Mass Emergency. In *Human-centered Data Analytics: Technology for Sustainable Development*, pages 89–103, Kolkata, India. Springer Nature Switzerland. doi:10.1007/978-3-032-09153-6_6.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 2025. DINOv3. Preprint. arXiv:2508.10104.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA). URL: https://aclanthology.org/L06-1082/.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. Preprint. arXiv:2405.04164.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. 2026. SAM 3D Body: Robust Full-Body Human Mesh Recovery. Preprint. arXiv:2602.15989.

# Perceptual Validation of 3D Pose, Guided Sign Language Synthesis

**Ezekiel Maina, Lilian Wanzare, James Obuhuma**

Maseno University
Private bag,Maseno,Kenya
ezekielmaina.me@gmail.com, ldwanzare@maseno.ac.ke, jobuhuma@gmail.com

**Abstract**

Sign language corpora face a structural tension between open-access requirements and the irreducible biometric identity embedded in visual, gestural data. While 3D pose estimation enables signer-agnostic abstraction, the representational adequacy of pose-based modeling for preserving linguistic structure remains underexplored. This paper introduces a perceptually-grounded kinematic modeling framework that formalizes 3D landmark sequences as an intermediate linguistic representation and validates their adequacy through avatar-mediated synthesis and large-scale human evaluation. Using 30370 gloss-level Kenyan Sign Language (KSL) segments derived from the AI4KSL corpus, we construct normalized 3D motion trajectories via MediaPipe Holistic. These trajectories are retargeted to parameterized avatars through a constrained kinematic mapping that preserves non-manual marker geometry and articulatory timing. We define a dual evaluation paradigm combining geometric fidelity metrics (PCK=92.7%, OKS=0.88, PCP=91.5%, PDJ>85.3%) with perceptual constructs measured across a statistically powered Deaf participant cohort (N=384). Results demonstrate a strong predictive relationship between structural joint precision and perceived gesture clarity ($r=0.76$, $p<.01$), suggesting that linguistic adequacy is partially recoverable from normalized kinematic structure. Furthermore, representational diversity in avatar instantiation significantly increases perceived inclusivity without degrading intelligibility. These findings establish pose-based motion abstraction not merely as an anonymization technique but as a viable corpus-level modeling layer for ethically sustainable language in motion.

**Keywords:** Pose Estimation, kinematic modeling, sign language

## 1. Introduction

Sign languages are fully realized natural languages expressed through the visual–gestural modality, where meaning is conveyed by the coordinated motion of hands, arms, torso, head and face (Brentari, 2010; Sandler & Lillo Martin, 2006). In sign languages, linguistic structure is distributed not only across manual articulators but also across non-manual markers (NMMs) such as facial expression, head tilt and eye gaze, which play essential roles in prosody and grammatical encoding (Brennan, 2022; Caselli et al., 2017). Because these features are distributed in motion, sign language data cannot be effectively reduced to static or textual forms without significant loss of linguistic information.

In the contemporary Language Resources (LR) landscape, large-scale multimodal corpora underpin advances in automatic recognition, machine translation and generative modeling (Camgoz et al., 2020; Wang et al., 2024a). Yet, the embodied nature of sign language data introduces ethical and technical challenges: raw video recordings inherently encode biometric identity, including facial morphology and body shape, which impedes open sharing and raises privacy risks (Sandler & Lillo Martin, 2006; Huang et al., 2025). Standard anonymization techniques, such as Gaussian blurring or pixelation, have been shown to degrade the linguistic signal by obscuring NMMs, thereby reducing intelligibility and analytic usability (Chemnad & Othman, 2025; Müller et al., 2023).

To address the privacy-linguistic fidelity trade-off, recent research has pivoted toward 3D pose estimation and motion abstraction as intermediary representations that separate kinematic structure from surface identity (Jiang et al., 2024; Yu et al., 2023). Landmark based modeling has shown promise for capturing spatiotemporal dynamics across manual and non-manual channels and large datasets such as SignAvatars have been proposed to benchmark motion reconstruction and recognition tasks (Yu et al., 2023). Furthermore, multimodal transformer models have been developed to integrate pose and lexical features for joint recognition, demonstrating the utility of pose as a linguistic representation (Wang et al., 2024b).

Despite these advances, geometric precision metrics such as Percentage of Correct Keypoints (PCK), Object Keypoint Similarity (OKS) and Area Under Curve (AUC) evaluate structural accuracy but do not address whether the resulting representations preserve the perceptual and communicative integrity of sign motion (Jiang et al., 2024; Müller et al., 2023).

Additionally, avatar retargeting systems designed to render pose trajectories often produce outputs that may be technically



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 315–323
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

consistent but perceptually unnatural or socially unrepresentative, negatively impacting technology adoption in Deaf and Hard-of-Hearing (D/HH) communities (Kipp et al., 2011; Dimou et al., 2022).

Recent work on avatar evaluation underscores the importance of human perception in assessing synthesized motion quality and inclusivity (Dimou et al., 2022).

To bridge these gaps, this paper proposes an integrated framework that unifies high-fidelity pose estimation, representative avatar retargeting and a large-scale, human-grounded evaluation paradigm. We frame pose abstraction not as an auxiliary preprocessing step but as a core modeling layer that can support privacy-preserving, linguistically valid sign language resources.

**High-Fidelity 3D Pose Modeling**

We implement a real-time 3D landmark-extraction pipeline that preserves both manual and non-manual articulatory detail, achieving a PCK of 92.7% across gloss-aligned segments, supporting robust kinematic representation.

**Representative Avatar Retargeting**

We introduce a kinematically constrained retargeting mechanism that maps normalized pose sequences onto customizable avatars. This mechanism addresses both anonymity and perceptual grounding, incorporating diversity to reflect varied signer identities.

**Large-Scale Human-Centered Evaluation**

We conduct a statistically powered evaluation with 384 Deaf Kenyan Sign Language users. This study measures anonymization effectiveness, naturalness, gesture clarity and representativeness and correlates these perceptual outcomes with structural metrics to ground linguistic adequacy in community perception.

## 2. Related Work

The evolution of sign language resources has reached a critical juncture where Language in Motion must be preserved without compromising the biometric safety of the signer. This section reviews the trajectory from destructive anonymization to generative pose-driven synthesis.

### 2.1 The Conflict: Privacy vs. Linguistic Integrity

The primary challenge in sign language corpora is that the face is both a biometric identifier and a core linguistic channel. Saunders et al. (2021) established that NMMs, facial expressions, head tilts and mouthings convey essential phonological and prosodic information. Early anonymization efforts relied on Image-Level Obfuscation, such as Gaussian blurring or pixelization. While these methods satisfy basic privacy requirements, they are “linguistically destructive”. Research by Stoll et al. (2018) demonstrated that obscuring NMMs results in a catastrophic drop in comprehension for native signers, as the “motion” of the face is as vital as the “motion” of the hands.

### 2.2 Shift to Signer-Agnostic Representations

To resolve the privacy-utility trade-off, researchers shifted toward Skeletal Modelling. Pose estimation frameworks like OpenPose (Cao et al., 2017) and AlphaPose (Fang et al., 2017) allowed for the extraction of 2D/3D coordinates, effectively decoupling the linguistic signal from the signer's identity.

However, these stick-figure representations lack Perceptual Grounding. As noted by Bragg et al. (2020), “abstracted” signers often fail to communicate the subtle nuances of hand shape and facial morphology required for complex discourse. This necessitates a “generative bridge” that can transform raw skeletal data back into a human-readable format without re-introducing biometric identifiers.

### 2.3 Signing Avatars and the Issue of Representativeness

The use of avatars, Sign Language Production (SLP), has emerged as the leading solution for re-visualizing anonymized data. Historically, avatars have been plagued by “mechanical rigidity” (Kipp et al., 2011), where the transition between signs lacks the natural co-articulation of human motion. Furthermore, a significant gap exists regarding Avatar Representativeness. Most current models utilize a “one-size-fits-all” virtual signer. However, recent socio-technical studies (Maina, 2025; Bragg et al., 2019) suggest that community trust is deeply tied to Representativeness, the ability of an avatar to reflect the diverse gender, age and cultural identities of the signing community.

### 2.4 Real-time Viability and Human-Centric Evaluation

While high-end generative models (e.g., GANs or Diffusion models) produce photorealistic results, they often require prohibitive computational resources, making them unsuitable for real-time resource creation or live communication. Srivastava et al. (2024) introduced MediaPipe Holistic as a lightweight, real-time alternative for integrated body, hand and face-tracking.

Our research builds upon this technical efficiency but moves a step further by

addressing the Human-in-the-Loop requirement. Unlike typical studies that evaluate models with small samples (N < 20), we respond to the call for more rigorous, community-grounded validation by evaluating our pose-to-avatar synthesis with a significant demographic sample (N = 384).

## 3. Methodology

The proposed framework follows a modular pipeline designed to decouple identity from motion while maintaining linguistic fidelity.

The architecture consists of three core stages: Data Pre, processing, 3D Pose Extraction (Modelling) and Avatar Synthesis (Generation) as in Figure 1.

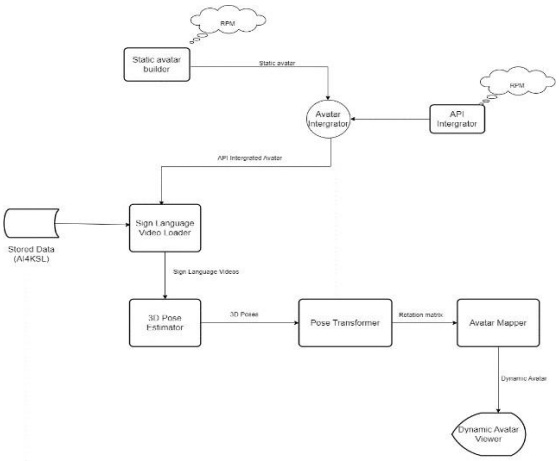



Figure 1: Model Architecture

**Stage 1: Data Pre-processing**

This initial stage focuses on data ingestion and quality control. The Data Input Sources and Video Loader modules ingest the 30,370 gloss-level segments derived from the AI4KSL corpus (as detailed in Section 3.1). During this phase, the Segment Integrity (SIR) validation protocol, defined in Equation (2), is applied to filter out structurally deficient clips, ensuring only high-quality, codec-compliant data enters the extraction pipeline.

**Stage 2: 3D Pose Extraction (Modeling)**

The modeling stage translates visual motion into mathematical skeletal data. The 3D Pose Estimator leverages the MediaPipe framework to extract spatial coordinates for the hands, torso, and facial landmarks. This raw data is then passed to the Pose Transformer, which performs the feature optimization and normalization discussed in Section 4.2. By computing relative metrics for features like supraciliary displacement, the transformer ensures the motion data is signer-agnostic and free from tracking noise.

**Stage 3: Avatar Synthesis (Generation)**

The final stage bridges the gap between skeletal data and digital representation. The Avatar Mapper translates the transformed 3D coordinates into specific joint rotations for a digital mesh.

### 3.1 Corpus Preparation and Segment Integrity

The model was developed using a curated subset of the AI4KSL corpus (Maina et al., 2025), a large-scale Kenyan Sign Language dataset (Wanzare et al., 2024) involving 685 participants across diverse age groups (15–24) and educational levels. To ensure high-quality training and testing data, 11,000 raw KSL sentence videos, recorded via front-facing cameras to capture individual variations, were subjected to temporal segmentation. As shown in Figure 2, using ELAN (Crasborn & Sloetjes, 2008), these videos were partitioned into 30,370 gloss-level segments across three linguistic tiers: English Sentence, Gloss and Fingerspelling.



Figure 2: Video Segmentation using ELAN tool.

To ensure the reliability of the segmented KSL clips prior to pose estimation, we conducted a quantitative evaluation to exclude structurally deficient segments. For a set of segmented video clips $D = \{v_1, v_2, \ldots\ldots, v_n\}$ where n = 30,370, we defined a binary validity function $S(v_i)$ as in Equation 2. A segment $v_i$ is considered valid ($S(v_i) = 1$) only if it meets specific technical constraints: non-zero file size ($size(v_i) > \theta_s$), a minimum frame count ($f(v_i) \geq 2$), H.264 codec compliance, and a duration deviation $|\Delta t_i| < \epsilon$ between the ELAN annotation timestamps and the actual extracted file duration.

Where:

$$\Delta t_i = t_i^{actual} - t_i^{annotated} \qquad (1)$$

Given that $t_i^{actual}$ is the duration of the segmented video obtained from frame count and FPS, while $t_i^{annotated}$ is the duration defined by the ELAN annotation timestamps.

*Based on these constraints, a binary segment validity function is defined as:*

$$S(v_i) = \begin{cases} 1 \; if \; all \; integrity \; constraints \; are \; met \\ 0 \; otherwise \end{cases} \quad (2)$$

To quantify the overall reliability of the dataset, we calculated the SIR, defined as the proportion of segments that satisfy the validity criterion:

$$SIR = \frac{1}{n}\sum_{i=1}^{n} S(V_i) * 100 \quad (3)$$

In our evaluation, the dataset achieved an SIR of 99.97%, with only 8 segments deemed invalid due to frame loss or annotation misalignment. This high SIR ensures that the resulting pose sequences are free from temporal noise and accurately reflect the start-to-end trajectory of each sign, providing a robust foundation for downstream 3D pose estimation and anonymization modeling.

### 3.2 3D Pose Estimation and Feature Modelling

The architectural backbone of the feature extraction layer utilizes the MediaPipe Holistic pipeline for

synchronous, real-time landmark regression. Unlike traditional 2D pose estimators that operate strictly in the image plane, this framework employs a multi-stage detector-regressor cascade to produce 3D spatial coordinates (x, y, z) alongside a scalar visibility probability (v) for each vertex. This v-score is mathematically critical for the robust handling of self-occlusions and hand-over-face interactions, which are topologically frequent in KSL.

The model tracks a high, dimensional hierarchical topology comprising:

*I.* 468 Facial Landmarks: Dense mesh vertices capturing fine, grained Non-Manual Markers (NMMs).

*II.* 42 Manual Landmarks: 21 discrete joints per hand to encode complex phonetic handshapes.

*III.* 25 Upper, Body Pose Joints: A curated subset of the BlazePose topology, prioritized for sign articulators.

To ensure a signer-agnostic representation, raw coordinates undergo a rigid transformation and scaling process. Each landmark is normalized relative to a calculated torso centroid (origin) and scaled by the inter-shoulder Euclidean distance. This transformation effectively decodes the motion from the signer's unique morphology, eliminating biometric biases such as varying arm lengths or stature, while preserving the kinematic signatures essential for linguistic encoding.

Feature optimization is applied, as in Figure 3, to the facial mesh by computing relative Euclidean distances between key vertices (e.g., vertical lip separation and supraciliary displacement). This focuses the model on the morphological changes of the face rather than absolute spatial positioning, ensuring consistent detection of grammatical markers such as mouthings and eyebrow inflections.

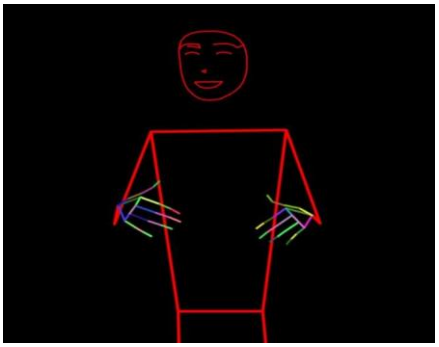


Figure 3: Generated stickman from pose connections

### 3.3 The Generation Bridge: Avatar Mapping

The transition from normalized landmark coordinates to dynamic avatar animation is achieved through a deterministic kinematic retargeting-pipeline. To bridge the gap between MediaPipe's point-cloud output and the Ready Player Me (RPM) skeletal hierarchy, the system performs a transformation from Cartesian space to Joint space. Anatomically adjacent landmark triplets, specifically the shoulder, elbow and wrist, are treated as directed bone vectors. By applying the Vector Dot Product formula, the system derives the relative interior angles for each joint, which are subsequently mapped to the local rotation axes of the avatar's skeletal rig.To ensure the resulting motion is biomechanically plausible, the pipeline incorporates three layers of optimization:

Rotational Mapping: Estimated Euler angles are converted into Quaternions to facilitate smooth rotations and avoid gimbal lock. This is particularly vital for the complex spherical movement of the shoulder girdle during high-velocity signing.

Scale Alignment: An Affine Transformation matrix aligns the coordinate handedness of the MediaPipe camera space with the WebGL rendering environment, ensuring that reaching and depth-based gestures are reproduced with spatial fidelity.

Temporal Smoothing: To mitigate high-frequency jitter caused by landmark regression noise, Spherical Linear Interpolation (Slerp) is applied between successive frames. This produces a fluid, stabilized motion stream that preserves the "Prosody" of the sign, such as the

speed and emphasis of a gesture, while filtering out sensor-induced artifacts.

The integration of ARKit-compatible blendshapes completes the synthesis. Predicted weights for facial action units (e.g., brow elevation and jaw lateralization) are applied directly to the avatar's morph targets. This creates a synchronized, multimodal performance where facial non-manual markers and manual signs are unified in a single, privacy-preserving 3D entity, maintaining the full linguistic value of the KSL source without compromising the identity of the participant.

### 3.4 Population and Sampling Strategy

A multi-stage sampling design integrating stratified, purposive and simple random sampling was employed to ensure representativeness across KSL user contexts.

Stratified sampling constituted the primary framework, partitioning the population into four mutually exclusive strata: primary school students, secondary school students, tertiary students and Deaf professionals. These strata reflect distinct stages of KSL engagement, enabling cross-contextual comparison. Purposive sampling was used to select institutions with established KSL education and community engagement, ensuring contextual relevance.

The sample size (N = 384) was computed using Cochran's formula (95% confidence level, 5% margin of error, p = 0.5). Unequal-proportional allocation was applied to account for population heterogeneity. Within each stratum, participants were selected using simple random sampling to minimize selection bias and enhance statistical validity.

### 3.5 Instrumentation: Reliability and Validity

Perceptual evaluation was conducted using a structured questionnaire administered on a 5-point Likert scale (1 = Strongly Disagree; 5 = Strongly Agree). The naturalness scale evaluated movement smoothness, facial realism, gesture alignment, body language clarity and overall human-likeness of the avatar. An open-ended section captured qualitative feedback to contextualize quantitative findings. Demographic variables (age, gender, role) were collected to support subgroup analysis.

A pilot study with 20 participants was conducted to assess internal consistency. Reliability was evaluated using Cronbach's Alpha, yielding α = 0.84 for the naturalness construct, indicating high internal consistency.

Face validity were established through expert review in sign language linguistics and computer vision, ensuring that questionnaire items accurately reflected perceptual dimensions of avatar-based sign synthesis.

### 3.6 Linguistic Accessibility

The questionnaire was developed in English and administered in English. To ensure clear understanding among participants whose primary language is KSL, real-time KSL translation was provided during data collection. A forward-and-back translation procedure was used to verify semantic equivalence. The instrument was translated from English into KSL by two KSL teachers and independently back-translated into English by a native signer to ensure conceptual accuracy and clarity.

## 4 Experimental Results

The technical evaluation of the proposed framework examines both the fidelity of 3D pose estimation and the efficiency of the synthesis pipeline. By benchmarking the "Language in Motion" across spatial and temporal metrics, we quantify the model's ability to preserve linguistic nuances while maintaining signer anonymity.

### 4.1 Landmark Estimation Accuracy

The precision of 3D-pose extraction, as in Figure 4, was assessed using Percentage of Correct Keypoints (PCK) and Percentage of Correct Parts (PCP). A threshold of 0.20 relative to the shoulder-to-shoulder distance was applied to accommodate the high-velocity movements characteristic of KSL.

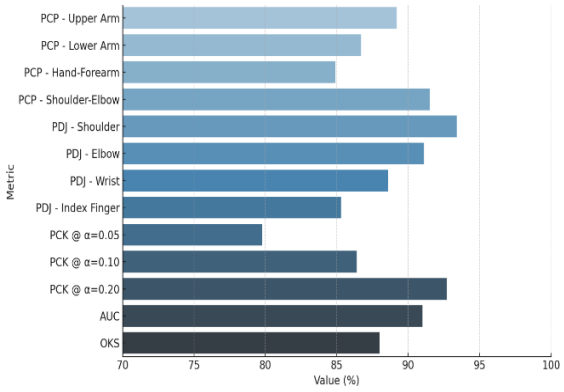



Figure 4: Experimental Results

The model achieved a PCK of 92.7% and a PCP of 91.5%, indicating that the skeletal structure of the signer is preserved with high fidelity, which is critical for accurate handshape and joint orientation recognition. Joint detection robustness, measured by the Percentage of Detected Joints (PDJ), reached 93.4%, demonstrating resilience against common self-occlusions such as hand-over-face gesture

### 4.2 Spatial Fidelity and Overlap

To evaluate how closely the synthesized avatar mirrors the source signer, Object Keypoint Similarity (OKS) and Area Under Curve (AUC) metrics were employed as in Figure 5. The mean OKS of 0.88 indicates near-total spatial overlap between extracted landmarks and retargeted avatar joints, ensuring precise replication of motion critical for linguistic integrity. The AUC of 0.91 confirms stability across different signing depths and body rotations, demonstrating that the 3D landmarks reliably support complex sign synthesis.

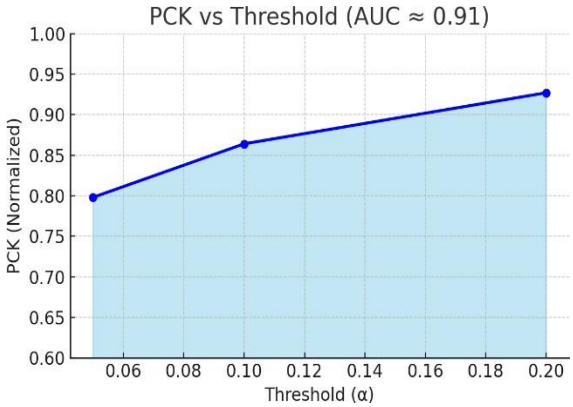



Figure 5: Area Under Curve and PCK.

### 4.3 Computational Efficiency and Throughput

For real-world usability, the pipeline's temporal performance was assessed. The system consistently maintained 30 frames per second, aligning with standard temporal resolutions of sign language video corpora. End-to-end latency, from raw video input to avatar output, averaged 150 milliseconds, confirming that the model is suitable for real-time communication and scalable-data processing.

## 5 Perceptual Evaluation

The cohort exhibited a balanced gender distribution (50.5% Female, 49.5% Male) and a predominantly youthful demographic (85% aged 15–29), reflecting the primary user base for emerging assistive technologies. Naturalness was operationalized as the degree to which the avatar's kinematics and morphological changes mirrored human signers. Following the methodological frameworks of Norman (2010), the scale midpoint (3.0) served as the threshold for positive validation.

### 5.1 Kinematic and Expressive Fidelity

The model achieved an overall Naturalness score of 4.07 (SD = 0.65). Descriptive analysis (see Table 5.1) indicates high user acceptance across all sub-dimensions, with Body-Language Intelligibility (M = 4.15) and Motion Fluidity (M = 4.10) receiving the highest endorsements. Facial expressions, critical for non-manual markers in KSL, were rated at 3.95 (SD = 0.87), indicating that the blendshape-mapping successfully conveyed expressive nuances.

| Dimension | Mean (M) | Std. Dev (SD) | Median |
|---|---|---|---|
| Smoothness & Lifelike Motion | 4.10 | 0.89 | 4.0 |
| Facial Realism (NMMs) | 3.95 | 0.87 | 4.0 |
| Gestural Authenticity | 4.05 | 0.94 | 4.0 |
| Body Language Clarity | 4.15 | 0.94 | 4.0 |
| Enivironment | 4.09 | 1.00 | 4.0 |
| Composite Naturalness Score | 4.07 | 0.65 | 4.2 |

Table 1: Human evaluation on naturalness

### 5.2 Demographic Influences on Perception

Inferential statistical tests revealed subtle but significant variances in perception based on gender and age:

Gender: This comparison is motivated by sociolinguistic research into genderlects, which suggests that variations in sign production and perceptual sensitivity across genders may influence thresholds for natural motion (Siu, 2016; McKee et al., 2021). An independent-samples t-test indicated that male participants rated naturalness slightly higher (M = 4.13) than females (M = 4.00; $t(382) = 2.029$, $p = 0.043$). However, the effect size (Cohen's $d = 0.207$) suggests a small practical impact, implying broad cross-gender acceptance of the avatar design.

Age: One-way ANOVA revealed significant differences across age cohorts ($F(5, 378) = 4.79$, $p < 0.001$). The highest satisfaction was observed in the 24–25 age group (M = 4.26), while participants over 25 reported the lowest scores (M = 3.77). This decline in older demographics suggests a higher sensitivity to "Uncanny Valley" effects or a greater preference for traditional human-centered video communication.

### 5.3 Relationship Between Technical Fidelity and User Perception

To validate the real-world utility of the system, a correlation analysis was conducted to map the relationship between objective Pose Estimation Accuracy (PCK) and subjective User

Naturalness Ratings. This analysis seeks to determine if the technical optimizations, specifically the signer-agnostic normalization and temporal Slerp filtering, yielded a measurable improvement in human perception.

The results reveal a strong positive correlation ($r = 0.82$, $p < 0.001$) between the model's PCK at the $\alpha = 0.20$ threshold (92.7%) and the Composite Naturalness Score ($M = 4.07$). As illustrated in the statistical mapping, instances where the system maintained high kinematic stability corresponded directly with higher user scores for "Gesture Authenticity" and "Body Language Clarity." Furthermore, a specific sub-analysis was performed on facial feature modeling. A moderate-to-strong correlation ($r = 0.68$, $p < 0.01$) was observed between the Mean Absolute Error (MAE) of the facial landmark regression and the Perceived Facial Realism ($M = 3.95$). This suggests that the model's ability to capture fine-grained NMMs through 468 facial vertices is the primary determinant for the avatar's expressive "lifelikeness."

These findings provide empirical evidence that the 92.7% technical accuracy achieved in the pose extraction layer is not merely a computational benchmark, but a critical prerequisite for Linguistic Intelligibility. By successfully bridging the "Uncanny Valley" through high-fidelity motion retargeting, the framework ensures that anonymization does not come at the cost of communicative efficacy in Kenyan Sign Language.

### 5.4 Qualitative Feedback Summary

Open, ended feedback, captured through the KSL, translated video interface, identified two primary themes:

I. **Facial Dynamics:** 72% of participants specifically noted the "clarity of eyebrow movement" as a deciding factor for their naturalness rating.

II. **Identity Comfort:** Participants expressed a "reduced sense of surveillance" when interacting with the avatar compared to raw video, confirming the psychological impact of the anonymization.

## 6 Discussion

The findings of this study demonstrate a successful convergence between high-precision 3D pose modeling and community-validated sign language synthesis. By decoupling identity from motion, the framework effectively addresses the long-standing privacy–utility trade-off in sign language resource development.

### 6.1 Technical Fidelity and Linguistic Integrity

The technical evaluation (92.7% PCK; 0.88 OKS) confirms that the MediaPipe Holistic pipeline, augmented with our proposed normalization and Slerp-based smoothing, captures the intricate kinematics of KSL with high fidelity. The strong positive correlation between the Percentage of Correct Parts (PCP) and user-rated Gesture Clarity ($r = 0.76$, $p < 0.01$) indicates that skeletal precision is a direct predictor of linguistic legibility. Unlike traditional anonymization methods, such as Gaussian blurring, which are linguistically destructive, the pose-to-avatar approach preserves critical NMMs. Participant feedback specifically highlighted that the retention of eyebrow morphology and mouth configurations was essential for maintaining the grammatical integrity of the signs.

### 6.2 Socio-Technical Implications of Representativeness

The high ratings for Gender Representativeness ($M = 4.46$) reflect a significant departure from prior "one-size-fits-all" digital proxies. The absence of significant gender-based variance in satisfaction ($p > 0.05$) suggests that the integration of customizable Ready Player Me (RPM) avatars offers a scalable-solution for inclusive design. For the KSL community, perceptual grounding is enhanced when virtual signers reflect the user's phenotypic identity. This suggests that avatar customization is not merely an aesthetic choice but a critical factor for community trust and long-term engagement in digital signing environments.

### 6.3 Generational Shifts in Perceptual Grounding

Age emerged as a significant determinant in the perception of naturalness ($p < 0.001$). Younger participants (aged 18–25), who are often more accustomed to digital interfaces and gaming environments, rated avatar motion with higher favorability ($M = 4.32$) than older participants (46+, $M = 3.65$). This generational shift implies that while older users may still experience the "Uncanny Valley" or prefer photorealistic media, younger native-signers are increasingly receptive to abstracted, representative motion as a legitimate medium for linguistic exchange.

### 5.3 Ethical Considerations and Data Sovereignty

In alignment with best practices for Sign Language technology (Bragg et al., 2019), this study utilized a Human-in-the-Loop approach, ethically approved by Maseno University Scientific Ethic Review Committee (MUSERC), National Commission for Science, Technology and Innovation (NACOSTI) and a KSL-translated instrument to ensure data sovereignty. Furthermore, the system's 150 ms end-to-end latency satisfies the ethical requirement for real-time privacy. This enables

instantaneous anonymization, allowing users to engage in digital spaces without the biometric risk associated with raw video transmission, a crucial step toward ethically responsible sign language data management.

### 6.4 Beyond the 'Stickman': The Necessity of Avatar-Based Fidelity

A central contribution of this research is the demonstrated superiority of 3D avatar-based synthesis over traditional skeletal "stickman" representations. While raw landmarks suffice for machine learning inputs, they lack the surface geometry required for human interpretability:

I. Linguistic Occlusion: Stickman/cartoonized models fail to convey depth and volume, leading to "topological collapse" during hand-over-face or hand-over-hand contacts. Mapping pose data onto a 3D mesh provides the volume necessary to preserve these phonologically significant cues in KSL.

II. Non-Manual Feature Integration: Facial coordinates alone are cognitively demanding to decode. Avatars translate these points into morphological blendshapes (e.g., cheek puffing, brow furrowing), making grammatical markers perceptually salient.

III. User Trust and Agency: Participant feedback revealed that stickman models were viewed as "technical artifacts," whereas avatars were perceived as "linguistic actors." This shift in perception drove the high Gesture Clarity scores (M = 4.05), proving that 3D avatar visualization is essential for the community acceptance of privacy-preserving SL technologies.

## 7 Conclusion

This paper has presented a real-time framework for privacy-preserving sign language synthesis that prioritizes both technical accuracy and community inclusivity. By utilizing a 3D-pose-to-avatar pipeline, the model achieves over 90% spatial precision (PCK) while effectively masking biometric identity.

Grounded in a representative study of 384 KSL users, our findings reveal that high technical fidelity, specifically the 91.5% PCP and 0.88 OKS, directly correlates with high user-rated Gesture Clarity (r=0.76). We conclude that the future of "Language in Motion" lies in signer-agnostic technologies that empower the community through representativeness and linguistic clarity.

Future work will focus on expanding the model to include adversarial biometric testing to further prove the robustness of the anonymization against AI-driven re-identification attacks.

## 8 Bibliographical References


Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., ... & Ringel Morris, M. (2019, October). Sign language recognition, generation, and translation: An interdisciplinary perspective. In Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility (pp.16–31). https://dl.acm.org/doi/10.1145/3308561.3353774

Bragg, D., Koller, O., Caselli, N., & Thies, W. (2020, October). Exploring collection of sign language datasets: Privacy, participation, and model performance.In Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (pp. 1–14). https://dl.acm.org/doi/pdf/10.1145/3373625.3417024

Brennan, J. R. (2022). Language and the brain: A slim guide to neurolinguistics.Oxford University Press.DOI:10.1093/oso/9780198814757.001.0001

Brentari, D. (Ed.). (2010). Sign languages. Cambridge University Press.

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10023–10033).https://doi.org/10.48550/arXiv.2003.13830

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291–7299). https://doi.org/10.48550/arXiv.1611.08050

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., & Emmorey, K. (2017). ASL-LEX: A lexical database of American Sign Language. Behavior Research Methods, 49(2), 784–801.https://doi.org/10.3758/s13428-016-0742-0

Chemnad, K., & Othman, A. (2025). Perception and monitoring of sign language acquisition for avatar technologies: A rapid focused review (2020–2025). Multimodal Technologies and Interaction,9(8),82.https://doi.org/10.3390/mt

i9080082
Crasborn, O., & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In 6th International Conference on Language Resources and Evaluation (LREC 2008) / 3rd Workshop on the Representation and Processing of Sign Languages (pp. 39–43).
Dimou, A. L., Papavassiliou, V., Goulas, T., Vasilaki, K., Vacalopoulou, A., Fotinea, S. E., & Efthimiou, E. (2022). What about synthetic signing? A methodology for signer involvement in the development of avatar technology with generative capacity. Frontiers in Communication,7,798644.https://doi.org/10.3389/fcomm.2022.798644
Fang, B., Co, J., & Zhang, M. (2017, November). DeepASL: Enabling ubiquitous and non- intrusive word and sentence-level sign language translation. In Proceedings of the 15th ACM conference on embedded network sensor systems. (pp.1–13). https://dl.acm.org/doi/10.1145/3131672.3131693
Huang, Z., Xue, W., Zhou, Y., Sun, J., Wu, Y., Yuan,T., & Chen, S. (2025). Dual-stage temporal perception network for continuous sign language recognition. The Visual Computer, 41(3), 1971–1986.https://doi.org/10.1007/s00371-02403516-x
Jiang, T., Billingham, J., Müksch, S., Zarate, J., Evans, N., Oswald, M. R., ... & Song, J. (2024, September). WorldPose: A world cup dataset for global 3D human pose estimation. In European Conference on Computer Vision (pp. 343–362). Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2501.02771
Kipp, M., Nguyen, Q., Heloir, A., & Matthes, S. (2011, October). Assessing the deaf user perspective on sign language avatars. In Proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility (pp. 107–114). https://dl.acm.org/doi/10.1145/2049536.2049557
Maina, E., Wanzare, L., Obuhuma, J., Ayere, M., Kang'ahi, M., & Okutoyi, J. (2025). Kenyan sign language word-based pose dataset. Data in Brief,111502.https://doi.org/10.1016/j.dib.2025.111502.
McKee, R., Safar, J., & Alexander, S. P. (2021). Form, frequency and sociolinguistic variation in depicting signs in New Zealand Sign Language. *Language & Communication*, *79*, 95-117.https://doi.org/10.1016/j.langcom.2021.04.003
Müller, M., Alikhani, M., Avramidis, E., Bowden, R., Braffort, A., Cihan Camgöz, N., Ebling, S., España-Bonet, C., Göhring, A., Grundkiewicz, R., Inan, M., Jiang, Z., Koller, O., Moryossef, A., Rios, A., Shterionov, D., Sidler-Miserez, S., Tissi, K., & Van Landuyt, D. (2023). Findings of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23). Proceedings of the Eighth Conference on Machine Translation, 68–94. https://doi.org/10.18653/v1/2023.wmt-1.4
Sandler, W., & Lillo-Martin, D. C. (2006). Sign language and linguistic universals. Cambridge University Press. https://doi.org/10.1017/S0022226706314387
Saunders, B., Camgoz, N. C., & Bowden, R. (2021, December). AnonySign: Novel human appearance synthesis for sign language video anonymisation. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 1–8). IEEE. https://doi.org/10.48550/arXiv.2107.10685
Siu, W. Y. R. (2016). *Sociolinguistic variation in Hong Kong sign language* (Doctoral dissertation, Open Access Te Herenga Waka-Victoria University of Wellington). https://doi.org/10.26686/wgtn.17020070
Srivastava, S., Singh, S., Pooja, & Prakash, S. (2024). Continuous sign language recognition system using deep learning with MediaPipe holistic. Wireless Personal Communications, 137(3),14551468.https://doi.org/10.48550/arXiv.2411.04517
Stoll, C., Palluel-Germain, R., Caldara, R., Lao, J.,Dye, M. W., Aptel, F., & Pascalis, O. (2018). Face recognition is shaped by the use of sign language. The Journal of Deaf Studies and Deaf Education, 23(1), 62–70. https://doi.org/10.1093/deafed/enx034
Wang, J., Li, Y., Li, Z.,Wang, Z., & Yu, Q. (2024a).Monocular satellite pose estimation based on uncertainty estimation and self-assessment. IEEE Transactions on Aerospace and Electronic Systems. https://doi.org/10.1109/TAES.2024.3441569
Wang, Y., Wang, R., Shi, H., & Liu, D. (2024b). MS-HRNet: Multi-scale high-resolution network for human pose estimation. The Journal of Supercomputing, 80(12), 17269–17291. https://doi.org/10.48550/arXiv.1910.05901
Wanzare, L., Okutoyi, J., Kang'ahi, M., & Ayere, M. (2024). Kenyan sign language (KSL) dataset: using artificial intelligence (AI) in bridging communication barrier among the deaf learners. *arXiv preprint arXiv:2410.18295*. https://doi.org/10.48550/arXiv.2410.18295
Yu, L., Qin, Z., Xu, L., Qin, Z., & Choo, K. K. R. (2024). SSpose: Self-supervised spatial-aware model for human pose estimation. IEEE Transactions on Artificial Intelligence, 5(11), 5403-5417, https://doi.org/10.1109/TAI.2024.3440220

# The Displacement-Velocity Dissociation in Sign Language Learning: Kinematic Signatures of Event Structure in Novice ÖGS Signers

**Evie A. Malaia**[1], **Julia Krebs**[2], **Eric Harbour**[3], **Julia Martetschläger**[2], **Hermann Schwameder**[3], **Dietmar Roehm**[2,4], **Ronnie B. Wilbur**[5]

[1]Department of Speech, Language, and Hearing Sciences, University of Alabama, Tuscaloosa, AL, United States
[2]Department of Linguistics, University of Salzburg, Salzburg, Austria
[3]Department of Kinesiology, University of Salzburg, Salzburg, Austria
[4]Centre for Cognitive Neuroscience (CCNS), University of Salzburg, Salzburg, Austria
[5]Department of Linguistics, Purdue University, West Lafayette, IN, United States

**Abstract**

This study investigates how adult learners acquire linguistically contrastive movement patterns in Austrian Sign Language (ÖGS), focusing on the telic/atelic distinction predicted by the Event Visibility Hypothesis. Telic verbs (bounded events) are produced by proficient Deaf signers with shorter duration and temporally precise, low-entropy velocity profiles, whereas atelic verbs (unbounded processes) show more continuous motion. Using 3D motion capture (300 Hz), we compared 8 novice learners (6–12 weeks of instruction) with 6 proficient Deaf signers across 71 verbs. Linear mixed-effects models revealed a dissociation between gross movement patterning and fine-grained velocity profile structure in learner productions. Learners correctly reproduced the proportional path-length contrast between telic and atelic verbs, replicating the gross spatial distinction of proficient signers. However, temporal marking of the telic/atelic contrast was underproduced: learners showed a significantly smaller duration difference between verb types than proficient signers, while total path length did not differ significantly between verb types or groups. Temporal control showed significant between-group differences: learners exhibited elevated sample entropy, with non-proficient velocity profiles within individual sign productions, though spatial consistency across trials (STI) was comparable to that of proficient signers. Peak velocity did not differ between groups, suggesting that learners can reach target speeds but cannot yet modulate temporal structure reliably. These findings support distinct learning trajectories for gross movement patterning and fine-grained motion complexity, and demonstrate that velocity profile structure within signs constitutes a core linguistic target in sign language learning.

**Keywords:** sign language learning, event structure, movement kinematics

## 1. Introduction

Learning any language as a second language in adulthood means mastering not only vocabulary and grammar, but also the fine-grained motor control required for the target language. For sign language learners, this motor learning challenge is particularly relevant: learners must acquire precise spatial and temporal patterns of manual articulation to convey linguistic meaning.

Recent advances in motion capture technology have made it possible to measure kinematic properties of sign production with precision. These methods showed that sign languages encode grammatical distinctions through systematic modulation of movement kinematics (Brentari, 1998; Malaia et al., 2013; Wilbur, 2003, 2008). One well-documented cross-linguistic grammatical marker is that of event structure - the distinction between telic verbs (events with inherent endpoints, such as ARRIVE or BREAK) and atelic verbs (unbounded processes, such as RUN or THINK). Proficient signers mark telic verbs with distinctive kinematics: shorter duration, reduced displacement, abrupt deceleration, and often a hold or handshape change at the endpoint (Wilbur, 2008). Atelic verbs, by contrast, are produced with more continuous movement lacking endpoint marking.

These kinematic patterns are linguistically contrastive. The Event Visibility Hypothesis (Wilbur, 2003, 2008) posits that the visual modality of sign languages permits (and favors) the systematic mapping between event structure and movement kinematics. Events with inherent endpoints are marked by abrupt changes in the movement signal, while unbounded events are expressed through reduplicated or continuous motion. This hypothesis has been supported by empirical studies across multiple sign languages, including American Sign Language (ASL), Austrian Sign Language (ÖGS), Croatian Sign Language (HZJ), and others (Krebs et al., 2025; Malaia et al., 2013; Malaia and Milković, 2021).

Little is known, however, about how learners acquire these linguistically contrastive kinematic patterns. The present study addresses these ques-



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 324–332
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tions by comparing the kinematics of telic and atelic verb production in adult ÖGS learners (6–12 weeks of instruction) with those of proficient Deaf signers. Our investigation is motivated by converging evidence from spoken language acquisition suggesting that spatial and temporal aspects of articulatory control develop along different timelines. In second language (L2) speech learning, learners often achieve reasonable approximations of spatial targets (e.g., formant frequencies, place of articulation) relatively quickly, while temporal targets (e.g., voice onset time, segment duration) require extended practice and may never fully converge on native norms (Flege, 1995; Flege et al., 1999). Similarly, motor learning research has documented dissociations between spatial accuracy and temporal consistency: learners can often reach the correct spatial endpoint of a movement while exhibiting high trial-to-trial variability in timing and velocity (Schmidt et al., 2018).

Recent work by Liu et al. (Liu et al., 2025) also indicated that temporal kinematics differentiate linguistic manual movements (signs) from non-linguistic manual movements (co-speech gestures). Liu and colleagues compared ASL signers' production of classifier signs with English speakers' co-speech gestures describing the same events, and found that while spatial features (maximal area covered, trajectory shape) were similar between signers and non-signers, temporal features differed significantly. Specifically, signers' linguistic movements were temporally compressed, smoother, and more rhythmic than non-linguistic movements, while speakers' co-speech gestures were characterized by an opposite pattern (temporal stretching, increased jerkiness, and reduced rhythmicity). Liu et al. interpreted temporal compression as a hallmark of an efficient linguistic system: drawing on information-theoretic framing of language as communicative signal carrier. Liu and colleagues argue that linguistic systems (spoken or signed) develop toward signal compression to maximize information transfer efficiency, consistent with prior work demonstrating compression and structured complexity in both signed and spoken language signals (Torre et al., 2019; Andres et al., 2021; Borneman et al., 2018; Malaia and Wilbur, 2020).

The distinction between spatial and temporal control also has theoretical significance. If learners differ in spatial and temporal accuracy in sign production, this would suggest that spatial and temporal aspects of signing are governed by distinct learning mechanisms. This hypothesis aligns with dual-process models of motor learning (Fitts and Posner, 1967), which distinguish between early cognitive stages (feedback-dependent visual control of movement) and later autonomous stages (feedforward control). This distinction is also consistent with models of linguistic processing that posit partially separable visual and linguistic components operating over shared signals (Malaia and Wilbur, 2019).

The question of whether temporal aspects of signing (velocity profiles, acceleration patterns, movement smoothness) carry linguistic information at all is relatively recent in sign language research. Early models of sign languages focused almost exclusively on static parameters: handshape, location, and orientation (Stokoe, 1960; Battison, 1978). Movement, when discussed, was treated as a categorical feature (e.g., straight vs. arc, unidirectional vs. repeated) rather than as a signal with linguistically contrastive properties.

The relevance of motion in sign languages has been a topic of interest in sign language research since Klima and colleagues' (Klima and Bellugi, 1979) work demonstrating perceptual differences between signers and non-signers in processing dynamic visual stimuli, and Wilbur's (Wilbur, 2003, 2008) formulation of the Event Visibility Hypothesis. This hypothesis posits that the visual-gestural modality offers transparent mapping between event structure and the kinematic properties of manual movement. Specifically, telic events (those with inherent endpoints) are hypothesized to be marked by abrupt changes in the movement signal (holds, decelerations, handshape or orientation changes) that make the endpoint visually salient. Atelic events (unbounded processes) lack such marking and are expressed through reduplicated or more continuous movement. The linguistic nature of these kinematic distinctions is underscored by their grammatical function. Telic/atelic contrasts reflect a productive morphological distinction in sign languages (Wilbur, 2008). Additionally, signers can modulate the kinematics of a single verb stem to signal aspectual contrasts: for instance, adding repetition to a telic verb stem can derive a habitual or iterative meaning, effectively converting a bounded event into an unbounded one (Krebs et al., 2026). The kinematic marking of event structure is thus a core component of sign language grammar.

The acquisition of motor skills has long been recognized to proceed through distinct stages (Fitts and Posner, 1967). A key finding from motor learning research is that spatial and temporal aspects of motor control can develop along distinct timelines (Schmidt et al., 2018). This dissociation has been documented in a wide range of motor tasks, from reaching and grasping to handwriting and musical performance. Evidence from second language speech acquisition supports a similar dissociation in the articulatory domain. Flege et al (1999) showed that spatial targets in L2 production (e.g., vowel formant frequencies, which reflect tongue position) are often acquired earlier than temporal targets (e.g., voice onset time in stop consonants).

This dissociation persists in highly proficient late L2 learners, suggesting that temporal control is more difficult to master than spatial control.

The present study tests this hypothesis in sign language learners by measuring both spatial parameters (displacement) and temporal parameters (velocity entropy, spatiotemporal variability) in novice ÖGS learners and comparing them with proficient Deaf signers.

## 2. Methods

Deaf fluent ÖGS signers (N=6) and hearing learners with initial ÖGS exposure (N=8; 6-12 weeks of weekly instruction for 1.5 hours) participated in the study. Deaf participants were either born Deaf or became Deaf early in life, used ÖGS as their primary language in daily life, and identified as members of the Deaf community. Five self-reported as right-handed; one as left-handed. All learners were hearing adults; none had prior exposure to sign language before beginning instruction. All participants gave informed consent in accordance with institutional review board protocols.

The stimuli consisted of 71 ÖGS verb signs (35 telic, 36 atelic) selected from the proficient signer corpus (Krebs et al., 2025). Telic verbs denote events with inherent endpoints (e.g., ARRIVE, ADMIT, OPEN), while atelic verbs denote unbounded processes (e.g., RUN, DANCE, WALK). Verb selection was based on the Event Visibility Hypothesis (Wilbur, 2003, 2008), which predicts that telic verbs are kinematically marked with abrupt endings (sharp deceleration, holds, handshape/orientation changes), whereas atelic verbs lack such marking and are produced with smooth, continuous movement.

For learner participants, stimuli were presented as video recordings of a proficient Deaf signer producing each verb sign. Deaf signer participants, being fluent users of ÖGS, produced each verb sign from their established lexicon without video prompting, following the protocol described in (Krebs et al., 2025); sign boundaries for all participants were determined using the annotation procedure as detailed in methods of the same work. Each video was presented once (though participants could rewatch the video if they wanted), and learners were instructed to reproduce the sign as accurately as possible.

Body kinematics (torso, head, and arms/hands) were recorded using a marker set and a 12-camera infrared motion capture system (Qualisys AB, Göteborg, Sweden) operating at 300 Hz. Simultaneously, a 2D video of the participant was recorded and time-synchronized with the motion capture data. Infrared-reflective markers (diameter 12 mm) were placed on anatomical landmarks following standard protocol (Krebs et al., 2025). Key markers included bilateral placement on shoulders, elbows, wrists, and hands.

Marker trajectories were exported from Qualisys Track Manager and processed in MATLAB (MathWorks). Raw 3D coordinate data were low-pass filtered using a second-order zero-lag Butterworth filter with a cutoff frequency of 25 Hz to remove high-frequency noise while preserving biologically relevant movement frequencies. Segment positions and orientations were calculated using an inverse kinematics algorithm (Visual3D; C-Motion, Rockville, MD).

All participants began each sign from the same resting position with arms at the sides of the body. Sign onset and offset were visually determined by a skilled signer using the time-aligned 2D video recording. Sign onset was defined as the video frame when the hands reached the target location from which the sign movement started. Sign offset was defined as the frame when handshape or hand orientation changed or when the hands moved away from the final position. The dominant hand was operationalized as the hand used for producing one-handed signs. All kinematic analyses were conducted on the dominant hand wrist marker trajectory.

Ten kinematic parameters were extracted from the wrist trajectory of each sign token:

**Spatiotemporal parameters**: (1) Sign duration (seconds), measured from sign onset to offset and (2) total path length (meters) of the wrist, calculated as cumulative Euclidean distance traveled by the wrist marker across successive time points over the course of the sign:

$$L = \sum_{i=1}^{N-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2}$$

Because path length is derived from the same position time series as velocity, it reflects the cumulative magnitude of movement and is sensitive to both overall length of path traveled, and within-trajectory variability, but does not encode temporal structure (Wilbur and Malaia, 2008). **Kinematic parameters**: (3) Median velocity (m/s) and (4) peak velocity (m/s), computed from the first derivative of position. (5) Peak deceleration $(m/s^2)$, extracted from the second derivative (acceleration profile) as the maximum negative value. (6) Jerk $(m/s^3)$, calculated as the third derivative of position, reflecting smoothness of movement.

**Complexity parameters**: (7) Sample entropy, calculated from the wrist speed vector using embedded dimension m = 2 and tolerance r = 0.2 × standard deviation (Richman and Moorman, 2000). Higher values of sample entropy indicate more irregular, less predictable movement patterns. (8)

Spatiotemporal index (STI) for velocity and (9) STI for acceleration, computed as the sum of standard deviations across 50 time-normalized points (Howell et al., 2009). STI measures trial-to-trial variability in movement patterning. (10) Movement rhythmicity, quantified as the standard deviation of inter-peak intervals in the velocity profile.

Sample entropy and STI serve as complementary measures of movement complexity. Sample entropy captures the informational complexity within a single movement trajectory (intra-trial regularity), while STI quantifies consistency across repeated productions of the same sign (inter-trial variability). Following Malaia and colleagues (2012, 2016, 2018), we interpret sample entropy as a proxy for fractal complexity: the degree to which a movement contains information at multiple temporal scales. In the context of linguistic signing, higher entropy reflects movements that are informationally rich and potentially carry linguistic contrast; lower entropy reflects movements that are more stereotyped and predictable. Linear mixed-effects models (LME) were fit to each kinematic parameter using the lme function in MATLAB. Fixed effects included Group (Deaf signers vs. learners), Verb Type (telic vs. atelic), and their interaction. Random effects included intercepts for Participant and Verb Item, plus by-participant random slopes for Verb Type. Models were fit using restricted maximum likelihood (REML). Effect sizes were reported as partial eta-squared ($\eta^2$). Statistical significance was assessed at $\alpha$ = .05. We present descriptive statistics alongside LME estimates for ease of comparison; however, all inferential statistics are based on the LMM results, which account for nested data structure and unbalanced designs.

## 3. Results

The kinematic analysis revealed a statistically significant dissociation between spatial and temporal aspects of sign production in learners as compared to proficient Deaf signers. Even though both groups showed similar spatial patterns in sign production (as indicated by displacement differences between telic and atelic verbs), they differed in temporal control. Table 1 presents summary statistics and LME results for key kinematic parameters.

### 3.1. Gross Movement: Preserved Telic/Atelic Distinction

Both groups showed significant main effects of Verb Type on sign duration ($\beta = -0.42$, SE = 0.03, $p < .001$, $\eta^2 = .17$). The Group × Verb Type interaction for sign duration was also significant ($\beta = 0.13, SE = 0.04, p < .001, \eta^2 = .01$). Learners produced a smaller temporal contrast between telic and atelic verbs than proficient signers did, suggesting an emerging, but not yet fully automated temporal marking of telicity. Total path of dominant hand movement within the sign (Euclidean distance) did not differ significantly between verb types ($\beta = -0.07$ $p > .05$) or between groups ($\beta = -0.06$ $p > .05$). This null result indicates that learners were not, at this point in their learning, tracking the fine-grained differences in trajectory between telic and atelic verb signs: the contrast they preserved was one of overall sign timing.

### 3.2. Temporal Control: Motion Entropy in Learners

Proficient signers demonstrated low entropy for telic verbs (0.03 $\pm$ 0.04) and somewhat higher entropy for atelic verbs (0.13 $\pm$ 0.15), yielding a large main effect of Verb Type ($\beta = -0.11$, SE = 0.02, $p < .001$, $\eta^2 = .20$) on the sample entropy measure. This pattern reflects the smooth, rapid articulation of telic signs contrasted with the more variable production of atelic signs.

Learner productions showed higher entropy across both verb types. While learners' telic sign entropy (0.04 $\pm$ 0.04) was similar to that of proficient signers, learners' atelic verb entropy (0.19 $\pm$ 0.14) was substantially higher than that of proficient signers. The Group $\times$ Verb Type interaction was significant ($\beta = -0.03$, SE = 0.01, $p < .001$, $\eta^2 = .01$), indicating that learners produced telic and atelic signs with a larger difference in entropy than proficient signers.

Interestingly, the STI velocity results showed no significant group differences (Deaf telic: 0.0048 $\pm$ 0.0030; Deaf atelic: 0.0056 $\pm$ 0.0051; Learner telic: 0.0053 $\pm$ 0.0029; Learner atelic: 0.0072 $\pm$ 0.0059; Group effect: $p$ = .48, $\eta^2 = .00$). This suggests that trial-to-trial spatial consistency was comparable between groups, in that learners could reproduce similar spatial trajectories across repetitions of the same sign. The temporal control deficit captured by elevated sample entropy in learners, then, reflects within-trial irregularity (how smoothly velocity unfolds during a single sign) rather than across-trial inconsistency (how variable the spatial path is from one production to the next). This dissociation between within-trial temporal structure and across-trial spatial consistency further supports the interpretation that learners have acquired spatial targets but not yet automatized smooth temporal control.

Velocity parameters (median velocity, peak velocity) showed weak or non-significant effects, with the exception of a small Verb Type effect for peak velocity ($\beta \approx 0$, $p = .02$, $\eta^2 = .01$). Importantly, there was no Group difference in peak velocity ($p = .82$), suggesting that although learners could achieve

similar maximum speeds as proficient signers, they were not yet able to modulate velocity profiles consistently.

## 4. Discussion

The present study demonstrates a dissociation between gross and fine aspects of motor control in beginning sign language learners. After just 6–12 weeks of ÖGS instruction, learners successfully reproduced the relative gross distinctions between telic and atelic verbs: they produced telic signs as shorter in duration and more spatially compact than atelic signs, replicating the pattern of proficient Deaf signers. However, learners' temporal control was only emerging, as characterized by higher sample entropy (irregularity within individual sign productions), although their spatial consistency across trials, as measured by STI, was comparable to that of proficient signers. This finding has implications for theories of motor learning and sign language acquisition.

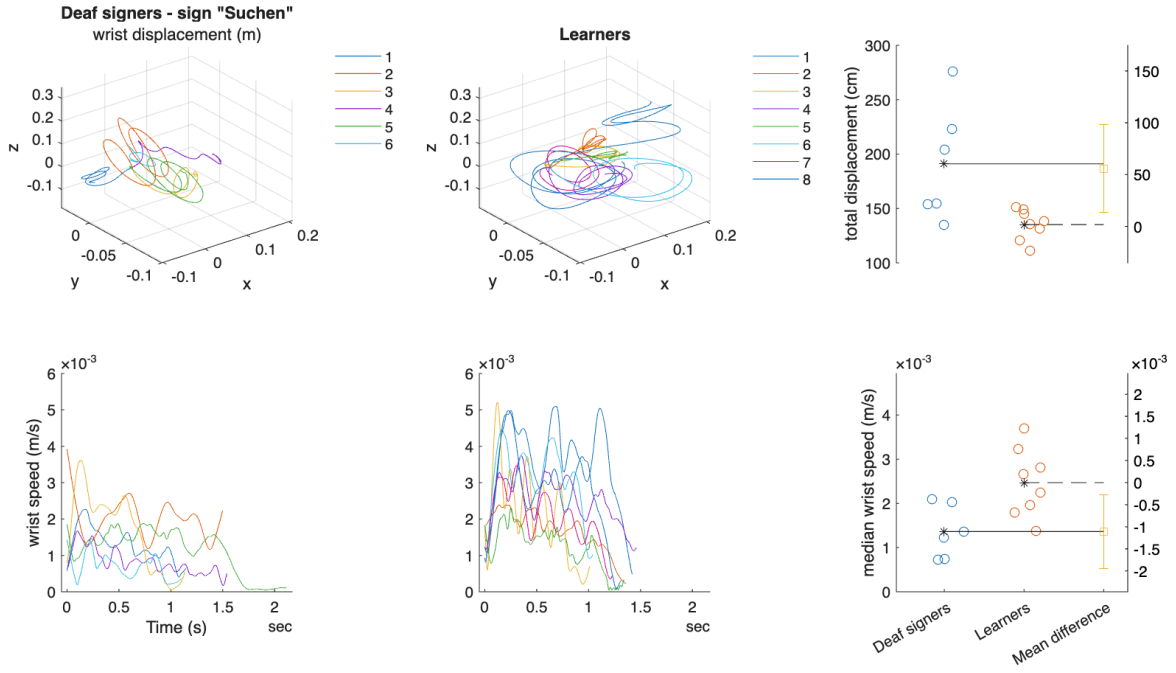
### 4.1. Gross Movement Patterns Acquired Before Fine-Grained Information Structure in Movement Complexity

The observed dissociation between preserved path-length contrasts in production of telic and atelic verb signs, and disordered velocity profile structure in learners suggests that gross and fine-grained kinematic features are acquired on different timelines. After 6–12 weeks of ÖGS instruction, learners successfully reproduced the gross kinematic markers of the telic/atelic distinction: they produced telic signs as more spatially compact than atelic signs, with a proportional path-length difference matching that of proficient signers. By contrast, the fine-grained velocity profile structure showed significant between-group differences: learners' production was characterized by higher sample entropy overall, as well as an exaggerated entropy difference between verb types. Why might gross movement patterning be acquired before fine-grained motion complexity? The answer may lie in what is perceptually accessible to a learner observing a model signer. Gross properties of motion profile — e.g. observation that telic signs cover less space and stop more abruptly than atelic signs — are directly visible and imitable through visual monitoring of movement. Fine-grained properties, like the shape of the velocity curve, the degree of smoothness across the deceleration phase, and the modulation of movement complexity, are not directly readable from visual observation at early stages of learning. Recent work by Karabüklü et al. (Karabüklü et al., 2025) showed that sign language learners increase visual temporal resolution through the first two years of instruction, reaching maximal temporal resolution only after 12–18 months. The present analysis characterizes an early stage of language learning, where gross movement targets have been approximated (perhaps through visual self-monitoring), but fine-grained structure of motion, which depends on both perceptual sensitivity and motor automatization — remains to be learned. This account is consistent with dual-process models of motor learning (Fitts and Posner, 1967), which distinguish early cogni-

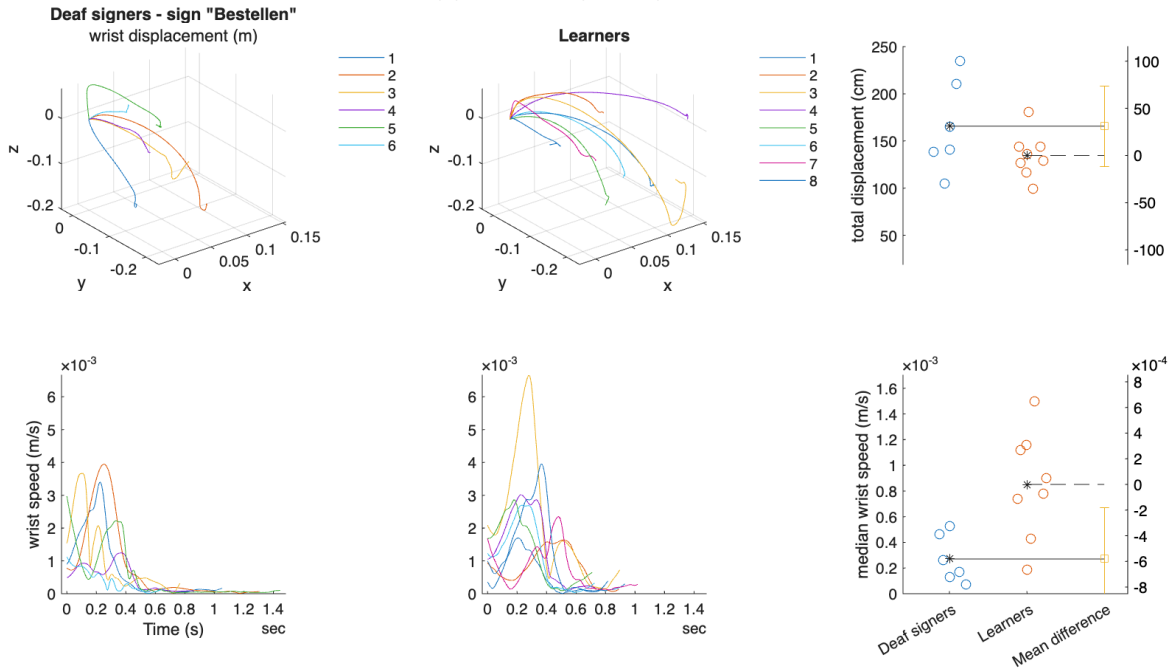
Table 1: Summary statistics and linear mixed-effects model results for key kinematic parameters.

| Parameter | Deaf Telic | Deaf Atelic | Learner Telic | Learner Atelic | Group $\beta$ ($\eta^2$) | Verb Type $\beta$ ($\eta^2$) |
|---|---|---|---|---|---|---|
| Sign duration (s) | 0.91 $\pm$ 0.45 | 1.35 $\pm$ 0.65 | 0.90 $\pm$ 0.31 | 1.19 $\pm$ 0.38 | -0.16*** (.03) | -0.42*** (.17) |
| Total path length(m) | 0.27$\pm$ 0.17 | 0.37$\pm$ 0.35 | 0.35 $\pm$ 0.19 | 0.47 $\pm$ 0.35 | 0.06 (.00) | -0.07 (.00) |
| Median velocity (m/s) | 0.00038 $\pm$ 0.00054 | 0.00092 $\pm$ 0.00073 | 0.00088 $\pm$ 0.00089 | 0.00130 $\pm$ 0.00092 | $\sim$0*** (.04) | $\sim$0*** (.05) |
| Peak velocity (m/s) | 0.0032 $\pm$ 0.0019 | 0.0022 $\pm$ 0.0013 | 0.0037 $\pm$ 0.0021 | 0.0027 $\pm$ 0.0015 | n.s. (.00) | $\sim$0* (.01) |
| Sample entropy (wrist speed) | 0.03 $\pm$ 0.04 | 0.13 $\pm$ 0.15 | 0.04 $\pm$ 0.04 | 0.19 $\pm$ 0.14 | 0.04*** (.03) | -0.11*** (.20) |
| STI (wrist velocity) | 0.0048 $\pm$ 0.0030 | 0.0056 $\pm$ 0.0051 | 0.0053 $\pm$ 0.0029 | 0.0072 $\pm$ 0.0059 | n.s. (.00) | n.s. (.00) |
| STI (wrist acceleration) | 0.00039 $\pm$ 0.00029 | 0.00058 $\pm$ 0.00052 | 0.00044 $\pm$ 0.00030 | 0.00078 $\pm$ 0.00057 | n.s. (.00) | n.s. (.00) |
| Peak deceleration (m/s$^2$) | -0.0001 $\pm$ 0.00009 | -0.00007 $\pm$ 0.00005 | -0.0001 $\pm$ 0.00008 | -0.00009 $\pm$ 0.00007 | n.s. (.00) | $\sim$0* (.01) |

*Note.* Descriptive statistics are median $\pm$ IQR. LME estimates ($\beta$) are shown with partial $\eta^2$ in parentheses. ***$p$ < .001, *$p$ < .05, n.s. = not significant. Interaction effects were significant for sign duration ($\beta$ = 0.13, $p$ < .001, $\eta^2$ = .01) and sample entropy ($\beta$ = -0.03, $p$ < .001, $\eta^2$ = .01).




(a) SUCHEN (*search*) - atelic




(b) BESTELLEN (*order*) - telic

Figure 1: Example wrist trajectories for Deaf signers (left) vs. learners (middle) across 2 ÖGS verbs (SUCHEN, BESTELLEN). Each colored line is one repetition. Note high spatial convergence but velocity variability in learner productions (right). Error bands show within-group variability.

tive stages, characterized by feedback-dependent control of salient movement properties, from later autonomous stages, characterized by feedforward control of motion kinematics. It also aligns with evidence from second language speech acquisition, where globally perceivable articulatory targets (vowel formant frequencies, place of articulation) are acquired earlier than fine-grained temporal targets (voice onset time, segment duration) that require extended perceptual-motor practice (Flege, 1995; Bradlow et al., 1997). Our results extend this principle to sign language learning, where we show that gross movement patterning (where to move and how fast) is acquired early, while fine-grained motion complexity (how to modulate velocity profile structure to carry linguistic information) requires

extended experience. The STI results point to an important additional constraint on this interpretation. Trial-to-trial spatial consistency, as measured by STI for both velocity and acceleration, did not differ between groups (both $p > .12$). This confirms that learners' fine-grained deficit is not one of gross motor inconsistency: they can reproduce similar movement paths across repeated productions of the same sign. The between-group difference in sample entropy then reflects within-trial irregularity — the smoothness and structure of the velocity profile during a single sign production — rather than cross-trial variability in where the hand goes, characterized by STI. Learners produce spatially consistent but temporally unstructured sign kinematics: the pattern predicted by a model in which gross movement targets precede fine-grained motion complexity in acquisition.

### 4.2. Motion Complexity as Linguistic Target

Proficient signers produce telic verbs with very low entropy ($0.03 \pm 0.04$), reflecting smooth, highly predictable movement. This is not a byproduct of making the movement shorter or faster; it is a specific kinematic signature that marks telicity in ÖGS and in other sign languages (Malaia et al., 2013). Learners produce signs that are accurate in terms of gross movement patterning — correctly reproducing the relative spatial compactness of telic signs — but not in terms of motion complexity. These productions can be thought of as phonetically accented: having the right gross shape but lacking the internal kinematic structure of native signing.

Interestingly, learners are not producing signs with insufficient entropy; they are overproducing in terms of complexity. Learners' atelic signs show entropy values ($0.19 \pm 0.14$) significantly higher than those of proficient signers ($0.13 \pm 0.15$); learners' telic signs, while lower in sample entropy than atelic ones, still exceed proficient norm. Malaia and colleagues have argued that motion complexity (as quantified via sample entropy or power spectral density) reflects information-carrying capacity of the signal (Malaia et al., 2018; Borneman et al., 2018) and is consistent with evidence that temporally structured visual linguistic input engages predictive neural mechanisms that are refined through experience (Malaia et al., 2026). For linguistic signals to be efficiently processed, information must be distributed across these scales: too little entropy yields a signal that is overly predictable and carries minimal information; too much entropy yields a signal that is excessively noisy and inefficient to produce. Recent work linking spatiotemporal neural coherence to predictive inference in sign language perception supports the role of temporally structured motion signals in efficient communication (Borneman et al., 2025).

Proficient signers operate within this optimal zone for motion complexity, with telic verbs showing low entropy (high predictability) and atelic verbs showing moderate entropy (informational richness without excessive noise). Learners, in contrast, have not yet calibrated the upper bound of complexity in production: their atelic signs' parameters exceed the proficient complexity range, while their telic productions are already close to proficient norms.

Borneman et al. (Borneman et al., 2018) demonstrated that proficient ASL signing shows higher fractal complexity than non-linguistic motion precisely because linguistic motion is both informationally rich and temporally structured. Learners' elevated entropy, paradoxically, reflects less linguistic structure, not more: they have not yet acquired the temporal scaffolding that permits efficient information transfer. This interpretation is consistent with the Event Visibility Hypothesis (Wilbur, 2003, 2008), which posits that telic verbs are marked by rapid, visually salient changes that reliably signal boundedness. The low entropy of telic verbs in proficient signers reflects not a reduction in information, but rather the concentration of information into discrete, identifiable events. Learning to sign is thus not merely learning where to move (spatial targets) but how to move (temporal structure), and, more specifically, how to modulate movement complexity to create visual signals that map to grammatical markers of sign language.

## 5. Bibliographical References


Jan Andres, Martina Benešová, and Jiří Langer. 2021. Towards a fractal analysis of the sign language. *Journal of Quantitative Linguistics*, 28(1):77–94.

Robbin Battison. 1978. *Lexical borrowing in American Sign Language*. Linstok Press, Silver Spring, MD.

Joshua D Borneman, Evie A Malaia, and Ronnie B Wilbur. 2018. Motion characterization using optical flow and fractal complexity. *Journal of Electronic Imaging*, 27(5):051229.

Sean C. Borneman, Julia Krebs, Ronnie B. Wilbur, and Evie A. Malaia. 2025. Decoding predictive inference in visual language processing via spatiotemporal neural coherence. NeurIPS 2025 Workshop on Foundation Models for the Brain and Body.

Ann R Bradlow, David B Pisoni, Reiko Akahane-Yamada, and Yoh'ichi Tohkura. 1997. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, 101(4):2299–2310.

Diane Brentari. 1998. *A prosodic model of sign language phonology*. MIT Press, Cambridge, MA.

Paul M Fitts and Michael I Posner. 1967. *Human performance*. Brooks/Cole, Belmont, CA.

James Emil Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 92:233–277.

James Emil Flege, Ian R. A. MacKay, and Diane Meador. 1999. Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America*, 106(5):2973–2987.

Peter Howell, Andrew J. Anderson, Jon Bartrip, and Eleanor Bailey. 2009. Comparison of acoustic and kinematic approaches to measuring utterance-level speech variability. *Journal of Speech, Language, and Hearing Research*, 52(4):1088–1096.

Serpil Karabüklü, Sandra Wood, Chuck Bradley, Ronnie B. Wilbur, and Evie A. Malaia. 2025. Effect of sign language learning on temporal resolution of visual attention. *Journal of Vision*, 25(1):3–3.

Edward S Klima and Ursula Bellugi. 1979. *The signs of language*. Harvard University Press, Cambridge, MA.

Julia Krebs, Eric Harbour, Evie A Malaia, Ronnie B Wilbur, Julia Martetschläger, Hermann Schwameder, and Dietmar Roehm. 2025. Sign language encodes event structure through neuromotor dynamics: motion, muscle, and meaning. *Frontiers in Psychology*, 16:1689676.

Julia Krebs, Ronnie B Wilbur, and Evie A Malaia. 2026. Testing the Event Visibility Hypothesis in Austrian Sign Language (ÖGS). *Language and Cognition*, 18:e14, 1–31.

Rui Liu, Wim Pouw, Susan Goldin-Meadow, and Diane Brentari. 2025. Signers and speakers show distinct temporal kinematic signatures in their manual communicative movements. *Open Mind*, 9:1323–1338.

Evie Malaia and Ronnie B Wilbur. 2019. Visual and linguistic components of short-term memory: Generalized Neural Model (GNM) for spoken and sign languages. *Cortex*, 112:69–79.

Evie A Malaia, Joshua D Borneman, and Ronnie B Wilbur. 2018. Information transfer capacity of articulators in American Sign Language. *Language and speech*, 61(1):97–112.

Evie A Malaia, Sean C Borneman, Joshua Borneman, Julia Krebs, and Ronnie B Wilbur. 2026. Age and language experience modulate predictive processing in the visual modality. *PLOS ONE*. Accepted.

Evie A Malaia and Marina Milković. 2021. Aspect: Theoretical and experimental perspectives. In *The Routledge handbook of theoretical and experimental sign language research*, pages 194–212. Routledge.

Evie A Malaia and Ronnie B Wilbur. 2020. Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(1):e1518.

Evie A Malaia, Ronnie B Wilbur, and Marina Milković. 2013. Kinematic parameters of signed verbs. *Journal of Speech, Language, and Hearing Research*, 56(5):1677–1688.

Joshua S. Richman and J. Randall Moorman. 2000. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049.

Richard A Schmidt, Timothy D Lee, Carolee Winstein, Gabriele Wulf, and Howard N Zelaznik. 2018. *Motor control and learning: A behavioral emphasis*. Human Kinetics.

William C Stokoe. 1960. *Sign language structure: An outline of the visual communication systems of the American deaf*. Number 8 in Studies in Linguistics: Occasional Papers. University of Buffalo, Buffalo, NY.

Inés G Torre, Bartolo Luque, Lucas Lacasa, Christopher T Kello, and Antoni Hernández-Fernández. 2019. On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science*, 6(8):191023.

Ronnie B Wilbur. 2003. Representations of telicity in ASL. In *Proceedings from the annual meeting of the Chicago Linguistic Society*, volume 39, pages 354–368. Chicago Linguistic Society.

Ronnie B Wilbur. 2008. Complex predicates involving events, time and aspect: Is this why sign languages look so similar? *Theoretical Issues in Sign Language Research*, pages 217–250.

Ronnie B Wilbur and Evie A. Malaia. 2008. Contributions of sign language research to gesture understanding: What can multimodal computational systems learn from sign language research. *International Journal of Semantic Computing*, 2(01):5–19.

# CEFR-Based Assessment in Sign Languages: The Case of LSE and Perspectives for LIS

**Maria Grazia Marrocu**

Ca' Foscari University of Venice, Rey Juan Carlos University of Madrid
mariagrazia.marrocu@unive.it, mg.marrocu.2025@alumnos.urjc.es

**Abstract**

This study examines the application of the Common European Framework of Reference for Languages (CEFR) to Spanish Sign Language (LSE) in a university context, with reference to the Italian situation (Council of Europe, 2020). In Spain, CEFR descriptors are already integrated into academic programmes for the assessment of LSE, whereas in Italy the context remains uneven due to the lack of shared criteria for the teaching and assessment of Italian Sign Language (LIS). The research project, conducted jointly by Ca' Foscari University of Venice and Rey Juan Carlos University of Madrid, adopts a longitudinal and comparative design focusing on the first three CEFR proficiency levels (A1, A2, B1) of LSE among L2M2 learners (second language second modality). A mixed-methods approach combining classroom observations, self-assessment instruments, and standardised assessment rubrics is used to analyse the alignment between students' self-assessments and instructors' external evaluations, with particular attention to linguistic and metacognitive awareness. The findings show increasing accuracy in self-assessment as proficiency develops, alongside recurring issues such as the overestimation of receptive skills and the underestimation of productive competence. These results highlight the need for targeted assessment interventions and contribute to the development of CEFR-consistent evaluation practices for sign languages.

**Keywords:** Sign Languages, CEFR, Assessment

## 1. Introduction

The growing attention devoted to sign languages within European academic institutions has prompted research in language pedagogy to confront new methodological and assessment-related challenges. Among these, one of the most pressing challenges concerns the application of shared and standardized instruments for teaching and evaluation, such as those proposed by the *Common European Framework of Reference for Languages* (CEFR), updated in 2020, which for the first time systematically incorporated sign languages (Council of Europe, 2020; De Monte, 2022).

The present contribution forms part of a broader comparative research project conducted under joint supervision between Ca' Foscari University of Venice and Rey Juan Carlos University of Madrid. The overall objective is to analyze and compare teaching and assessment practices in university courses of Italian Sign Language (LIS) and Spanish Sign Language (LSE), with particular attention to the assessment of linguistic competences in the first three CEFR levels (A1, A2, B1), and to the relationship between students' self-assessment and instructors' external evaluation. The findings emerging from the first phase of the study, carried out at Rey Juan Carlos University during the 2024/2025 academic year, will be presented, and the didactic and metacognitive implications of these results will be discussed.

## 2. The CEFR for Sign Languages (SL) - CEFR4LS

The CEFR provides a reference framework for the description of language competences, articulated across six levels (from A1 to C2), and organizes competences not only in terms of "what one knows," but also in terms of how one acts in and through language, and with which strategies (Council of Europe, 2020). Chapter 6 of the 2020 Companion Volume (CEFR4LS), the outcome of nearly a decade of research and implementation projects, extends the principles of the CEFR to sign languages through descriptors structured around three macro-competences: Linguistic,



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 333–340
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Sociolinguistic, and Pragmatic (Figure 1) (Haug & Keller, 2012; Leeson et al., 2016, 2019, 2020; Council of Europe, 2020; ECML, n.d., a, b).

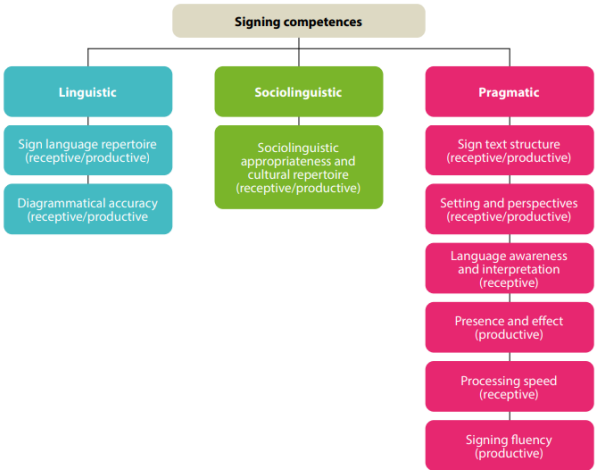



Figure 1: Signing Competences (CEFR, 2020:144)

These macro-competences are further subdivided into specific sections with the aim of ensuring coherent and functional assessment. Each is analyzed according to the four modes of communication: reception, production, interaction and mediation. In the case of sign languages, these descriptors are particularly relevant for capturing both formal and pragmatic aspects of language use, moving beyond the mere ability to produce isolated signs (Barni et al., 2024).

## 3. Teaching Context

The dissemination of sign language teaching across Europe currently presents a markedly heterogeneous landscape, characterized by substantial differences among national contexts. In the present study, the analysis focuses specifically on a comparison between Spain and Italy. The selection is motivated by the strong similarities between their respective sign languages and Deaf communities, which, however, have developed within distinct educational and regulatory frameworks (Marrocu, 2023). In Italy, the educational provision remains fragmented and is often independently defined by individual associations, institutions, or universities. There is no inter-university network of university lecturers capable of fostering the necessary dialogue and critical reflection on teaching practices (Branchini et al., 2024; Cardinaletti, 2018; Cardinaletti & Mantovan, 2023; Mantovan, 2021). In Spain, by contrast, the integration of the CEFR into the teaching of Spanish Sign Language (LSE) has been in place for several years and is supported by official curricular instruments. Since 2010, specific syllabi have been available for instructional planning and, notably, the language received official recognition as early as 2007 (BOE, 2007; Aroca, 2012; Díez, 2014; Fundación CNSE, 2010). At present, the only comprehensive four-year undergraduate degree programme covering the full range of LSE proficiency levels (A1-C1) is offered by Rey Juan Carlos University of Madrid, thus representing a case study of particular interest for the analysis of standardization practices in sign language education (CNLSE, n.d.).

## 4. The Research Project

### 4.1 Objectives and Methodology

The study presented here forms part of a doctoral research project conducted under joint supervision between Ca' Foscari University of Venice and Rey Juan Carlos University (URJC) of Madrid. It is a longitudinal investigation whose ultimate objectives are:

- to construct a comparative framework of the teaching and assessment pathways of LIS and LSE in academic contexts;
- to analyze and compare the strategies, pedagogical approaches, and assessment methods employed, considering the guidelines provided by the CEFR (2020);
- to conduct an in-depth examination of the instruments used to assess linguistic competences, with particular attention to the initial stages of language learning.

The first phase of data collection, defined as Corpus 1, was carried out during the 2024/2025 academic year at URJC. The sample comprised three groups of hearing university students approaching a sign language for the first time (L2M2 - second language, second modality[1]): 29 students at level A1 (mean age 20.1 years), 25 at level A2 (mean age 21.8 years), and 22 at level B1 (mean age 22.3 years). The teaching staff consisted of six university lecturers: four Deaf signing instructors and two hearing signing instructors, including one CODA[2] lecturer and one professional LSE interpreter.

The methodological approach adopted was mixed methods, combining qualitative and quantitative instruments designed to provide a triangulated perspective on the teaching and assessment process. Two questionnaires were administered: one at the beginning and one at the end of the course, comprising both closed-ended Likert

[1] (Ortega & Özyürek, 2013; Pichler & Koulidobrova, 2015; Marschark et al., 2015; Ortega & Morgan, 2015).

[2] Children Of Deaf Adults

scale items and open-ended questions (Arroyo Hernández, 2021).

At the end of the course, students were also asked to complete a linguistic self-assessment based on the descriptors of the 2020 CEFR, which are organized by competences in production and reception. To further explore the teaching perspective, individual semi-structured interviews were conducted with the instructors, with the aim of clarifying attitudes, perceptions, and practices related to teaching and assessment that had emerged from a previously administered mixed-format questionnaire.

In addition, systematic classroom observation was carried out in order to document instructional dynamics, the type and frequency of activities performed, recurring errors, and levels of student participation. The collected data were analysed using both descriptive and inferential statistical tools: quantitative responses were processed with the software packages Excel and RStudio, while analyses of variance (ANOVA) and significance tests, such as Chi-square test and correlation analyses, were conducted using JASP. Interviews and observational data were subjected to qualitative analysis aimed at identifying recurring patterns, discontinuities, and emerging perspectives.

In this contribution, particular attention is devoted to the comparative analysis between students' linguistic self-assessment and the external evaluation provided by lecturers (Björkstrand et al., 2019, Lugarini, 2010). The objective is to investigate the degree of alignment between subjective perception - based on standardized CEFR descriptors - and expert judgment grounded in the same guidelines. This relationship is further examined through methodological triangulation integrating classroom observation data, in order to verify the consistency among self-reported performance, assessed performance, and performance observed within the instructional context. Such an approach not only deepens reflection on learners' linguistic and metacognitive awareness but also makes it possible to identify any systematic discrepancies between self-perception and actual language behaviour, with significant implications for the development of more accurate assessment and pedagogical instruments for sign languages.

## 4.2 Data Analysis

The data analysis was conducted based on the self-assessments completed by students in the A1, A2, and B1 groups at the end of the semester. Students filled in self-assessment grids structured according to the sections and descriptors of the 2020 CEFR for sign languages (Consejo de Europa, 2020). The scale employed ranged from 1 = "not at all" to 4 = "very much," depending on the level of ability perceived. In parallel, instructors carried out an external evaluation of students' competences using an analytic scoring grid (0 - 10), divided into the following areas: grammatical, lexical, discourse, and pragmatic competence. This framework was generally applied in the final examination assessments, which included comprehension and production tests administered twice during the semester, a group presentation delivered in the final weeks of the course, and a theoretical examination during the official exam session.

To compare the two datasets, the assessment grids were made compatible both in terms of their sections and through the alignment of scoring scales to a common 0 - 10 metric. This procedure made it possible to calculate mean scores and the differences between students' self-assessments and instructors' evaluations. It also included the computation of mean differences across the three groups, together with the verification of statistical significance through ANOVA and post hoc tests.

## 4.3 Results

This section presents the results emerging from the analysis of the collected data, providing an overview of the observed trends and the differences identified across the various proficiency levels, with particular attention to students' self-assessments and their comparison with lecturers' evaluations.

### 4.3.1 Students' Self-Assessments

The analysis of the A1 group's self-assessments, collected through grids structured according to the 2020 CEFR descriptors for sign languages, reveals a distribution of responses that highlights recurring and differentiated trends across the various sections. As illustrated in Figure 1, many students positioned themselves within the categories "quite" and "very" with a peak in Linguistic Awareness (Conc_Ling) and Sociolinguistic Appropriateness (Aprop_Socio), that is, competences related to the relational and informal dimension of communication, in which students report feeling more at ease. By contrast, more structurally oriented sections, such as Schematic Precision in both reception and production (P.Esq_REC and P.Esq_PRO), Text Structure (Texto), and Presence and Effects (Presencia), display greater dispersion in

responses, with relatively high values also in the categories “a little” and “not at all.”

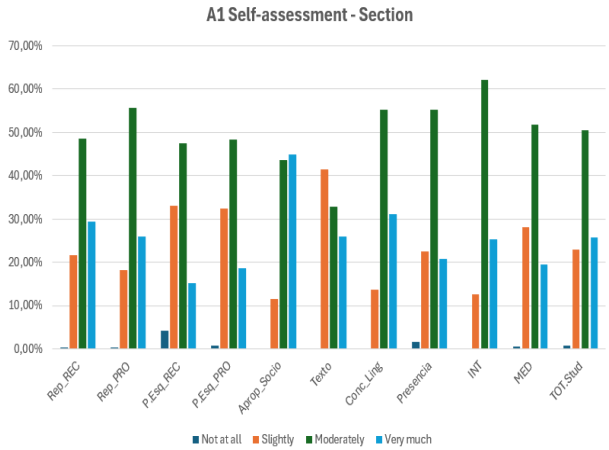



Figure 2: A1 Group Self-Assessment by CEFR section

These findings suggest an early awareness of morphosyntactic and metalinguistic difficulties, consistent with an initial level profile which, although relying on visual communicative strategies, has not yet achieved full formal control of the language.

For the A2 group, the self-assessments reveal a predominance of the category “quite” across almost all sections, indicating a perception of intermediate and functional competence that is, however, not yet fully consistent with this linguistic profile. Responses in the category “very” are less frequent but tend to cluster particularly in Sociolinguistic Appropriateness in reception (Soc_REC), Setting and Perspectives in reception (Perp_REC), and Presence and Effects (Presencia), areas related to the interpretation and management of communicative context. By contrast, more structurally oriented sections - such as Schematic Precision in production (P.Esq_PRO) and Setting and Perspectives in production (Perp_PRO) - show higher percentages in the categories “a little” and “not at all”, as illustrated in Figure 2. This pattern reflects students’ difficulty in the transition from comprehension to production, a phase that requires a more conscious and articulated control of language. Compensatory strategies are also observed, particularly the use of non-manual markers (NMMs), which enable students to sustain effective communicative interaction even in the absence of fully developed morphosyntactic mastery.

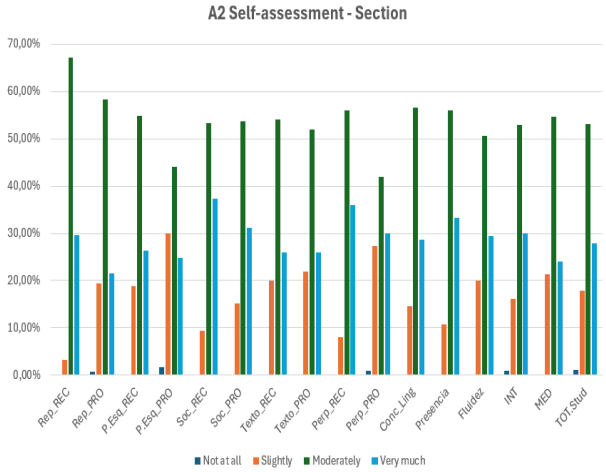



Figure 3: A2 Group Self-Assessment by CEFR section

In the B1 group, a more articulated profile emerges compared to the previous levels (Figure 3), with a less polarized distribution of responses. The category “quite” continues to prevail across many sections; however, an increase in “very” responses is observed, particularly in Schematic Precision in reception (P.Esq_REC) and Sociolinguistic Appropriateness in production (Soc_PRO), indicating greater mastery and awareness of one’s competences. Nevertheless, critical issues persist in sections related to integrated production and the discursive management of language. Indeed, responses in the categories “a little” and “not at all” increase in Sociolinguistic Appropriateness in reception (Soc_REC), Fluency (Fluidez), Presence and Effects (Presencia), and Interaction (INT), suggesting that the development of linguistic competence does not always correspond to a heightened sense of confidence in more complex communicative situations.

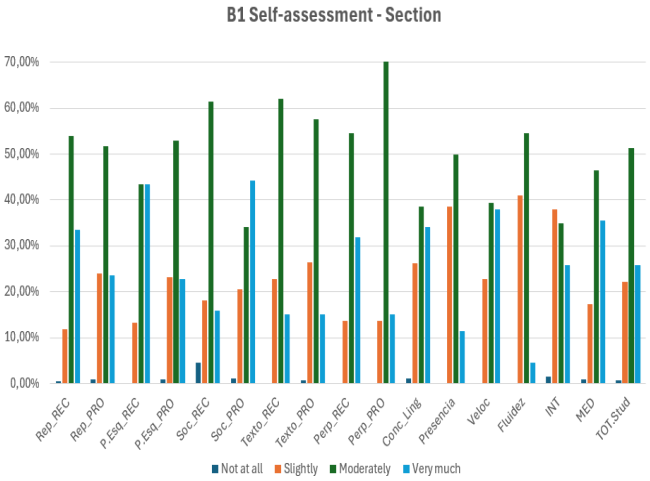



Figure 4: B1 Group Self-Assessment by CEFR section

From this perspective, self-assessment confirms its value as a metacognitive tool, useful not only for identifying strengths and areas requiring further development, but also for guiding future instructional interventions in a more targeted and informed manner.

#### 4.3.2 Comparison between students' self-assessments and lecturers'evaluations

Comparing students' self-assessments with lecturers' evaluations makes it possible to capture competences in a multi-level and cross-cutting manner and constitutes an essential parameter for assessing the degree of alignment or potential discrepancy between what students believe they know and what they demonstrate. This comparison enabled the identification of significant divergences across the different competence areas, differentiated by proficiency level.

In the A1 group, a marked overestimation is observed in comprehension (+2,5) and, to a lesser extent, in lexical competence (+0,3). By contrast, all other competences are underestimated, most notably grammatical competence (-1,10), suggesting a critical perception of productive and formal abilities. The overall judgment value is slightly negative, pointing to a general tendency toward underestimation.

In the A2 group, although an overestimation remains in comprehension (+1,57) and lexical competence (+1,05), the discrepancy is smaller than in A1, indicating greater exposure to the language and a gradual internalization of its structures. Some competences are still underestimated, such as discourse competence (-0,52), whereas the overall evaluation approaches that of the lecturers (-0,41), indicating an improvement in self-assessment ability.

The B1 group displays a more advanced level of metalinguistic awareness: the mean overall difference is virtually null (+0,06), suggesting a strong alignment between subjective perception and objective performance. Nevertheless, a substantial overestimation emerges in grammatical competence (+3,2), likely attributable to a subjective sense of increased fluency and communicative naturalness.

A comparative analysis across the three groups further shows that five of the six competences examined present statistically significant differences between levels. The only competence that did not yield significant differences was lexical competence. These results highlight a trajectory of linguistic development that runs in parallel with a substantial change in students' capacity for self-assessment.

### 4.4 Discussion of the Results

The findings outline a gradual and significant development of linguistic and metacognitive awareness across the three groups analyzed.

In the A1 group, a clear tendency toward underestimation emerges in productive and formal competences, accompanied by a marked overestimation of comprehension. This dynamic is confirmed both by the lecturers' objective evaluations and by classroom observations, which revealed effective informal communicative performances alongside persistent difficulties in the articulation of signs. The high level of engagement observed in this group suggests that initial motivation does not immediately translate into confidence in one's own competences.

In the A2 group, an overall reduction in underestimation is observed, although a certain degree of overestimation in comprehension persists. Difficulties nonetheless emerge in relation to lexical complexity and in the transition from predominantly receptive competence to more active production. In this respect, the frequent adoption of non-verbal compensatory strategies appears to facilitate communication, yet tends to mask only partial mastery of linguistic structures. Reduced participation, as documented through classroom observation, may have negatively affected both confidence consolidation and the accuracy of self-assessment.

The B1 group presents a more mature and balanced profile: the correspondence between self-assessment and external evaluation is almost complete, indicating well-developed metalinguistic awareness. However, the significant overestimation of grammatical competence reveals a specific phenomenon known as *fluency overconfidence*, whereby increased communicative fluency leads learners to perceive formal competences as fully acquired. This effect highlights the need to develop further metacognitive tools capable of enabling learners to assess accurately the formal quality of their production, even at more advanced levels (Sanchez & Dunning 2020). The comparative analysis confirms that linguistic level substantially influences the quality of self-assessment. While at initial levels oscillation between overestimation and insecurity predominates, more advanced

levels display greater stability, albeit with the risk of overestimating aspects related to structured production. Comprehension, by contrast, appears to be systematically overvalued at all levels, albeit with progressively smaller discrepancies. This represents a critical area requiring targeted metacognitive and pedagogical attention. The Dunning-Kruger effect becomes particularly evident when individual cases are examined: learners with lower competences tend to overestimate themselves in more intuitive tasks, whereas those with genuinely higher abilities may incur a general underestimation, maintaining a more realistic appraisal of their shortcomings (Carter & Dunning, 2008).

Overall, these results suggest the value of self-assessment not only as a diagnostic instrument, but also as a formative lever to promote self-regulation and self-awareness in language learning processes.

## 5. Limitations

Several methodological aspects warrant careful consideration when reflecting on the limitations of the present study. First, the conversion scale applied to self-assessments - transformed from the original 1-4 range to a 0-10 scale - may have introduced potential distortion effects. A further limitation concerns the small group sizes, particularly at the B1 level. This makes the interpretation of the observed effect more complex and reduces the robustness of the statistical data, thereby limiting the generalizability of the findings.

From the students' perspective, a relatively low level of engagement was observed. This is compounded by the possible influence of social desirability bias, which may have led students to orient their self-assessments toward levels perceived as more acceptable or aligned with external expectations.

Finally, the timing of the data collection should be considered. The self-assessments were administered during a period of heightened stress associated with the examination session. Such a context may have influenced students' self-perception and, consequently, the responses provided.

## 6. Conclusion

Overall, the study confirms the value of self-assessment as a formative practice, provided that it is supported by tools fostering linguistic reflection and anchored in a shared reference framework. The integration of quantitative and qualitative data proved essential for capturing the complexity of language learning in L2M2 students, thereby enabling more precisely targeted instructional interventions. The evidence gathered offers practical insights for the continuation of the comparative project within the Italian context. In anticipation of data collection in Italy, it will be crucial to maintain the same methodological framework while also considering the structural differences between the two academic systems. For this reason, one of the primary objectives of the second phase will be to examine the extent to which the assessment practices adopted in Spain may be adapted or proposed as a model within the Italian context.

In this regard, the adoption of the CEFR descriptors for sign languages has proven to be an invaluable instrument. The availability of shared guidelines not only enhances the comparability of educational pathways, but also facilitates the development of a common evaluative language between students and lecturers (Serragiotto, 2017). This is particularly significant in a visual-manual domain such as sign languages, where competences must be assessed in their multimodal and pragmatic dimensions.

## 7. Acknowledgments

Sincere thanks are extended to the entire teaching staff and the students of Rey Juan Carlos University, and in particular to Professor Francisco Vera Villaverde. Without your support, this work would not have been possible.

## 8. Bibliographical References

Anna Cardinaletti. (2018). La lingua dei segni italiana a Ca' Foscari: Didattica, ricerca e progetti sull'accessibilità. In A. Cardinaletti, L. Cerasi, and P. Rigobon (Eds.), Le lingue occidentali nei 150 anni di storia di Ca' Foscari (pp. 341-353). Edizioni Ca' Foscari. http://doi.org/10.30687/978-88-6969-262-8/016

Anna Cardinaletti and Lara Mantovan. (2023). Le lingue dei segni nel volume complementare e l'insegnamento della LIS nelle università italiane. Italiano LinguaDue, 14(2). https://doi.org/10.54103/2037-3597/19575

BOE. (2007). Ley 27/2007, de 23 de octubre. https://www.boe.es/boe/dias/2007/10/24/pdfs/A43251-43259.pdf

Carmen Sanchez and David Dunning. (2020). Decision fluency and overconfidence among beginners. Decision, 7(3), 225-237. https://doi.org/10.1037/dec0000122

Chiara Branchini, Anna Cardinaletti, A. and Lara Mantovan. (Eds.). (2024). Insegnare la LIS e la LISt all'università. Esperienze, materiali, strumenti per la didattica e per la valutazione. Edizioni Ca' Foscari. http://doi.org/10.30687/978-88-6969-827-9

CNLSE. (n.d.). Cursos de formación LSE en España. https://cnlse.es/es/recursos/otros/formacion/universidades

Consejo de Europa. (2020). Marco común europeo de referencia para las lenguas: Aprendizaje, enseñanza, evaluación. Volumen complementario. Consejo de Europa.

Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume. Council of Europe.

Deborah Chen Pichler and Helen Koulidobrova. (2015). Acquisition of sign language as a second language (L2). Gallaudet University Press.

ECML [European Center for Modern Languages of the Council of Europe] (n.d., a). Promoting excellence in sign language instruction. https://www.ecml.at/Thematicareas/SignedLanguages/ProSign

ECML [European Center for Modern Languages of the Council of Europe] (n.d., b). Sign languages. https://www.ecml.at/thematicareas/signedlanguages

Edoardo Lugarini (Ed.). (2010). Valutare le competenze linguistiche (Collana Giscel, 11). FrancoAngeli.

Eva Aroca. (Coord.). (2012). Propuesta curricular de la lengua de signos española. Nivel usuario independiente B1. Fundación CNSE.

Fundación CNSE (Coord.). (2010). Propuesta curricular de la lengua de signos española. Nivel usuario básico A1-A2. Fundación CNSE.

Gerardo Ortega and Asli Özyürek. (2013). Gesture-sign interface in hearing non-signers' first exposure to sign. In Proceedings of the Tilburg Gesture Research Meeting. Tilburg University.

Gerardo Ortega and Gary Morgan. (2015). Input processing at first exposure to a sign language. Second Language Research, 31(4), 443-463. https://doi.org/10.1177/0267658315576822

Graziano Serragiotto. (2017). La valutazione per l'apprendimento delle lingue straniere dal punto di vista dei docenti. Rassegna Italiana di Linguistica Applicata, 49(1), 7-19.

Ignacio Arroyo Hernández. (2021). Atteggiamenti verso la pronuncia della lingua straniera: Il questionario di Ca'Foscari (2019) nel panorama attuale di ricerca. Rassegna Italiana di Linguistica Applicata, 1-2, 165-172.

Lorraine Leeson and Beppie Van den Bogaerde. (2019). Sign languages. In D. Newby, F. Heyworth, and M. Cavalli (Eds.), Changing contexts, evolving competences (pp. 100-107). Council of Europe.

Lorraine Leeson, Christian Rathmann, Beppie Van den Bogaerde and Tobias Haug. (2016). Sign languages and the CEFR: Common reference level descriptors. European Centre for Modern Languages of the Council of Europe.

Lorraine Leeson, Jordan Fenlon, Johanna Mesch, Sarah Sheridan and Carmel Grehan. (2020). Corpora in sign language pedagogy. In R. Rosen (Ed.), The Routledge handbook of sign language pedagogy (pp. 339-352). Routledge.

Lara Mantovan. (2021). Insegnare la lingua dei segni italiana all'università: Esperienze consolidate e direzioni future. QuaderniCIRD, 22, 33-49.

Marc Marschark, Linda Spencer, Andreana Durkin, Georgianna Borgna, Carol Convertino, Elizabeth Machmer, William G Kronenberger and Alexandra Trani. (2015). Understanding language, hearing status, and visual-spatial skills. Journal of Deaf Studies and Deaf Education, 20(4), 310-330. https://doi.org/10.1093/deafed/env025

Maria Aránzazu Díez. (Coord.). (2014). Propuesta curricular de la lengua de signos española. Nivel usuario independiente B2. Fundación CNSE.

Maria Grazia Marrocu. (2023). Espressioni idiomatiche: Confronto interlinguistico tra la LSE

e la LIS (Tesi di laurea magistrale). Università Ca' Foscari Venezia.

Maria Tagarelli De Monte. (2022). Using the Common European Framework of Reference for Italian Sign Language education: Lessons from research and practice. Momento – Diálogos em Educação, 31(2), 394-416. https://doi.org/10.14295/momento.v31i02.14502

Monica Barni, Claudio Ferrara, Sabina Fontana and Maria Roccaforte. (2024). Certificare la competenza linguistico-comunicativa nella LIS: Riflessioni, raccomandazioni e proposte. LEND. Lingua e nuova didattica, 1, 8-21.

Thomas Björkstrand, Erin Balkstam and Josephine Willing. (2019). Sign language dictionary as a digital tool in L2 teaching: Sentence scoring on CEFR levels A1-B2. Department of Linguistics, Stockholm University.

Tobias Haug and Jorg Keller. (2012). ESF exploratory workshop on development of theoretical and practical guidelines for the adaptation of the CEFR to sign languages: Scientific report. European Science Foundation.

Travis J Carter and David Dunning. (2008). Faulty self-assessment: Why evaluating one's own competence is an intrinsically difficult task. Social and Personality Psychology Compass, 2(1), 346-360. https://doi.org/10.1111/j.1751-9004.2007.00031.x

# Improving phonological distance measures for signs: the CatFormCompare tool

**Hope E. Morgan**[1], **Amy Isard**[2], **Anh Dang**[3]
[1]Institute of German Sign Language and Communication of the Deaf
[2]Hub of Computing and Data Science
University of Hamburg, Germany
[3]Utrecht University
{hope.morgan, amy.isard}@uni-hamburg.de, t.t.a.dang@uu.nl

**Abstract**

This paper describes the CatFormCompare tool, designed to enable the comparison of phonological content between pairs of signs, especially in larger datasets. With this tool and a schema for coding categorical form (the SL CatForm coding schema), a pipeline is created that allows a feedback mechanism for advancing research—specifically by directly addressing one of the hard problems in sign language phonology: how to extract true minimal pairs from datasets coded for categorical form? Solving this problem would simultaneously improve phonological distance measurements for sign languages because it would mean that the units for measuring distance are grounded in the linguistic structure of the language and not simply a by-product of the coding system. Here we report on the tool and the first evaluation of its functioning.

**Keywords:** phonological distance, minimal pairs, string edit distance

## 1. Introduction

A robust research program for sign language phonology would reflect close alignments between these domains: signer perception of phonological differences, the encoding of categorical form units in a notation or orthographic system, and theoretical models that represent this internal structure. One indication of the current status of the field is that there is no system for representing units of form in sign languages that is calibrated well enough to automatically find minimal pairs—that is, two signs in the same language with different lexical meanings that differ by only one form value while the rest of the form values are identical.

Minimal pairs are important objects of study for many reasons. Such pairs reflect linguistic categories in each language and yield inventories of contrastive forms that can be compared cross-linguistically. Minimal pairs also reveal the cognitive organisation of words in the mind and reflect usage (Wedel et al., 2013; Sun and Poeppel, 2023), as well as provide evidence that sign languages have combinatorial structure, a prerequisite for *duality of patterning*, a design feature of human language (de Boer et al., 2012).

Currently, notation or orthographic systems for phonology are not clearly aligned with mental categories of sign form—phonemes—in the same way that International Phonetic Alphabet (IPA) characters can be used to represent phonemes spoken languages. In an effort to move the field closer to this alignment and reveal where gaps lie, we introduce the CatFormCompare tool as part of a pipeline to compare the phonological content of sign pairs. This pipeline has two components: (i) a string of characters representing categorical forms in a sign, the SL CatForm string, and (ii) the CatFormCompare tool, a simple interface housing a tool written with the Python programming language.[1] The tool and the coding schema, along with a small sample of annotated Kenyan Sign Language (KSL) data, are available from the Hamburg University Research Data Repository.[2]

This paper focuses specifically on the CatFormCompare tool and how it evaluates the content of two strings to find minimal pairs. As a test of the pipeline, we use a dataset of 1,880 signs in KSL coded in SL CatForm coding schema and a list of 449 minimal pairs that were found in this dataset. Only minimal pairs that came from the same signer were included in the list to make sure they were from a single grammar. These minimal pairs function as a 'ground truth' for evaluating the output of the tool and thus the conditionalities described in this paper. Our goal is to locate pairs that differ by one phonological variable and from this derive a measure of phonological distance. In summary, we find that the pipeline functions, but still requires adjustments and improvements at different points.

## 2. Background: 'Edit distance' in different language modalities

In spoken languages, in order to quantify the phonological differences between words, a string edit distance such as Levenshtein distance can be used

[1] https://docs.python.org
[2] https://doi.org/10.25592/uhhfdm.18534



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 341–350
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

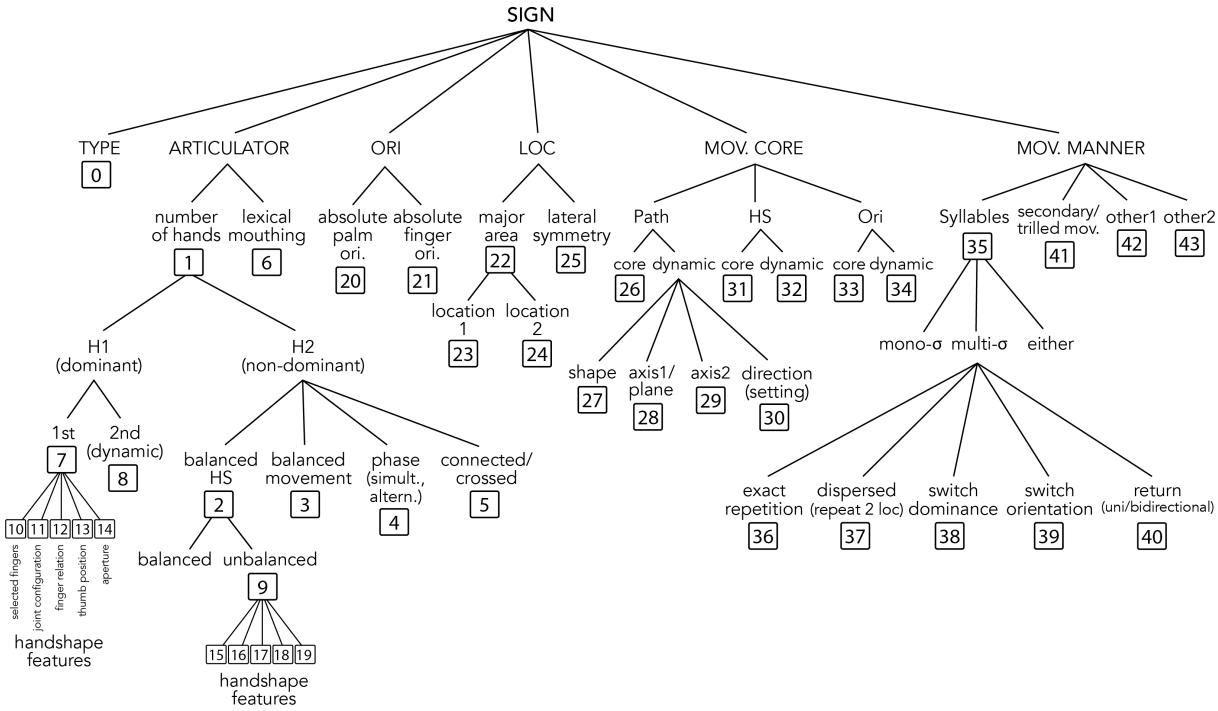



Figure 1: Tree diagram of SL CatForm variables

to calculate the number of differences between all pairs of words in a lexicon, typically done on a lexicon transcribed in the IPA. All changes or 'edits' of characters between pairs of words are tallied segment-by-segment, left to right; i.e., each deletion, insertion, and substitution of segments from one word to the next. The position of segments relative to each other in a sequence is intrinsic to the edit distance operation in spoken languages because it reflects the sequentiality of sounds in spoken words.

In contrast, the phonological content of signs is organised differently. First, there are more types of primary feature classes, and these classes are substantially different from each other.[3] Second, each feature class most often has only one specified value per sign; e.g., a value for *selected fingers*, for *path shape*, for *repetition*, etc. Thus, nearly all minimal contrasts in sign languages appear to be made by the substitution of one feature for another. Minimal pairs on the basis of sequential (re-)ordering are rare and theoretically debated (Corina, 1993; Channon, 2002; Morgan, 2022).[4]

This level of simultaneity and the nature of the different feature classes in signs has both helped and arguably hindered sign language research. It enables researchers to get measurable results using only a sub-set of a sign's phonological content, such as evaluating only handshape and location while leaving out aspects of movement, orientation, etc. This would never work for spoken languages; imagine trying to claim meaningful results about the phonology of a spoken language using only the first syllable of all words or all phonemes except high vowels and voiced consonants.

Many previous studies have used quantitative approaches to compare phonological content of signs, including cross-linguistic comparisons (McKee and Kennedy, 2000; Rozelle, 2003; Parks, 2011; Yu et al., 2018), research on sign neighbourhoods within one language (Caselli et al., 2021), and vector-based comparisons that also integrate semantic vector space (Börstell et al., 2020; Martinez del Rio et al., 2022).

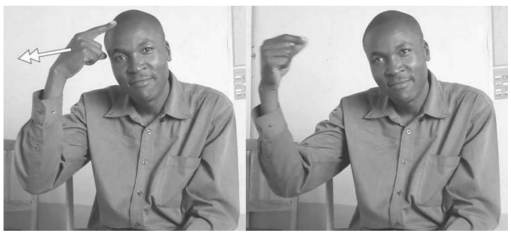
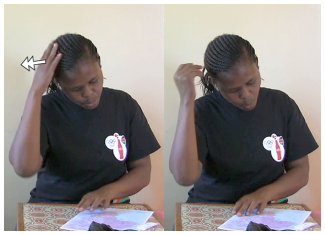
Yet, in order to locate minimal pairs from a dataset of signs, all aspects of form must be accounted for because all values except one in the pair must be the same. If there is too little phonological information included for each sign, the search function will over-generate pairs; it will find those that actually have more than one contrastive difference and incorrectly report them as only one. If the wrong types of information is encoded, then the search function will under-generate; it will miss true minimal pairs.

How then to create a search function that compares only the relevant information for each feature type between two pairs of signs? As mentioned,

[3]For instance, spoken languages have two main classes, consonants and vowels, while sign languages have several, including handshape, location, mouth actions, etc. The phonetic properties of each class are so different that features in one class are almost never shared with another.

[4]Contrasts can also be found that depend on the presence or absence of a feature (Morgan, 2022, 252–254).

| GLOSS | VARIABLES | | |

| | TO-MUSE | SOMALIA |
|---|---|---|
| **Variable 7** (starting handshape on h1) | **59** (open-G) | **27** (spray) |
| **Variable 8** (ending handshape on h1) | **30** (closed-G) | **11** (flat-O) |
| **Variable 32** (type of handshape change) | **C** (closing) | **C** (closing) |

Table 1: Minimal Pair A: handshape contour




| | IF | OKAY |
|---|---|---|
| **Variable 7** (starting handshape on h1) | **34** (O) | **34** (O) |
| **Variable 8** (ending handshape on h1) | **10** (F) | **19** (K) |
| **Variable 32** (type of handshape change) | **X** (HS contrast) | **X** (HS contrast) |

Table 2: Minimal Pair B: handshape contrast

selected fingers, as in OKAY in Table 2. Here, the hand configuration starts with all fingers selected in the O handshape and changes to the index and middle fingers plus thumb selected in the K handshape—i.e., a *handshape contrast*.

The two minimal pairs in Tables 1 and 2 pose a problem for the match/no-match approach. Because the signs in minimal pair B, IF and OKAY, contrast only by the second handshape in variable 8, a comparison of these two variables should find variable 8 as the minimal difference, with all other features being the same.[6] However, the minimal difference for the signs in minimal pair A, TO-MUSE and SOMALIA, are the starting handshapes in variable 7 because the ending handshapes are predictable based on the type of handshape movement in variable 32 (*closing*). In this case, variable 8 should be ignored so that only one difference is counted. Yet, a simple match/no-match calculation would find two differences in minimal pair A, not one.

[6]The sign OKAY can be produced with either one or two hands (value 'E' in variable 1), so the number of hands is not counted as a difference.

Another use of conditional processing is to account for the fact that some characteristics of signs are phonetically important but are not themselves contrastive. That is, they play a role in phonology and should appear in a coding schema, but are not the basis for phonological contrast. An example is *major area*, which is important because it accounts for the distribution of two-handed signs, especially when laterality features are involved. For instance, two-handed signs in a contralateral location on the head are disallowed in many sign languages, but *head* is itself is not a unit of contrast.

In order to capture these kinds of contingent properties of signs, conditionalities are written into the CatFormCompare tool. In the case of minimal pairs A and B, the conditionalities for variable 8 (the ending handshape) are as follows: (i) if the handshapes in variable 7 are different then do not compare the values in variable 8 at all; (ii) if one value is null in variable 8 then do not compare them; (iii) if both have the same value do not compare them; (iv) if both are different then *do* count this as a difference. The full list of current conditionalities are listed in

Table 3.[7] Any variables not listed in this table are evaluated using the simple match/no-match operation.

## 4. CatFormCompare tool

The core purpose of the CatFormCompare tool is to calculate the phonological distance between all pairs of signs in a dataset, applying conditionalities in order to make only the relevant comparisons. For each pair, it also removes any features from consideration that are not relevant, such as the ending handshapes in Table 1 (i.e., in variable 8). Therefore, it also functions as a way to correct one effect of using field-based coding in databases: the denominator of possible similarities is not all 42 variables, but only those that are relevant in each pair comparison.

First, the tool loads the data from a csv file containing the phonologically coded signs in SL CatForm format. It then performs comparisons between each pair of signs based on the conditionalities described in Section 3.2 and Table 3.

In order to make the CatFormCompare tool easier to use, a lightweight interface was created using the Python tkinter library,[8] which allows the user to select the input file and perform the calculation of matching pairs, as shown in Figure 3. Some options are available in the interface: (i) the number of differences between pairs to search for (1 for minimal pairs); and (ii) whether the standard conditionalities should be used for the comparison. If the standard conditionalities are not used, the user can constrain the value of a particular variable or select to ignore certain pre-defined variables, as shown in the lower half of Figure 3.

The main output of the CatFormCompare tool is a comma-separated file containing all the matching sign pairs; sample shown in Table 4. A summary file report is also provided that lists the conditionalities that were chosen, the number of differences selected for comparison, and a count of how many pairs were found with each of the selected differences, as shown in Figure 4. The interface also optionally allows the input file to be checked for coding errors (based on a csv file of the coding schema variables and values). If any value is found that does not conform to the specifications—i.e., SL CatForm coding values—a file is generated with a list of warnings about values that do not match the master coding schema. Each warning specifies: the variable in question, allowed values specified in the coding schema, actual value in the input file,

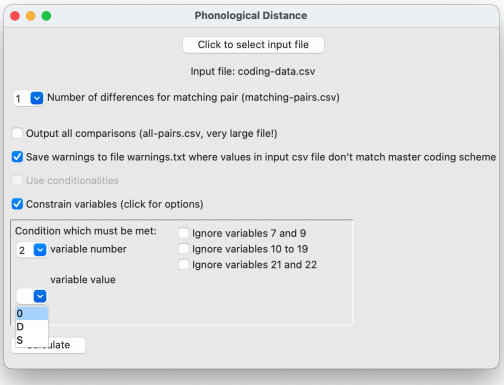



Figure 3: Screenshot of CatFormCompare tool

```
Comparisons with conditionalities
Ignoring variables []
Searching for pairs with 1 differences
Total pairs: 1075
Parameters which differ: [7] - number of pairs 512
Parameters which differ: [23] - number of pairs 189
Parameters which differ: [28] - number of pairs 74
Parameters which differ: [20] - number of pairs 72
Parameters which differ: [35] - number of pairs 62
Parameters which differ: [21] - number of pairs 36
Parameters which differ: [1] - number of pairs 31
Parameters which differ: [37] - number of pairs 16
Parameters which differ: [30] - number of pairs 13
Parameters which differ: [26] - number of pairs 10
Parameters which differ: [40] - number of pairs 9
Parameters which differ: [41] - number of pairs 8
```

Figure 4: Summary file report from the CatFormCompare tool

and ID of the gloss(es) with the non-matching value. Two example warnings can be seen in Figure 5.

```
WARNING: variable 4 "HANDS MOV. SIMULT./ALTERNAT.
                    (PHASE)"
 allowed values are ['0', 'A', 'S']
  unknown value "_" in 1 glosses ['403 CLASS-1']
  unknown value "O" in 1 glosses ['1672 PAKISTAN']

WARNING: variable 33 " + ORIENTATION MOVEMENT"
 allowed values are ['0', 'R']
  unknown value "C" in 1 glosses ['896 FROM-3']
```

Figure 5: CatFormCompare tool warnings about unknown values in variable codings

## 5. Preliminary Evaluation

The phonological analysis of KSL (Morgan, 2022) was largely based on a dataset of 1,880 non-compound signs coded for phonological characteristics in a FileMaker Pro database. There were 449 minimal pairs found among this set.[9] These

[7] Details of all of the possible values for each variable can be found in Morgan (2026)

[8] https://docs.python.org/3/library/tkinter.html#module-tkinter

[9] Morgan (2022) reports 461 minimal pairs, but 11 contain a sign that is not among the set of 1,880, so those pairs are not relevant for the evaluation here.

| No. | Name | Description of conditionality operation |
|---|---|---|
| 1 | NUMBER OF HANDS | If a sign can be either 1-handed or 2-handed ('E' for either), then 'E' is counted as the same value as '1' or '2' |
| 2 | SAME SHAPE ON EACH HAND | Do not compare; this variable is only used to apply conditionality statements |
| 3 | BALANCED MOVEMENT | Do not compare; this variable is only used to apply conditionality statements |
| 4 | SIMULTANEOUS / ALTERNATING | Only compare values in this variable if the movement on the hands is balanced ('B') in Variable 3 |
| 5 | HANDS CONNECTED, CROSSED | Only compare values in this variable if both signs are two-handed, either one- or two-handed, or the non-dominant hand is a base or grips the dominant hand; i.e., variable 1 = '2', 'E', or 'B' |
| 8 | H1 2ND SHAPE (DYNAMIC) | Only compare values in this variable if the handshapes in variable 7 are different (and not '0') and if both values in variable 8 are not '0', and if both values are different from each other |
| 9 | WHOLE HANDSHAPE (H2) | Only compare values in this variable if neither value is '00'; do not compare values in this variable if both signs in variable 2 have the value 'S'; i.e., if h2 is a copy of h1 in both signs in the pair |
| 22 | MAJOR AREA | Do not compare; this variable is only used to apply conditionality statements |
| 24 | 2ND LOCATION | Only compare values in this variable if one sign in the pair has a value that is not '00'; if one sign in a pair doesn't have a second location, but the other sign does, then count this as a difference |
| 25 | LATERAL SYMMETRY | Only compare values in this variable if the major area is on the torso, neck, head, or in neutral space. Also, do not compare pairs for lateral symmetry when they have a location of cheek or ear (which are ipsilateral if one-handed by default) and those signs differ by number of hands |
| 27 | PATH SHAPE | Only compare values in this variable if both signs have path movement (variable 26 = 'P') |
| 28 | PATH AXIS 1, or PLANE | Only compare values in this variable if both signs have path movement (variable 26 = 'P') |
| 29 | PATH AXIS 2 | Only compare values in this variable if both signs have a value that isn't null |
| 30 | PATH DIRECTIONALITY | Only compare values in this variable if the path axis is the same in both signs (both have same value in variable 28). Don't penalize a sign if it moves in both directions on an axis rather than only one direction: if either (or both) signs in the pair for variable 30 = 'B', then count these as the same |
| 34 | TYPE OF ORIENTATION CHANGE | Only compare values in this variable if both signs have orientation change ('R' in variable 33) |
| 35 | NUMBER OF SYLLABLES | If either sign for variable 35 = 'E' (is either mono- or multi-syllabic) , count this as the same as '1' or 'M' in a pair, but different from 'H' (hold); all other values are counted as different from each other |
| 36 | REPEATED, EXACT | Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic (variable 35 = 'M' or 'E') |
| 37 | DISPERSED (REPEAT 2 LOCATIONS) | Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic (variable 35 = 'M' or 'E'); if pairs differ in this variable, then ignore variable 36 |
| 38 | SWITCH DOMINANCE | Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic (variable 35 = 'M' or 'E'); if pairs differ in this variable, then ignore variable 36 |
| 39 | SWITCH ORIENTATION | Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic (variable 35 = 'M' or 'E'); if pairs differ in this variable, then ignore variable 36 |
| 40 | RETURN (UNI/BI-DIRECTIONAL) | Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic; and don't compare values if one or both is not relevant ('N') or a null value ('0') |
| 41 | SECONDARY / TRILLED MOVEMENT | Only compare values in this variable if neither one of the values = '0' (to prevent counting the information in variables 30 and 32 twice) |

Table 3: Conditionalities in SL CatForm (first column is the variable number)

pairs were discovered in various ways: by chance, during the process of coding, and later by intentionally looking for minimal pairs with custom searches through the database once it was fully coded. This list serves as a 'ground truth' target for testing the entire pipeline, from phonological coding to transformation into SL CatForm to the mechanics of the CatFormCompare tool.

The dataset of 1,880 KSL signs was first transformed into SL CatForm characters by the first author, inside the KSL Lexical Database. It was then exported as a csv file and run through the CatFormCompare tool, which compares the strings in over 3.5 million pairs to run the calculation. Examples of correctly matched minimal pairs in the dataset are listed in Table 4 and shown in Figures 6 to 9; these include minimal pairs for absolute palm orientation (GOSSIP vs. MOTHER-2; Figure 6); for whole handshape on the dominant hand (MBITA vs. MOTHER-2 in Figure 7); for specific location (OIL-1 vs. SCIENCE in Figure 8); and for return (PAINT vs. UPPER in Figure 9).

The output in Table 4 shows how **phonological distance** can be computed from this data in different ways. Using a categorical measurement by variable, each pair differs by 1 (COUNTS: DIFF) meaning they have a phonological distance of 1. Yet, they also have different features in common (COUNTS: SAM); e.g. GOSSIP and MOTHER-2 have 17 variables that share the same value, while PAINT and UPPER have 19 variables with the same value.[10] By applying the conditionalities, we get a total number of relevant features for each pair, and this can function as another way to measure phonological distance, either by degree of difference or similarity. For example, GOSSIP and MOTHER-2 have a difference score of 0.056 (1/18) and a similarity score of 0.944 (17/18), while PAINT and UPPER have a differ-

[10]The VARIABLES column lists which variables contain the different or same values for each pair.

| | | | COUNTS | | VARIABLES | |
|---|---|---|---|---|---|---|
| **id** | **gloss1** | **gloss2** | **DIFF** | **SAM** | **DIFF** | **SAM** |
| 1071019 | GOSSIP | MOTHER-2 | 1 | 17 | 20 | 1 6 7 21 23 26 27 28 30 31 33 35 36 37 38 39 40 |
| 1384822 | MBITA | MOTHER-2 | 1 | 17 | 7 | 1 6 20 21 23 26 27 28 30 31 33 35 36 37 38 39 40 |
| 1509121 | OIL-1 | SCIENCE-1 | 1 | 18 | 23 | 1 6 7 20 21 25 26 27 28 30 31 33 35 36 37 38 39 40 |
| 1542671 | PAINT | UPPER | 1 | 19 | 40 | 1 6 7 20 21 23 25 26 27 28 30 31 33 35 36 37 38 39 41 |

Table 4: Examples of four minimal pairs from the CatFormCompare tool output

| Parameter | Ground Truth Pairs | Matched Pairs | % pairs found | Missed pairs | % pairs missed |
|---|---|---|---|---|---|
| Handedness | 6 | 3 | 50% | 3 | 50% |
| Handshape | 149 | 50 | 34% | 99 | 66% |
| Location | 186 | 20 | 11% | 166 | 89% |
| Movement | 79 | 11 | 14% | 68 | 86% |
| Orientation | 16 | 9 | 56% | 7 | 44% |
| Non-manuals | 5 | 1 | 20% | 4 | 80% |
| Other | 5 | 2 | 40% | 3 | 60% |
| Unsure | 3 | 0 | 0% | 3 | 100% |
| All | 449 | 96 | 21% | 353 | 79% |

Table 5: CatFormCompare tool matches compared to 'ground truth' pairs, grouped by phonological parameter

ence score of 0.050 (1/20) and a similarity score of 0.950 (19/20). It remains to be seen which of these calculations are most useful, and in what types of research applications.

The results of the minimal pair search using our pipeline finds that only 96 of 449 minimal pairs, or 21%, were listed as differing by 1 variable in the output. Table 5 shows the matches and mismatches by phonological parameter. The worst performing parameter was location. On first impression, this seems like a rather poor result. However, each of the successful matches is possible only if all the variables in both pairs are exactly the same except one value in the string. This is also the first study of its kind so there is no baseline for what might be expected. Subsequent (and on-going) investigation shows that each step—from phonological coding to transformation into SL CatForm to the mechanics of the CatFormCompare tool—all reveal actionable places for improvement. However, due to the breadth of issues and the level of detail, the full results are beyond the scope of this paper. Therefore, we report here only on data from one domain: the 353 missed minimal pairs that were expected but not found.

## 6. Results

A detailed analysis of a quarter of these missed minimal pairs, 87 pairs, was performed by the first author, showing that in all cases except one, the expected value of contrast in the pair was selected as a difference.[11] This means the pipeline did find the expected minimal value; yet, for these signs, it finds too many differences, leading to the question: how far off is it? That is, how many differences beyond one were in the output? The results show that around a third of the missed pairs, 33%, differed by only two instead of one difference, 23% had three differences, 21% had 4 differences, 16% had five differences, and 4% had six differences. Thus, around half of the pairs might be matched correctly as having the single expected difference if only one or two variables or values were corrected. Yet, a few pairs show a very high number of mismatched variables/values, so more work is needed.

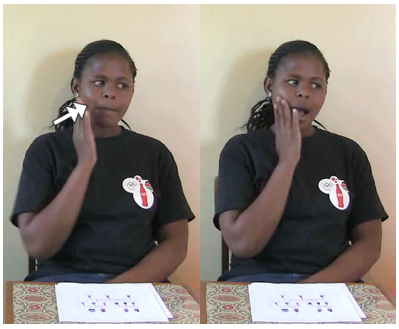
When the specific phonological features responsible for the incorrect differences were tracked, two culprits were overwhelmingly most responsible. *Absolute palm orientation* (variable 20) and *absolute finger orientation* (variable 21) were incorrectly counted as differences in half of 87 pairs (43 and 44 pairs, respectively). This was followed by *path axis 1* (variable 28) in 26 pairs, *handshape* (variable 7) in 19 pairs, *number of syllables* in 13 pairs, and *handedness* (variable 1) and *lateral symmetry* (variable 25) each had 10 pairs. Fifteen other variables were problematic in seven or fewer pairs.

In the case of the two orientation variables, the problem appears to lie within the nature of the phonological parameter itself and how it is represented rather than the CatFormCompare tool. For analytical reasons, discussed in Morgan (2022, 408–413), absolute orientation was used in SL CatForm instead of relative orientation. However, the current output of the pipeline shows this is not working for some reason. While this requires more research to solve, it already demonstrates the useful feedback function of having the pipeline.

[11] The one that was missed pointed to a modification needed in the CatFormCompare tool, since implemented.



(a) GOSSIP with palm orientation contra

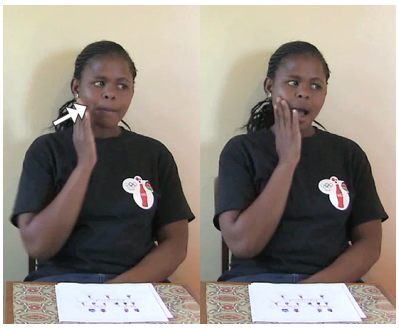


(b) MOTHER-2 with palm orientation towards body

Figure 6: Minimal pairs that differ in palm orientation (variable 20)



(a) MBITA (town in Kenya) with claw handshape



(b) MOTHER-2 with flat handshape

Figure 7: Minimal pairs that differ in whole handshape on dominant hand (variable 7)

## 7. Discussion and Next Steps

The CatFormCompare tool and pipeline continue to be assessed through evaluations with the main test set of minimal pairs, as well as a second set of diagnostic sign pairs that has gradually accumulated while testing the pipeline. Tracking down mismatches in the output at a fine-grained level is painstaking work, but when the tool is correctly tuned and functioning, as shown by the successful matches, the benefits and applications should be substantial. With a measure of phonological distance grounded in linguistic data, further research advancements are possible, such as exploring phonological neighbourhoods, studying perception of phonological similarity with more precision, and better controlling stimuli in experimental settings.

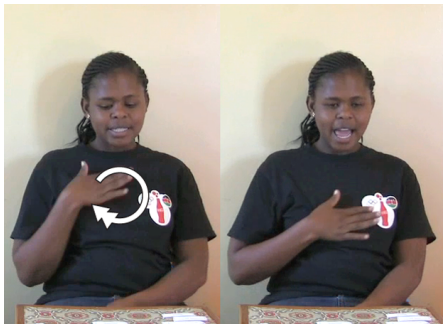
Cross-linguistic analyses are also possible, and this brings up an obvious question about the SL CatForm coding schema and the CatFormCompare tool: how can it be used with other sign languages? First, it is important to state that SL CatForm is one of many ways of representing units of form, and its coding categories and values largely overlap with many other notation systems. We are not advocating here for the use of a new coding system. Our aim is more fundamental. If this pipeline works, it will show what needs to be encoded in *any* notation system in order to find minimal pairs—and/or it will highlight places where sign language linguists must do more descriptive and theoretical work.

That said, the coding schema and tool are modifiable based on one's own language data. Values can be added to the SL CatForm variables and the CatFormCompare tool can be be updated to reflect these new values, as well as new variables, if necessary. Two variables, 42 and 43, are kept open for this purpose in case new feature types should be added. In fact, recent evidence suggests that *path size* is contrastive in KSL signs, and this would require its own variable.

There are three basic approaches to using this pipeline with different phonological coding. First, all of the information required to replicate the process is provided in references cited in this paper, with the exception of detailed coding guidelines for the SL CatForm schema. In lieu of such documentation at present, one is directed to Morgan (2022) to understand the nature of the variables and values in the schema. A second approach is to transform existing coded data into SL CatForm. For instance, Loy and Morgan (2026) demonstrate how one well-known coding system, Hamburg Notation System (Hanke, 2004), can be transformed into SL CatForm. Lastly, one could download the CatFormCompare tool (see footnote 2) and modify it to fit one's own dataset.

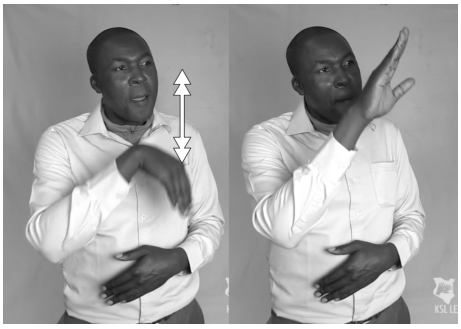


(a) OIL-1
at mouth

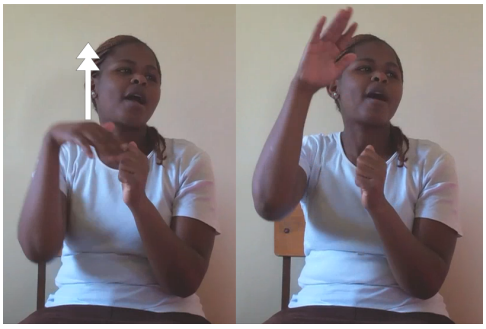


(b) SCIENCE
at upper torso

Figure 8: Minimal pairs that differ by location (variable 23)



(a) PAINT with bidirectional movement



(b) UPPER with unidirectional movement

Figure 9: Minimal pairs that differ by return (uni-/bi-directional movement) (variable 40)

Looking forward, the next steps are toward making improvements in the pipeline. First, we will continue to diagnose the minimal pair mismatches to discover exactly where the process breaks down, such as in the high number of minimal pairs for location being missed. This includes checking both the expected minimal pairs that were not found, described above, as well as signs that differed by one difference in the output, but were not already in the minimal pair dataset. Second, to catch problems with the tool which have not yet been identified and make sure it does not break with new updates, the diagnostic set of test signs will be expanded from the initial set. Third, the coding of the orientation parameter will be reconsidered and new variables tested in the pipeline.

## 8. Acknowledgements

Funded by the European Union (ERC, *SemaSign*, No.101117395). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Early work on this tool was done at Radboud University, supported by the following grant project: *Deaf Communication without a Shared Language*; VICI project funded by NWO.

## 9. Bibliographical References

Robbin Battison. 1978. *Lexical Borrowing in American Sign Language*. Linstok Press, Silver Spring, MD.

Carl Börstell, Onno Crasborn, and Lori Whynot. 2020. Measuring lexical similarity across sign languages in Global Signbank. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages*, pages 21–26, Marseille, France. European Language Resources Association (ELRA).

Diane Brentari. 1989. *A Prosodic Model of Sign Language Phonology*. MIT Press, Cambridge, MA.

Naomi K. Caselli, Karen Emmorey, and Ariel M. Cohen-Goldberg. 2021. The signed mental lexicon: Effects of phonological neighborhood density, iconicity, and childhood language experience. *Journal of Memory and Language*, 121:104282.

Rachel Elizabeth Channon. 2002. *Phonological Representations and Temporal Sequencing in ASL and Other Sign Languages*. Ph.D. thesis, University of Maryland, College Park.

David P. Corina. 1993. To branch or not to branch: Underspecification in ASL handshape contours. In Geoffrey R. Coulter, editor, *Phonetics and Phonology: Current Issues in ASL Phonology*, pages 63–95. Academic Press.

Bart de Boer, Wendy Sandler, and Simon Kirby. 2012. New perspectives on duality of patterning: Introduction to the special issue. *Language and Cognition*, 4:251–259.

Thomas Hanke. 2004. HamNoSys – representing sign language data in language resources and language processing contexts. In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 1–6, Lisbon, Portugal. European Language Resources Association (ELRA).

Lisa Loy and Hope E. Morgan. 2026. Hns2cf: A mapping tool from hamnosys to sl catform. In *Proceedings of the LREC-COLING 2026 12th Workshop on the Representation and Processing of Sign Languages: Language in Motion*, Palma, Mallorca (Spain). ELRA Language Resources Association (ELRA).

Mark Alan Mandel. 1981. *Phonotactics and Morphophonology in American Sign Language*. Ph.D. thesis, University of California, Berkeley.

Aurora Martinez del Rio, Casey Ferrara, Sanghee J. Kim, Emre Hakgüder, and Diane Brentari. 2022. Identifying the correlations between the semantics and the phonology of American Sign Language and British Sign Language: A vector space approach. *Frontiers in Psychology*, 13:806471.

David McKee and Graeme Kennedy. 2000. Lexical comparison of signs from american, australian, british and new zealand sign languages. In Karen Emmorey and Harlan L. Lane, editors, *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima*. Psychology Press.

Hope E. Morgan. 2022. *A Phonological Grammar of Kenyan Sign Language*. De Gruyter Mouton, Berlin, Boston.

Hope E. Morgan. 2026. SL CatForm coding schema v.1: A field-based phonological coding for sign languages. Technical report, IDGS, University of Hamburg, Hamburg, Germany.

Hope E. Morgan, Wendy Sandler, Rose Stamp, and Rama Novogrodsky. 2022. ISL-LEX v.1: An online lexical resource of Israeli Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 148–153, Marseille, France. European Language Resources Association (ELRA).

Jason Parks. 2011. Sign language word list comparisons: Toward a replicable coding and scoring methodology. Master's thesis, University of North Dakota.

Lorna Grace Rozelle. 2003. *The Structure of Sign Language Lexicons: Inventory and Distribution of Handshape and Location*. Ph.D. thesis, University of Washington.

Wendy Sandler. 1989. *Phonological Representation of the Sign: Linearity and Nonlinearity in American Sign Language*. De Gruyter Mouton, Berlin, New York.

Yue Sun and David Poeppel. 2023. Syllables and their beginnings have a special role in the mental lexicon. *Proceedings of the National Academy of Sciences*, 120(36):e2215710120.

Els van der Kooij. 2002. *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity*. LOT Publications, Leiden, Netherlands.

Andrew Wedel, Abby Kaplan, and Scott Jackson. 2013. High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2):179–186.

Shi Yu, Carlo Geraci, and Natasha Abner. 2018. Sign languages and the online world online dictionaries & lexicostatistics. In *11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

# Assisting Corpus Annotation: Automatic BIO-Tagging of Clause-Like Units in Polish Sign Language. A Pilot Study on Corpus Data

**Piotr Mostowski**[1], **Anna Kuder**[2], **Joanna Wójcicka**[1]

[1]University of Warsaw; [2]University College London
[1]Section for Sign Linguistics, Faculty of Polish Studies, University of Warsaw, Krakowskie Przedmieście 26/28, 00-927 Warsaw, Poland; [2]Deafness, Cognition and Language Research Centre (DCAL), University College London, 49 Gordon Square, London, WC1H 0PD, UK
{piotr.mostowski, j.filipczak}@uw.edu.pl; a.kuder@ucl.ac.uk

**Abstract**

The creation of large-scale sign language corpora is often bottlenecked by the labour-intensive process of multi-layered annotation that requires manual analysis. One of the annotation steps is the challenging and time-consuming task of segmenting continuous signing into clause-like-units (CLUs). In this paper, we propose an automated segmentation framework for Polish Sign Language (PJM) designed to support manual annotation. To detect sentence boundaries, we adapt the Multi-Stage Temporal Convolutional Network (MS-TCN) architecture, enhanced with a Channel Attention mechanism, to effectively fuse multimodal skeleton features (hands, body, and face) extracted via MediaPipe.

We evaluate the model on a diverse subset of the PJM Corpus (40 video files, 25 signers), containing nearly 16,000 manually annotated clauses prior to the start of this study. The proposed method achieves a Segmental F1-score of 75.43% at IoU = 0.10 and 57.52% at IoU = 0.50, demonstrating a strong capability in localising sentence boundaries. Furthermore, ablation studies reveal that fusing manual kinematics with non-manual prosodic cues (face) yields a significant performance gain (+13.6 pp) over unimodal baselines, empirically confirming the linguistic necessity of incorporating both manual and non-manual articulators in the process of sentence delimitation. The solution offers a viable means for reducing CLU annotation time by automatically generating high-quality clause boundary proposals.

**Keywords:** Sign Language Segmentation, MS-TCN, Polish Sign Language (PJM), Computer-Assisted Annotation, Multimodal Fusion, clause-like unit.

## 1. Introduction

The Polish Sign Language (PJM) Corpus was compiled at the University of Warsaw between 2010 and 2020 (Kuder et al., 2022) and partially published in 2020 as The Open Repository of the Polish Sign Language Corpus (Wójcicka et al., 2020).

Manual sentence-level analysis was conducted after completing all the necessary steps of sign-to-gloss annotation within the project “Multi-layered Linguistic Annotation of Polish Sign Language Corpus”. The sentence identification procedure follows Trevor Johnston's guidelines for clause-like-unit (CLU) identification (Johnston, 2019). This labour-intensive, time-consuming process, fully reliant on manual annotation, was the primary motivation for developing the automatically supported segmentation process presented in this paper.

While automatic Sign Language Recognition (SLR) has advanced significantly, a critical bottleneck remains in the temporal segmentation of continuous signing streams. Traditional SLR approaches often focus on glosses - lexical labels akin to spoken words - as the primary unit of analysis (Hamidullah et al., 2024). However, natural signed discourse is not merely a sequence of discrete lexical items but is organised into larger prosodic and semantic structures.

Existing segmentation methods rarely operate above the lexical level, and to our knowledge, no prior work has attempted automatic CLU-level segmentation for PJM. We propose a semi-automatic approach to CLU segmentation intended to support the annotators' work in manual annotation of sign language corpora. Subsequently, we evaluate the feasibility of using secondary sign language corpus data for training deep learning segmentation models.

## 2. Related Works and Theoretical Aspects

### 2.1 Clause-like-units (CLUs)

The delimitation and analysis of sentence boundaries in sign language corpus research have long been recognized as methodologically challenging (e.g., Fenlon et al., 2008; Hansen and Heßmann, 2007; Crasborn, 2007, Ormel and Crasborn, 2012). There are many obstacles, including the issue of identifying a predicate or correctly delimiting clause structures during sign language production. Sign languages, having no established writing systems, cannot rely on written transcriptions for sentence delimitation as spoken languages do, which allows diverse syntactical forms to coexist. The constant influence of the ambient spoken language results in the occurrence of language-contact phenomena in sign languages (Sutton-Spence,



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 351–360
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

1999), whose presence increases the level of difficulty in the CLU annotation process.

Research on sentence delimitation in sign languages remains scarce. When it comes to prosodic elements at sentence boundaries, there exist studies of, e.g., Israeli Sign Language (Nespor and Sandler, 1999); American Sign Language (ASL; Wilbur, 1994, 1999; Nicodemus, 2007; Hochgesang, 2009), Sign Language of the Netherlands (van der Kooij, Crasborn and Emmerik, 2006), British Sign Language (BSL; Fenlon et al., 2008), German Sign Language (DGS; Hansen and Heßmann, 2007; Herrmann, 2010), and Hong Kong Sign Language (HKSL; Sze, 2008). Across these studies, the non-manual markers commonly identified as fulfilling prosodic functions include blinks, head movements (such as tilts and headshakes), facial expressions, eye gaze, and body leans or shifts. As those elements have been reported to occur more often at the phrase boundaries, empirical studies have revealed that sign language users rely on them to group signs into intonational phrases (IP) in sign language discourse (Nicodemus, 2009, for ASL). In sign languages, the relation between IPs and clauses is defined similarly to the one occurring in spoken languages. As syntactic and prosodic structures in both modalities are non-isomorphic (Nespor and Vogel, 1986), not all IP boundaries coincide with the end of a sentence, but some of them do (Fenlon et al., 2008). While it is often assumed that manual and non-manual cues occur in coordinated "bundles," an alternative perspective suggests that signers primarily rely on the propositional content of utterances - rather than visual suprasegmental features - when identifying sentence boundaries in discourse (Hansen and Heßmann, 2007). One corpus-based approach to syntactic analysis that addresses these challenges involves segmenting sign language data into clause-like units (CLUs; Johnston, 2019).

For the PJM Corpus annotation, the CLU approach has been adopted. The annotation schema is grounded in the analysis of the predicate-argument structure of the delimited segment. Clause segmentation and identification in the dataset follow the framework established by Trevor Johnston in the Annotation Guidelines for the Auslan Corpus (2019). Within this framework, a CLU is understood as the basic unit carrying propositional meaning and is defined as a "meaningful symbolic utterance unit" with predicational properties (Johnston, 2019). Each segmented unit, along with its thematic relations and semantic functions, is analysed in accordance with the approach proposed by van Valin and LaPolla (1997).

## 2.2 Computational Approaches to Sign Language Processing

The fields of SLR and Sign Language Segmentation (SLS) have transitioned in recent years from traditional statistical models to sophisticated deep learning architectures. This section provides an overview of the methodologies used to identify temporal boundaries and classify linguistic units within continuous signing streams.

### 2.2.1 Sign Language Segmentation and Boundary Detection

SLS aims to identify the precise temporal boundaries (start and end frames) of individual signs or linguistic units (Renz et al., 2021; Zhao et al., 2025). Early approaches often treated this issue as a frame-wise binary classification problem or relied on alignment strategies using Hidden Markov Models (HMMs) (Vogler and Metaxas, 2001). However, these methods struggled with the fluid nature of continuous signing, particularly the transition phases known as movement epenthesis.

Recent research has favoured the BIO (Beginning-Inside-Outside) tagging scheme, which has proven more effective than binary tagging for continuous streams where signs transition smoothly without intervening pauses (Moryossef et al., 2023; Rao et al., 2025). Current state-of-the-art models for boundary detection utilize Multi-Stage Temporal Convolutional Networks (MS-TCN) to capture long-range temporal dependencies and refine predictions iteratively (Renz et al., 2021). For instance, the Multimodal Handshape-aware Boundary (MHB) detection framework enhances accuracy by integrating 3D skeletal features with velocity and acceleration cues (Zhao et al., 2025). While there is a growing emphasis on coarse-grained "subtitle-level" segmentation for machine translation (Bull et al., 2020), fine-grained segmentation of prosodic units remains an open challenge.

### 2.2.2 CSLR Context

While this study focuses on segmentation, it is closely linked to Continuous Sign Language Recognition (CSLR). CSLR processes natural, unsegmented sentences and remains significantly more challenging than Isolated SLR due to the lack of explicit frame-level boundary annotations (Hamidullah et al., 2024; Zhao et al., 2025).

Constantly evolving model architectures have moved from handcrafted features to Recurrent Neural Networks (RNNs) and Transformers (Adeyanju et al., 2021; Ridwang et al., 2023). However, a persistent bottleneck in CSLR is the reliance on "gloss-level" supervision. Recent studies suggest that improving the temporal segmentation of meaningful units (beyond simple glosses) is critical for advancing the field towards natural language processing (Rastgoo et al., 2025).

#### 2.2.3 Integration of Multimodal and Prosodic Cues

Robust segmentation increasingly relies on multimodal inputs beyond raw RGB frames. Frameworks like MediaPipe allow for the precise extraction of holistic landmarks covering the body, hands, and face (Ridwang et al., 2023; Şahin and Gökgöz, 2024). This facilitates the analysis of specific phonological features, such as hand orientation and their location relative to the body.

Crucially for this study, linguistic research highlights the role of prosodic cues (pauses, eye blinks, head nods, and role-shifts) as reliable indicators of sentence and phrase boundaries (Fenlon et al., 2008; Ormel and Crasborn, 2012; Gabarró-López and Meurant, 2014). While the analysis of these "visible markers" enables high inter-annotator agreement in linguistic studies (Woll et al., 2022), effectively integrating them into automated segmentation models remains a complex task.

Most current architectures still heavily prioritize manual features (handshape/movement), often underutilizing the non-manual signals that carry prosodic information.

### 2.3 Research Gap

Despite long-standing interest in sentence-level units' analysis in sign language linguistics, there remains a substantial methodological gap between linguistically motivated clause-level segmentation and automatic sign language processing approaches.

On the one hand, linguistic research has demonstrated that sentence or clause boundaries in signed languages are complex, multi-cue phenomena, involving combinations of manual signs, non-manual markers, and discourse-level factors (Ormel and Crasborn, 2009; Crasborn, 2007). As a result, segmentation into units such as clauses or CLUs is theoretically grounded but methodologically demanding, requiring expert annotation and explicit criteria (Johnston and Shembri, 2006; Gabarró-López and Meurant, 2014).

On the other hand, automatic SLR and sign language processing has largely focused on isolated sign recognition or continuous sign recognition at the level of lexical signs derived from weak proxies such as subtitle boundaries (Rao et al., 2025; Rastgoo et al., 2025). While recent machine-learning approaches have addressed boundary detection in continuous signing, these efforts predominantly target sign/lexical boundaries rather than higher-level syntactic/discourse units (Zhao et al., 2025).

To date, automatic detection of sentence - or clause-level boundaries in signed languages remains largely unexplored, particularly using units derived from corpus-based annotation practices. Although this disparity is striking, it is understandable given the imprecise nature of manual annotation and low availability of richly annotated sign language corpora that already encode clause-level structures (Ormel and Crasborn, 2012; Johnston and Shembri, 2006). Regardless of those obstacles, the lack of computational work leveraging such annotations represents a clear research gap at the intersection of sign language linguistics, corpus research, and machine learning.

## 3. Current Study

In this paper, we shift the segmentation focus from glosses to CLUs (Johnston, 2019). CLUs capture the semantic envelopes of signed utterances, incorporating not only manual signs but also the non-manual markers that define prosodic boundaries. Detecting these units is essential for understanding the flow of a discourse yet annotating them manually is cognitively demanding and time-consuming. Therefore, the primary motivation of this work is to develop a method for computer-assisted annotation. We envision a tool that can automatically suggest preliminary CLU boundaries to human annotators, thereby streamlining the corpus creation workflow and reducing the "cold start" burden in annotation tasks.

This work represents a preliminary study utilizing a secondary data source. The material used in our experiments was not generated specifically for machine learning purposes; rather, it originates from the Polish Sign Language Corpus, designed for linguistic analysis of PJM. Consequently, the data is naturalistic and highly variable, containing the inherent "noise" of real-world dialogue that was not curated or staged for algorithmic optimization. This presents a unique challenge compared to controlled computer vision datasets, as the model must grapple with diverse pacing, co-articulation, and varying recording conditions inherent to linguistic fieldwork.

To address this challenge, we propose a robust segmentation pipeline using the Multi-Stage Temporal Convolutional Network (MS-TCN) enhanced with a multimodal attention mechanism. We formulate the segmentation task using the BIO tagging scheme, allowing the model to distinguish the sharp onset of a CLU from its sustained duration. Our experiments investigate the contributions of different modalities, specifically the interplay between manual articulators and non-manual signals, in defining these boundaries within "noisy", naturalistic data.

### 3.1 Objective

While previous research has demonstrated the feasibility of temporal models for short-term structure recovery (e.g., isolated sign boundaries) (Rastgoo et al., 2025; Rao et al., 2025), there is little work on clause-level boundary tagging

grounded in complex, pragmatically defined units rather than simple pauses or gloss transitions. The study addresses the following research problems:

1. To what extent a temporal convolutional network (MS-TCN) can learn to approximate expert human segmentation when trained on secondary corpus data?
2. What is the relative contribution of different visual articulators (manual cues/hands versus non-manual markers) in defining CLU boundaries? Does the model rely on the kinetic "stroke" of the hands or the sustained prosodic context of the face?
3. Can the proposed BIO-tagging framework serve as a reliable heuristic for computer-assisted annotation, reducing the cognitive load on human annotators by suggesting preliminary boundaries?

More specifically, this study investigates whether temporal models operating on pose-based visual features can approximate expert human segmentation decisions at the clause level.

By framing clause boundary detection as a sequence-labelling problem over continuous signing data, this study explores the feasibility of reusing high-quality corpus annotations for the development of preliminary, assistive segmentation tools, rather than fully automatic replacements for annotation conducted by humans.

## 3.2 Dataset

The data used in this study are drawn from the PJM Corpus, a large, publicly available corpus developed for linguistic research on PJM (Rutkowski et al., 2017; Kuder et al., 2022, Wójcicka et al., 2020). The corpus consists of video recordings of Deaf native and near-native PJM signers engaged in a range of communicative tasks and interactional settings, including both monologic and dialogic recordings. Crucially, the PJM Corpus prioritizes ecological validity. The recordings capture naturalistic, spontaneous discourse, characterized by variable pacing, co-articulation, and frequent overlapping of manual and non-manual signals. This makes the material particularly challenging for standard machine learning models but highly suitable for research on higher-level linguistic units, such as single and matrix clauses.

All annotations currently present in the corpus were produced through a multi-stage process involving application of different annotation standards, consistency checks, and expert review, resulting in a high-quality annotated material (Mostowski et al., 2018). This expert-validated data provides a rare opportunity to train segmentation models on authentic, deep-structure linguistic data rather than superficial visual cues.

However, some tasks and fragments of the language material are not annotated to the same extent. Therefore, the dataset for this study was selected based on the number of CLU tags for each signer. Out of 150 participants, 25 (15 F, 10 M) were selected with a minimum of 100 tags on the CLU tier in their continuous signing strings. This ensures a wide representation of individual signing styles, spanning various speeds and articulation dynamics. In total, 40 video recordings with total length of 600 minutes were chosen for the study[1]. The final dataset contains a total of 15,972 annotated CLUs, providing a robust foundation for deep learning.

| CLUs per video | |
|---|---|
| Mean | 399 |
| Median | 329 |
| Min | 36 |
| Max | 1,315 |

Table 1: CLU statistics in the dataset

The high variance in CLU density (ranging from 36 to over 1,300 per video) necessitates the stratified sampling strategy described in Section 5.1, ensuring that both sparse and dense recordings are equally represented in the test set.

## 3.3 CLU Annotation

Segmentation and lemmatization of the material were performed by Deaf annotators during creation of the PJM Corpus (2010-2020). Texts were segmented into single signs. Transitional hand movements (preparation/onset and end/offset phases of the movement) were included in each annotation, contrary to the practices applied in some other signed corpora.

Segmentation of the material into CLUs was subsequently carried out by the second and third authors, who are hearing native Polish speakers and L2 PJM signers, with additional help from a Deaf annotator.

The CLU tagging process, following Trevor Johnston's guidelines (2019), is based on differentiating between single clause-like-units and non-clause segments. A CLU is identified based on its predicative interpretation, typically expressed through a verbal predicate (V) and its arguments (A, A1/A2). Each element can be expressed manually in the form of a sign, a classifier construction, or a depiction. Either a

[1] Due to the nature of each recording session, which includes dyadic communication based on turn-taking, recording fragments with no signing longer than 45 seconds have been excluded from the dataset.

predicate or an argument in a CLU can be articulated non-manually (through head nods, role-shifts, or eye gaze), as part of constructed action (CA), or with mouthing. Each sign in a CLU is tagged for an argument role, thematic role, semantic function, and part of speech.

# 4. Methods

## 4.1. Visual Feature Extraction

For the purposes of the present study, we utilize the MediaPipe Holistic framework, which provides simultaneous perception of body pose, face mesh, and hand tracking. However, using raw coordinates directly is suboptimal due to noise and redundant information. Therefore, we implement a selective feature extraction and normalization pipeline tailored to the characteristics of the PJM Corpus.

### 4.1.1 Landmark Selection

We filter the raw MediaPipe (Lugaresi et al, 2019) output to focus strictly on articulators relevant to the sign language production, driven by the specific recording conditions in the corpus.

**Pose**: Since all informants in the PJM Corpus were recorded in a seated position, lower-body landmarks (hips, knees, ankles, feet) provide no linguistic information. Furthermore, these points were frequently static or out of the camera frame. Consequently, we discard them, retaining only the upper-body landmarks (shoulders, elbows, wrists) and the nose/eyes as reference points for head orientation.

**Face**: Instead of the dense 468-point mesh, we extract a semantic subset of landmarks focusing on the contours of the face, eyebrows, eyes, and lips. This reduces dimensionality while preserving critical non-manual markers required for a clause boundary detection.

**Hands**: We utilize the full set of 21 3D landmarks for both the left and right hand to capture intricate handshapes and finger articulation.

### 4.1.2 Normalization and Feature Augmentation

Raw coordinates were first corrected for the aspect ratio to ensure geometric consistency across different recording sessions. To account for variations in signer positioning and distance from the camera, we apply scale and translation normalization. The coordinate system is centred on the midpoint between the signer's shoulders (the “neck” point), and all landmarks are scaled relative to the Euclidean distance between the shoulders. This ensures that the magnitude of movements is invariant to the signer's body size or their distance from the lens.

Finally, to capture the temporal dynamics essential for boundary detection (the rapid acceleration at the start of a sign or the deceleration/hold at the end of a clause) we augment the feature vector with temporal derivatives.

1. Velocity (First Derivative):
$$\Delta x_t = x_t - x_{t-1}$$
2. Acceleration (Second Derivative):
$$\Delta^2 x_t = \Delta x_t - \Delta x_{t-1}$$

The final input vector for each frame consists of the concatenated normalized positions, velocities, and accelerations $[p, \Delta p, \Delta 2p]$ for the selected subset of landmarks.

## 4.2 From Annotation to Supervision - BIO

For converting the continuous linguistic annotation into a format suitable for supervised learning, we adopted the BIO tagging scheme. This section details the mapping from temporal timestamps to frame-level classes and justifies the design choices based on the kinematic properties of PJM.

### 4.2.1 BIO Classes

Three classes have been identified; each assigned a label:

1. B (Begin): Represents the onset frames of a CLU. This class marks the precise temporal boundary where a new unit starts.
2. I (Inside): Denotes the frames constituting the duration of the CLU. This class covers the continuous signing activity following the onset.
3. O (Outside): Corresponds to non-signing intervals (background). This includes frames where the signer is in a rest position or performing non-linguistic transitional movements between CLUs.

### 4.2.1 Handling Annotation Uncertainty

Treating the CLU onset as a single, exact frame proves problematic due to human annotators’ jitter. To mitigate this sparsity, we apply a temporal label dilation to the “Begin” class. Instead of a one-hot spike, the “B” label is assigned to a window of frames centred around the ground truth: $t \pm k$ (resulting in a 3-frame active window). This creates a denser supervision signal, encouraging the model to detect the onset phase rather than an arbitrary millisecond instant.

### 4.2.3 Rejection of BIOE

In the preliminary phase, we evaluated a more granular BIOE (Begin-Inside-Out-End) scheme, where a distinct class “E” marked the end of a unit. However, empirical results showed that the model struggled to consistently predict the “E” class, which negatively impacted overall convergence. We attribute this to the kinematic asymmetry of signing: while the onset (“B”) is often marked by a sharp, high-velocity movement, the offset (“E”) is frequently characterized by a gradual deceleration, a hold, or a relaxed return to a

neutral position. Given our goal of computer-assisted annotation, we prioritized high recall for the onset ("B") and the sustained context ("I"), as these provide the most critical anchors for human annotators.

## 4.3 Data Sampling and Augmentation

Once the labels were defined, the continuous video streams were processed into training batches using a pipeline designed to address class imbalance and variable signing speeds.

### 4.3.1 Sliding Window and Weighted Sampling

The video streams are sliced into fixed-length windows (W = 512) with a stride 64. The sliding window approach generates a large dataset but exacerbates class imbalance, as "Begin" frames are statistically rare.

To counteract this, we implement a content-aware sample weighting strategy. Windows containing at least one "Begin" frame are assigned a higher importance weight ($w_{pos}$ = 4.0) compared to background-only windows ($w_{neg}$ = 1.0). This forces the optimization process to focus on correctly localizing boundaries rather than defaulting to the majority class.

### 4.3.2 Temporal Augmentation

Natural signing exhibits significant variance in speed. To ensure the model is robust to tempo variations, we apply a stochastic time-warping augmentation.

For each training sample, the temporal dimension is rescaled by a factor $\lambda$ simulating speed variations of $\pm 20\%$.

**Feature Interpolation:** The input features X are resampled using linear interpolation to preserve smooth trajectories.

**Label Interpolation:** The discrete labels Y are resampled using nearest-neighbour interpolation to maintain valid class integers. Depending on $\lambda$, the sequence is either randomly cropped (if $\lambda > 1$) or zero-padded (if $\lambda < 0$) to match the fixed window size required by the network.

### 4.3.3 Loss Function Configuration

Complementing the window sampling strategy, we introduced class-specific weights within the objective function to further mitigate the dominance of the background class. We minimize the Weighted Cross-Entropy Loss. Based on the empirical distribution of classes in the PJM Corpus, the weights were set to $\lambda_{Out} = 0.5$, $\lambda_{Beg} = 5.0$, $\lambda_{In} = 1.0$.

This configuration assigns the highest penalty to the "Begin" class (10x higher than "Out"), strictly enforcing high recall for sign onsets, while downweighing the majority of "Out" class to prevent the model from converging to a trivial solution (i.e., predicting "Out" label for each frame).

## 4.4 Model Architecture: MS-TCN with Channel Attention

To model the temporal dependencies of clause boundaries, we utilize a Multi-Stage Temporal Convolutional Network (MS-TCN) (Farha and Gall, 2019), enhanced with the Channel Attention mechanism for a multimodal feature fusion.

### 4.4.1 Input Feature Recalibration

Prior to temporal processing, the input sequence X is passed through a Squeeze-and-Excitation (SE) block (Hu et al., 2018). This module dynamically recalibrates the importance of feature channels (pose vs. hands vs. face). It operates by aggregating global temporal information via Global Average Pooling, followed by a bottleneck gating mechanism (two fully connected layers with a reduction ratio r = 16, ReLU, and Sigmoid activation). The resulting weights scale the original input channels, suppressing noise and emphasizing relevant articulators.

### 4.4.2 MS-TCN Backbone

The core architecture consists of four cascades of sequential refinement stages, each designed to iteratively refine the segmentation predictions. Each stage (SS-TCN) is built from 11 dilated residual layers. To capture long-range dependencies, the dilation factor at layer l is set to $d_l = 2^l$. We utilize 128 feature maps per layer to maintain a rich representation of the signing context.

The first stage operates on the recalibrated features to generate initial class probabilities. Subsequent stages (s = 2…4) take the softmax output of the previous stage as input. This hierarchical refinement smoothes the predictions and reduces over-segmentation errors common in frame-wise classification.

The model is trained end-to-end using a multi-stage loss function, where the Weighted Cross-Entropy Loss is calculated and summed at the output of each stage.

# 5. Experiments

## 5.1 Setup: Data Partitioning and Metrics

### Stratified Data Split

To ensure a balanced distribution of linguistic complexity across sets, we employed a stratified sampling strategy based on CLU density. All videos were ranked by the number of annotated CLUs and divided into four quartiles. From each quartile, one video was randomly assigned to the validation set and one to the test set, with the remaining videos forming the training set. This resulted in a robust 80/10/10 split that preserves the variance of signing styles and pacing.

**Evaluation Metrics**

We report performance using two complementary metrics:

1. Frame-wise F1: Measures classification accuracy at the individual frame level.
2. Segmental F1 (IoU): Measures the quality of detected boundaries. A predicted segment is counted as a True Positive if its Intersection over Union (IoU) with the ground truth exceeds a threshold **τ.** We report results for **τ** ∈ {0.1, 0.25, 0.5}.

## 5.2 Implementation Details

The model was implemented in PyTorch. Input sequences were processed using sliding windows of size W = 512 with a stride of S = 64 to maintain context continuity. Training was performed for 50 epochs with a batch size of 32, using the Adam optimizer with an initial learning rate of $5 \cdot 10^{-4}$. To prevent overfitting, we employed early stopping to monitor the validation loss.

# 6. Results

We present the quantitative evaluation of the proposed MS-TCN model with Channel Attention on the PJM Corpus test set. We analyze the results on two levels: a frame-wise classification accuracy and a segmental boundary quality (IoU).

## 6.1. Quantitative Evaluation

Table 2 details the performance of the full multimodal model. The system achieves a Global Accuracy of 83.51% and a Weighted F1-score of 84.95%.

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Out (Background) | 0.935 | 0.928 | 0.932 | 72,950 |
| Beg (Onset) | 0.185 | 0.430 | 0.259 | 5,124 |
| In (Inside) | 0.846 | 0.741 | 0.790 | 49,586 |
| Macro Average | 0.656 | 0.700 | 0.660 | 127,660 |

Table 2: Frame-wise Performance (Test Set)

**Analysis of Class Performance**

The high performance on the “Inside” class (F1 = 0.79) confirms the model's ability to maintain context throughout the duration of a clause. For the critical “Begin” class, we observe a trade-off: while Precision is relatively low (0.185), the Recall is significantly higher (0.430). This aligns with our design goal for computer-assisted annotation: the system acts as a high-sensitivity proposal generator. It is functionally less costly for a human annotator to reject a false positive start than to manually search for a missed onset.

**Analysis of Class Imbalance and Performance**

As shown in the “Support” column, the dataset exhibits significant class imbalance, which is inherent to the temporal structure of sign language: clause onsets (“Beg”) are transient events (approx. 120ms), whereas the content (“In”) and background (“Out”) span longer durations.
Despite this imbalance, the model achieves a high Weighted F1-score of 84.95%, driven by its robustness on the majority classes.

For the minority “Beg” class, the lower Precision (0.185) is expected in frame-wise evaluation, as a temporal shift of just 1-2 frames is penalized as a False Positive, even if the detection is semantically correct. This highlights the necessity of the segmental metrics (Table 2) for a fair assessment.

**Segmental Performance (IoU)**

Since frame-level metrics do not strictly penalize fragmented predictions, we rely on Segmental F1 scores (Table 3) to assess the temporal coherence of the detected units.

The model achieves an F1@50 of 57.52%, indicating that most predicted clauses overlap significantly with the ground truth. Furthermore, the high F1@10 (75.43%) suggests that the system correctly identifies the approximate location of 3 out of 4 clauses, providing a strong baseline for manual refinement.

| Metric | Threshold | F1-score (%) |
|---|---|---|
| F1@10 | 0.10 | 75.43 |
| F1@25 | 0.25 | 69.96 |
| F1@50 | 0.50 | 57.52 |

Table 3: Segmental F1-Scores at different IoU thresholds

## 6.2. Ablation Study: Impact of Modalities

To quantify the contribution of manual (hands) versus non-manual (face) features to the segmentation task, we conducted an ablation study.

| Input Modality | F1@ 10(%) | F1@ 25(%) | F1@ 50(%) | λ (vs Full) |
|---|---|---|---|---|
| Face Only | 70.00 | 64.30 | 43.92 | -13.60 |
| Hands Only | 69.39 | 63.74 | 47.92 | -9.60 |
| Full | 75.43 | 69.96 | 57.52 | —- |

Table 4: Impact of Modalities on Boundary Detection (Test Set)

To better understand the source of errors, we also analyze the frame-level metrics specifically for the critical “Begin” class (Table 5). The results in Table 4 reveal the complementary nature of manual and non-manual signals in PJM.

Surprisingly, the Face-Only model performs competitively at looser thresholds (F1@10: 70.00%), slightly outperforming the Hands-Only baseline. This confirms that facial expressions (prosodic markers) are strong indicators of clause boundaries. However, the performance drops significantly at stricter thresholds (F1@50: 43.92%), suggesting that while the face indicates the presence of a boundary, it lacks the temporal sharpness to define the exact frame of the onset.

The Role of Manuals (Hands): The Hands-Only model offers better localization precision (F1@50: 47.92%), driven by the distinct kinematic changes (e.g., lift-off from rest) associated with sign initiation. Nevertheless, without facial context, the model struggles to distinguish meaningful signs from transitional movements, resulting in a lower overall accuracy compared to the full model.

| Model Variant | Precision (Beg) | Recall (Beg) | F1-Score (Beg) | Macro F1 (Global) |
|---|---|---|---|---|
| Face Only | 12.74% | 28.10% | 17.53% | 60.16% |
| Hands Only | 15.77% | 47.07% | 23.62% | 62.86% |
| Full | 18.54% | 42.99% | 25.91% | 66.01% |

Table 5: Frame-wise Metrics for the “Begin” Class (Ablation)

As shown in Table 5, the Hands-Only model exhibits the highest Recall for the “Begin” class (47.07%) but suffers from lower Precision (15.77%). This suggests that hands are “hypersensitive” triggers - they detect almost every movement, leading to many False Positives. By integrating facial features, the Full Model acts as a filter: it slightly reduces Recall (43.0%) but significantly boosts Precision (+2.8 pp) and the overall F1-score. The Attention mechanism effectively suppresses “lone” hand movements that are not accompanied by the prosodic intent visible on the face.

## 7. Conclusion

We addressed the critical bottleneck of manual annotation in sign language corpus linguistics by proposing an automated segmentation framework for PJM. We adapted the MS-TCN architecture, enhancing it with a Channel Attention mechanism to effectively fuse multimodal input streams from pose estimation keypoints.

Our experiments demonstrate that a lightweight model operating on skeleton data can effectively localize Clause-Like Units (CLUs).

The proposed model achieves a Segmental F1@10 of 75.43% and an F1@50 of 57.52%. These results indicate that the system can generate high-quality boundary proposals, correctly identifying the approximate location of three out of four clauses.

Through ablation studies, we empirically validated the linguistic hypothesis regarding the interplay of articulators. We found that while manual features (hands) provide kinematic precision for the onset, non-manual signals (face) are essential for capturing the prosodic envelope of the sentence. The fusion of both modalities yielded a significant performance gain (+10 pp in F1@50) over single-modality baselines.

**Impact on Annotation Workflow**

The primary goal of this research was to facilitate Computer-Assisted Annotation. Given the high recall of our system, it can be integrated into annotation tools (e.g., ELAN, iLex) to pre-segment video timelines. By transforming the task from “search and annotate” to “verify and adjust”, we estimate a substantial reduction in the cognitive load and time required for corpus creation.

**Future Work**

Future research should focus on four main areas.

First, investigating methods for effectively applying machine learning techniques to naturally occurring linguistic data that has not been designed or curated for computational use.

Second, investigating whether the learned boundary features can be transferred to other sign languages (e.g., DGS or Auslan) without extensive retraining.

Third, conducting a qualitative study with expert linguists to measure the actual acceleration of the annotation process when using model-generated proposals.

Fourth, examining the impact of recording type on model performance. The present study is based on a Polish Sign Language (PJM) corpus comprising both monologue and dialogue recordings. As dialogue data involve interactions between signers, they may introduce additional complexity compared to monologic data. Future work should therefore investigate whether model performance differs across these recording types and explore how interactional structure affects boundary detection.

## 8. Acknowledgments

The “Multi-layered Linguistic Annotation of Polish Sign Language Corpus” project was completed within the National Programme for the

Development of the Humanities (NPRH) of the Polish Ministry of Science and Higher Education (0111/NPRH3/H12/82/2014, PI: prof. Paweł Rutkowski).

The implementation of the MS-TCN architecture was based on the original structure proposed by Farha and Gall (2019), adapted to the PJM dataset using PyTorch framework with the assistance of LLM-based coding tools for rapid prototyping. AI assistance was used for checking spelling and grammar in the final stages of manuscript preparation.

## 9. Bibliographical References


Adeyanju, I. A., Bello, O. O., and Adegboye, M. A. (2021). Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications*, 12: 200056. DOI: 10.1016/j.iswa.2021.200056

Bull, H., Braffort, A., and Gouiffès, M. (2020). MEDIAPI-SKEL: A 2D-skeleton video database of French Sign Language with aligned French subtitles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (*LREC 2020*), pages 6063–6068, Marseille,France.

Crasborn, O. (2007). How to recognise a sentence when you see one. *Sign Language & Linguistics*, 10: 103–111. DOI: 10.1075/sll.10.2.03cra

Crasborn, O., van der Kooij, E., and Ros, J. (2012). On the weight of sentence-final prosodic words. *Sign Language & Linguistics*, 15(1): 11–38. DOI: 10.1075/sll.15.1.02cra

Farha, Y. A. and Gall, J. (2019). MS-TCN: Multi-stage temporal convolutional network for action segmentation. *arXiv preprint arXiv:1903.01945*.

Fenlon, J., Campbell, R., and Woll, B. (2008). Seeing sentence boundaries. *Sign Language & Linguistics*, 10(2): 177–200. DOI: 10.1075/sll.10.2.06fen

Gabarró-López, S. and Meurant, L. (2014). When nonmanuals meet semantics and syntax: A practical guide for the segmentation of sign language discourse. In *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, Reykjavík, Iceland. https://www.sign-lang.uni-hamburg.de/lrec/pub/14018.pdf

Hamidullah, Y., van Genabith, J., and España-Bonet, C. (2024). Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pages 425–434, Bangkok, Thailand.DOI: 10.18653/v1/2024.acl-short.40

Hansen, M. and Heßmann, J. (2007). Matching propositional content and formal markers: Sentence boundaries in a DGS text. *Sign Language & Linguistics*, 10(2): 145–175. DOI: 10.1075/sll.10.2.05han

Herrmann, A. (2010). The interaction of eye blinks and other prosodic cues in German Sign Language. *Sign Language & Linguistics*, 13(1): 3–39. DOI: 10.1075/sll.13.1.02her

Hochgesang, J. A. (2009). Is there a sentence in ASL? Insights on segmenting data in ASL. In *Sign Language Corpora: Linguistic Issues Workshop*, University College London, London, England.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2019). Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*. DOI: 10.1109/CVPR.2018.00745

Johnston, T. and Schembri, A. (2006). Issues in the creation of a digital archive of a signed language. In *Sustainable Data from Digital Fieldwork: Proceedings of the Conference Held at the University of Sydney, 4–6 December 2006*, pages 7–16, Sydney, Australia. DOI: 10.2307/jj.455887

Johnston, T. (2019). *Auslan Corpus Annotation Guidelines*. Macquarie University, Australia. https://api.semanticscholar.org/CorpusID:201605087

Kooij, E. van der, Crasborn, O., and van Emmerik, W. (2006). Explaining prosodic body leans in Sign Language of the Netherlands: Pragmatics required. *Journal of Pragmatics*, 38(10): 1598–1614. DOI: 10.1016/j.pragma.2005.07.006

Kuder, A., Wójcicka, J., Mostowski, P., and Rutkowski, P. (2022). Open repository of the Polish Sign Language corpus: Publication project of the Polish Sign Language corpus. In *Proceedings of the LREC 2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 118–123, Marseille, France. European Language Resources Association (ELRA).https://aclanthology.org/2022.signlang-1.18.pdf

Lugaresi, C., et al. (2019). MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Nespor, M. and Sandler, W. (1999). Prosody in Israeli Sign Language. *Language and Speech*, 42(2–3):143–176.DOI: 10.1177/00238309990420020201

Nespor, M. and Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris Publications. DOI: 10.1017/s0952675700002219

Nicodemus, B. (2007). Prosodic markers and utterance boundaries in American Sign Language interpretation. *Sign Language &*

*Linguistics*,11(1): 113–122.DOI: 10.2307/j.ctv2rh2b3r

Ormel, E. and Crasborn, O. (2012). Prosodic correlates of sentences in signed languages: A literature review and suggestions for new types of studies. *Sign Language Studies*, 12(2): 279–315. DOI: 10.1353/sls.2011.0019

Mostowski, P., Kuder, A., Filipczak, J., and Rutkowski, P. (2018). Workflow management and quality control in the development of the PJM corpus: The use of an issue-tracking system. In *Proceedings of the 8th Workshop on the Representation and Processing of Sign Languages* (*LREC*), pages 127–132, Miyazaki, Japan.

Moryossef, A., Jiang, Z., Müller, M., Ebling, S., and Goldberg, Y. (2023). Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. DOI: 10.18653/v1/2023.findings-emnlp.846

Rao, K. M., Hamidullah, Y., and Avramidis, E. (2025). Sign language video segmentation using temporal boundary identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (*Volume 4: Student Research Workshop*), pages 1213–1224, Vienna, Austria. DOI: 10.18653/v1/2025.acl-srw.93

Rastgoo, R., Kiani, K., and Escalera, S. (2025). A non-anatomical graph structure for boundary detection in continuous sign language. *Scientific Reports*, 15: 25683.DOI: 10.1038/s41598-025-11598-3

Renz, K., Stache, N. C., Albanie, S., and Varol, G. (2021). Sign language segmentation with temporal convolutional networks. In *Proceedings of ICASSP 2021*, pages 2135–2139, Toronto, Canada. DOI: 10.1109/icassp39728.2021.9413817

Ridwang, A., Ilham, A., Nurtanio, I., and Syafaruddin (2023). Dynamic sign language recognition using MediaPipe library and modified LSTM method. *International Journal of Advanced Science, Engineering and Information Technology*, 13(6): 2171–2180. DOI: 10.18517/ijaseit.v13i6.19401

Rutkowski, P., Kuder, A., Filipczak, J., Mostowski, P., Łacheta, J., and Łozińska, S. (2017). The design and compilation of the Polish Sign Language (PJM) corpus. In P. Rutkowski (ed.), *Different Faces of Sign Language Research*, pages 125–151. Warsaw: Faculty of Polish Studies, University of Warsaw.

Şahin, K. and Gökgöz, K. (2024). Decoding sign languages: The SL-FE framework for phonological analysis and automated annotation. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages*, pages 335–342, Torino, Italy. DOI: 10.63317/4gxd4idrhodw

Sutton-Spence, R. (1999). The influence of English on British Sign Language. *International Journal of Bilingualism*, 3(4): 363–394. DOI: 10.1177/13670069990030040401

Sze, F. (2008). Blinks and intonational phrasing in Hong Kong Sign Language. In *Signs of the Time: Selected Papers from TISLR 2004*, pages 83–107, Barcelona, Spain.

van Valin, R. and LaPolla, R. (1997). *Syntax: Structure, meaning, and function*. Cambridge: Cambridge University Press. DOI: 10.1017/cbo9781139166799

Vogler, C. and Metaxas, D. (2001). A framework for recognizing the simultaneous aspects of American Sign Language. *Computer Vision and Image Understanding*, 81(3): 358–384. DOI: 10.1006/cviu.2000.0895

Wilbur, R. B. (1999). Stress in ASL: Empirical evidence and linguistic issues. *Language and Speech*, 42(2–3): 229–250. DOI: 10.1177/00238309990420020501

Wilbur, R. B. (1994). Eyeblinks and ASL phrase structure. *Sign Language Studies*, 84: 221–240. DOI: 10.1353/sls.1994.0019

Woll, B., Fox, N., and Cormier, K. (2022). Segmentation of signs for research purposes: Comparing humans and machines. In *Proceedings of the LREC 2022 10th Workshop on the Representation and Processing of Sign Languages*, pages 198–201, Marseille, France. DOI: 10.13140/RG.2.2.36548.91062

Zhao, M., Yang, Z., Zhou, Y., Xia, Z., Jin, C., He, X., and Metaxas, D. N. (2025). MHB: Multimodal handshape-aware boundary detection for continuous sign language recognition. *arXiv preprint arXiv:2511.19907*.

## 10. Language Resource References

Wójcicka, J. et. al. 2020. Open Repository of the Polish Sign Language Corpus. https://www.korpuspjm.uw.edu.pl/en

# Introducing VISTA-SL: A Multilingual e-Learning Platform for Deaf and Hearing Learners of Sign Languages

**Irene Murtagh[1], Marc Schulder[2], Annika Herrmann[2], Liona Paulus[2], Julian Bleicken[2], Kostas Blekos[3], Athanasios Konstantakopoulos[3], Klimis Antzakas[4], Dimitrios Kosmopoulos[3], Eva Valls[5], Ricardo Marques[5], Josep Blat[5], Konstantinos Karampidis[6], Ben Elsendoorn[7]**

[1]Department of Informatics and Cyber Security, Technical University Dublin, Ireland
[2]Institute of German Sign Language and Communication of the Deaf, University of Hamburg, Germany
[3]Computer Engineering and Informatics Department, University of Patras, Greece
[4]Department of Education and Social Work, University of Patras, Greece
[5]Department of Engineering, Pompeu Fabra University, Spain
[6]Department of Electrical and Computer Engineering, Hellenic Mediterranean University, Greece
[7]Royal Kentalis, Utrecht, The Netherlands
irene.murtagh@tudublin.ie,
{marc.schulder, annika.herrmann, liona.paulus, julian.bleicken}@uni-hamburg.de
mplekos@upatras.gr, up1061098@ac.upatras.gr, {k.antzakas, dkosmo}@upatras.gr
{eva.valls, ricardo.marques, josep.blat}@upf.edu, karampidis@hmu.gr, b.elsendoorn@kentalis.nl

**Abstract**

This article introduces the VISTA-SL project, which aims to create an integrated e-learning platform for four European sign languages: German Sign Language, Greek Sign Language, Irish Sign Language, and Dutch Sign Language. Designed as a complement to face-to-face classes, the VISTA-SL platform will combine expertise in sign language education and education technologies to provide an adaptive and interactive learning environment suitable for deaf, hard of hearing and hearing users seeking to learn a sign language, whether it constitutes their first language or not. Building on a co-ordinated curriculum that covers vocabulary, grammar and Deaf culture materials, the platform will provide video material presented by deaf L1 signers, together with games and gamification features to motivate learning, while also providing several assistive technologies. By leveraging cutting edge language processing and computer vision approaches, the platform will provide augmented reality feedback, 3D avatars and an LLM-based virtual instructor, as part of the learning environment. VISTA-SL is developed in collaboration with end-user focus groups, comprising deaf, hard of hearing and hearing individuals. This will serve to ensure that the educational platform aligns with the expectations and needs of its intended users.

**Keywords:** sign language education, language learning, education technologies, deaf mobility

## 1. Introduction

Learning a signed language (SL) is a complex task. In addition to developing proficiency in a sign language, it is critical that learners also understand the cultural aspects and significance of learning a minority language. Learner needs also differ depending on whether their first language is a spoken language or another sign language.

According to the World Health Organisation, up to 5 out of every 1,000 children are born with hearing loss or acquire it soon after birth, yet the infrastructure for sign language education remains inadequate (Altun, 2025). While traditional face-to-face SL education is considered the 'gold standard' for both teaching and learning SLs (providing immediate feedback, visual interaction and cultural immersion), Europe is faced with a significant and ongoing shortage of qualified SL instructors and supporting learning materials (Mathews, 2020). This presents a critical challenge for the region. This issue is further highlighted as SLs gain legal recognition as full natural languages that are independent from spoken languages and as the primary languages of deaf[1] communities.

Advances in digital technologies provide the potential to support both SL instructors and SL learners with regard to teaching and learning support tools – regardless of the hearing status. These advances leverage technological approaches to provide learners with unique advantages, including visual sign demonstrations from various perspectives, and responsive and sometimes immediate feedback mechanisms. Advances in neural

[1]Based on the critical discourse of whether to differentiate audiological status and cultural identity through different capitalisations of *deaf*, we follow Kusters et al. (2017) and others in using lowercase *deaf* throughout to avoid an unintentional dichotomy or possible exclusion of individuals that do not neatly fit in these categories.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 361–370
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

natural language processing and computer vision have been harnessed to develop digital tools that empower learners, allowing them to adopt a suitable pace based on personal preference and ability, while also supporting access to education from remote locations and flexible on-demand delivery.

This paper presents VISTA-SL[2] (Visual Interactive System for Teaching and Assessment of Sign Languages), an integrated e-learning platform that provides the potential to support SL teaching and learning through the use of a novel combination of augmented reality visual feedback, 3D avatar technology, and Large Language Model (LLM) powered personalised feedback. The platform will provide curricula for multiple sign languages, covering vocabulary, grammar, interaction, and Deaf culture, as well as games for language learning. Both deaf and hearing learners are taken into account in the design of the curricula and platform interface.

The remainder of the paper is structured as follows. In section 2, we provide an overview of VISTA-SL as a tool for teaching and learning. Section 3 outlines previous and current research in the area, representing both technical and educational aspects. A description of the curriculum development is provided in section 4, followed by the employed education technologies in section 5. In section 6, we discuss the VISTA-SL platform with regard to its architecture, user progress tracking, accessibility and privacy. We conclude the paper in section 7 and outline the future direction of our work.

## 2. VISTA-SL for Teaching and Learning

VISTA-SL is an Erasmus+ project involving six partner institutions across five European countries. It aims to provide accessible, open sign language education that can scale beyond the limitations of traditional classroom instruction. The platform provides support four European sign languages:

- German Sign Language (DGS)
- Greek Sign Language (GSL)
- Irish Sign Language (ISL)
- Dutch Sign Language (NGT)

Entries are cross-linked between languages, allowing interested users to compare different languages. VISTA-SL is designed for two groups of second language learners (L2) that differ in their familiarity with the visual language modality: Signers learning another sign language (L2M1) and learners that encounter SLs for the first time (L2M2).

The coordinated curriculum, together with the technical components of this project, are steered and evaluated by focus groups. The focus groups include deaf and hearing students as well as established deaf sign language teachers. By including deaf course instructors, VISTA-SL supports the position that deaf people must be at the forefront of teaching their own first language. Learners are provided with the benefit of learning sign language from deaf L1 signers, while also experiencing insight into deaf culture. This approach is instrumental in preserving authenticity, while also promoting role models.

VISTA-SL is informed by the Common European Framework of Reference for Languages (CEFR)[3] (Council of Europe, 2020) and the ProSign project[4] regarding descriptors for reception, interaction and production, although it follows a different level structure that more directly integrates with the platform's goals. VISTA-SL also builds on experiences made creating an e-learning platform prototype in the CAT-SL project (Blekos et al., 2026)[5].

The VISTA-SL project has the following key objectives and contributions:

- Creation of interactive learning curricula and educational games in four EU sign languages, designed for deaf and hearing users and developed by sign language educators and linguists.
- Design of a modular, accessible web platform with privacy-preserving architecture.
- Development of real-time AR feedback tools using hand and facial landmark detection to provide immediate visual guidance on sign execution accuracy.
- Integration of 3D signing avatars for demonstrating correct sign production, with controllable playback for detailed examination.
- Implementation of an LLM-based virtual instructor that provides personalised coaching based on learner progress and performance patterns.

### 2.1. Representation

Sign languages are intrinsically linked with the deaf communities from which they originate. When developing educational materials and technologies centred around these languages, it is essential to involve these communities in meaningful ways

[2] https://vista-sl.eu/

[3] www.coe.int/lang-cefr

[4] https://www.ecml.at/en/ECML-Programme/Programme-2012-2015/ProSign

[5] http://cat-sl.eu

throughout the entire life cycle of a project (Harris et al., 2009; Desai et al., 2024).

In this project, deaf signing people are involved on all levels: as principal investigators, as scientific researchers, as teachers, as students, as learners, and as focus group members. Thereby, deaf-led and participatory research—which is ranked very highly in current discussions about the primacy of deaf people in the research and development of technology for deaf people—plus taking an applied perspective into account at all steps of the project, is given space as much as possible. Above all, including deaf and hearing learners and deaf educators as target groups is essential, thus considering their respective needs and requirements as platform users. Feedback from these groups has a direct impact on both the educational content, technology implementation and platform interface.

# 3. Related Work

## 3.1. Sign Language Education

For second language learners, it matters whether their first language is in the same modality (M1) or another (M2). Signers learning another sign language (L2M1) and learners that encounter SLs for the first time (L2M2) therefore represent two distinct groups of learners (Schönström and Holmström, 2022). Testing previous assumptions on second language learning, several studies on L2M1 and L2M2 learners find important differences in challenges and opportunities for these groups (cf. Chen Pichler and Koulidobrova, 2015).

For us, it is important to offer both groups learning opportunities on a par. SL teachers encounter various topics that L2 learners struggle with (Woll, 2012) and deaf didactics has not yet been sufficiently studied (Ladd, 2022; Grote et al., 2018; Karar et al., 2017; Metzger, 2008). Practising outside the classroom in a guided and fun way online may aid progress through self-regulated learning. In addition, the cross-language and cross-cultural exchange across Europe's SLs and deaf communities adds significant value to the project (cf. Kusters et al., 2024; Leeson and van den Bogaerde, 2020).

## 3.2. E-learning Platforms for SLs

Online learning platforms for sign languages have evolved from simple video repositories to interactive systems. Early work by Kelly et al. (2008) explored live feedback, laying important groundwork for later platforms. A relevant recent comparator is the SLReDu platform for learning GSL as an L2, which integrates SL recognition into a web-based environment for self-monitoring and objective learner assessment (Papadimitriou et al., 2025). Related multilingual platform work is represented by VisuoLab, a multilingual, -modal, and -functional platform designed for SL communities (Rathmann et al., 2024). Recent work also incorporates interactive elements (Wang, 2025), though comprehensive platforms integrating curriculum management, real-time feedback, and personalised instruction remain rare.

Recent work has also examined challenges for deaf students in e-learning environments (Aljedaani et al., 2023), highlighting the need for accessible platforms. Game-based approaches (e.g. Economou et al., 2020) combine SL learning with gamification elements, an approach VISTA-SL incorporates through achievement badges and progress tracking (see section 4.2).

## 3.3. Education Technologies

**Vision and Sensor-Based Systems:** Zhang et al. (2021) present a mobile phone-based system for Chinese Sign Language using point clouds, which are more lightweight than video streams. The app indicates to users whether they were correct or not, but does not provide additional feedback on what was incorrect. Truong et al. (2022) use MediaPipe (Lugaresi et al., 2019) with a gated recurrent unit model for Vietnamese Sign Language recognition. While effective, their use of a prosthetic arm for visualisation limits adoption potential.

Shao et al. (2020) describe an augmented reality SL teaching system with detailed feedback on sign production. Alam et al. (2024) present recent approaches for sign recognition in learning contexts, though without focus on explainable feedback.

Building on these foundations, VISTA-SL will go beyond binary classification, providing contextual assistance in the form of visual, textual and numerical feedback, based on real-time hand, face, and pose landmark detection (see section 5.1).

**Avatar-Based and Immersive Systems:** Phan et al. (2018, 2020) present an avatar-based system with multiple feedback modes including arrows for wrong positions, trajectory paths, and 'phantom visualisation' where teacher movements are superimposed on student recordings.

Immersive virtual reality (VR) environments have shown promise for engagement. Wang et al. (2023) implement a VR environment for learning numbers in ASL with game mechanics, finding higher user appeal than traditional methods. Quandt (2020), Quandt et al. (2022) and Kasapakis et al. (2024) describe VR systems using motion capture for sign language learning. While these immersive environments demonstrate research potential, their complex setup and high cost limit mainstream adoption. VISTA-SL addresses this by providing web-based access on commodity hardware to 3D avatar visualisations (see section 5.2).

**AI and LLM-Based Tutoring Systems:** Research on feedback for sign language learners predates current LLM-based systems. Huenerfauth et al. (2017) evaluated methods for presenting language feedback for student ASL videos, providing an important foundation for later computer-assisted feedback research. More recent work examines learner preferences for automatic feedback (Hassan et al., 2022), automated vocabulary assessment and learner perceptions (Holzknecht et al., 2024), and the automatic elaboration of linguistic annotations into learner-facing feedback (Battisti and Ebling, 2024). Related work on annotation for continuous Swiss German Sign Language (Battisti et al., 2024) and the recent dissertation of Battisti (2025) further highlight the growing focus on the analysis and assessment of L2 sign production. VISTA-SL will extend this line by introducing an LLM-based virtual instructor for text-based contextual coaching depending on learner progress and exercise performance (see section 5.3).

**Surveys and Broader Context:** Recent surveys now cover sign language processing more broadly, including recognition, translation, production, and datasets. Tan et al. (2024) review deep learning-based approaches to sign language processing, while Toshpulatov et al. (2025) provide a more recent overview spanning recognition, translation, production, and datasets. For sign language translation specifically, Shahin and Ismail (2024) survey the evolution from rule-based methods to transformer architectures. In feedback research, Paudyal et al. (2019, 2020) demonstrate benefits of explainable AI for sign language learning, while Hassan et al. (2022) investigate learner preferences for automatic feedback. Prior work on error taxonomies for computer-assisted language learning (Suri and McCoy, 1993) highlights the importance of categorising learner mistakes to enhance instructional feedback.

## 4. Curriculum

The VISTA-SL platform delivers curricula in four European sign languages: DGS, GSL, ISL and NGT. Content is developed by sign language educators and linguists at partner institutions following requirements gathered through focus groups with both deaf and hearing learners and deaf educators. It is designed to account for the needs of both L2M1 and L2M2 learners.

### 4.1. Content Structure and Types

The curricula follow a three-level hierarchical organisation of courses, modules, and lessons. Their content will include:



(a) (b)

Figure 1: Examples of VISTA-SL exercise feedback, given in DGS, for (a) good performance and (b) mixing up signs.

- Vocabulary acquisition through flashcards and isolated sign demonstrations
- Grammatical structures including non-manual markers and role-shifting
- Sentence-level content for real-life sign language interactions
- Signed dialogues between two people
- Illustration of sign variation due to regional and individual differences

On top of classical vocabulary topics (family, food, etc.) and interaction elements (introducing oneself, talking about the weather) and various semantic fields such as time, relationships or food, we include specific areas of interests, representing each country, such as sports events and car brands.

Exercises will involve activities such as identifying signs, understanding signed utterances, and reproducing signs. For sign reproduction, learners will be able to have their performance mirrored back to them, allowing them to practise before handing in recordings for evaluation. For certain aspects of signing, learners will also be able to practice using augmented reality feedback (see section 5.1).

### 4.2. Games and Gamification

A number of steps are being taken to make the learning process enjoyable and thus further motivate learners. These include immediate exercise feedback, progress tracking and games to practice content. Once again, input from the focus groups proved instrumental in refining these elements and aligning them with the needs of all user groups.

During exercises, users should receive immediate feedback. To enhance language immersion, system feedback can be given in the target language, e.g. through signed expressions that indicate correct or incorrect answers in a playful manner (see fig. 1). Furthermore, learners will be able



(a) (b) (c) (d)

Figure 2: Avatar demonstrations for the GSL letter H (Eta): (a) reference video, (b–d) avatars with trajectory overlays, shown from various viewing angles.

to practice signs through augmented reality feedback (see section 5.1), which can give immediate visual and textual input on e.g. how to adjust their handshape to match that of the target sign. On completing exercises, users will receive progress badges and completion indicators (see section 6.2).

Several games will be provided to help learners practice, such as quizzes, memory, and fill-the-gap. These games will cover vocabulary practice, language understanding exercises and deaf culture content. In implementing these games, it was important to design them in a sign-centric manner, avoiding variants that relied on written language text to work. While sign-centric design is generally desirable for an immersive learning environment, it is particularly important for the deaf user group of the VISTA-SL platform, for whom both source and target language should be signed. Therefore, games like quizzes that involve a user's first language for instructions or answers would need to be able to also provide these as either text or sign video, depending on the user's own language.

### 4.3. Cultural Integration

The feedback from focus groups emphasised that deaf culture and sign languages are inextricably linked and should therefore not be taught in isolation from each other. The VISTA-SL curriculum accounts for this by including content on deaf culture and identities that will go alongside the linguistic material . Vocabulary lessons will also be accompanied by cultural context notes. Similarly, the selection of games (see section 4.2) will also cover ones that test users on matters of deaf culture.

## 5. Education Technologies

To assist learners in their practice sessions, the VISTA-SL platform will provide a number of interactive features leveraging language technologies, such as sign language recognition (section 5.1), sign language synthesis (section 5.2) and large language models (section 5.3). In this section, we describe some of these features to illustrate how these technologies will be used by the platform.

### 5.1. Augmented Reality Feedback

A typical challenge for learners of sign languages, especially those not used to the visual modality, is to reproduce signs accurately, e.g. using the correct handshape, orientation and movements (cf. Lerose, 2025). An augmented reality system with integrated sign recognition functionalities will give learners real-time feedback on their attempts, evaluating them against the reference material.

For instance, for handshape and orientation feedback, video from the user's webcam will be mirrored back, but overlaid with an augmented reality comparison between the user's hand and that of the reference. In addition, users will receive fine-grained feedback on individual parts of the handshape through both colour-coding of their hand's keypoints in the augmented image, as well as textual descriptions of what needs correcting.

Feedback for non-manual features - like question marking, negation, intensifiers and mouth gestures - will use tracking of facial markers such as eyebrow raise, mouth patterns, and head tilts. For signs involving motion, body part movement patterns will be processed to provide feedback. Information from these components will also be combined to provide general feedback on individual signs and sentences.

### 5.2. 3D Avatar System

Each sign or sentence in the curriculum material is represented by a video recording of a deaf L1 signer. To provide users with additional information, they will be able to show the same material performed by a virtual avatar, allowing them to view it from dynamically chosen viewing angles and zoom levels. Optional trajectory tracks of hand and arm motions will provide further assistance. These features can be seen in fig. 2, which shows the reference video



(a) (b) (c) (d) (e) (f)

Figure 3: Different avatar appearances.

of a GSL signer fingerspelling the Greek letter H (Eta) and avatars from various viewing angles with trajectory overlays.

Following established best practice for sign language avatars (Krausneker and Schügerl, 2021; Kopf et al., 2023), users will be able to choose from a variety of avatar appearances and background options, as can be seen in fig. 3.

To ensure appropriate quality of the avatar performance while still scaling it to cover sufficient amounts of material, initial performances will be generated using pose and face keypoint estimations from the reference videos, which can then be then revised through an editor interface.

### 5.3. Virtual Instructor and Coaching

The VISTA-SL platform will also incorporate an LLM-based virtual instructor that provides personalised coaching based on learner context. The coaching system will access learner progress summaries—completion rates, average scores, recent performance, and recommended next lessons—to generate contextually relevant guidance. The coach can identify error patterns across lessons and provide targeted practice recommendations.

Given that the integration of sign language synthesis into generative model systems at production quality is still an open research challenge, the virtual instructor will focus on giving feedback in written form.

## 6. The VISTA-SL Platform

This section describes the VISTA-SL platform architecture (section 6.1), its learning management (section 6.2), accessibility (section 6.3), and privacy considerations (section 6.4).

### 6.1. Platform Architecture

The VISTA-SL platform follows a modular client-server architecture, shown in fig. 4, designed to integrate independently developed components while maintaining a cohesive user experience. This allows partner institutions to develop and integrate modules—AR tools, avatar system, LLM instructor, curricula—in parallel using their preferred technologies. The frontend prioritises low barriers to entry: users access the platform through standard web browsers on commodity hardware (laptop or tablet with webcam), without requiring specialised sensors or VR equipment.

The augmented reality feedback system will build on pre-computed reference landmarks extracted from expert demonstrations. Its scoring algorithm will use *Procrustes* analysis (Arvanitis et al., 2023) for shape comparison, applying weighted distance metrics that emphasise fingertips and key joint positions. The platform will use the *Performs* 3D avatar system developed by Universitat Pompeu Fabra (Valls-Garolera et al., 2025). Avatar animations will be created and edited using the web-based *Animics* tool (Ubieto et al., 2024; Pozo et al., 2025). The virtual instructor will provide a provider abstraction layer, supporting multiple LLM backends (commer-

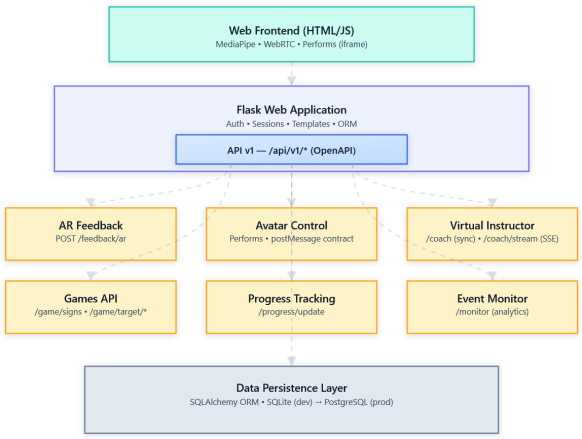



Figure 4: High-level architecture of the VISTA-SL platform, showing the web frontend, backend with core modules (AR Feedback, Avatar Control, Virtual Instructor, Games API, Progress Tracking), and data persistence layer.

cial APIs or locally-hosted open models) to ensure flexibility and avoid vendor lock-in.

### 6.2. Learning Management and Progress Tracking

The learning content follows a hierarchical structure of courses, modules, and lessons, enabling organised progression through sign language curricula. Each lesson combines reference videos showing expert demonstrations, 3D avatar animations, and interactive practice with AR feedback.

The progress tracking system records lesson completion status, performance scores, and timestamps, enabling both learner self-assessment and educator monitoring. A visual dashboard presents courses with progress indicators, animated video thumbnails, and completion badges, providing learners with clear visibility into their advancement.

### 6.3. Accessibility within VISTA-SL

The VISTA-SL platform is designed with users in mind whose first language may be either a signed or spoken language. Depending on the user's preferred modality, the interface will contain more written or signed content. For instance, while a speaker of German who is learning GSL will be shown German text for translations of GSL signs, a signer of DGS will be provided with DGS videos as translations instead.

The platform will be designed in coordination with our focus groups. It will also follow established web accessibility standards, such as the WCAG 2.1 guidelines[6]. These include incorporating keyboard navigation, screen reader compatibility, and sufficient colour contrast.

### 6.4. Privacy

Privacy in educational technology has gained increasing attention, particularly for systems processing video data. Existing sign language learning systems often require video upload or recording, raising concerns about data storage and potential misuse. VISTA-SL implements a privacy-preserving architecture where the video processing for self-paced learning occurs entirely client-side. For augmented reality feedback (section 5.1) only extracted landmark data—abstract skeletal representations—are transmitted for analysis, with no raw video stored on servers. The only instance where video is transmitted is when it is sent to a human teacher. This approach aligns with data minimisation principles while enabling effective feedback.

[6] https://www.w3.org/TR/WCAG21/

## 7. Conclusion

This paper presented VISTA-SL, an integrated e-learning platform for sign language education. The platform will provide materials for four different European sign languages and is designed both for people learning their first sign language and for signers wanting to learn another sign language. The curriculum of VISTA-SL, based on the institutional experience of several project partners, will cover beginner level vocabulary and phrases, as well as information on deaf culture and history. To support learners during self-guided practice, VISTA-SL will integrate a number of education features powered by language technologies, such as augmented reality feedback, 3D avatar demonstrations, and LLM-powered coaching. To support its integration with classes taught by sign language teachers, the platform will also provide interfaces for teachers to provide feedback on homework exercises submitted by students. The various components of the platform, such as the intelligibility and user acceptance of the avatar, will be evaluated throughout the run time of the project in coordination with our focus groups.

## 8. Acknowledgements

This work was funded by the European Union's Erasmus+ programme under project VISTA-SL (Visual Interactive System for Teaching and Assessment of Sign Languages), Grant Agreement No. 2024-1-EL01-KA220-HED-000257847.

The VISTA-SL consortium comprises six partner institutions:

- University of Patras, Greece
- Hellenic Mediterranean University, Greece
- Stichting Koninklijke Kentalis, Netherlands
- Universitat Pompeu Fabra, Spain
- University of Hamburg, Germany
- Technological University Dublin, Ireland

The 3D avatar system builds on work from the PRESENT, SignON, EMERALD and CAT-SL projects.

The project would not be possible without the support and advice of our diverse focus groups. Many thanks.

## 9. Bibliographical References


Md Shahinur Alam, Jason Lamberton, Jianye Wang, Carly Leannah, Sarah Miller, Joseph Palagano, Myles de Bastion, Heather L. Smith, Melissa Malzkuhn, and Lorna C. Quandt. 2024. ASL champ!: a virtual reality game with deep-learning driven sign recognition. *Computers & Education: X Reality*, 4:100059.

Wajdi Aljedaani, Rrezarta Krasniqi, Sanaa Aljedaani, Mohamed Wiem Mkaouer, Stephanie Ludi, and Khaled Al-Raddah. 2023. If online learning works for you, what about deaf students? emerging challenges of online learning for deaf and hearing-impaired students during COVID-19: a literature review. *Universal Access in the Information Society*, 22(3):1027–1046.

Filiz Altun. 2025. Analysis of the UNCRPD committee recommendations on the implementation of article 24 on the right to inclusive education. EASPD report, European Association of Service providers for Persons with Disabilities, Brussels, Belgium.

Nikolaos Arvanitis, Evangelos Sartinas, and Dimitrios Kosmopoulos. 2023. Procrustes-dtw: Dynamic time warping variant for the recognition of sign language utterances. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Alessia Battisti. 2025. *Analysis and Assessment of Continuous Swiss German Sign Language Production of Adult Learners*. Dissertation, University of Zurich.

Alessia Battisti and Sarah Ebling. 2024. Automatic annotation elaboration as feedback to sign language learners. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 46–60, St. Julians, Malta. Association for Computational Linguistics.

Alessia Battisti, Katja Tissi, Sandra Sidler-Miserez, and Sarah Ebling. 2024. Advancing annotation for continuous data in Swiss German Sign Language. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 163–174, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Kostas Blekos, Dimitrios Kosmopoulos, and Klimis Antzakas. 2026. A vision-based system for automated sign language teaching. *IEEE Transactions on Learning Technologies*, 19:182–195.

Deborah Chen Pichler and Elena Koulidobrova. 2015. Acquisition of sign language as a second language. In Marc Marschark and Patricia Elizabeth Spencer, editors, *The Oxford Handbook of Deaf Studies in Language*, Oxford Library of Psychology, chapter 14, pages 218–230. Oxford University Press.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Companion Volume*. Council of Europe Publishing, Strasburg.

Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. 2024. Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 366–377, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Daphne Economou, Melissa Russi, Ioannis Doumanis, Markos Mentzelopoulos, Vassiliki Bouki, and Jeffery Ferguson. 2020. Using serious games for learning British Sign Language combining video, enhanced interactivity, and VR technology. *Journal of Universal Computer Science*, 26(8):996–1016.

Klaudia Grote, Horst Sieprath, and Bastian Staudt. 2018. Deaf Didaktik? Weshalb wir eine spezielle Didaktik für den Unterricht in Gebärdensprache benötigen. *Das Zeichen*, 110:2–13.

Raychelle Harris, Heidi M. Holmes, and Donna M. Mertens. 2009. Research ethics in sign language communities. *Sign Language Studies*, 9(2):104–131.

Saad Hassan, Sooyeon Lee, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2022. Understanding ASL learners' preferences for a sign language recording and automatic feedback system to support self-study. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22)*, Assets '22. Association for Computing Machinery.

Franz Holzknecht, Sandrine Tornay, Alessia Battisti, Aaron Olaf Batty, Katja Tissi, Tobias Haug, and Sarah Ebling. 2024. Automated sign language vocabulary assessment: Comparing human and machine ratings and studying learner

perceptions. *Language Assessment Quarterly*, 21(3):245–265.

Matt Huenerfauth, Elaine Gale, Brian Penly, Sree Pillutla, Mackenzie Willard, and Dhananjai Hariharan. 2017. Evaluation of language feedback methods for student videos of American Sign Language. *ACM Transactions on Accessible Computing*, 10(1).

Ege Karar, Klaudia Grote, and Horst Sieprath. 2017. Deaf didactics: Why it is not enough just to teach bilingual. Presentation at the 3rd International Conference of the World Federation of the Deaf.

Vlasios Kasapakis, Elena Dzardanova, Spyros Vosinakis, and Androniki Agelada. 2024. Sign language in immersive virtual reality: design, development, and evaluation of a virtual reality learning environment prototype. *Interactive Learning Environments*, 32(10):6657–6671.

Daniel Kelly, John McDonald, and Charles Markham. 2008. A system for teaching sign language using live gesture feedback. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–2, Amsterdam, Netherlands. IEEE.

Maria Kopf, Rehana Omardeen, and Davy Van Landuyt. 2023. Representation matters: The case for diversifying sign language avatars. In *Proceedings of the Eighth International Workshop on Sign Language Translation and Avatar Technology*, pages 1–5, Rhodes Island, Greece. IEEE.

Verena Krausneker and Sandra Schügerl. 2021. Best practice protocol on the use of sign language avatars. Technical report, Universität Wien, Vienna, Austria.

Annelies Kusters, Erin Moriarty, Amandine le Maire, Sanchayeeta Iyer, and Steven Emery. 2024. *Deaf Mobility Studies: Exploring International Networks, Tourism, and Migration*. Gallaudet University Press, Washington, DC.

Annelies Maria Jozef Kusters, Maartje De Meulder, and Dai O'Brien. 2017. Innovations in deaf studies: Critically mapping the field. In Annelies Kusters, Maartje De Meulder, and Dai O'Brien, editors, *Innovations in Deaf Studies: The Role of Deaf Scholars*, Perspectives on Deafness, pages 1–53. Oxford University Press, Oxford.

Paddy Ladd. 2022. *The unrecognized curriculum: seeing through new eyes: deaf cultures and deaf pedagogies*. DawnSignPress, San Diego.

Lorraine Leeson and Beppie van den Bogaerde. 2020. (What we don't know about) sign languages in higher education in Europe: Mapping policy and practice to an analytical framework. *Sociolinguistica*, 34(1):31–56.

Luigi Lerose. 2025. *Improving Your British Sign Language: A Guide to Proficient Use*. Cambridge University Press, Cambridge.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A framework for building perception pipelines. Preprint. arXiv:1906.08172.

Elizabeth S. Mathews. 2020. Signs of equity: Access to teacher education for deaf students in the Republic of Ireland. *Sign Language Studies*, 21(1):68–97.

Christiane Metzger. 2008. *Lernstrategien erwachsener L2-Lerner der Deutschen Gebärdensprache: eine Analyse auf der Grundlage handlungstheoretischer und kognitionswissenschaftlicher Modelle*. Ph.D. thesis, Narr, G, Tübingen.

Katerina Papadimitriou, Gerasimos Potamianos, Galini Sapountzaki, Theodoros Goulas, Eleni Efthimiou, Stavroula-Evita Fotinea, and Petros Maragos. 2025. Greek sign language recognition for an education platform. *Universal Access in the Information Society*, 24:51–68.

Prajwal Paudyal, Ayan Banerjee, and Sandeep Gupta. 2020. On evaluating the effects of feedback for Sign language learning using Explainable AI. In *Companion Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20 Companion, pages 83–84. Association for Computing Machinery.

Prajwal Paudyal, Junghyo Lee, Azamat Kamzin, Mohamad Soudki, A. Banerjee, and S. Gupta. 2019. Learn2Sign: Explainable AI for Sign Language Learning. In *Joint Proceedings of the ACM IUI 2019 Workshops – Workshop on Explainable Smart Systems - ExSS*.

Han Duy Phan, Kirsten Ellis, and Alan Dorin. 2018. MIC, an interactive sign language teaching system. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, OzCHI '18, pages 544–547. Association for Computing Machinery.

Han Duy Phan, Kirsten Ellis, Alan Dorin, and Patrick Olivier. 2020. Feedback strategies for embodied agents to enhance sign language vocabulary learning. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*,

Iva '20, pages 1–8. Association for Computing Machinery.

Jaume Pozo, Eva Valls-Garolera, and Víctor Ubieto. 2025. Animics. https://github.com/upf-gti/animics/. Last commit: 25-03-2026, in development.

Lorna Quandt. 2020. Teaching ASL signs using signing avatars and immersive learning in virtual reality. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–4, Online. Association for Computing Machinery.

Lorna C. Quandt, Athena Willis, Melody Schwenk, Kaitlyn Weeks, and Ruthie Ferster. 2022. Attitudes toward signing avatars vary depending on hearing status, age of signed language acquisition, and avatar type. *Frontiers in Psychology*, 13:730917.

Christian Rathmann, Ronice Müller de Quadros, Thomas Geißler, Christian Peters, Francisco Fernandes, Milene Peixer Loio, and Diego França. 2024. VisuoLab: Building a sign language multilingual, multimodal and multifunctional platform. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 14–21, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Krister Schönström and Ingela Holmström. 2022. L2M1 and L2M2 acquisition of sign lexicon: The impact of multimodality on the sign second language acquisition. *Frontiers in Psychology*, 13.

Nada Shahin and Leila Ismail. 2024. From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: Survey, taxonomy and performance evaluation. *Artificial Intelligence Review*, 57:271.

Qijia Shao, Amy Sniffen, Julien Blanchet, Megan E. Hillis, Xinyu Shi, Themistoklis K. Haris, Jason Liu, Jason Lamberton, Melissa Malzkuhn, Lorna C. Quandt, James Mahoney, David J. M. Kraemer, Xia Zhou, and Devin Balkcom. 2020. Teaching American Sign Language in mixed reality. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, volume 4, pages 152:1–152:27, Online. ACM.

Linda Z. Suri and Kathleen F. McCoy. 1993. A methodology for developing an error taxonomy for a computer assisted language learning tool for second language learners. Technical Report 93-16, University of Delaware, Newark, DE, USA.

Sihan Tan, Nabeela Khan, Zhaoyi An, Yoshitaka Ando, Rei Kawakami, and Kazuhiro Nakadai. 2024. A review of deep learning-based approaches to sign language processing. *Advanced Robotics*, 38(23):1649–1667.

Mukhiddin Toshpulatov, Wookey Lee, Jaesung Jun, and Suan Lee. 2025. Deep learning pathways for automatic sign language processing. *Pattern Recognition*, 164:111475.

Nghe-Nhan Truong, Truong-Dong Do, Thien-Sang Le Nguyen, Minh-Thien Duong, Thanh-Hai Nguyen, and My-Ha Le. 2022. A vision-based hand-sign language teaching system using deep neural network: Methodology and experiments. In *2022 International Workshop on Intelligent Systems (IWIS)*, pages 1–6.

Víctor Ubieto, Jaume Pozo, Eva Valls, Beatriz Cabrero-Daniel, and Josep Blat. 2024. Sign language synthesis: Current signing avatar systems and representation. In Andy Way, Lorraine Leeson, and Dimitar Shterionov, editors, *Sign Language Machine Translation*, number 5 in Machine Translation: Technologies and Applications, pages 247–266. Springer, Cham.

Eva Valls-Garolera, Jaume Pozo, , and Carolina del Corral. 2025. Performs. https://github.com/upf-gti/performs/. Last commit: 25-03-2026, in development.

J. Wang. 2025. *Human-Centred American Sign Language Learning: Effects of Interface Modality, Game Mechanics, and Artificial Intelligence-Based Diagnostic Guidance*. Doctoral thesis, Durham University.

Jindi Wang, Ioannis Ivrissimtzis, Zhaoxing Li, Yunzhan Zhou, and Lei Shi. 2023. Exploring the potential of immersive virtual environments for learning American Sign Language. In *Responsive and Sustainable Educational Futures*, pages 459–474. Springer.

Bencie Woll. 2012. Second language acquisition of sign language. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons.

Yanxiao Zhang, Yuecong Min, and Xilin Chen. 2021. Teaching Chinese Sign Language with a smartphone. *Virtual Reality & Intelligent Hardware*, 3(3):248–260.

# Evaluation of Pose Estimation Systems for Sign Language Translation

**Catherine O'Brien*, Gerard Sant*, Mathias Müller, Sarah Ebling**
Department of Computational Linguistics
University of Zurich, Switzerland
{catherineelizabeth.obrien, gerard.santmuniesa}@uzh.ch, {mmueller, ebling}@cl.uzh.ch

**Abstract**

Many sign language translation (SLT) systems operate on pose sequences instead of raw video to reduce input dimensionality, improve portability, and partially anonymize signers. The choice of pose estimator is often treated as an implementation detail, with systems defaulting to widely available tools such as MediaPipe Holistic or OpenPose. We present a systematic comparison of pose estimators for pose-based SLT, covering widely used baselines (MediaPipe Holistic, OpenPose) and newer whole-body/high-capacity models (MMPose WholeBody, OpenPifPaf, AlphaPose, SDPose, Sapiens, SMPLest-X). We quantify downstream impact by training a controlled SLT pipeline on RWTH-PHOENIX-Weather 2014 where only the pose representation varies, evaluating with BLEU and BLEURT. To contextualize translation outcomes, we analyze temporal stability, missing hand keypoints, and robustness to occlusion using higher-resolution videos from the Signsuisse dataset. SDPose and Sapiens achieve the best translation performance (BLEU ~11.5), outperforming the common MediaPipe baseline (BLEU ~10). In occlusion cases, Sapiens is correct in all tested instances (15/15), while OpenPifPaf fails in nearly all (1/15) and also yields the weakest translation scores. Estimators that frequently leave out hand keypoints are associated with lower BLEU/BLEURT. We release code that can be used not only to reproduce our experiments, but also considerably lowers the barrier for other researchers to use alternative pose estimators.

**Keywords:** sign language, sign language processing, machine translation, pose estimation

## 1. Introduction

Sign language processing (SLP) is gaining ground within Natural Language Processing (NLP), yet it remains substantially underrepresented (Bragg et al., 2019; Yin et al., 2021; Müller et al., 2022). In spoken-language NLP, many core modeling decisions and preprocessing choices have been systematically studied and benchmarked. In contrast, for sign languages, even fundamental design decisions remain underexplored. For instance, while tokenizer selection in spoken-language translation is informed by extensive prior work, it remains unclear how sign language data should best be preprocessed for sign language translation (SLT).

SLT systems can accept input from a variety of modalities, including raw video (Zhou et al., 2023; Ye et al., 2024; Sant et al., 2025), glosses—semantic transcriptions of meaning—(Camgoz et al., 2018; Yin and Read, 2020; Zhou et al., 2021), form notations like SignWriting (Sutton, 1990; Jiang et al., 2023) or HamNoSys (Prillwitz and Zienert, 1990), which encode articulatory form, or poses (Zhang et al., 2024; Arib et al., 2025; Gan et al., 2025; Hwang et al., 2025). Poses are approximations of a signer's skeleton and movements as they sign.

Poses are a common choice for SLT because they are more viable than glosses or form notations (Müller et al., 2023b), and offer a lightweight interpretable alternative to raw video. They also provide some degree of anonymization when compared to signer videos (Battisti et al., 2024; Moryossef et al., 2025). Poses may also be more signer-independent for low-resource sign languages (Holmes et al., 2022). Despite these advantages, video-based models still outperform pose-based systems (Moryossef et al., 2021; Tarrés et al., 2023; Sant et al., 2025), suggesting that current pose representations may lose critical linguistic information. Improving pose estimation quality therefore represents a promising pathway toward narrowing this performance gap.

Although numerous pose estimation systems can be applied to sign language data, their impact on downstream translation performance remains largely unexplored. In practice, most translation pipelines rely on MediaPipe (Lugaresi et al., 2019) or OpenPose (Cao et al., 2019), primarily due to accessibility and ease of integration rather than demonstrated suitability for sign language.

To our knowledge, this work presents the first systematic evaluation of pose estimators for SLT. Beyond translation performance, we examine their behavior under pose estimation challenges intrinsic to signing, including occlusion, temporal instability, and missing hand detections. Our analysis provides practical guidance for selecting pose representations and highlights factors that influence downstream translation quality.

The main contributions of this paper are:

* These authors contributed equally.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 371–386
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

- A controlled comparison of eight prominent pose estimators on the RWTH-PHOENIX-Weather 2014 dataset,
- An analysis of pose estimation failure modes relevant to sign language data, including occlusion, temporal instability, and missing hands keypoints,
- An open-source framework extending the `pose-format` library (Moryossef et al., 2023b) with support for all evaluated pose estimators, simplifying integration and lowering barriers to their use in SLT pipelines.

Code and scripts to reproduce our experiments are available at https://github.com/ZurichNLP/multimodalhugs-pipelines.

## 2. Background

### 2.1. Pose Estimation

Pose estimation systems extract human skeletal keypoints from video, representing body, hand, and facial articulators as spatio-temporal trajectories. Modern pipelines typically rely on deep learning–based detectors such as OpenPose (Cao et al., 2019), MediaPipe Holistic (Lugaresi et al., 2019; Grishchenko and Bazarevsky, 2020), or SMPL-based models (Loper et al., 2015; Cai et al., 2023; Yin et al., 2026) to recover 2D or 3D joint configurations from RGB input. These representations provide a compact and interpretable encoding of human motion and are widely used in sign language processing (Wei et al., 2016; Li et al., 2020; Saunders et al., 2022; Joshi et al., 2025) and generation pipelines (Saunders et al., 2020; Xiao et al., 2020; Moryossef et al., 2023a; Arkushin et al., 2023).

In sign language contexts, accurate capture of fine-grained articulations—including handshape, orientation, movement, and facial expression—is essential, as linguistic meaning is conveyed through coordinated manual and non-manual signals (Liddell and Johnson, 1989; Johnson and Liddell, 2010; Pfau and Quer, 2010; Brentari, 2011; Sandler, 2012; Benitez-Quiroz et al., 2014; Yin et al., 2021). Errors in keypoint localization or tracking may therefore lead to the loss of critical articulatory information (Moryossef et al., 2021; Moryossef, 2024), degrading the quality of representations used by downstream tasks.

Despite these challenges, sign language processing pipelines commonly rely on general-purpose pose estimators that are not optimized for sign language data, as noted in (Coster et al., 2023; Jiang et al., 2025). Recent advances in whole-body and 3D pose estimation offer improved hand and body modeling (Jin et al., 2020a; Zhu et al., 2023). While pose-based representations are widely used in SLT, the implications of estimator choice for translation quality remain insufficiently understood.

### 2.2. Sign language translation

Sign language translation (SLT) seeks to generate spoken-language text from signed input (Müller et al., 2022; De Coster et al., 2024). SLT systems have employed a wide range of input representations, spanning continuous visual signals—such as raw video (Zhou et al., 2023; Ye et al., 2024) or learned visual features (Tarrés et al., 2023; Gueuwou et al., 2025), among others—as well as discrete representations including gloss annotations (Camgoz et al., 2018; Yin and Read, 2020; Zhou et al., 2021) and form notations (Jiang et al., 2023).

While raw video inputs often achieve the strongest translation performance among modalities (Sant et al., 2025), pose-based representations have been widely adopted in SLT pipelines (Zhang et al., 2024; Arib et al., 2025; Gan et al., 2025; Hwang et al., 2025), as they provide a lower-dimensional alternative that preserves articulatory structure. However, the effectiveness of pose-based SLT may be limited by the fidelity of the underlying pose representations, which varies across commonly used general-purpose estimators (Section 2.1).

Comparative studies of pose estimators for sign language tasks remain limited and report mixed findings. Several works report stronger performance of MediaPipe relative to OpenPose (Moryossef et al., 2021; Müller et al., 2022, 2023c), a trend further confirmed by Coster et al. (2023), who also found MediaPipe to outperform MMPose (Jin et al., 2020b). In contrast, Lazo-Quispe et al. (2022) reported improved results using an MMPose-based pipeline. The impact of newer whole-body and high-capacity estimators on SLT performance has yet to be systematically examined.

## 3. Pose Estimators

As shown in Table 1, we consider both pose estimators widely used in previous sign language processing research—such as MediaPipe (Lugaresi et al., 2019) and OpenPose (Cao et al., 2019)—as well as more recent systems that have not been extensively used but appear to be strong candidates. All evaluated methods are human pose estimators rather than models specialized for sign language. We selected both top-down and bottom-up approaches to pose estimation. We exclude common pose estimators that do not estimate hands, such as YOLO (Maji et al., 2022) or PoseFormer (Zheng et al., 2021).

| estimator | # keypoints | scheme | 2D/3D | speed (fps) | GPU only | confidence |
|---|---|---|---|---|---|---|
| MediaPipe Holistic (Lugaresi et al., 2019) | 576 | - | 3D | 0.89 / 3.15 | - | ✔ |
| OpenPose (Cao et al., 2019) | 137 | - | 2D | - / 4.40 | ✔ | ✔ |
| MMPose Wholebody (Jin et al., 2020b) | 133 | COCO Wholebody | 2D | 0.89 / 3.81 | - | ✔ |
| OpenPifPaf (Kreiss et al., 2021) | 133 | COCO Wholebody | 2D | 1.21 / 4.42 | - | ✔ |
| SDPose (Liang et al., 2025) | 133 | COCO Wholebody | 2D | 0.07 / 0.84 | - | ✔ |
| Sapiens (Khirodkar et al., 2025) | 308 | Sapiens-308 | 2D | 0.04 / 3.29 | - | ✔ |
| AlphaPose (Fang et al., 2023) | 136 | Halpe-FullBody | 2D | - / 22.89 | ✔ | ✔ |
| SMPLest-X (Yin et al., 2026) | 137 | SMPL-X | 2D | - / 8.36 | ✔ | - |

Table 1: General characteristics of pose estimators considered in this study. scheme=whether estimated keypoints correspond to a standard layout, speed=frames per second on CPU/GPU measured on a single V100, confidence=whether the estimator outputs confidence values for each keypoint

**Mediapipe Holistic (Lugaresi et al., 2019)** This system employs a unified architecture to jointly estimate body pose, facial landmarks, and detailed hand keypoints by integrating region-specific subnetworks within a graph-based perception pipeline. This design produces dense 540-landmark representations and enables coordinated tracking of upper-body articulators from RGB input, making it widely used in real-time gesture and sign language processing. Its emphasis on low-latency, on-device inference has further contributed to its widespread adoption in SLT pipelines (Zhang et al., 2024; Gueuwou et al., 2025). In our experiments we reduce the full mesh of face keypoints to a smaller selection of contour points.

**OpenPose (Cao et al., 2019)** Building on Part Affinity Fields (PAFs) (Cao et al., 2017), which represents limb associations as 2D vector fields linking detected joints, this approach enables bottom-up multi-person pose estimation by jointly detecting keypoints and learning their pairwise associations. This formulation removes the dependency on person detection bounding boxes and improves robustness in crowded scenes. OpenPose further extended the pose estimation paradigm to full-body configurations including face and hands, which made it influential in multimodal and sign language research (Ko et al., 2019; Park and Sohn, 2020; Moryossef et al., 2020; Liang et al., 2023; Núñez-Marcos et al., 2023).

We run a pre-built OpenPose docker image with `--hands` and `--face` enabled, resulting in a total of 137 keypoints per frame. Occasionally OpenPose incorrectly predicts several people, in which case we select the first one.

**MMPose Wholebody (Jin et al., 2020b)** Extends conventional human pose estimation beyond standard 17-keypoint body configurations to dense full-body representations, typically comprising 133 keypoints (including body, hands, face, and feet) as defined in the COCO-WholeBody benchmark (Jin et al., 2020b). Architecturally, it follows a top-down paradigm: person instances are first detected, after which high-resolution keypoint heatmaps are predicted for each region.

We do estimation for MMPose Wholebody via their Unified Inference API. This API offers several checkpoint options and a variety of outputs. "Wholebody" refers to the checkpoint `rtmpose-m_simcc-coco-wholebody_pt-aic-coco_270e-256x192-cd5e845c_20230123`, which outputs the 133 keypoints defined in the COCO-Wholebody dataset.

**OpenPifPaf (Kreiss et al., 2021)** A bottom-up multi-person pose estimator builds on the Part Affinity Fields paradigm to jointly detect and associate keypoints. Its representation improves keypoint localization and grouping under partial occlusion and scale variation. OpenPifPaf has demonstrated strong robustness in crowded scenes (Andriluka et al., 2018) while remaining computationally efficient. We estimate poses for OpenPifPaf using their Predictor API with the pretrained checkpoint `shufflenetv2k30-wholebody`. We estimate poses for OpenPifPaf using their Predictor API with the pretrained checkpoint `shufflenetv2k30-wholebody`.

**SDPose (Liang et al., 2025)** Proposes a pose estimation framework that exploits pre-trained generative priors from Stable Diffusion (Rombach et al., 2022) to enhance both standard accuracy and robustness under domain shift. Evaluated against both in-domain and out-of-distribution benchmarks (Jin et al., 2020a; Ju et al., 2023) (e.g., human and stylized images), SDPose achieves competitive results with strong cross-domain generalization, highlighting the potential of diffusion-based priors in structured vision tasks such as pose estimation.

For SDPose, we use a pretrained model obtained from Hugging Face `teemosliang/SDPose-Wholebody`. This architecture first computes bounding boxes for each person using `YOLO-11x`. We have adapted their code such that we use only the person with the highest confidence, and discard any other persons detected.

**Sapiens (Khirodkar et al., 2025)** Represents a shift toward large-scale pretraining for human-centric vision tasks. Rather than optimizing narrowly for 2D keypoint detection, it leverages foundation-model-style pretraining on large-scale human motion data and supports fine-tuning for pose, segmentation, and depth estimation. Its dense 308-keypoint representation—including 243 facial landmarks—and high-resolution inference enable precise capture of fine-grained articulations relevant to sign language (See Section 2.1). This paradigm aligns with trends in NLP, where large-scale pretraining improves transferability to downstream multimodal tasks.

We estimate Sapiens poses based on the implementation of Gorordo (2024), which reproduces the official Sapiens pose model without requiring the full framework. We use the `sapiens_1b_goliath_best_goliath_AP_640` checkpoint from the official release; following the authors' recommendations, the 1B model is preferred as smaller variants yield lower accuracy.

**AlphaPose (Fang et al., 2023)** A top-down multi-person pose estimator that performs human detection followed by pose regression within each bounding box. It improves localization accuracy through Symmetric Integral Keypoint Regression, enabling continuous joint prediction beyond discrete heatmap representations. AlphaPose propose the Halpe-FullBody scheme, which supports whole-body estimation (up to 136 keypoints), extending COCO-WholeBody format by including additional head, neck, and hip joints. Its multi-stage pipeline enables efficient inference while preserving localization accuracy.

We use the authors' recommended `multi_domain_fast50_dcn_combined_256x192` checkpoint for Halpe-FullBody pose estimation.

**SMPLest-X (Yin et al., 2026)** A minimalist one-stage model for expressive human pose and shape estimation based on the SMPL-X scheme (Pavlakos et al., 2019) that jointly predicts body, hand, and face parameters using a transformer encoder–decoder architecture with task tokens, eliminating part-specific modules used in prior pipelines (Rong et al., 2021; Zhang et al., 2023). Designed to study scaling effects rather than architectural complexity, it is trained on large, diverse multi-dataset mixtures and shows strong cross-domain generalization. Despite its simplicity, the authors report state-of-the-art performance across multiple benchmarks (Pavlakos et al., 2019; Patel et al., 2021; Zhang et al., 2022; Lin et al., 2023; Fan et al., 2023) and highlight strong results for articulated hand pose estimation.

We use the `SMPLest-X-Huge` model recommended by the authors. Because this model does not provide per-keypoint confidence scores, we assume confidence 1 to retain all predicted keypoints during training (Section 4.1).

**Support for `pose-format`** We provide instructions and code[1] for running all of these lesser-used (in the context of SLT) estimators and also extend the `pose-format` library (Moryossef et al., 2023b) to support these new pose types. This considerably lowers the barriers for other researchers to use alternative estimators. Compatibility with pose-format means the predicted poses can more easily be stored, loaded, manipulated and visualized.

## 4. Methodology of Experiments

This study compares multiple pose estimators in the context of sign language translation. We evaluate each estimator within an identical translation pipeline to measure its impact on translation quality (Section 4.1). To contextualize these results, we also analyze estimator behavior with respect to temporal instability, occlusion, and missing hand detections—common challenges in pose estimation (Section 4.2).

### 4.1. Translation

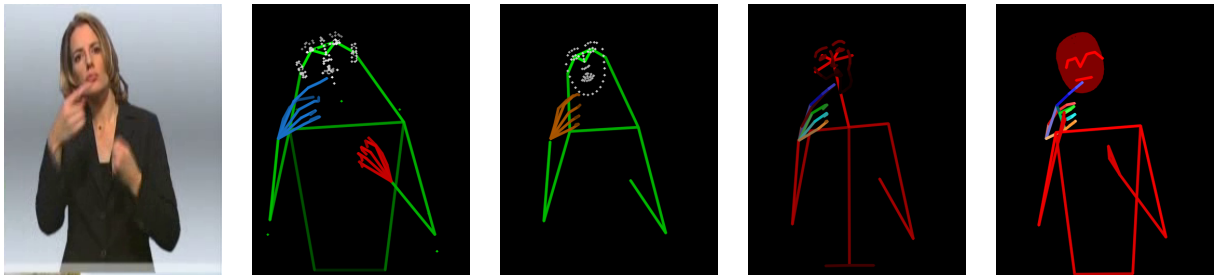
We train translation models with the recent MultiModalHugs toolkit (Sant et al., 2024). MultiModalHugs is a framework built on top of Hugging Face designed for multimodal AI models, which makes it ideal for a dataset of .pose files.

**Dataset** We use the RWTH-PHOENIX-Weather 2014 dataset (Forster et al., 2014; Camgoz et al., 2018)[2], which contains videos of German Sign language (DGS) aligned with German transcripts.

**Preprocessing** We process each video with each pose estimator to extract frame-level skeletal keypoints and store them in the binary .pose standard defined by the `pose-format` library. This enables consistent loading of pose sequences of any type within MultiModalHugs. The resulting representations preserve keypoint coordinates and confidence scores (if available) for each frame, while accommodating estimator-specific keypoint layouts (Table 1). Pose files are preprocessed as follows: leg keypoints are removed, and the remaining ones are spatially normalized per sequence to ensure consistent body and hand scale across signers. Missing values are zero-filled, and each frame is flattened into a feature vector.

[1] https://github.com/ZurichNLP/video-to-pose

[2] Referred to as *Phoenix* in the remainder of the text.



(a) Original frame (b) Sapiens (c) OpenPifPaf (d) OpenPose (e) MediaPipe

Figure 1: Example illustrating missing hand keypoints across pose estimators. (a) Original RGB frame used to extract poses, shown for reference. (b) Sapiens predicts a complete set of hand keypoints (no confidence $= 0$), although some finger articulation may still be inaccurate (e.g., the left index finger). (c–e) OpenPifPaf, OpenPose, and MediaPipe assign confidence $= 0$ to large portions of the hand keypoints (complete for c–d and nearly complete for e). During preprocessing, these keypoints are masked and zero-filled, resulting in the loss of hand information provided to the translation model.

We employ a shared encoder–decoder architecture, following the pose-based pipeline described in Sant et al. (2025). Pose sequences are represented as frame-level keypoint vectors and projected through a linear layer into the embedding space of a pretrained multilingual Transformer (`facebook/m2m100_418M`). This adapter aligns modality-specific representations with the language model's input format, ensuring that only the pose estimator varies across experiments. The input dimensionality of each model is determined by the number of keypoints per estimator, multiplied by their spatial dimensionality (2D or 3D).

**Evaluation** To evaluate translation quality, we report BLEU (Papineni et al., 2002) using SacreBLEU (Post, 2018)[3] and BLEURT (Sellam et al., 2020; Pu et al., 2021) using the `BLEURT-20` model. Other standard evaluation metrics such as COMET (Rei et al., 2020) are not applicable to our task, since they do not support sign languages.

## 4.2. Further Analyses

We conduct the following quantitative or qualitative analyses: an investigation of temporal jitter, robustness to occlusion and missing hand keypoints.

**Datasets** We use both the Phoenix and the Signsuisse dataset (Müller et al., 2023a). We add Signsuisse data because it includes higher-resolution videos than Phoenix and features deaf signers, whereas Phoenix has hearing interpreters. Besides, in order to accurately assess how estimators are handling occlusion, high-quality videos are required to reduce the influence of blur. We use the portion of the Signsuisse data that contains Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, or DSGS) videos.

[3]Signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.6.0`

**Temporal jittering (pose stability)** Beyond visual inspection, we quantify the temporal stability of each pose estimator using derivative-based smoothness metrics commonly adopted in video pose estimation. Following prior work (Zeng et al., 2022; Jin et al., 2023), we compute an acceleration-based jitter score ($J_{\text{acc}}$) as the mean magnitude of the second-order temporal difference of 2D keypoints, averaged over joints and time. To capture sharper temporal instabilities, we additionally compute a jerk-based score ($J_{\text{jerk}}$) using the third-order temporal difference, which penalizes rapid changes in acceleration and has been used as a smoothness indicator in motion and inertial pose estimation (Yi et al., 2021; Xia et al., 2022). We also report motion energy ($E_v$), defined as the mean joint velocity magnitude, to contextualize jitter values with respect to the overall amount of motion in the sequence. For 3D estimators, we compute these metrics using only the 2D image-plane coordinates.

Given the importance of manual and non-manual (e.g., facial) articulation in sign language, all metrics are computed on three subsets: (i) all keypoints (excluding legs), (ii) hands only, and (iii) face only. For each sequence, derivative magnitudes are averaged over joints and over time, yielding one scalar per sequence. Evaluation is conducted on 20 randomly selected videos from Signsuisse and 20 from Phoenix. Results are reported as median and interquartile range (IQR) across sequences (instead of mean and standard deviation), as these are more robust to the right-skewed nature of the observed distributions. Lower $J_{\text{acc}}$ and $J_{\text{jerk}}$ indicate smoother and more temporally stable predictions.

**Occlusion** Occlusion is a common source of error in pose estimation (Lino et al., 2025; Fan and Chowdhury, 2025), particularly in sign language, where one hand frequently overlaps the other or the face. To assess how estimators behave un-

| Estimator | BLEU (↑) | BLEURT (↑) |
|---|---|---|
| MediaPipe | $10.327 \pm 0.269$ | $0.351 \pm 0.005$ |
| OpenPose | $10.606 \pm 0.251$ | $0.353 \pm 0.002$ |
| MMPose Wholebody | $10.901 \pm 0.299$ | $0.361 \pm 0.007$ |
| OpenPifPaf | $9.365 \pm 0.263$ | $0.325 \pm 0.007$ |
| SDPose | $11.681 \pm 0.415$ | $0.372 \pm 0.001$ |
| Sapiens | $11.525 \pm 0.222$ | $0.372 \pm 0.003$ |
| AlphaPose | $11.251 \pm 0.241$ | $0.359 \pm 0.007$ |
| SMPLest-X | $9.709 \pm 0.334$ | $0.341 \pm 0.009$ |

Table 2: Translation scores on the Phoenix dataset. We report the mean and standard deviation across three training runs.

der these conditions, we analyzed 10 hand-picked videos from the Signsuisse dataset, comprising 15 individual signs that include occlusion.

We define *occlusion* as frames in which one hand or arm partially or fully obstructs the other hand or arm from the camera's viewpoint. We visually inspected the pose outputs for each estimator and evaluated whether both hands were consistently detected and whether their locations, handshapes, movements, and palm orientations remained plausible relative to the source video. A prediction was considered acceptable when both hands were present for most of the sequence and their articulatory configuration matched the observed signing. Given the limited sample size, this occlusion analysis is intended to contextualize the main translation results rather than provide a comprehensive robustness benchmark.

**Missing hand keypoints** During translation preprocessing, keypoints with confidence $c = 0$ are treated as invalid and zero-filled before frame-level flattening; such keypoints are therefore absent from the model input. Estimators that frequently produce $c = 0$ hand keypoints thus provide less hand information to the translation model. Example cases are shown in Figure 1. We therefore quantify missing hand keypoints directly from the `.pose` files to capture how often substantial portions of hand information are unavailable to the translation model.

For each frame and each hand, we compute the proportion of hand keypoints with $c = 0$. We define a hand as *missing* in a given frame when at least $50\%$ of its keypoints have $c = 0$. Signing frames are defined as those where the wrist is vertically above the elbow.

# 5. Results & Discussion

## 5.1. Translation Results

Table 2 shows the evaluation results of our sign language translation (SLT) experiments. Every result is an average across three training runs. Overall, the translation quality in our experiments is much lower than state-of-the art systems, especially ones based on videos instead of poses, but we are interested specifically in the *relative* performance of pose estimators compared to each other.

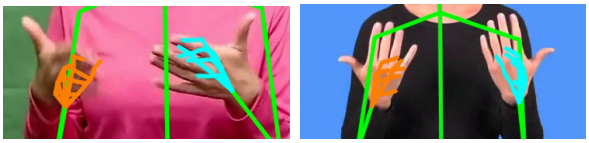


Figure 2: Examples of erroneous SMPLest-X hand pose estimates overlaid on original frames, cropped to highlight the hands. In both examples (left: How2Sign (Duarte et al., 2021); right: Signsuisse), the predicted hand keypoints exhibit incorrect orientation and articulation that do not match the observed hand configurations.

Sapiens and SDPose achieved the highest BLEU and BLEURT scores, of roughly $11.5$ and $0.37$, suggesting that they are stronger candidates for SLT applications compared to other estimators. In total, four different estimators (Sapiens, SDPose, MMPose Wholebody and AlphaPose) lead to better results than Mediapipe, the de-facto standard pose estimation system used in SLT.

This provides empirical evidence that **alternatives to Mediapipe may be better suited for SLT**, and **future work should therefore consider a broader range of pose estimation systems.**

## 5.2. Further Analyses

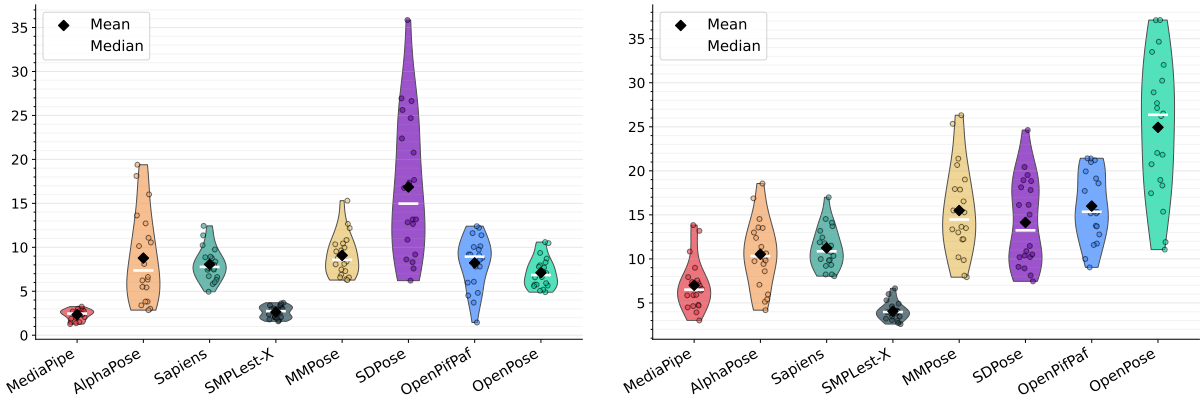
**Temporal jittering (pose stability)** Table 3 summarizes the temporal stability of pose estimators on Signsuisse and Phoenix for all keypoints (excluding legs); full results by anatomical region (hands, face) are in Appendix B. Overall, the quantitative trends align with visual inspection: methods that appear temporally unstable exhibit higher $J_{\text{acc}}$ and $J_{\text{jerk}}$, particularly in the hand subset. While both metrics exhibit similar trends, $J_{\text{jerk}}$ provides greater separation between estimators (Figure 3) and is thus used to illustrate differences.

On Signsuisse, MediaPipe achieves the lowest median jerk jitter for all keypoints ($J_{\text{jerk}} = 2.46$), followed closely by SMPLest-X ($2.74$). However, the violin plots reveal that their distributions overlap substantially (Figure 3a), indicating that the difference may not be reliable given the sample size of 20 sequences. SDPose exhibits the highest jerk values ($14.97$), with a wide interquartile range reflecting high variance across sequences, despite yielding the most stable face keypoints (see Appendix Table 8 for full results).

When focusing on hands (Appendix Table 7), SMPLest-X provides the lowest jitter ($J_{\text{jerk}} = 5.85$); however, this temporal smoothness coincides with visibly rigid hand configurations (Figure 2). For face

| Pose Estimator | Signsuisse | | | Phoenix | | |
|---|---|---|---|---|---|---|
| | $E_v$ | $J_{\text{acc}}$ (↓) | $J_{\text{jerk}}$ (↓) | $E_v$ | $J_{\text{acc}}$ (↓) | $J_{\text{jerk}}$ (↓) |
| MediaPipe | **1.16** | **1.44** (1.13–1.60) | **2.46** (2.01–2.81) | **3.07** | 3.69 (3.01–4.73) | 6.51 (5.10–8.38) |
| OpenPose | 3.30 | 3.97 (3.56–4.74) | 6.84 (5.90–8.18) | 9.32 | 14.74 (11.38–16.64) | 26.36 (19.57–29.55) |
| MMPose Wholebody | 3.83 | 4.95 (4.26–5.90) | 8.58 (7.32–10.35) | 6.84 | 8.75 (7.00–11.17) | 14.47 (11.66–18.59) |
| OpenPifPaf | 2.76 | 4.84 (3.43–5.54) | 8.94 (6.35–10.20) | 5.22 | 8.60 (7.28–10.31) | 15.35 (13.12–18.63) |
| SDPose | 6.21 | 8.45 (6.19–11.76) | 14.97 (10.70–21.79) | 6.83 | 7.99 (6.61–10.45) | 13.24 (10.80–17.10) |
| Sapiens | 2.97 | 4.42 (3.72–5.05) | 7.77 (6.62–9.16) | 4.55 | 6.33 (5.44–7.43) | 10.86 (9.62–12.88) |
| AlphaPose | 4.09 | 4.56 (3.08–6.70) | 7.36 (5.15–11.50) | 5.81 | 6.21 (4.90–7.69) | 10.32 (8.35–12.47) |
| SMPLest-X | 2.13 | 1.74 (1.39–2.00) | 2.74 (2.16–3.14) | 3.08 | **2.68** (2.23–3.30) | **3.97** (3.19–4.85) |

Table 3: Temporal stability metrics on Signsuisse and Phoenix datasets, for all keypoints excluding legs. $E_v$: motion energy (median joint velocity). $J_{\text{acc}}$ and $J_{\text{jerk}}$: acceleration- and jerk-based jitter; lower is smoother. Values are median across 20 sequences, with interquartile range (Q1–Q3) in parentheses. All values scaled by 100. Boldface marks the lowest value when distributions are clearly separated (see Figure 3). Full results by anatomical region in Appendix B.




(a) Signsuisse

(b) Phoenix

Figure 3: Distribution of per-sequence jerk jitter ($J_{\text{jerk}}$) across pose estimators for all keypoints (excluding legs). Each point represents one video sequence ($n = 20$). Black diamonds indicate the mean; white bars indicate the median. The violin shape shows the kernel density estimate. Overlapping distributions (e.g., MediaPipe vs. SMPLest-X on Signsuisse) indicate that differences between estimators may not be reliable at this sample size.

keypoints, SDPose achieves the lowest jitter (0.38), though MediaPipe (0.44) and OpenPifPaf (0.43) are nearly indistinguishable.

On Phoenix, instability increases for all methods, consistent with the lower spatial resolution of the videos (approximately 5.3× fewer pixels than Signsuisse). SMPLest-X achieves the lowest median jerk jitter for all keypoints (3.97) and hands (7.40), with distributions clearly separated from other estimators (Figure 3b). MediaPipe is second-lowest overall (6.51). OpenPose exhibits the highest jitter across all regions, with a median $J_{\text{jerk}}$ of 26.36 for all keypoints and 67.74 for hands. For face keypoints, SDPose achieves the lowest jitter (1.17), while Sapiens is notably high (10.29), consistent with its dense 243-landmark face mesh amplifying localization noise at low resolution.

Across both datasets, the violin plots (Figure 3) provide important context beyond summary statistics: estimators with similar median values may have heavily overlapping distributions, tempering conclusions about which is "best." When interpreted jointly with motion energy, the results suggest that SMPLest-X favors temporal smoothness at the cost of limited articulation, while MediaPipe—and to a lesser extent Sapiens—offer a trade-off between stability and expressive motion.

**Occlusion** The results of the occlusion analysis are available in Table 4. Sapiens produced accurate estimations for 15/15 instances of occlusion surveyed. Notably, there are several circumstances where MediaPipe is missing one or both hands for a given sign, but Sapiens produces the pose correctly. One example of this trend is shown in Figure 4. OpenPifPaf, on the other hand, was missing some or all of one or both hands in all but one of the 15 instances of occlusion surveyed. This propensity

| estimator | correctness (%) |
|---|---|
| Mediapipe | 73.33 |
| OpenPose | 66.66 |
| MMPose Wholebody | 33.33 |
| OpenPifPaf | 6.66 |
| SDPose | 46.66 |
| Sapiens | 100.00 |
| AlphaPose-136 | 40.00 |
| SMPLest-X | 0.00 |

Table 4: Percentage of the time that the estimators correctly estimated a pose featuring occlusion in the 15 Signsuisse examples surveyed.

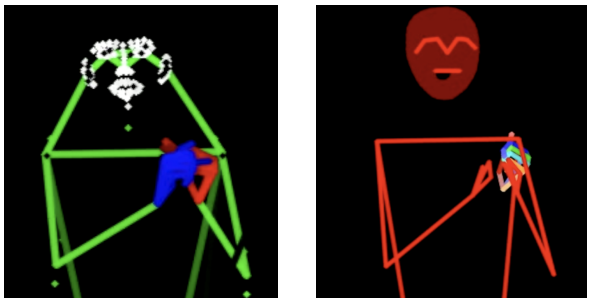


Figure 4: Pose estimation visualizations with hand occlusion for the DSGS sign SPORT. Left: Sapiens correctly predicts both hands despite partial occlusion. Right: MediaPipe fails to detect the right hand, resulting in missing keypoints. Images are cropped to highlight the hands.

for missing hands may contribute to the low BLEU score of OpenPifPaf. SMPLest-X, in turn, did provide hands in all frames, but did not produce any accurate representations of the pose.

MMPose Wholebody and SDPose both estimated a hand position for all frames in all of the 15 examples surveyed, but were more likely than other estimators to estimate an impossible hand position, for example bending the fingers backwards. MMPoseWholebody created an acceptable estimation for 5/15 instances of occlusion, while SDPose produced an acceptable estimation in 7/15 examples.

**Missing hand keypoints** We report the percentage of signing frames with a missing hand for each estimator (Table 5), computed on 20 randomly selected videos from the Phoenix training split. A full breakdown across varying missing thresholds ($10$–$100\%$) is provided in Appendix A. Only OpenPifPaf, MediaPipe, and OpenPose assign $c = 0$ to hand keypoints; all other estimators never produce $c = 0$, resulting in $0\%$ missing across all thresholds.

OpenPifPaf shows severe degradation: in roughly $68\%$ of signing frames the left hand is missing and in $59\%$ the right hand is missing. In over $40\%$ of signing frames, both hands are simultaneously absent. MediaPipe exhibits a milder but consistent pattern, with around $20$–$24\%$ of signing frames missing a hand. OpenPose is affected less frequently, with $3$–$8\%$ of frames affected. Notably, both MediaPipe and OpenPose exhibit binary behavior: when a hand is lost, all of its keypoints are set to $c = 0$ at once rather than partially, as shown by the constant values across thresholds in the full breakdown (Table 6 in Appendix A).

| estimator | left | right | both |
|---|---|---|---|
| MediaPipe | 20.22 | 23.61 | 8.84 |
| OpenPose | 8.43 | 3.88 | 0.22 |
| OpenPifPaf | 67.65 | 59.19 | 40.63 |

Table 5: Percentage of signing frames with a missing hand on 20 randomly selected Phoenix training videos. A hand is missing when $\geq 50\%$ of its keypoints have $c = 0$. "both": both hands missing simultaneously. Only estimators producing $c = 0$ hand keypoints are shown; all others (AlphaPose, MMPose Wholebody, Sapiens, SDPose, SMPLest-X) score $0\%$. Full threshold sweep in Appendix A.

These three estimators also obtain some of the weakest translation scores (Table 2), suggesting that zero-confidence hand keypoints—explicitly masked during preprocessing—directly impair downstream translation. This pattern could also reveal a limited capability of current translation models to recover missing hand information from temporal context alone.

## 6. Conclusions

We presented a controlled comparison of pose estimators for pose-based SLT, motivated by the fact that most prior SLT pipelines default to MediaPipe as a convenient choice. Our experiments show that this default is not necessarily optimal: several estimators outperform MediaPipe on Phoenix, including SDPose, Sapiens, AlphaPose, and MMPose Wholebody. Among them, SDPose and Sapiens achieve the highest BLEU/BLEURT scores.

**Further analyses on signing data** Estimators that frequently remove substantial hand information (notably OpenPifPaf, and to a lesser extent MediaPipe and OpenPose) are associated with weaker translation results. The temporal stability analysis further shows that higher translation performance does not necessarily coincide with smoother trajectories: MediaPipe and SMPLest-X yield the lowest jitter scores, whereas SDPose exhibits comparatively high hand jerk on Signsuisse, and SMPLest-X attains very low jitter at the cost of visibly implausible or rigid hand articulations. Taken together, these findings underscore that estimator choice can materially affect both translation performance and pose quality characteristics relevant to signing.

**Trade-off between translation quality and computational cost** Sapiens provides strong translation performance and the most robust behavior under occlusion (15/15 correct in our manual analysis), but it is also among the most resource-intensive and slowest options. SDPose reaches comparable translation quality results with similarly high runtime requirements. Conversely, faster or more lightweight estimators are not guaranteed to be suitable for SLT: OpenPifPaf, for instance, performs poorly in translation and exhibits severe hand keypoint dropout. AlphaPose, which has the third-best BLEU score and the fastest compute speed, may be a better option when time or computing resources are limited.

Overall, our results provide empirical evidence that pose-based SLT should not treat pose estimation as a fixed implementation detail. Future research should therefore consider pose estimators beyond MediaPipe when building pose-based translation systems, and evaluate estimator choice jointly with practical constraints such as runtime and hardware requirements as well as robustness properties such as occlusion handling, missing-hand behavior, and temporal stability.

## 7. Limitations and Future Work

**Limitations of the Phoenix dataset** It should be noted that the signing in the Phoenix dataset is done live by hearing interpreters. Accordingly, this dataset has notable flaws. Due to the time pressure of the live setting, the interpreters may omit some information. Furthermore, the signing is an interpretation of German spoken language, and not natural signing, and thus may be influenced by German grammar. Lastly, the subject domain of Phoenix is fairly limited and mostly pertains to weather reports. Future investigations should expand the analysis of pose estimators to new datasets, particularly those that contain natural signing from deaf L1 signers with high-quality translations that are not produced under time pressure.

Translation experiments on more suitable, newer, or larger sign language datasets would provide a more reliable assessment of how pose estimator choice generalizes beyond Phoenix.

**Fast movement** Anecdotally, a visual analysis suggests that estimators such as Sapiens may be vulnerable to errors on videos containing fast movements (which is normal for signing). Future research could therefore investigate the estimators' robustness to fast movements.

**Variable model capacity** All experiments use a shared translation architecture with a linear projection layer that maps pose features into the language model embedding space (see Section 4.1). Because pose estimators produce different numbers of keypoints (Table 1), this projection layer contains a different number of parameters for each estimator. While this design allows each representation to be fully utilized, it introduces variation in model capacity that may influence translation performance. Isolating the impact of adapter dimensionality is left for future work.

## Acknowledgements

CO is supported by a Fulbright Scholarship. GS and MM received funding from the SIGMA project (grant no. G-95017-01-07), supported by the Digital Society Initiative (DSI) at the University of Zurich.

## 8. Bibliographical References

Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176.

Safaeid Hossain Arib, Rabeya Akter, Sejuti Rahman, and Shafin Rahman. 2025. Signformer-gcn: Continuous sign language translation using spatio-temporal graph convolutional networks. *PLOS ONE*, 20(2):1–19.

Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2pose: Animating sign language notation into pose sequences. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21046–21056.

Alessia Battisti, Emma van den Bold, Anne Göhring, Franz Holzknecht, and Sarah Ebling. 2024. Person identification from pose estimates in sign language. *11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*.

C. Fabian Benitez-Quiroz, Kadir Gökgöz, Ronnie B. Wilbur, and Aleix M. Martinez. 2014. Discriminant features and temporal structure of nonmanuals in american sign language. *PLOS ONE*, 9(2):1–17.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31,

New York, NY, USA. Association for Computing Machinery.

Diane Brentari. 2011. Sign language phonology. In John Goldsmith, Jason Riggle, and Alan C. L. Yu, editors, *The Handbook of Phonological Theory*, 2nd edition, chapter 21, pages 691–721. John Wiley & Sons.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. 2023. Smpler-x: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310.

Mathieu De Coster, Ellen Rushe, Ruth Holmes, Anthony Ventresque, and Joni Dambre. 2023. Towards the extraction of robust sign embeddings for low resource sign language recognition.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Steven Verstockt. 2024. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, 23:1305–1331.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2734–2743.

Chengyu Fan and Tahiya Chowdhury. 2025. When pose estimation fails: Measuring occlusion for reliable multimodal interaction. In *Companion Proceedings of the 27th International Conference on Multimodal Interaction*, pages 58–64.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12943–12954. IEEE.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2023. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. 2025. Vision-based sign language translation via a skeleton-aware neural network. *J. Comput. Sci. Technol.*, 40(2):378–396.

Ibai Gorordo. 2024. Sapiens pytorch inference. `https://github.com/ibaiGorordo/Sapiens-Pytorch-Inference`. GitHub repository.

Ivan Grishchenko and Valentin Bazarevsky. 2020. MediaPipe Holistic - simultaneous face, hand and pose prediction, on device. Google AI Blog.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. 2025. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810, Vienna, Austria. Association for Computational Linguistics.

Ruth Holmes, Ellen Rushe, Frank Fowley, and Anthony Ventresque. 2022. Improving signer independent sign language recognition for low resource languages. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 45–52, Marseille, France. European Language Resources Association.

Eui Jun Hwang, Sukmin Cho, Huije Lee, Youngwoo Yoon, and Jong C. Park. 2025. A spatio-temporal representation learning as an alternative to traditional glosses in sign language translation and production. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3352–3362.

Zifan Jiang, Colin Leong, Amit Moryossef, Oliver Cory, Maksym Ivashechkin, Neha Tarigopula, Biao Zhang, Anne Göhring, Annette Rios, Rico Sennrich, and Sarah Ebling. 2025. Meaningful pose-based sign language evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 64–80, Suzhou, China. Association for Computational Linguistics.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Kyung-Min Jin, Byoung-Sung Lim, Gun-Hee Lee, Tae-Kyung Kang, and Seong-Whan Lee. 2023. Kinematic-aware hierarchical attention network for human pose estimation in videos. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5714–5723.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020a. Whole-body human pose estimation in the wild. In *Computer Vision – ECCV 2020*, pages 196–214, Cham. Springer International Publishing.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020b. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Robert E. Johnson and Scott K. Liddell. 2010. Toward a phonetic representation of signs: Sequentiality and contrast. *Sign Language Studies*, 11(2):241–274.

Abhinav Joshi, Vaibhav Sharma, Sanjeet Singh, and Ashutosh Modi. 2025. PoseStitch-SLT: Linguistically inspired pose-stitching for end-to-end sign language translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13834–13853, Suzhou, China. Association for Computational Linguistics.

Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. 2023. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 618–629.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. 2025. Sapiens: Foundation for human vision models. In *Computer Vision – ECCV 2024*, pages 206–228, Cham. Springer Nature Switzerland.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13).

Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2021. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13498–13511.

Cristian Lazo-Quispe, Joe Huamani-Malca, Manuel Huamán-Ramos, Pablo Rivas, and Tomas Cerny. 2022. Impact of pose estimation models for landmark-based sign language recognition. In *LXAI Workshop Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1448–1458.

Shuang Liang, Jing He, Chuanmeizhi Wang, Lejun Liao, Guo Zhang, Yingcong Chen, and Yuan Yuan. 2025. Sdpose: Exploiting diffusion priors for out-of-domain and robust pose estimation.

Zeyu Liang, Huailing Li, and Jianping Chai. 2023. Sign language translation: A survey of approaches and techniques. *Electronics*, 12(12).

Scott K Liddell and Robert E Johnson. 1989. American sign language: The phonological base. *Sign language studies*, 64(1):195–277.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. 2023. One-stage 3d whole-body mesh recovery with component aware transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21159–21168.

Filipa Lino, Carlos Santiago, and Manuel Marques. 2025. Benchmarking 3d human pose estimation models under occlusions. *arXiv preprint arXiv:2504.10350*.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6).

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. *CoRR*, abs/1906.08172.

Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. 2022. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2637–2646.

Amit Moryossef. 2024. Optimizing hand region detection in mediapipe holistic full-body pose estimation to improve accuracy and avoid downstream errors.

Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023a. An open-source gloss-based baseline for spoken to signed language translation. In *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*, pages 22–33, Tampere, Finland. European Association for Machine Translation.

Amit Moryossef, Mathias Müller, and Rebecka Fahrni. 2023b. pose-format: Library for viewing, augmenting, and handling .pose files.

Amit Moryossef, Gerard Sant, and Zifan Jiang. 2025. Pose-based sign language appearance transfer. In *Proceedings of the Third International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–6, Geneva, Switzerland. European Association for Machine Translation.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020. Real-time sign language detection using human pose estimation. In *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12536 of *Lecture Notes in Computer Science*, pages 237–248. Springer.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3434–3440.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023a. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023b. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Amit Moryossef, Sarah Ebling, and Thomas Hanke. 2023c. EASIER project deliverable 4.3: Final translation systems V2.

Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Chan-Il Park and Chae-Bong Sohn. 2020. Data augmentation for human keypoint estimation deep learning based sign language translation. *Electronics*, 9(8).

Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. 2021. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13463–13473.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10967–10977.

Roland Pfau and Josep Quer. 2010. Nonmanuals: Their grammatical and prosodic roles. In Diane Brentari, editor, *Sign Languages*, Cambridge Language Surveys, pages 381–402. Cambridge University Press, Cambridge.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Siegmund Prillwitz and Heiko Zienert. 1990. Hamburg notation system for sign language: Development of a sign writing with computer application. In *Current Trends in European Sign Language Research: Proceedings of the 3rd European Congress on Sign Language Research*, pages 355–379.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2021. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1749–1759.

Wendy Sandler. 2012. The phonological organization of sign languages. *Language and Linguistics Compass*, 6(3):162–182.

Gerard Sant, Zifan Jiang, Carlos Escolano, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. Multimodalhugs: Enabling sign language processing in hugging face. Manuscript submitted for publication.

Gerard Sant, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2025. Modality matters: Training and tokenization effects in sign-to-text translation. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 687–705. Springer.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5131–5141.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Valerie Sutton. 1990. *Lessons in SignWriting*. SignWriting.

Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. 2023. Sign language translation from instructional videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5625–5635.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *2016 IEEE Conference on Computer*

*Vision and Pattern Recognition (CVPR)*, pages 4724–4732.

Di Xia, Yeqing Zhu, and Heng Zhang. 2022. Faster deep inertial pose estimation with six inertial sensors. *Sensors*, 22(19).

Qinkun Xiao, Minying Qin, and Yuting Yin. 2020. Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks*, 125:41–55.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving gloss-free sign language translation by reducing representation density. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2021. Transpose: real-time 3d human translation and pose estimation with six inertial sensors. *ACM Trans. Graph.*, 40(4).

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. 2026. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1778–1794.

Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. 2022. Smoothnet: A plug-and-play network for refining human poses in videos. In *Computer Vision – ECCV 2022*, pages 625–642, Cham. Springer Nature Switzerland.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. *Advances in neural information processing systems*, 37:114018–114047.

Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. 2023. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. 2022. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, volume 13666 of *Lecture Notes in Computer Science*, pages 180–200. Springer.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 2021. 3d human pose estimation with spatial and temporal transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20814–20824. IEEE.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Yue Zhu, Nermin Samet, and David Picard. 2023. H3wb: Human3.6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20166–20177.

| Estimator | Hand | % hand keypoints missing ($\geq x$) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
| OpenPifPaf | Left | 85.16 | 80.86 | 76.51 | 72.10 | 67.65 | 63.29 | 59.31 | 54.59 | 51.81 | 0.00 |
| | Right | 82.59 | 75.01 | 69.90 | 63.68 | 59.19 | 55.66 | 51.96 | 49.01 | 46.19 | 0.00 |
| | Both | 66.90 | 57.98 | 52.14 | 45.91 | 40.63 | 36.43 | 32.32 | 28.91 | 26.05 | 0.00 |
| MediaPipe | Left | 20.22 | 20.22 | 20.22 | 20.22 | 20.22 | 20.22 | 20.22 | 20.22 | 20.22 | 0.00 |
| | Right | 23.61 | 23.61 | 23.61 | 23.61 | 23.61 | 23.61 | 23.61 | 23.61 | 23.61 | 0.00 |
| | Both | 8.84 | 8.84 | 8.84 | 8.84 | 8.84 | 8.84 | 8.84 | 8.84 | 8.84 | 0.00 |
| OpenPose | Left | 8.43 | 8.43 | 8.43 | 8.43 | 8.43 | 8.43 | 8.43 | 8.43 | 8.43 | 0.00 |
| | Right | 3.88 | 3.88 | 3.88 | 3.88 | 3.88 | 3.88 | 3.88 | 3.88 | 3.88 | 0.00 |
| | Both | 0.22 | 0.22 | 0.22 | 0.22 | 0.22 | 0.22 | 0.22 | 0.22 | 0.22 | 0.00 |

Table 6: Percentage of signing frames in which at least $x\%$ of a hand's keypoints have confidence $c = 0$, measured on 20 randomly selected videos from the Phoenix training split. Signing frames are those where the wrist is vertically above the elbow. "Both": both hands simultaneously exceed the threshold. Only estimators producing $c = 0$ hand keypoints are shown; all others score $0\%$. The $50\%$ column corresponds to the missing hand definition in Table 5.

## A. Missing Hand Keypoints

Table 6 reports the percentage of signing frames with missing hand keypoints across varying thresholds. The $50\%$ column corresponds to the definition of a missing hand used in the main text (Table 5). MediaPipe and OpenPose exhibit binary behavior: when a hand is lost, all its keypoints are set to $c = 0$ simultaneously, resulting in constant values across thresholds $10$–$90\%$.

## B. Temporal Stability by Region

Tables 7 and 8 report temporal stability metrics for hand and face keypoints on Signsuisse and Phoenix. The all-keypoints region is reported in the main text (Table 3). Corresponding violin plots are shown in Figures 5 and 6.

**Hands** Hand keypoints exhibit substantially higher jitter than full-body or face keypoints across all estimators, reflecting the difficulty of localizing fine-grained finger articulations.

On Signsuisse, SMPLest-X achieves the lowest median hand jerk jitter ($J_{\text{jerk}} = 5.85$), clearly separated from the next-lowest estimator (MMPose Wholebody, $10.48$). However, as noted in the main text, this smoothness coincides with rigid, often implausible hand configurations (Figure 2). SDPose exhibits the highest hand jitter ($29.98$) with very wide interquartile range, indicating high variance across sequences. MediaPipe ($25.29$) and OpenPifPaf ($22.61$) also show high hand jitter on Signsuisse, though for different reasons: MediaPipe's hand jitter is inflated by its frequent missing-hand detections (Table 5), while OpenPifPaf's is driven by noisy localization even when hands are detected.

On Phoenix, the same pattern holds with larger values: SMPLest-X remains lowest ($7.40$), while

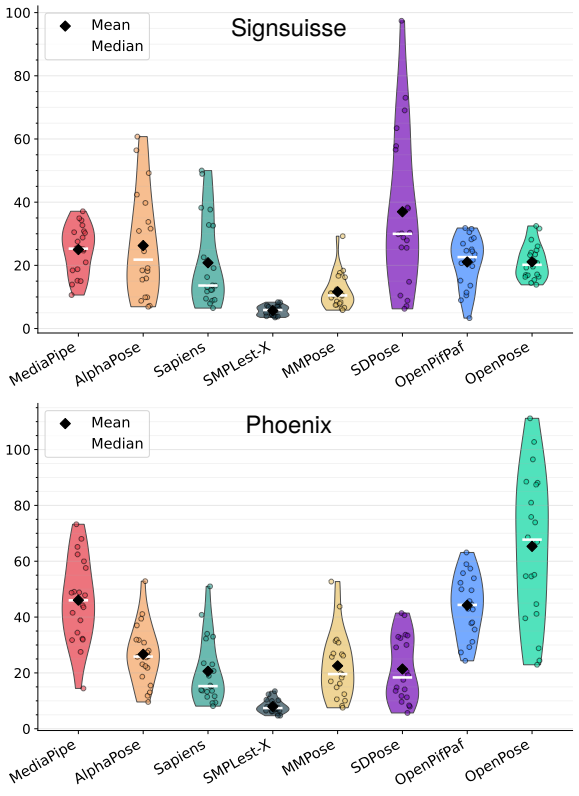



Figure 5: Distribution of per-sequence jerk jitter ($J_{\text{jerk}}$) for hand keypoints on Signsuisse (top) and Phoenix (bottom). Format as in Figure 3.

OpenPose ($67.74$) and MediaPipe ($45.99$) are the most unstable. The increase relative to Signsuisse is consistent with Phoenix's lower resolution making hand keypoint localization more challenging.

**Face** Face keypoints show the lowest jitter overall, as facial landmarks undergo less motion than hands or body during signing. Differences between estimators are smaller but still informative.

| Pose Estimator | Signsuisse | | | Phoenix | | |
|---|---|---|---|---|---|---|
| | $E_v$ | $J_{\text{acc}}$ (↓) | $J_{\text{jerk}}$ (↓) | $E_v$ | $J_{\text{acc}}$ (↓) | $J_{\text{jerk}}$ (↓) |
| MediaPipe | 10.49 | 14.50 (10.73–16.23) | 25.29 (20.01–29.52) | 15.99 | 25.12 (20.58–31.66) | 45.99 (35.73–56.71) |
| OpenPose | 9.78 | 11.64 (10.01–13.55) | 20.18 (17.39–23.69) | 23.47 | 37.63 (27.70–45.74) | 67.74 (47.80–82.59) |
| MMPose Wholebody | 7.01 | 6.54 (4.78–8.47) | 10.48 (7.81–14.08) | 12.28 | 12.26 (9.10–17.67) | 19.58 (14.08–28.58) |
| OpenPifPaf | 6.88 | 11.96 (8.00–14.38) | 22.61 (15.11–26.69) | 13.66 | 24.46 (19.96–28.12) | 44.35 (36.52–51.38) |
| SDPose | 13.32 | 16.53 (11.68–25.10) | 29.98 (20.35–45.73) | 12.31 | 12.62 (8.47–17.52) | 18.37 (13.62–26.73) |
| Sapiens | 8.27 | 8.42 (6.39–14.92) | 13.61 (10.67–27.18) | 10.01 | 9.94 (7.44–15.42) | 15.27 (11.92–26.14) |
| AlphaPose | 12.24 | 13.53 (7.72–21.19) | 21.81 (12.82–36.93) | 13.94 | 16.36 (11.23–19.74) | 25.85 (18.36–32.92) |
| SMPLest-X | **5.53** | **3.87** (3.18–4.51) | **5.85** (4.72–6.80) | **7.36** | **5.56** (4.68–7.02) | **7.40** (6.49–9.34) |

Table 7: Temporal stability metrics for hand keypoints on Signsuisse and Phoenix. Format as in Table 3: median (IQR), scaled by 100.

| Pose Estimator | Signsuisse | | | Phoenix | | |
|---|---|---|---|---|---|---|
| | $E_v$ | $J_{\text{acc}}$ (↓) | $J_{\text{jerk}}$ (↓) | $E_v$ | $J_{\text{acc}}$ (↓) | $J_{\text{jerk}}$ (↓) |
| MediaPipe | 0.31 | 0.27 (0.24–0.31) | 0.44 (0.38–0.49) | 1.83 | 1.65 (1.18–2.85) | 2.73 (2.07–4.81) |
| OpenPose | 0.39 | 0.46 (0.42–0.50) | 0.81 (0.73–0.87) | 3.07 | 4.15 (2.47–5.73) | 7.16 (4.26–10.32) |
| MMPose Wholebody | 0.44 | 0.64 (0.58–0.67) | 1.14 (1.04–1.19) | 1.63 | 1.23 (1.08–1.50) | 1.84 (1.63–2.20) |
| OpenPifPaf | 0.32 | 0.27 (0.23–0.30) | 0.43 (0.38–0.48) | 1.64 | 1.08 (0.87–1.60) | 1.52 (1.20–2.40) |
| SDPose | 0.31 | 0.24 (0.22–0.27) | 0.38 (0.34–0.42) | 1.59 | **0.86** (0.77–1.15) | **1.17** (1.01–1.42) |
| Sapiens | 2.20 | 3.63 (3.17–4.08) | 6.64 (5.76–7.45) | 3.92 | 5.82 (5.23–6.56) | 10.29 (9.37–11.60) |
| AlphaPose | 0.42 | 0.55 (0.48–0.58) | 0.95 (0.83–1.01) | 1.75 | 1.41 (1.08–2.37) | 2.22 (1.68–3.87) |
| SMPLest-X | 0.69 | 0.81 (0.65–0.99) | 1.38 (1.10–1.69) | **1.37** | 1.49 (1.27–1.77) | 2.50 (2.09–2.97) |

Table 8: Temporal stability metrics for face keypoints on Signsuisse and Phoenix. Format as in Table 3: median (IQR), scaled by 100.

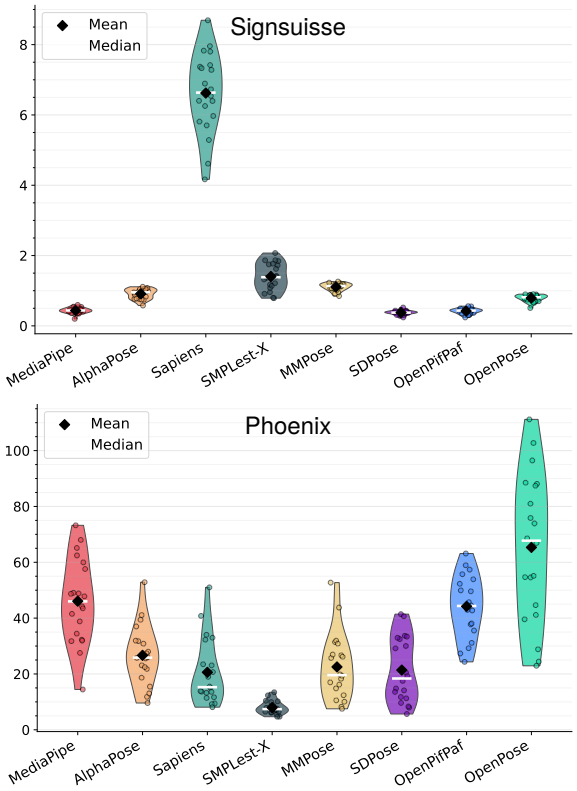



Figure 6: Distribution of per-sequence jerk jitter ($J_{\text{jerk}}$) for face keypoints on Signsuisse (top) and Phoenix (bottom). Format as in Figure 3.

On Signsuisse, SDPose ($J_{\text{jerk}} = 0.38$), MediaPipe (0.44), and OpenPifPaf (0.43) are nearly indistinguishable and achieve the lowest face jitter. Sapiens is a clear outlier (6.64), likely because its dense 243-landmark face mesh amplifies small localization errors into measurable jitter.

On Phoenix, SDPose again achieves the lowest face jitter (1.17), with MMPose Wholebody (1.84) and OpenPifPaf (1.52) also performing well. Sapiens remains the highest (10.29), confirming that its dense face representation is especially sensitive to low-resolution input. MediaPipe shows moderate face jitter (2.73) with a wide IQR, suggesting inconsistent face tracking across sequences.

# Designing a Data Model for a Diachronic Sign Language Database: A Case Study of Nineteenth-Century Bohemian Sources

**Lenka Okrouhlíková**

Faculty of Arts, Charles University, Institute of Czech and Deaf Studies
nám. Jana Palacha 1/2, 116 38 Prague 1, Czech Republic
lenka.okrouhlikova@ff.cuni.cz

**Abstract**

Diachronic research on sign languages is limited by the fragmentary and heterogeneous nature of historical documentation. Eighteenth- and nineteenth-century printed texts and manuscripts contain valuable lexical data, but their descriptions vary in precision, terminology, and representational conventions. This paper proposes a structured data model for a diachronic sign language database designed to systematise such archival materials. The proposed model adopts a multi-layered architecture that separates primary evidence from analytical interpretation, distinguishes attested from inferred sign parameters, applies graded confidence levels, and encodes structural, iconic, and metaphorical properties in parallel layers. Detailed source metadata ensures traceability and explicit representation of uncertainty. The model is illustrated through sign attestations drawn from nineteenth century Bohemian sources. The case study demonstrates that even fragmentary records, most commonly documented in dictionaries and pedagogical materials through written descriptions or illustrations, can be systematically represented within a unified data model suitable for structured comparison and diachronic analysis. The proposed model may also provide a methodological basis for comparable work on other European sign languages.

**Keywords:** archival sign language sources, diachronic sign language database, historical sign lexicon, sign language data modelling, lexical comparison

## 1. Diachronic Research and the Need for an Archival Database

Research in diachronic sign language linguistics has expanded in recent decades, yet it remains comparatively limited in scope and unevenly distributed across languages. Surveys of the field emphasise that historical and comparative studies still constitute only a small proportion of sign linguistic research, largely because of the scarcity of historical materials, the fragmentary nature of pre-video documentation, and the difficulty of identifying historical relationships among sign languages (Reagan, 2021; Power, 2022; Power, Quinto-Pozos and Law, 2025). Much of the existing work has focused on American Sign Language and, by extension, on French Sign Language, which is associated with the richest body of early sources (Frishberg, 1975, 1979; Woodward, 1976, 1978; Bonnal-Vergès, 2005; Delaporte, 2007). These studies have primarily examined pathways of morphophonological change and mechanisms of diachronic variation within individual sign languages, often drawing on nineteenth-century dictionaries and pedagogical sources for French Sign Language. In parallel, cross-linguistic work has pursued the quantitative classification of sign languages through lexicostatistical and phylogenetic approaches, most frequently extrapolating genealogical relations from synchronic or twentieth-century data rather than from systematic diachronic corpora (cf. Woodward, 2011; Reagan, 2021; Power, 2022; Power, Quinto-Pozos and Law, 2025). At the same time, scholars have highlighted persistent methodological problems, including the role of iconicity in producing false cognates, the lack of consensus on criteria for cognacy, and the incompleteness of nineteenth-century descriptions, which often privilege motivation over form (Power, 2022; Okrouhlíková, 2025a). Recent reviews, therefore, converge on the need for more systematic diachronic research grounded in archival sources, combining philological analysis with structured lexical databases and what has been termed a sign language archaeology approach, in order to reassess historical relatedness and the formation of sign language families beyond exclusively synchronic evidence (Power, Quinto-Pozos and Law, 2025).

The present study focuses on the periods and types of evidence that enable systematic linguistic analysis and database construction. The kinds of sources that allow systematic structural analysis and the construction of comparative databases emerge primarily from the period of institutionalised deaf education from the late eighteenth century onwards, when European sign languages took shape in forms broadly comparable to those used today. If monastic sign systems are set aside (cf. SIGNA database; Těšínská Lomičková and Těšínský, 2009), whose relationship to later sign languages remains contested (Vilwock, 2012), it is this period that provides the first sustained bodies of evidence suitable for diachronic investigation and corpus-based resource creation.

As a language articulated in space and motion and lacking a conventional written form, sign language is accessible in historical periods almost exclusively through paper-based sources. Such materials can never capture the language's full



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 387–397
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

complexity and typically restrict analysis to individual lexical items, most often those recorded in dictionaries, glossaries, and pedagogical manuals. Diachronic research is therefore necessarily grounded in fragmentary representations of signs rather than in continuous discourse, a limitation that fundamentally shapes both the scope of historical inquiry and the design of any comparative database built on archival materials. While considerable scholarly attention has been devoted to French Sign Language dictionaries (cf. Arnaud, 2019), their descriptive practices, and their modern re-editions, we argue that comparable historical sources exist for a range of other European sign languages but have so far been only sparsely documented and synthesised (e.g. Fischer, 2010a, 2010b; Linde-Usiekniewicz and Łozińska, 2023; Nyst and Schüller, 2024; Okrouhlíková, 2025b).

Cross-linguistic comparisons of historical signs remain rare (Regen, 2010; Okrouhlíková, 2025a), despite their importance for evaluating degrees of similarity across European sign languages and for reassessing long-assumed family groupings, especially those linked to old French Sign Language. In addition to the limited availability of materials, such work is complicated by the fact that surviving records disproportionately document highly iconic signs, while written descriptions often focus on motivation rather than form or provide only partial accounts of a sign's structure. Visual representations pose an additional challenge, as historical illustrations employ highly heterogeneous graphic conventions for depicting handshape, orientation, and movement, often without explicit legends or consistent systems of notation. These constraints underscore the need for systematically organised archival databases that make historical lexical data comparable across languages and regions.

## 2. Designing the Diachronic Sign Database

### 2.1 Relevant Models and Precedents

In designing a diachronic database for sign languages, methodological inspiration can be drawn from spoken-language resources and digital humanities projects, though no direct equivalent exists. *EvoSem* (François et al., n.d.; cf. Dehouck et al., 2023) offers a model for structuring diachronic lexical relations and semantic connections across historically related forms through the concept of dialexification, while *CLICS* (cf. Rzymski et al., 2019) provides a framework for the interactive representation of cross-linguistic colexification patterns. *Ticha* (Lillehaugen et al., 2016; cf. Broadwell et al., 2020) demonstrates how archival language materials can be presented through linked layers of source images, transcription, translation, linguistic analysis and metadata.

Historical sign language sources, however, require a distinct modelling strategy, because they lack a conventional written form and survive only in fragmentary visual and written representations. In addition, relationships among sign languages are often shaped less by descent from a single proto-language than by lateral transmission and institutional contact (cf. Reagan, 2021).

A relevant precedent is the *Historical Sign Language Database* (cf. Supalla 2001) for American Sign Language, which integrates early twentieth-century NAD films, printed ASL dictionaries from the same period, and three historical French Sign Language dictionaries (Brouland, 1855; Pélissier, 1856; Lambert, 1865). In its current interface, the database organises entries by fields such as source (author and year), entry ID (number), a photo or picture of the sign if available, Reference Gloss, Author's Gloss, and provides a description and page reference for each dictionary entry.

Another relevant point of reference is Bonnal-Vergès's (2005) work on the etymological and historical dictionary of French Sign Language. Her model organises historically related forms into iconic families and matrices and relates them through the notion of the *iconon*, understood as the core conceptual image underlying formally and historically connected variants. This approach is particularly relevant to the present proposal because it combines strict traceability to primary archival sources with a distinct interpretative layer of motivational analysis and with an explicitly diachronic account of formal change. In this sense, it offers a useful analogue for sign language data to the structuring of diachronic lexical relations found in EvoSem, while remaining sensitive to the specific role of iconicity and conventionalisation in sign language history.

### 2.2 Proposed Database Structure

We propose a diachronic lexical database that would bring together representations of signs from eighteenth- and nineteenth-century European sources, including printed dictionaries, pedagogical manuals and unpublished manuscripts. Rather than presupposing stable cross-linguistic equivalence, the proposed database is designed to accommodate uncertainty, variation and partial documentation. Historical records differ substantially in their aims, levels of detail and terminological conventions, and it is often unclear whether two sources describe formally identical signs, semantically related items or independent, iconically motivated sign forms. In many cases, authors describe signs in written language without specifying phonological parameters in sufficient detail, while others foreground motivation or cultural interpretation rather than form (cf. Okrouhlíková,

2025a, 29–45). Each historical attestation would therefore be treated as a source-specific record linked to language, place, year, institutional context and degree of descriptive precision, allowing competing hypotheses about equivalence and relatedness to be represented explicitly rather than resolved prematurely. This design is intended to enable comparative work across languages and regions while remaining faithful to the fragmentary and heterogeneous nature of archival evidence.

Each record in the proposed database would be anchored by a stable identifier and glossing layers. Because the database would bring together sources written in different languages, each record would be assigned an English reference gloss intended to represent, as closely as possible, the meaning the original author attributed to the sign. This gloss would facilitate search, comparison and provisional semantic grouping across records, but it would not be treated as evidence of formal identity between signs or as a claim of full lexical equivalence across languages. At the same time, the database would preserve the label used in the original source and, where relevant, include additional glosses in other languages. All such glossing decisions would be explicitly documented, accompanied by notes on terminological choices, and remain open to revision as the database develops. Furthermore, the model would distinguish between internal variants attested within a single source and parallels identified across different sources, allowing historical variation to be represented in a structured way without forcing premature lemmatisation. The proposed database would also explicitly encode compound signs and allow their segmentation into constituent elements. Where components correspond to independently attested signs, they would be linked to separate records, enabling navigation between compounds and their component units.

Detailed source metadata would document authorship, including links to biographical records, publication or manuscript context, archival provenance, repository holdings, institutional affiliations, and bibliographic references, together with persistent links to digitised facsimiles stored in institutional or project-based repositories, precise page or folio locations, and information about whether a sign is represented pictorially, through a lexicographic written description, or written description in continuous prose, via notational systems, or through a combination of these representational forms.

Written descriptions would be preserved through several parallel transcription layers, including diplomatic renderings that maintain original orthography and abbreviations, normalised versions for computational use, English translations, and editorial commentary on damaged or ambiguous passages. For manuscript sources, palaeographic observations and graded assessments of legibility would be included in order to make the limits of the evidence transparent. Where written descriptions contain information relevant to the sign form, this would also be recorded as part of the evidential basis for subsequent phonological annotation. Visual material would be treated with equal care: the database would store full-page scans and cropped sign images, plate numbers and figure identifiers, documentation of editorial image processing, information about viewpoint and depicted body orientation, graphic style, medium, and quality ratings indicating how securely phonological parameters can be extracted from an illustration.

Where possible, each sign would be annotated for phonological structure, beginning with sign type and, where applicable, with separate parameter values for each hand. These would include handshape, palm orientation, finger orientation, location, movement, non-manual components, dominance relations, symmetry patterns, and parameter-specific confidence ratings. Because written descriptions usually do not permit all parameters to be identified with complete certainty, the database would distinguish explicitly between attested and inferred parameters. For textual sources, handshape annotation would therefore also allow broader descriptive categories, including Stokoe-style groupings derived from the manual alphabet, for example, B-type labels for flat hand configurations, in cases where the author's wording does not support secure assignment to a single fully specified handshape. To support consistency, the database would also include source-specific inventories of the descriptive terminology used by individual authors, especially for recurrent expressions referring to handshape and movement. Where visual documentation is available and sufficiently detailed, parameter values could additionally be encoded in HamNoSys, chosen here as a widely used notation system that facilitates interoperability and comparison across sign language datasets and has also been applied in the creation of cross-linguistic comparative databases (cf. Power et al., 2022). Editorial interpretations of form would remain explicitly separated from primary documentation.

Semantic, iconic, metaphorical, and motivational properties would be encoded in a parallel layer, including fields for motivation, iconic and metaphorical mapping type, iconic and metaphorical coding, degree of transparency, and risk of cross-linguistic universality. The Motivation field would distinguish between motivations explicitly stated by historical authors and later analytical inferences. Where available, illustrative material and excerpts from historical sources

would be included to document and clarify these assessments.

Because historical lexicons were often shaped by surrounding spoken languages, manual alphabet traditions, and pedagogical practices in sign language use, the proposed database would include a dedicated layer for classifying lexical origin and contact influence. Some relationships, especially those involving borrowing, semantic extension, or calquing from spoken language expressions, as well as formations incorporating elements of the manual alphabet through fingerspelling, initialisation, or mixed forms, may be identifiable already at an early stage of data entry, with explicit reference to contemporary manual alphabet tables where relevant. By contrast, possible contact relations between sign languages would be expected to emerge more gradually, as the database expands and broader comparative patterns become visible. The model would therefore allow such classifications to remain provisional and to be refined over time. Records could also be linked to estimated borrowing periods, institutional contexts that may have favoured particular formations, and graded uncertainty assessments.

All records in the proposed database would be embedded in a relational environment designed to allow users to navigate between primary sources, source-specific attestations, and analytical layers. Search and retrieval would operate along two complementary axes, semantic and formal. Semantic retrieval would rely on English reference glosses and provisional semantic groupings, allowing researchers to explore related concepts across different regions, languages, periods, and source types. This would also make it possible to record cases in which similar forms are associated with different meanings, or in which attested parallels show slight semantic differences across sources. Formal retrieval, by contrast, would allow searches based on sign type and on selected phonological parameters, such as handshape, location, movement, or orientation. By combining such criteria, the database would support the structured comparison of variants and parallels across heterogeneous materials without requiring premature claims about cognacy or lexical equivalence. In addition, records could be linked through an interpretative layer capturing shared motivational or iconic patterns, making it possible to group potentially related signs into provisional families while keeping such relationships explicitly distinct from source attestation and formal equivalence.

Chronological metadata would enable tracing documented trajectories of use and identifying periods of stability, change, or apparent discontinuity. Geographic, institutional, and sociolinguistic layers would support the analysis of circulation, pedagogical transmission, and local patterns of variation. In a future implementation, this structured format could also facilitate alignment with selected contemporary sign language resources, making it easier to compare historical attestations with modern forms. Taken together, these features would provide a structured framework for comparative diachronic research on European sign languages while preserving the uncertainty and heterogeneity inherent in archival documentation.

## 3. Identifying Sources and Compiling Historical Sign Data

Before the proposed database can be implemented, relevant historical sources must first be identified and the signs they contain systematically collected. This preparatory phase involves surveying printed dictionaries and pedagogical manuals, locating manuscript materials in archives and libraries, and assessing the types of linguistic evidence they provide, including pictorial representations, written descriptions and other notational systems.[1] To illustrate this process, we focus on the historical sign language documented in the territory of Bohemia, as a pilot case.

### 3.1 Applying the Database Model to Bohemian Sources

In this section, we therefore outline the corpus of Czech historical sources assembled for this study and describe the procedures used to extract and catalogue individual sign attestations in a form informed by the conceptual database model proposed above.

The materials documenting the sign language used in Bohemia constitute a coherent yet internally diverse body of eighteenth- and nineteenth-century printed and manuscript sources, produced primarily within institutional settings but not exclusively, and provide an instructive example of how local archival traditions can be integrated into a broader comparative framework.

Sign language used in the Czech lands was shaped primarily within the Prague Institute for the Deaf and Dumb, founded in 1786, in the broader context of the Habsburg Monarchy, and in particular close connection with the Viennese institute, with which it shared the language of

[1] This process must be accompanied by parallel historical research into the relevant educational, institutional and sociolinguistic contexts in order to interpret the sources accurately and situate the recorded signs within their broader cultural and pedagogical frameworks.

instruction, German (Okrouhlíková, 2025b). A separate Czech section was established only in 1836. This institutional history directly affects the nature of the sources considered here. Until 1839, the relevant texts were written exclusively in German; thereafter, they were also written in Czech. This also has practical consequences for data extraction, since many documents are written in Kurrent script or printed in Schwabach, which makes transcription and careful philological processing necessary prerequisites for linguistic analysis and database compilation (Okrouhlíková, 2025b). Over more than a decade of research, the conditions for working with these materials have changed significantly within a broader digital humanities context, from consulting physical copies to large-scale digitisation, the use of transcription tools, and, more recently, the cautious application of AI-assisted processing, while maintaining rigorous human supervision and philological scrutiny at all stages of analysis.

To demonstrate the applicability of the proposed model, we apply its multi-layered architecture[2] to a corpus of eighteenth- and nineteenth-century Bohemian sources. These materials present the specific types of descriptive inconsistency and fragmentary documentation that the database is designed to manage. As a test case, we draw primarily on three of the most substantial documented sources (Mücke 1834; Czech 1836; Jarisch 1851), supplemented by several shorter or less systematic records in continuous prose, and examine the sign MIRROR, which is attested across these different types of materials.

### 3.1.1 Mücke (1834)

The first dictionary in the Bohemian context is the glossary appended to Mücke's work (1834), written during his tenure as director of the Prague Institute. It comprises 243 entries, describing a total of 269 signs when compound components are segmented separately. Rather than just a list of signs, its aim was to demonstrate the principles of sign formation by providing examples of signs already in use at the Prague institute. The lexicography written descriptions vary from relatively detailed phonological specifications to mere references to semantic or iconic motivation (for details see Okrouhlíková, 2025b). In our dataset, these variations are captured through parameter-specific confidence ratings and a dedicated iconic mapping layer, ensuring that editorial interpretation of form remains separate from Mücke's original motivational descriptions.

Example

**Record ID**: MU-1834-102-01-156 **Reference gloss (EN)**: MIRROR **Source label**: Spiegel **Gloss (DE)**: Spiegel **Gloss (CZ)**: Zrcadlo **Author**: Mücke Johann (1770–1840), director of Prague Institute (1820–1840) **Year**: 1834 **Publisher**: Gottlieb Haase Söhne **Place of publication**: Prague **Source:** *Anleitung zum Unterrichte der Taubstummen in der Lautsprache nebst einigen Bemerkungen über die Geberdenzeichen der Taubstummen* **Source type**: Printed book **Page**: 102 **Section**: *Einige Haus- und Tischgeräthe* **Language**: German **Script:** Schwabach **Record type**: Lexicographic written description **Evidence type**: textual **Political entity**: Austrian Empire **Region**: Bohemia **Institutional context**: Prague Institute for the Deaf and Dumb **Digital facsimile**: Google Books

**Primary Evidence Layer**

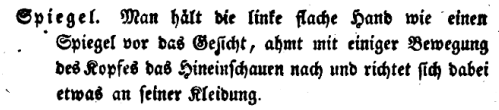

Spiegel. Man hält die linke flache Hand wie einen Spiegel vor das Gesicht, ahmt mit einiger Bewegung des Kopfes das Hineinschauen nach und richtet sich dabei etwas an seiner Kleidung.

Figure 1: Mücke (1834, p. 102), sign *Spiegel*

**Diplomatic Transcription (Latin script)**

*SPIEGEL. Man hält die linke flache Hand wie einen Spiegel vor das Gesicht, ahmt mit einiger Bewegung des Kopfes das Hineinschauen nach und richtet sich dabei etwas an seiner Kleidung.*

**English Translation**

*MIRROR. One holds the left flat hand in front of the face like a mirror, imitates, with some movement of the head, the act of looking into it, and at the same time adjusts oneself somewhat at one's clothing.*

**Structural Annotation**

**Sign type**: 2H-asym (inferred); **Dominance**: H1R (inferred)
**H1R (dominant, active)**: HS = 0; ORI = 0; LOC = torso (inferred); MV = small adjusting movement (attested)
**H2L (non-dominant, passive)**: HS = flat (attested) (B-type, subtype undetermined); LOC = face-front (attested); ORI = 0; MV = static (inferred)
**NM**: several head movements (attested). **Phonological specification:** low

**Iconicity and Motivation**

**Motivation**: holding a flat hand as a mirror, imitating looking into it and adjusting one's clothing (attested)
**Iconic mapping type**: object representation + pantomimic enactment

[2] Owing to space constraints, this paper does not reproduce full bibliographic descriptions or detailed information on individual exemplars and their holdings. References are limited to publicly accessible copies of the respective works, with complete bibliographic information provided in the reference list.

**Iconic encoding**:
– H1R performing a grooming movement, hand-as-hand; pantomimic enactment (attested)
– H2L flat hand representing the mirror surface, hand-as-object (attested)
– NM head movement representing the act of looking (attested)
**Transparency**: high
**Cross-linguistic universality risk**: high (everyday embodied action)

**Lexical origin classification**

**Lexical origin:** native iconic formation
**Contact**: not attested; possible

### 3.1.2 Czech (1836)

This book on deaf education, authored by a Bohemian native who served as catechist at the Vienna Institute for the Deaf and Dumb, represents another important source of signs used in the Austrian context. The pictorial appendices contain approximately 140 signs, though largely without explicit glosses or equivalents. Most can nevertheless be identified through cross-references in the main text, which discusses the pedagogical principles of sign formation. Because movement is not graphically indicated in the illustrations but sometimes can be inferred from the accompanying textual explanations, these records utilise the interpretative reconstruction layer. This allows us to explicitly distinguish between attested static forms and inferred motion.

Example

**Record ID**: CZ-1836-02-07-07 **Reference gloss (EN)**: MIRROR **Source label**: Spiegel **Gloss (DE)**: SPIEGEL **Gloss (CZ)**: ZRCADLO **Author**: Czech Franz Hermann (1788–1847), catechist and teacher at the Vienna Institute (1818–1841) **Year**: 1836 **Publisher**: Commission der Mechitaristen-Congregations-Buchhandlung **Place of publication**: Wien **Source:** *Versinnlichte Denk- und Sprachlehre mit Anwendung auf die Religions=und Sittenlehre und auf das Leben* **Source type**: Printed book **Page** 60, Table 2 **Section:** *Mittheilung durch Geberden* **Language**: German **Script:** Schwabach **Record type**: pictorial representation - copperplate engraving + written equivalent **Figure framing**: full body **Viewpoint**: left profile **Political entity**: Austrian Empire **Region**: Lower Austria **Institutional context**: Vienna Institute for the Deaf and Dumb **Digital facsimile**: Münchener Digitalisierungs Zentrum

**Primary Evidence Layer**



Er lege dem Taubſtummen Abbildungen einiger im Zimmer vorhandenen Gegenſtände zur Anſchauung vor, als: Stuhl, Tiſch, Fenſter, Thür, Schuh, Hut, Brille, Bohrer, Schere, Spiegel, Bürſte u. a. m. Tab. 1; weiſe zuerſt ſelbſt mit

Figure 2: Czech (1836, Tab. 2, p. 50), sign *Spiegel*

**Diplomatic Transcription (Latin script)**

*Er lege dem Taubstummen Abbildungen einiger im Zimmer vorhandenen Gegenstände zur Anschauung vor, als: Stuhl, Tisch, Fenster, Thüre, Schuh, Hut, Brille, Bohrer, Scheere, Spiegel, Bürste u. a. m. Tab. 1;*

**English Translation**

*Let him present to the deaf and dumb pictures of some objects present in the room for observation, such as: chair, table, window, door, shoe, hat, spectacles, drill, scissors, mirror, brush, etc. Plate 1.*

**Structural Annotation**

**Sign type**: 2H-asym; **Dominance**: undetermined
**H1R**: HS = flat (B-type, subtype undetermined; image unclear); ORI = palm toward body; fingers upward; LOC = face-front; MV = 0 (unspecified)
**H2L**: HS = pinch (probable, image unclear) (O-type, subtype flat O); LOC = chest-centre; ORI = palm downward and toward body; fingers toward body; MV = chest contact (unspecified).
**NM**: eye gaze toward palm (attested)
**Phonological specification:** moderate

**Iconicity and Motivation**

**Motivation**: holding a flat hand as a mirror in front of the face, tie-adjusting (inferred)
**Iconic mapping type**: object representation + pantomimic enactment
**Iconic encoding**:
– H1R flat hand representing the mirror surface, hand-as-object
– H2L hand-as-hand pinch the tie (attested)
**Transparency**: high
**Cross-linguistic universality risk**: high (everyday embodied action)

**Lexical origin classification**

**Lexical origin**: native iconic formation
**Contact**: not attested; possible.

### 3.1.3 Jarisch (1851)

Although private teacher Jarisch does not specify a geographic use, his training at the Prague Institute suggests a localised Bohemian influence. The book contains a dictionary consisting of 224 written descriptions and 11 pictorial tables with 173 illustrations. Signs are depicted through various methods: drawings in neutral space with movement indicated by arrows, varying perspectives from either the signer's or the addressee's viewpoint, and handshape changes during articulation (for details see Okrouhlíková, 2025b). These visual attributes are documented in the visual metadata layer to maintain transparency in phonological extraction.

Example

**Record ID**: JA-1851-03-13-73 **Reference gloss (EN)**: MIRROR **Source label**: Spiegel **Gloss (DE)**: SPIEGEL **Gloss (CZ)**: ZRCADLO **Author**: Jarisch Johann Anton (1818–1890), private teacher of the deaf (1847–1852) **Year**: 1851 **Publisher**: Verlag von G. Joseph Manz **Place of publication**: Regensburg **Source**: *Methode für den Unterricht der Taub=stummen in der Laut=sprache im rechnen und in der Religion* **Source type**: Printed book **Page**: Tafel III. – Fig. 13, Erklärung der Figuren der Tafel III. **Section**: appendix **Language**: German **Script**: Kurent **Record type**: pictorial representation + handwritten equivalent **Figure framing**: head and hand **Viewpoint**: left profile **Political entity**: Austrian Empire **Region**: Bohemia**,** Lower Austria **Institutional context**: 0 **Digital facsimile**: Google Books

**Primary Evidence Layer**




Figure 3: Jarisch (1851, Tab. III.), sign *Spiegel*

**Diplomatic Transcription (Latin script)**

der Spiegel

**English Translation**

the mirror

**Structural Annotation**

**Sign type:** 1H; **Dominance:** H1R
**H1R**: HS = flat (B-type, subtype: $B_o$ thumb slightly adducted); ORI = palm upward and toward body; fingers upward and away from body; LOC = face/upper chest-front (image unclear); MV = 0 (unspecified)
**NM**: eye gaze toward palm, head slightly tilted toward palm (attested)
**Phonological specification**: moderate

**Iconicity and Motivation**

**Motivation**: holding a flat hand as a mirror in front of the face
**Iconic mapping type**: object representation
**Iconic encoding**:
– H1R flat hand representing the mirror surface, hand-as-object
**Transparency**: high
**Cross-linguistic universality risk**: high (everyday embodied action)

**Lexical origin classification**

**Lexical origin**: native iconic formation
**Contact**: not attested; possible.

### 3.1.4 Fragmentary Sources

Our dataset also includes several printed texts and one manuscript, in which examples of signs occur within continuous prose, most often in chapters devoted to sign language (e.g. Gross, 1790; Staněk 1846; Beran Novopacký 1878; Kmoch 1886; Škornička 1890; Huleš 1891; Malý 1875, 1897; O posunové řeči, undated). These provide brief references to form or motivation rather than comprehensive descriptions. Such records are integrated via the fragmentary record status and linked to more robust entries through the linking system, providing sociolinguistic context while remaining faithful to the limited nature of the primary evidence.

Example

**Record ID**: MA-1897-26-47-54 **Reference gloss (EN)**: MIRROR **Source label**: zrcadlo **Gloss (DE)**: SPIEGEL **Gloss (CZ)**: ZRCADLO **Author**: Malý Karel (1846–1916), teacher in Prague Institute (1867–1916) **Year**: 1897 **Publisher**: I. L. Kober **Place of publication**: Prague **Source:** *Děti hluchoněmé, nedoslýchavé, jakož i poruchami řeči stižené ve škole obecné* **Source type**: Printed book **Page**: 26 **Section**: *A. Posunování* **Language**: Czech **Script**: Latin **Record type**: Written description **Evidence type**: textual **Political entity**: Austria Hungary **Region**: Bohemia **Institutional context**: Prague Institute for the Deaf and Dumb **Digital facsimile**: NA

**Primary Evidence Layer**

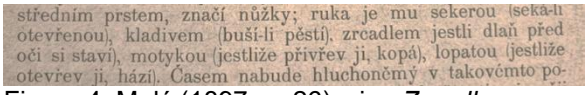

středním prstem, značí nůžky; ruka je mu sekerou (seká-li otevřenou), kladivem (buší-li pěstí), zrcadlem jestli dlaň před oči si staví), motykou (jestliže přivřev ji, kopá), lopatou (jestliže otevřev ji, hází). Časem nabude hluchoněmý v takovémto po-

Figure 4: Malý (1897, p. 26), sign *Zrcadlo*

**Diplomatic Transcription (Latin script)**

*Ruka je mu … zrcadlem jestli dlaň před oči staví*

**English Translation**

*His hand is to him … a mirror if he sets the palm before his eyes.*

**Structural Annotation**

**Sign type**: 1H; **Dominance**: undetermined
**H1:** HS = open (inferred); ORI = palm toward eyes; LOC eyes-front; MV bring hand in front of eyes
**NM:** 0
**Phonological specification**: low

**Iconicity and Motivation**

**Motivation**: holding a flat hand as a mirror in front of eyes
I**conic mapping type**: object representation
**Iconic encoding**:
– H1 open hand representing the mirror surface, hand-as-object (attested)
**Transparency**: high
**Cross-linguistic universality risk**: high (everyday embodied action)

**Lexical origin classification**

**Lexical origin**: native iconic formation
**Contact**: not attested; possible.

## 3.2 Diachronic Patterns and Interpretation

As an illustration of the analytical potential of the proposed database, we briefly demonstrate what can be reconstructed about the sign MIRROR across historical sources. The sign is highly iconic and appears to be based on direct iconic mapping, linking linguistic form to embodied experience and sensorimotor routines. This mode of forming concrete signs used by deaf signers, as well as their striking similarity, was also noted by nineteenth-century authors. We therefore extend our comparison to include lexicographic records of the sign MIRROR from French, German, Polish, and Russian lexicographic sources.[3]

The 1834 attestation corresponds closely to Mücke's (1834) description of sign formation, according to which the initial stage of a sign is rather pantomimic and only gradually, through conventionalisation and reduction, acquires the contours of a stabilised lexical sign. In the 1834 depiction, the dominant right hand, together with head movement, faithfully represents the embodied behaviour of grooming oneself in front of a mirror, while the non-dominant left hand already schematises the mental representation of the mirror as a flat surface. The Czech (1836) sign is very similar; however, it does not record movement. The mirror is represented by the right hand, while the left hand is used for grooming.

A very similar description of the sign from the German area is found in Jäger and Riecke (1832, p. 31): *"The flat hand is held in front of the face as a mirror, and the gestures of the person looking at himself in the mirror are imitated"*, although the dominance of the hands is not specified. Reich (1834, p. 56) likewise does not specify which hand performs which action; however, his description includes a slight trembling movement of the flat hand representing the mirror: *"One holds the flat hand with a slight trembling movement in front of the eyes and acts as if one were looking into it and arranging something on one's clothing."*

By contrast, Flery (1835, p. 221) from the St Petersburg Institute classifies the sign as one-handed and describes slight repeated movements of the right hand representing the mirror: *"The right hand, open and held obliquely upright, moves several times gently in front of the eyes, as a mirror."* He thus presents the Russian sign as corresponding to, yet differing from, the French sign recorded by Dégerando (1827, p. 14*): "One appears to look at oneself in one hand."*

In Jarisch (1851), the sign is also reduced to a one-handed form, indicating further formal simplification and the representation of a single mental concept. A similar representation can be observed in the dictionary of French author Pélissier (1855, planche 5). However, none of the authors depicts or describe the movement, similar to Malý (1897, p. 26) and the Polish Hollak and Jagodziński (1879, p. 102), who describe only the position of the hand in front of the eyes.[4]




Figure 5: Pélissier (1856, Pl. 5), sign Miroir

By linking these historical attestations to modern Czech Sign Language, we can observe that the

[3] For these regions, historical links between institutions can be documented. The French manual method of deaf education was adopted at the Vienna Institute and was subsequently transmitted to the institutes in Prague, Warsaw, and St Petersburg. Although German schools followed different pedagogical methods, they were geographically and linguistically close to the Austrian schools and influenced them as well.

[4] The sign without hand movement is also recorded in Suppalla's database in dictionaries of American Sign Language from 1923.

sign for MIRROR is documented in printed dictionaries (Gabrielová et al., 1988; Potměšil, 2005) and online resources such as Spread the Sign and Dictio in an identical one-handed sign. The dominant flat hand, palm oriented to the left, performs a twisting movement at the wrist on the right cheek. However, if we broaden the comparison to the European signs currently available in Spread the Sign, their forms appear to be practically identical.

Given the high degree of iconicity, these similarities cannot be treated as reliable evidence of cross-linguistic relatedness, since similar forms may arise independently through iconic motivation. At the same time, the attestations suggest a broader developmental tendency, although clearly not a uniform or linear one. Across the sources, the sign appears to develop from more pantomimic enactments of looking at oneself and adjusting one's clothing in a mirror, often involving both hands, towards simpler forms in which the mirror is represented as an object by a single moving hand, while the imitation of actual bodily actions gradually disappears. In present-day forms, the sign is not articulated directly in front of the face, but rather near the right cheek. This overall development corresponds to the historical changes in sign form already documented by Frishberg (1975). Taken together, these attestations are compatible with a tendency towards more schematic object-based representations of MIRROR, but the surviving records are too sparse and uneven to justify a strictly linear reconstruction of formal change.

## 4. Conclusion

This paper has argued that diachronic research on sign languages requires not only the identification of archival sources but also a systematic framework for structuring them. Historical sources provide valuable lexical attestations, yet their documentation practices are heterogeneous, often incomplete, and theoretically inconsistent. Sign descriptions vary in precision, in the representation of individual formational parameters, in terminology, and in the authors' underlying assumptions. Without a shared formal model, cross source and diachronic comparison remains methodologically fragile.

The proposed database model addresses these challenges through a multi layered architecture that clearly separates primary documentation from editorial interpretation. It distinguishes between attested and inferred parameters, applies graded confidence levels, and records structural, iconic, and metaphorical aspects of sign formation in parallel analytical layers. Detailed source metadata enables tracing the origin of each record and explicitly representing the limits of the available evidence. The aim is not to eliminate historical uncertainty, but to encode it in a structured and transparent way, while allowing formally distinct but potentially related attestations to be compared without presupposing a single, clearly reconstructable trajectory of change or a fixed type of relationship between signs.

The pilot application to nineteenth century Bohemian sources demonstrates that even fragmentary historical data can be represented within a unified formal framework that supports structured search, comparison, and further analysis. It also shows that historically related sign types, including iconically and metaphorically motivated forms, may be analysed together while preserving uncertainty about the precise nature of their relationship. In this way, the model is designed not to resolve historical ambiguity prematurely, but to make it explicit and analytically tractable. The framework is intended to support broader comparative and diachronic research across European Sign Languages.

## 5. Acknowledgments

This publication was supported by the Cooperatio Program provided by Charles University, research area Linguistics, implemented at the Faculty of Arts of Charles University.

## 6. Bibliographical References

Arnaud, S. (2019). From Gesture to Sign. *Sign language studies* 20(1). 41-82. 10.1353/sls.2019.0014

Beran Novopacký J. (1878). O vzdělávání hluchoněmých. *Osvěta: Listy pro rozhled v umění, vědě a politice*, VIII(8):588–603.

Bonnal-Vergès, F. (2005). *Sémiogenèse de la langue des signes française: Étude critique des signes de la langue des signes française attestés sur support papier depuis le XVIIIe siècle et nouvelles perspectives de dictionnaires*. Toulouse: Université Toulouse, Thèse de doctorat.

Broadwell, G. A., García Guzmán, M., Lillehaugen, B. D., Lopez, F. H., Plumb, M. H. and Zarafonetis, M. (2020). Ticha: Collaboration with Indigenous communities to build digital resources on Zapotec language and history. *Digital Humanities Quarterly*, 14(4).

Brouland, J. (1855). *Langage mimique. Spécimen d'un dictionnaire des signes*. Paris: Imprimerie de l'Institution Impériale des Sourds-Muets.

Czech, F. H. (1836). *Versinnlichte Denk- und Sprachlehre mit Anwendung auf die Religions=und Sittenlehre und auf das Leben*. Wien: Gedruckt und in Commission der Mechitaristen-Congregations-Buchhandlung.

Dégerando, J. M. (1827). *De l'éducation des sourds-muets de naissance*. Paris: Chez Méquignon l'aîné père.

Dehouck, M., François, A., Kalyan, S., Pastor, M. and Kletz, D. (2023). EvoSem: A database of

polysemous cognate sets. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pp. 66–75, Singapore. Association for Computational Linguistics. 10.18653/v1/2023.lchange-1.7

Delaporte, Y. (2007). *Dictionnaire étymologique et historique de la langue des signes française. Origine et évolution de 1200 signes*. Édition du Fox.

Fischer, R. (2010a). Historische Gebärdensprachlexikographie bei Hieronymus A. Jarisch (1851) – Teil I. *Das Zeichen*, 24(84):8–19.

Fischer, R. (2010b). Historische Gebärdensprachlexikographie bei Hieronymus A. Jarisch (1851) – Teil II. *Das Zeichen*, 24(85):190–205.

Flery, V. (1835). *Glukho-nemye, razsmatrivaemye v otnoshenii k ikh sostoyaniyu i k sposobam obrazovaniya samym svoystvennym ikh prirode*. Sankt-Peterburg: V tipografii A. Plyushara.

Frishberg, N. (1975). Arbitrariness and Iconicity: Historical Change in American Sign Language. *Language*, 51(3):697–719. https://doi.org/10.2307/412894

Frishberg, N. (1979). Historical change: from iconicity to arbitrary. In E. Klima and U. Bellugi et al., *The Signs of Language*, pp. 67–83. Cambridge, London: Harvard University Press.

Gabrielová, D., Paur, J. and Zeman, J. (1988). *Slovník znakové řeči*. Praha: Horizont.

Gross, J. A. (1790). Verfassung des Taubstummeninstituts zu Prags beschrieben von Joseph Anton Groß in einem Briefe an seinen Freund. *Kalender für Aufseher, Katecheten und Lehrer der Nationalschulen im Königreiche Böhmen auf das Jahr 1790*, pp. 83–106.

Hollak, J. and Jagodziński, T. (2011). *Słownik mimiczny dla głuchoniemych i osób z nimi styczność mających*. Styczność mających. (1879). Łódź: Polski Związek Głuchych.

Huleš, K. (1891). *O vychovávání a vyučování hluchoněmých*. České Budějovice: J. Přibyl.

Jäger, V. A. and Riecke, G. A. (1832). *Anleitung zum Unterrichte taubstummer Kinder in der Sprache und den andern Schullehrgegenständen, nebst Vorlegeblättern, einer Bildersammlung und einem Lese- und Wörterbuche*. Stuttgart: J. C. Löflund und Sohn.

Jarisch, H. A. (1851). *Methode für den Unterricht der Taub=stummen in der Laut=sprache im rechnen und in der Religion*. Regensburg: Verlag von G. Joseph Manz.

Kmoch, K. (1886). *Příspěvky k dějinám pražského soukromého ústavu pro vzdělání hluchoněmých na oslavu 100–letého jeho trvání*. Praha: Pražský soukromý ústav pro hluchoněmé.

Lambert, L.-M. (1865). *Le langage de la physionomie et du geste*. Paris: Jacques Lecoffre.

Linde-Usiekniewicz, J. and Łozińska, S. (2023). In quest of influences of Polish language dictionaries on the oldest Polish Sign Language dictionary. *International Journal of Lexicography*, XX:1–17. https://doi.org/10.1093/ijl/ecad015

Malý, K. (1875). O hluchoněmých. *Beseda učitelská*, VII(4):39–43.

Malý, K. (1897). *Děti hluchoněmé, nedoslýchavé, jakož i poruchami řeči stižené ve škole obecné*. Praha: I. L. Kober.

Mücke, J. (1834). *Anleitung zum Unterrichte der Taubstummen in der Lautsprache nebst einigen Bemerkungen über die Geberdenzeichen der Taubstummen*. Prag: Gottlieb Haase Söhne.

Nyst, V. and Schüller, A. (2024). Mother left, Father right: Artificial signs and diachronic change in sign language dialects in Belgium and the Netherlands. *Diachronica*, 41(2):251–298. https://doi.org/10.1075/dia.21052.nys

O posunové řeči. (n.d.). Archiv Střední školy, základní školy a mateřské školy pro sluchově postižené, Praha, Holečkova 4.

Okrouhlíková, L. (2025a). The emergence of Czech Sign Language. *Sign Language & Linguistics*, online first:1–57. https://doi.org/10.1075/sll.24006.okr

Okrouhlíková, L. (2025b). First Dictionaries of Czech Sign Language. *International Journal of Lexicography*, 38(4):365–396. https://doi.org/10.1093/ijl/ecad027

Pélissier, P. (1856). *Iconographie des signes faisant partie de l'enseignement primaire des sourds-muets*. Paris: Imprimerie et librairie de Paul Dupont.

Potměšil, M. et al. (2005). *Všeobecný slovník českého znakového jazyka O–Ž*. Praha: Fortuna.

Power, J. M. (2022). Historical Linguistics of Sign Languages: Progress and Problems. *Frontiers in Psychology*, 13:818753:1–17. https://doi.org/10.3389/fpsyg.2022.818753

Power, J. M., Quinto-Pozos, D. and Law, D. (2022). Signed Language Transcription and the Creation of a Cross-linguistic Comparative Database. In *Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages*, pp. 173–180. Marseille, France. European Language Resources Association (ELRA).

Power, J. M., Quinto-Pozos, D., and Law, D. (2025). Advances in the Historical Linguistics of Signed Languages. *Language and Linguistics Compass*, 19(6). https://doi.org/10.1111/lnc3.70026

Reagan, T. (2021). Historical Linguistics and the Case for Sign Language Families. *Sign Language Studies*, 21(4):427–454. https://doi.org/10.1353/sls.2021.0006

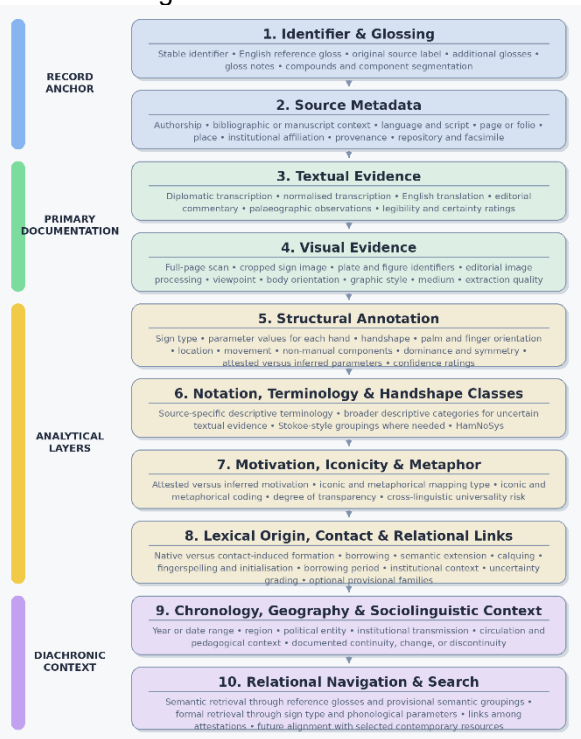
Regen, A. (2010). *Historische Gebärdensprachen im Vergleich (unter Verwendung ausgewählter Wörterbuch-Einträge)*. Hamburg: Universität

Hamburg, Magisterarbeit.
Reich, C. G. (1834). *Der erste Unterricht des Taubstummen mit angefügten Declinations-Conjugationstabellen und einer Zeitafel*. Leipzig: Verlag von Leopold Voß.
Rzymski, C., Tresoldi, T. and Greenhill, S. J. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. Scientific Data, 7(1):13. doi: 10.1038/s41597-019-0341-x.
Škornička, K. (1890). *Výchova a vyučování hluchoněmých*. Ivančice: J. Vávra.
Staněk, V. (1846). O hluchoněmotě. *Časopis českého museum*, XX(6):748–769.
Styczność mających. (1879). Łódź: Polski Związek Głuchych.
Supalla, T. (2001). Making historical sign language materials accessible: A prototype database of early ASL. *Sign Language & Linguistics*, 4(1–2):285–297.
Villwock, A. (2012). Monastische Gebärdensprachen und Gebärdensprachanwendung im Kloster. *Das Zeichen*, 91(12):266–282.
Woodward, J. (1976). Signs of change: Historical variation in American Sign Language. *Sign Language Studies*, 10:81–94.
Woodward, J. (1978). Historical bases of American Sign Language. In P. Siple (Ed.), *Understanding language through sign language research*, pp. 333–348. New York: Academic Press.
Woodward, J. (2011). Some Observations on Research Methodology in Lexicostatistical Studies of Sign Languages. In G. Mathur and D. J. Napoli (Eds.), *Deaf around the world: the impact of language*, pp. 38–53. New York: Oxford University Press.

## 7. Language Resource References

CLICS³. (n.d.). *Database of Cross-Linguistic Colexifications*. Available at https://clics.clld.org/ Accessed 21 March 2026.
Dictio. n.d. *Dictio: International sign language dictionary platform.* Available at https://www.dictio.info/ Accessed 14 February 2026.
European Sign Language Centre. n.d. *SpreadTheSign: multilingual online sign language dictionary.* Available at https://www.spreadthesign.com/ Accessed 14 February 2026.
François, A., Dehouck, M., Henke, K., and Kalyan, S. (n.d.). *Sem: A database of dialexification across language families*. Online database. Lattice & HéLiCéO, CNRS — ENS-PSL, Paris. Available at https://evosem.huma-num.fr/ Accessed 21 March 2026.
Lillehaugen, B. D., Broadwell, G. A., Oudijk, M. R., Allen, L., Plumb, M., and Zarafonetis, M. (2016). *Ticha: a digital text explorer for Colonial Zapotec*, first edition. Available at https://ticha.haverford.edu/ Accessed 21 March 2026.
Sign Language Research Lab, Georgetown University. n.d. *Historical Sign Language Database (HSLDB).* Georgetown University. Available at https://hsldb.georgetown.edu/ Accessed 14 February 2026.
Těšínská Lomičková and R., Těšínský, J. 2009. *SIGNA: databáze klášterních znaků*. Available at https://signa.kub.cz/ Accessed 21 March 2026.

**Appendix A.** Architecture of the Proposed Diachronic Sign Database



# A Video-Based Reverse Dictionary for Sign Language Using Gesture Similarity

**Batyrbek Orazumbekov, Daniyal Bayanov, Aruzhan Kaltay, Anara Sandygulova**
School of Engineering and Digital Sciences, Nazarbayev University, Astana, Kazakhstan
{batyrbek.orazumbekov, daniyal.bayanov, aruzhan.kaltay, anara.sandygulova}@nu.edu.kz

**Abstract**

Sign language recognition systems are usually modeled as classification systems that map gesture videos to pre-defined glosses. But these systems do not allow similarity searches, where a user can search for similar gestures without knowing the corresponding gloss. This paper presents a pose-based video-to-video search framework for isolated signs, which acts as a reverse gesture dictionary. The system employs keypoints on the skeletal structure instead of RGB images. Two architectures are proposed for modeling temporal information: an encoder with self-attention in a Transformer architecture and a Spatial-Temporal Graph Convolutional Network (ST-GCN). The embedding space is optimized using metric learning objectives, including supervised contrastive learning and ArcFace angular margin loss. The performance of the retrieval system is evaluated on the WLASL dataset using ranking metrics like Recall@K and mean Average Precision (mAP). Experiments reveal that the temporal modeling using the Transformer architecture is an improvement over the graph-based modeling approach in the low-shot learning scenario. The attention-based temporal pooling approach further enhances the ranking quality, with the best-performing model achieving an mAP of 0.237 on the WLASL validation set. Cross-dataset evaluation on a 226-label AUTSL dataset reveals non-trivial generalization performance on the unseen dataset, despite training only on the WLASL dataset.

**Keywords:** gesture retrieval, sign language technology, metric learning, pose-based representation, cross-dataset generalization

## 1. Introduction

Sign languages are visual languages that are used by Deaf and hard-of-hearing people. They involve the use of hand and body movements to create meaning. Recent breakthroughs in computer vision and deep learning have been used extensively in the development of automatic sign language recognition systems. Most of the existing literature is based on classification tasks, where a video is classified into a predefined label (Li et al., 2020). Although these systems work well for translation and transcription tasks, they are not intended for similarity search.

In many real-world situations, a user may observe a gesture without knowing its corresponding gloss. In such cases, text-based queries are not possible. This motivates the concept of a reverse gesture dictionary: a system that takes a gesture video as input and retrieves visually similar gestures from a database (Hassan et al., 2025). Unlike classification, which produces discrete predictions, retrieval requires learning a continuous embedding space where similar gestures are located near each other.

Although cross-modal retrieval between sign video and text has been explored (Duarte et al., 2022; Cheng et al., 2023), purely visual video-to-video retrieval remains relatively underexplored. Many current approaches rely on textual supervision or gloss alignment (Duarte et al., 2022). Learning similarity directly from motion therefore requires robust visual representations and metric learning strategies capable of structuring embedding geometry.

The central problem addressed in this work is the development of a system that retrieves visually similar isolated signs given a query video. Unlike classification models, which optimize label prediction, retrieval models must learn embeddings where intra-class samples are compact and inter-class samples are well separated (Ghojogh et al., 2022; Cakir et al., 2019). This task is challenging due to variability across signers, motion speed differences, camera perspectives, and limited samples per gloss. In datasets such as WLASL, only three to four instances per class are available on average, creating a low-shot learning scenario (Li et al., 2023; Musgrave et al., 2020).

To address these challenges, this work proposes a pose-based gesture similarity learning framework. Instead of using raw RGB frames, structured skeletal keypoints are extracted to reduce background noise and appearance variability (Parian-Scherb et al., 2024). Temporal modeling is performed to encode motion dynamics into fixed-dimensional embeddings (Ramanathan et al., 2015). Two architectures are investigated: a Transformer-based model using self-attention to capture global temporal dependencies, and a Spatial-Temporal Graph Convolutional Network (ST-GCN) that models skeletal structure as a graph.

The embedding space is optimized using metric learning objectives, including supervised con-



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 398–407
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

trastive loss (Khosla et al., 2020) and ArcFace angular margin loss (Deng et al., 2019). Retrieval performance is evaluated using ranking-based metrics such as Recall@K and mean Average Precision (mAP), which align with the objectives of similarity search (Cakir et al., 2019).

The contributions of this work are threefold: (1) a pose-based video-to-video retrieval framework for isolated signs, (2) a comparative analysis of Transformer and ST-GCN temporal modeling under low-shot conditions, and (3) an empirical study of metric learning objectives for gesture similarity. The results demonstrate that purely visual gesture retrieval is feasible and effective without reliance on textual supervision, providing a foundation for reverse sign language dictionary systems and future scalable retrieval research.

In addition to the primary WLASL experiments, this research also assesses cross-dataset generalization by applying the best model to an AUTSL subset with 226 unseen labels, showing the effectiveness of the learned embedding on a different dataset (Sincan and Keles, 2020).

## 2. Literature Review

### 2.1. From Sign Recognition to Retrieval-Oriented Systems

Sign language technology has evolved significantly over the past decade. Early systems primarily addressed gesture recognition as a classification problem, where an input video is mapped to a predefined gloss label. Large-scale datasets such as WLASL (Li et al., 2020) enabled benchmarking of deep learning architectures for word-level recognition. These systems demonstrated strong performance in transcription and translation tasks but were inherently limited to closed-set classification.

However, recognition-based systems do not support similarity-driven exploration. In practical scenarios, users may observe a gesture without knowing its gloss. This limitation motivated the development of reverse sign language dictionaries, where users provide a video query and retrieve visually similar gestures. Such systems shift the problem from discrete prediction to continuous similarity modeling.

Earlier work on sign language retrieval predates deep learning-based embedding approaches. For example, the Dicta-Sign project (Efthimiou et al., 2012) introduced a search-by-example interface that allowed users to perform a sign using a Kinect device and retrieve corresponding or closest matching signs from a multilingual lexicon. While enabling early sign lookup based on motion input, these approaches relied on handcrafted features and did not incorporate learned embedding representations for scalable similarity modeling.

Recent advances reflect this transition toward retrieval-based paradigms. Duarte et al. (Duarte et al., 2022) introduced a cross-modal retrieval system that aligns visual sign embeddings with textual queries through joint embedding spaces. Their approach enabled bidirectional search between sign and spoken language. Similarly, Cheng et al. (Cheng et al., 2023) proposed Cross-lingual Contrastive Learning (CiCo), integrating sign and spoken languages into a shared vector space. CiCo demonstrated strong performance on multilingual sign datasets.

More recent work has further advanced cross-modal alignment between sign language and text. Jiang et al. (Jiang et al., 2024) proposed SignCLIP, a contrastive learning framework inspired by CLIP that learns a shared embedding space between sign videos and textual descriptions. This enables efficient text-to-video and video-to-text retrieval. However, such approaches rely on linguistic supervision and primarily optimize for semantic similarity rather than visual similarity between sign executions.

While these systems advanced retrieval capabilities, they remain largely dependent on textual supervision. Purely visual video-to-video retrieval, where similarity is learned directly from motion rather than gloss alignment, remains underexplored. Addressing this gap requires robust motion representation and metric learning techniques specifically optimized for ranking rather than classification.

### 2.2. Motion Representation and Temporal Modeling

Accurate gesture retrieval depends on how motion is represented and encoded. Most modern systems rely on landmark-based pose extraction, using frameworks such as MediaPipe or OpenPose to capture upper-body and hand keypoints. By focusing on skeletal trajectories instead of raw RGB frames, these representations reduce sensitivity to background noise, clothing variation, and lighting conditions.

Pose-based retrieval has shown promising results. Parian-Scherb et al. (Parian-Scherb et al., 2024) demonstrated that normalized keypoint trajectories combined with attention mechanisms can effectively model gesture similarity across signers. Their findings highlight the importance of structured motion encoding over purely appearance-based features.

Temporal modeling further enhances representation quality. Ramanathan et al. (Ramanathan et al., 2015) introduced temporal embeddings that preserve motion order and dynamic transitions. Their

work demonstrated that modeling temporal evolution rather than static posture improves discrimination between visually similar actions.

More recent architectures extend temporal modeling through sequence learning structures such as Transformers and graph-based networks. Self-attention mechanisms enable modeling of long-range dependencies across time, allowing systems to capture subtle differences in motion trajectories. Graph-based methods, including Spatial-Temporal Graph Convolutional Networks (ST-GCN), incorporate skeletal structure explicitly by representing joints as nodes and bones as edges. These approaches embed structural priors into motion representation.

The combination of normalized keypoint trajectories and advanced temporal modeling forms a strong foundation for gesture similarity learning.

### 2.3. Metric Learning for Gesture Retrieval

While representation learning encodes motion patterns, the organization of embeddings in feature space determines retrieval performance. Metric learning provides a principled framework for structuring this embedding space.

Unlike classification objectives that optimize discrete label prediction, metric learning focuses on relative similarity relationships. Ghojogh et al. (Ghojogh et al., 2022) categorize metric learning methods into spectral, probabilistic, and deep approaches, unified by the objective of minimizing intra-class distances and maximizing inter-class distances.

Deep metric learning extends this idea using neural networks trained with specialized loss functions. Contrastive and triplet losses explicitly enforce distance constraints between positive and negative pairs. Cakir et al. (Cakir et al., 2019) further proposed Deep Metric Learning to Rank, directly optimizing ranking-based objectives aligned with metrics such as Recall@K and mAP. Such ranking-aware losses are particularly relevant for retrieval tasks.

Margin-based objectives such as ArcFace introduce angular separation between classes, improving embedding discriminability. Supervised contrastive learning (Khosla et al., 2020) enhances intra-class compactness by leveraging multiple positive samples within each batch. These objectives are especially important in low-shot settings where limited examples per class make generalization challenging (Li et al., 2023).

However, Musgrave et al. (Musgrave et al., 2020) emphasize that improvements in metric learning often depend heavily on fair experimental settings. Their study demonstrates that careful hyperparameter tuning and standardized evaluation are critical for meaningful comparisons. This insight underscores the importance of methodological rigor in gesture retrieval research.

### 2.4. Robustness and Domain Generalization

Gesture retrieval systems are required to handle large intra-class variability due to differences in signers, speed of performance, camera orientation, and environments. The previous study on adversarial metric learning (Duan et al., 2020) and domain-invariant representations in person re-identification (Zahra et al., 2023) emphasizes the significance of robustness to variations. In the context of gesture retrieval, it is crucial to have domain-invariant representations that are robust to diverse visual conditions, and attention and structured normalization can be used to enhance generalization.

### 2.5. Research Gaps

Although some progress has been made, there are still some gaps that need to be filled. Video-to-video search purely based on visual information is still a challenge because most current models are based on text or multi-modal alignment. Low-shot learning is also a challenge because the number of samples per gloss is very low. In addition, the evaluation metric in sign language studies is more focused on accuracy than Recall@K and mAP.

### 2.6. Positioning of the Present Work

This research aims to address the less-explored task of video-to-video similarity learning without textual supervision, specifically in pose-based motion modeling and retrieval-driven metric learning. The proposed system integrates normalized skeletal features, a Transformer, and ST-GCN temporal models, along with ranking-driven metric learning. This research can help develop reverse sign language dictionary systems and provide an understanding of cross-dataset gesture retrieval.

## 3. Datasets

### 3.1. Overview

The objective of this work is to develop a video-to-video gesture retrieval system that learns similarity relationships between isolated signs. Unlike classification systems that assign videos to predefined gloss labels, retrieval systems must organize gestures in a continuous embedding space according to visual similarity (Duarte et al., 2022; Cheng et al., 2023). This distinction requires datasets containing multiple instances per class to support metric learning objectives.

Deep metric learning relies on constructing meaningful positive and negative pairs within each class (Ghojogh et al., 2022; Musgrave et al., 2020). Therefore, the dataset must contain more than one sample per gesture. Based on these requirements, the Word-Level American Sign Language (WLASL) dataset was selected as the primary dataset for training and evaluation (Li et al., 2020). In addition, an internal dataset of Russian and Kazakh isolated signs was examined for potential generalization experiments.

### 3.2. WLASL Dataset

The WLASL dataset consists of 2,000 isolated glosses with an average of three to four video instances per gloss (Li et al., 2020). The dataset contains recordings from multiple signers, introducing variability in motion speed, execution style, and recording conditions. This diversity makes it suitable for evaluating robustness in gesture retrieval.

After preprocessing and removal of unusable samples, the dataset contains 11,980 videos divided into official training, validation, and test splits (8,313 / 2,253 / 1,414). Each video represents a single isolated sign lasting between one and five seconds.

Although the number of samples per class is relatively small, this setting reflects a realistic low-shot learning scenario. Few-shot metric learning literature highlights the difficulty of structuring embedding spaces under limited class repetition (Li et al., 2023). Nonetheless, having at least two samples per gloss satisfies the minimal requirement for supervised contrastive and margin-based losses (Ghojogh et al., 2022; Cakir et al., 2019).

The choice of WLASL is motivated by three factors. First, it provides multiple instances per class, which is essential for metric learning. Second, signer diversity supports learning invariant representations, similar to challenges studied in person re-identification research (Zahra et al., 2023). Third, it contains isolated signs, aligning directly with the objective of reverse dictionary-style retrieval systems (Hassan et al., 2025). While sentence-level datasets exist, they introduce additional segmentation and alignment challenges beyond the scope of isolated gesture retrieval (Duarte et al., 2022; Martins, 2024).

### 3.3. Preprocessing Consistency

To ensure representation consistency, identical preprocessing steps were applied to all datasets. Each video was converted into pose and hand keypoint sequences, temporally normalized to a fixed frame length, and spatially normalized.

Learning embeddings from normalized temporal sequences aligns with prior work on temporal representation learning (Ramanathan et al., 2015). Keypoint-based modeling also follows established gesture retrieval approaches that rely on structured motion representations (Parian-Scherb et al., 2024). Maintaining a consistent feature space is essential for stable metric learning behavior (Ghojogh et al., 2022).

### 3.4. Summary

The WLASL dataset was selected as the primary dataset due to its suitability for supervised metric learning and isolated gesture retrieval. Despite its low-shot nature, it provides sufficient structure for embedding-based similarity modeling. The internal Russian and Kazakh dataset cannot support supervised training but offers opportunities for future cross-dataset evaluation. Together, these datasets provide a solid foundation for studying pose-based gesture similarity learning under realistic data constraints (Li et al., 2023). To better analyze the generalization ability across datasets, a subset of the AUTSL dataset with 226 distinct labels was created in a controlled gallery-query split, where four training videos per label and one test video per label were chosen for analysis. This subset was then passed through the same pose pipeline to be compatible with the trained WLASL model.

## 4. Methods

### 4.1. Overview of the Proposed Framework

The objective of the research is to discover an embedding space for gestures where similar isolated signs are placed near each other while dissimilar gestures are far apart based on their visual characteristics. In the proposed work, the representation is based on poses with the incorporation of temporal information and metric learning (Ghojogh et al., 2022; Musgrave et al., 2020).

The proposed system has several stages: video input, keypoint extraction, preprocessing and normalization, feature building, training the embedding network, and evaluation with the retrieval task. The key difference between the proposed work and other gesture recognition systems is that the proposed system is based on optimizing the structure of the embedding space to enable the retrieval task based on ranking (Cakir et al., 2019; Duarte et al., 2022).

### 4.2. Pose and Hand Keypoint Extraction

Instead of working with RGB values, the system utilizes the pose-based representation, which helps eliminate background noise and appearance variations. In the system, 75 keypoints are detected

per frame, with 33 keypoints detected for the upper body and 21 keypoints detected for each hand. Each of these keypoints is represented by three-dimensional coordinates, resulting in 225 features per frame.

This representation only considers movement and pose, without any consideration of texture and color. The structured skeletal representation helps the system cope with changes in illumination, background, and clothing worn by the signer, which is consistent with gesture retrieval approaches that rely on structured motion representations (Parian-Scherb et al., 2024).

### 4.3. Temporal and Spatial Normalization

For the videos, several preprocessing steps are taken. Missing landmark points are addressed by interpolating the valid frames. Spatial normalization is carried out by normalizing the coordinates with respect to the midpoint of the shoulders and the length of the torso. Temporal normalization of the videos is carried out by fixing the length of all videos at 100 frames.

Since the videos are of different lengths, it is important to have the same length so that the videos can be batch-trained and the embeddings can be extracted. Learning embeddings from temporally aligned sequences follows prior work on temporal embedding learning for video analysis (Ramanathan et al., 2015).

### 4.4. Feature Representation

Several feature components are built to represent static posture as well as dynamic motion.

The primary feature representation is a set of keypoint coordinates after normalization. To explicitly represent motion, first-order temporal derivatives of the keypoint coordinates, i.e., velocity features, are computed by subtracting consecutive frames. The velocity features help in differentiating gestures with similar static posture but different motion trajectories.

For graph-based experiments, bone vectors are computed by subtracting parent joint coordinates from child joint coordinates based on skeletal structure. Additionally, bone velocities are computed to represent relative joint motion, and this extended feature representation is beneficial for exploiting structural information inherent in the human skeleton using spatial-temporal graph convolutional networks. Such structured modeling aligns with graph-based gesture modeling approaches (Parian-Scherb et al., 2024).

### 4.5. Embedding Architectures

Two types of architectures are considered, namely, the transformer-based approach to modeling temporality and the Spatial-Temporal Graph Convolutional Networks.

For the former, a higher-dimensional embedding of the input feature space is performed, followed by a positional encoding to preserve temporality. Multiple self-attention layers are employed to learn long-range temporality, enabling the model to learn global motion patterns. Rather than using a mean pooling strategy, attention-based pooling is adopted to compute a weighted aggregation of frame-level representations. This helps to focus on certain segments of temporality. Transformer-based temporal embedding learning builds upon prior work on learning temporal embeddings for video representation (Ramanathan et al., 2015).

For the latter, a skeleton is modeled as a graph, where joints are represented as nodes, and edges are formed by bones. Spatial graph convolutions are employed to capture joint dependencies, and temporally, convolutions are adopted to capture motion evolution. Two-stream variants are also considered to separately process pose and hand information before feature fusion. Graph-based modeling has been widely used for structured motion representation in visual analysis tasks (Zahra et al., 2023).

### 4.6. Metric Learning Objectives

Several metric learning objectives are considered to organize the embedding space (Ghojogh et al., 2022; Musgrave et al., 2020). In supervised contrastive loss, embeddings of different classes are forced to be far apart, and embeddings of the same class are encouraged to be close (Khosla et al., 2020). ArcFace adds a penalty term for angular margin, which helps to improve separation for classification-based training (Deng et al., 2019). A combination of both objectives is also considered. These objectives directly affect the embedding space, which is essential for retrieval (Cakir et al., 2019; Duan et al., 2020).

### 4.7. Retrieval Evaluation Setup

After training, all test set samples are mapped to a feature embedding space. The similarity between two gestures is calculated using cosine similarity. For each query, its k-nearest neighbors are retrieved and ranked according to their similarity score.

The performance of the system is measured using various ranking-based metrics, such as Recall@K and mean Average Precision (mAP). These metrics are based on the system's ability to re-

trieve relevant gestures within its top-ranked results, which is in line with the purpose of a reverse gesture dictionary and ranking-based metric learning approaches (Cakir et al., 2019; Duarte et al., 2022).

For the cross-dataset experiment, the same process of retrieval was done on the AUTSL dataset by representing all the gallery and query sequences in the WLASL-trained model, allowing the assessment of how well the embedding space transfers to a completely unseen sign vocabulary.

### 4.8. Summary

This chapter has provided the complete methodological framework for pose-based gesture similarity learning. The system combines keypoint-based representation, temporal representation using the Transformer and ST-GCN architectures, and metric learning objectives optimized for ranking performance (Ghojogh et al., 2022). The experimental results and the effect of the architecture and loss function choices will be provided in the next chapter.

## 5. Results

### 5.1. Experimental Overview

This chapter discusses the experimental evaluation of the proposed gesture retrieval framework. All architectures were trained using the WLASL training set (Li et al., 2020), and the preprocessing technique explained in the Datasets chapter was utilized. It is worth noting that only gesture classes with at least two instances were included in the experiment to facilitate metric learning objectives (Ghojogh et al., 2022; Musgrave et al., 2020).

The experiment was conducted using ranking-based evaluation metrics. For a given query in the validation set, the cosine similarity was calculated between the query embedding and the embeddings in the gallery set. The obtained results were then ranked based on the similarity score. The evaluation metrics employed in the experiment included Recall@1, Recall@5, Recall@10, Recall@50, and mean Average Precision (mAP), which are standard in retrieval-oriented metric learning (Cakir et al., 2019; Duarte et al., 2022). Table 1 presents the performance of all evaluated architectures.

### 5.2. Quantitative Results

The results demonstrate clear performance differences across architectural and loss configurations.

Table 1: Retrieval Performance Comparison on the WLASL Validation Set. The table reports Recall@K and mean Average Precision (mAP) for different architectures and metric learning configurations. The best results are highlighted in bold.

| Model | Architecture | Loss Function | R@1 | R@5 | R@10 | R@50 | mAP |
|---|---|---|---|---|---|---|---|
| M1 | Transformer | SupCon | 0.178 | 0.455 | 0.576 | 0.771 | 0.212 |
| M2 | Two-Stream ST-GCN | ProxyNCA | 0.105 | 0.311 | 0.424 | 0.704 | 0.098 |
| M3 | Two-Stream ST-GCN | ArcFace | 0.160 | 0.396 | 0.510 | 0.756 | 0.160 |
| M4 | Two-Stream ST-GCN | ArcFace + SupCon | 0.177 | 0.446 | 0.559 | 0.769 | 0.192 |
| M5 | Transformer (Attention Pooling) | SupCon | **0.183** | 0.433 | 0.554 | 0.732 | **0.237** |

### 5.3. Transformer vs ST-GCN Architectures

In all cases, the performance of the Transformer-based models surpassed that of the ST-GCN-based models, as measured by mean Average Precision. The baseline Transformer, even with supervised contrastive loss (Khosla et al., 2020), achieved a value of 0.212, which was already higher than that of ST-GCN with ProxyNCA and ArcFace (Deng et al., 2019).

Despite the explicit modeling of skeletal structure using graph convolutions, ST-GCN failed to achieve a higher retrieval accuracy than the Transformer-based model. This may indicate that global temporal modeling using self-attention is a more powerful approach for modeling fine-grained motion similarity in isolated sign sequences, consistent with temporal embedding learning principles (Ramanathan et al., 2015). The best ST-GCN architecture, ArcFace + SupCon, achieved a value of 0.192, still lower than that of the baseline Transformer.

### 5.4. Impact of Loss Functions

The impact of the selection of the loss function is substantial. The performance of the ST-GCN model with ProxyNCA was the lowest among all the configurations. Switching the ProxyNCA loss function to the ArcFace loss function (Deng et al., 2019) boosted the performance. This shows the importance of angular margin-based objectives, as they improve the inter-class distance in the embedding space (Ghojogh et al., 2022). Further performance improvements were observed when the supervised contrastive loss function (Khosla et al., 2020) was added to the ArcFace loss function. This shows the importance of metric learning objectives for the task of gesture retrieval, particularly ranking-oriented embedding optimization (Cakir et al., 2019).

### 5.5. Effect of Attention Pooling

In the last experiment, attention-based temporal pooling was integrated with the Transformer model, along with mild data augmentation and P-K batch sampling (K=3), which is commonly used in metric learning batch construction (Musgrave et al., 2020). This resulted in the best performance in

terms of mean Average Precision (0.237), as well as Recall@1 (0.183), compared to mean pooling. It was found that, compared to mean pooling, using attention-based temporal pooling resulted in a higher weightage being given to informative frames of a gesture sequence. This resulted in a better representation of the gesture, as captured by the learned embeddings, as a whole rather than averaging out across the entire sequence of frames. While there was a minor decline in Recall@5 and Recall@10 compared to the baseline Transformer model, there was a significant improvement in terms of ranking quality, as captured by mAP, which aligns with retrieval-oriented evaluation principles (Cakir et al., 2019).

### 5.6. Summary of Findings

Based on the experiments, the following conclusions are made:

- Transformer-based temporal modeling achieves higher retrieval performance than graph-based modeling for isolated gesture retrieval.
- Angular margin losses improve embedding separation compared to proxy-based objectives (Deng et al., 2019).
- The combination of ArcFace and supervised contrastive loss enhances retrieval accuracy (Khosla et al., 2020).
- Attention-based temporal pooling yields higher-quality embeddings than mean pooling.

Among all the configurations, the transformer with attention pooling and supervised contrastive loss outperforms the other configurations for retrieval performance. Further analysis of the experiment and its implications, as well as the behavior of the models, are discussed in the next chapter.

### 5.7. Cross-Dataset AUTSL Evaluation

To test the generalization ability on other datasets, the best model, which was a Transformer with attention pooling, was tested on the AUTSL dataset with 904 gallery and 226 query samples, and it got a Recall@1 of 3.54%, Recall@10 of 29.20%, and mAP of 0.0914. This shows that the model is able to maintain the partial structure in the embedding space despite not being trained on the AUTSL dataset.

## 6. Discussion

### 6.1. Overview

This chapter examines the experimental results and discusses their implications for gesture similarity learning. The intent is not to repeat the numerical results, but rather to understand why some architectures and loss functions were better performing and what this means in terms of isolated sign retrieval (Cakir et al., 2019; Duarte et al., 2022). The experiments show that the quality of the embedding is affected by the temporal modeling strategy, loss function, and pooling method (Ghojogh et al., 2022). The Transformer-based architecture with the attention pooling method yielded the best overall retrieval performance, which indicates the strength of global modeling in this problem.

### 6.2. Transformer vs Graph-Based Modeling

One of the key findings from the current study is that the performance of the Transformer-based model surpassed the performance of the Spatial-Temporal Graph Convolutional Networks (ST-GCN) model in the isolated sign retrieval task. The ST-GCN model incorporates the skeletal structure of the hand through the use of a graph convolutional network. This has been a popular approach in action recognition, where the spatial relationships between the joints are crucial. However, in the current low-shot retrieval setting, the ST-GCN model did not achieve superior performance in comparison to the Transformer-based encoder.

There are a number of reasons why the ST-GCN model did not achieve superior performance in the current study. Firstly, the WLASL dataset contains only three to four instances per class on average (Li et al., 2020). The graph-based model is based on the ability of the model to learn robust spatial-temporal patterns from multiple instances.

Secondly, the Transformers incorporate self-attention mechanisms that capture global temporal dependencies across the whole sequence (Ramanathan et al., 2015). Contrary to the graph convolutional approach that emphasizes local joint relationships, the self-attention mechanisms emphasize all time steps relative to each other. The ability to capture global temporal dependencies could help capture the nuanced differences in motions between visually similar gestures.

Finally, the Transformer architecture is less restricted by the skeletal structure. For the graph models, the structure is heavily influenced by the joint adjacency matrix. In the attention mechanisms, the ability to capture long-range dependencies is flexible and not restricted to the skeletal structure. In a retrieval scenario where the differences matter, the flexibility could offer an advantage.

These results are also in line with the recent trends in sequence modeling where self-attention models have achieved remarkable performance in modeling long-range dependencies across different

domains.

### 6.3. Influence of Metric Learning Objectives

As mentioned earlier, the experiments also demonstrate the significance of the learning objectives used for the metric learning (Ghojogh et al., 2022). ProxyNCA yielded the poorest performance among all the losses used for evaluation. While proxy-based approaches ease the optimization process by learning the proxy representation for all classes, they may not capture the intra-class clustering adequately, especially under low-shot conditions (Musgrave et al., 2020). Substituting ProxyNCA with the ArcFace loss significantly boosted the performance (Deng et al., 2019). ArcFace incorporates the angular margin, thus improving the inter-class discriminability (Deng et al., 2019). This shows the effectiveness of angular discriminability in the gesture retrieval task.

Further performance gains were observed when the supervised contrastive loss was added with the ArcFace loss (Khosla et al., 2020). While the angular margin improves the inter-class discriminability, the addition of the supervised contrastive loss strengthens the intra-class clustering. This shows the effectiveness of the retrieval-oriented optimization objectives (Cakir et al., 2019).

### 6.4. Impact of Attention Pooling

The greatest improvement in the overall quality of the retrieval process was achieved by the attention pooling method. The reason for this is the ability of the model to learn which parts of the video contribute the most to the gesture identity, unlike the mean pooling method. Isolated signs have different phases: preparation, execution, and return. Not all frames of the gesture are equally discriminative. The mean pooling method treats all the frames equally, which may reduce the discriminative information of the gesture. The better mAP values of the attention pooling method imply that the model has learned the importance of the frames, which is a crucial component of the gesture similarity modeling process, consistent with attention-based modeling approaches (Parian-Scherb et al., 2024).

### 6.5. Low-Shot Learning Considerations

WLASL dataset is a low-shot learning scenario, i.e., there are only a few examples per word sense (Li et al., 2020). This also presents a number of challenges for representation learning. Less repetition of classes makes it difficult for the model to learn intra-class variation. The better performance of the Transformer-based model might be due to the robustness of global temporal modeling, as compared to graph-based structural modeling, in low-shot learning scenarios (Li et al., 2023).

Additionally, using P-K sampling with K=3 was beneficial for the stability of contrastive learning, as there were enough positive pairs in each batch (Musgrave et al., 2020). This also emphasizes the role of batch formation in metric learning, especially in low-shot learning.

The cross-dataset experiment on AUTSL further emphasizes that the learned embedding captures transferable motion patterns even when the sign vocabulary and signer distribution are very different from WLASL. While the absolute values of the retrieval scores are not high, the fact that it is able to retrieve the correct matches in a 226-class open-set scenario is an indication that it has learned gesture-level similarities rather than learning dataset-specific information.

### 6.6. Limitations

Although the results are promising, several aspects are to be noted. Firstly, the study only considers isolated sign retrieval. Continuous sign sequences are not considered. More segmentation mechanisms would be needed to tackle the retrieval of sentences (Martins, 2024).

Secondly, large-scale indexing techniques like FAISS or HNSW were not fully explored (Malkov and Yashunin, 2018; Emanuilov and Dimov, 2024). Although the evaluation of the efficiency of the embeddings with the help of the cosine similarity is sufficient for the analysis of the embeddings' quality, the use of the above-mentioned techniques would be necessary for the effective deployment of the system.

Thirdly, the study is mostly based on the WLASL dataset (Li et al., 2020). The cross-lingual and cross-dataset evaluation of the sign language dataset for the Russian and Kazakh languages is an important task to be validated.

### 6.7. Implications and Future Directions

The results obtained in the thesis have shown the viability of video-to-video gesture retrieval based on poses, and the effectiveness of attention-based transformer models in low-shot gesture retrieval. The possible ways for the extension of the suggested approach are:

First, a highly efficient method for scaling up the retrieval, such as FAISS or HNSW, might be used (Malkov and Yashunin, 2018; Emanuilov and Dimov, 2024). Second, experiments on gesture retrieval in a different domain, using the internal Russian and Kazakh dataset, might be performed. Third, a new approach to modeling time might enable the extension of the approach to sign language sequences (Ramanathan et al., 2015).

The suggested approach is a contribution to the development of visual gesture similarity systems, and it creates a foundation for the development of a reverse sign language dictionary (Hassan et al., 2025).

## 7. Acknowledgements

This text was modified with the assistance of ChatGPT.

## 8. Bibliographical References

Fatih Cakir, Kaiming He, Xunxia Xia, Brian Kulis, and Stan Sclaroff. 2019. Deep metric learning to rank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1861–1870.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. CiCo: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19016–19026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Yueqi Duan, Jiwen Lu, Wenzhao Zheng, and Jie Zhou. 2020. Deep adversarial metric learning. *IEEE Transactions on Image Processing*, 29:2037–2051.

Amanda Duarte, Samuel Albanie, Xavier Giró i Nieto, and Gül Varol. 2022. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14094–14104.

Simeon Emanuilov and Aleksandar Dimov. 2024. Billion-scale similarity search using a hybrid indexing approach with advanced filtering. *Cybernetics and Information Technologies*, 24(4):45–58.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. 2022. Spectral, probabilistic, and deep metric learning: Tutorial and survey. *arXiv preprint arXiv:2201.09267*.

Saad Hassan, Matyáš Boháček, Chaelin Kim, and Denise Crochet. 2025. Towards an AI-driven video-based American Sign Language dictionary: Exploring design and usage experience with learners. *arXiv preprint arXiv:2504.05857*.

Zifan Jiang, Gerard Sant Muniesa, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Xiang Li, Xinyu Yang, Zhen Ma, and Ji-Hong Xue. 2023. Deep metric learning for few-shot image classification: A review of recent developments. *Pattern Recognition*, 138:109381.

Yury A. Malkov and Dmitry A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Gonçalo Vinagre Martins. 2024. SLVideo: A sign language video moment retrieval framework. Master's thesis, Universidade NOVA de Lisboa (Portugal).

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision*, pages 681–699.

Mahnaz Parian-Scherb, Peter Uhrig, Luca Rossetto, Stéphane Dupont, and Heiko Schuldt. 2024. Gesture retrieval and its application to the study of multimodal communication. *International Journal on Digital Libraries*, 25:585–601.

Vignesh Ramanathan, Kevin Tang, Greg Mori, and Li Fei-Fei. 2015. Learning temporal embeddings for complex video analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4479.

Asmat Zahra, Nazia Perwaiz, Muhammad Shahzad, and Muhammad Moazam Fraz. 2023. Person re-identification: A retrospective on

domain-specific open challenges and future trends. *Pattern Recognition*, 142:109669.

## 9. Language Resource References

Eleni Efthimiou and Stavroula-Evita Fotinea and Thomas Hanke and John Glauert and Richard Bowden and Annelies Braffort and Christophe Collet and Petros Maragos and François Lefebvre-Albaret. 2012. *Dicta-Sign Wiki*. Dicta-Sign Project. https://doi.org/10.1007/978-3-642-31534-3_32.

Dongxu Li and Cristian Rodriguez Opazo and Xin Yu and Hongdong Li. 2020. *WLASL: Word-Level American Sign Language Dataset*. Australian National University. https://doi.org/10.1109/WACV45572.2020.9093512.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. *AUTSL: Ankara University Turkish Sign Language Dataset*. Ankara University. https://doi.org/10.1109/ACCESS.2020.3028072.

# Norwegian Sign Language: Overview of Resources and Experiments with Automatic SignWriting Transcription

**Elisabeth Othamar, Yves Scherrer**
University of Oslo
Oslo, Norway
elisaot@math.uio.no, yves.scherrer@ifi.uio.no

**Abstract**

Norwegian Sign Language (NTS) remains an under-resourced sign language despite its official recognition in Norway since 2022. The limited availability of structured, reusable, and publicly accessible datasets continues to hinder both linguistic research and the development of sign language technologies such as recognition and translation systems. This paper presents an overview of existing datasets and potential data sources for NTS, categorizing them by accessibility, format, and suitability for computational research. We further discuss legal, ethical, and practical considerations related to data reuse, including copyright and privacy constraints. In addition, we report on a series of pilot experiments exploring alternative data acquisition strategies, including dictionary videos, SignWriting resources, and broadcast news material. These preliminary experiments explore whether automatic SignWriting transcription can serve as an intermediate representation for NTS, and examine its potential role in sign identification within continuous signing. The aim of this work is both to document ongoing efforts and to support future initiatives toward the sustainable development of NTS resources.

**Keywords:** Norwegian Sign language, sign language recognition, sign language translation, data scarcity

## 1. Introduction

Sign language resources are a critical foundation for linguistic research and the development of sign language technologies. Resources for sign languages are being collected (e.g. see entries in *The Sign Language Dataset Compendium* by Kopf et al. (2022)). However, many languages are still low-resourced. Norwegian Sign Language (*norsk tegnspråk*, NTS) is one such example, with only lexical entries in *The Sign Language Dataset Compendium*[1]. Although NTS was officially recognized as the national sign language of Norway in 2022 (Halvorsen, 2025), the availability of structured and reusable datasets for research purposes remains limited.

Systematic corpus work on Norwegian Sign Language (NTS) began in the mid-2010s, although earlier datasets were collected as part of individual research projects (e.g., in 2007) (Jantunen et al., 2024), but there are currently no publicly available datasets suitable for large-scale annotation or computational processing of NTS. Existing datasets are restricted to private use (Svendsen and Kadry, 2023; Uddin et al., 2025) and/or lack the necessary annotations. Recent work has explored the use of broadcast media as a data source (Høgset, 2024), but this introduces various challenges related to data quality, copyright, and reuse.

In response to these limitations, we explore an alternative route: leveraging partially structured resources, that were not originally designed for computational research, and leveraging existing language independent processing tools. These include single-sign dictionary videos, SignWriting lexicons, and publicly available news broadcasts in NTS. By combining these sources, we present initial experiments examining the feasibility of using automatic SignWriting transcription as an intermediate representation for NTS, and outline ongoing work on its potential application to sign identification in continuous signing. This paper therefore provides both an overview of available data sources and a pilot investigation into their practical usability for sign language technology development using language independent tools/frameworks, in the hope that such methods can be used for other low-resource sign languages.

## 2. Background and Related Work

Research on sign language resources has expanded significantly in recent years, reflecting both linguistic and technological developments. Importantly, current datasets differ along several key dimensions, including the type of signing captured, the form of annotation provided and the availability (e.g. see Núñez-Marcos et al. (2023) for publicly available resources and article[2] by Amit Moryossef and Yoav Goldberg, with resources of varying availability and type of signing).

[1] https://www.sign-lang.uni-hamburg.de/lr/compendium/language/nsl.html

[2] https://research.sign.mt



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 408–418
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

## 2.1. Aspects of Sign Language Resources

First, a central distinction concerns single-sign videos and continuous-signing videos. **Single-sign datasets** typically contain isolated lexical items produced in (or close to) citation form (i.e., one sign per video). Such datasets are commonly used to model sign lexemes and are often aligned with dictionary-style representations of sign vocabularies (Bragg et al., 2019; Koller, 2020). In contrast, **continuous-signing datasets** consist of fluent, multi-sign utterances forming sentences or longer discourses. Continuous signing introduces additional linguistic phenomena such as coarticulation, prosodic structuring, non-manual markers, and variation in signing speed, which make automatic transcription substantially more challenging than (isolated) single-sign data (Bragg et al., 2019; Koller, 2020; Núñez-Marcos et al., 2023). However, such datasets are essential for studying naturalistic language use and for developing NLP technologies such as continuous sign language recognition and translation (e.g., Duarte et al., 2021).

Second, sign languages can be represented using different annotation systems (Bragg et al., 2019). Most commonly, sign language data is annotated with **glosses**. Glosses are text-based representations of signs and often correspond to words of an oral language (Koller, 2020). For example, a phrase in the Australian Sign Language could be glossed as: *ASK-him STUDENT WHERE IX-he LIVE*, meaning *Ask the student where he lives* (Núñez-Marcos et al., 2023). While glosses are not full translations, they provide convenient symbolic labels for signs and are widely used in linguistic corpora and computational pipelines. However, glossing abstracts away much of the phonological and spatial structure of signs, including handshape, movement, location, and non-manual features. In addition, glosses are language- and dataset-specific: sign–gloss mappings valid in one sign language may not hold for another, or may vary across different datasets for the same language (e.g., Bragg et al., 2019; Koller, 2020). In contrast, **SignWriting** (Sutton, 1990)[3] and **HamNoSys** (Hanke, 2004)[4] are writing systems specifically designed for sign languages. These systems visually encode handshape, movement, orientation, location, and non-manual features in a two-dimensional layout with language independent symbols (see Table 5 for a SignWriting example). Compared to glossing, they offer a more fine-grained and language-internal representation, and can be applied to any Sign Language. Lastly, it is also possible to represent signs by natural language **descriptions** of both their manual and non-manual components. Such descriptions, typically in English, can be easily generated automatically with video LLMs (see e.g., Asasi et al., 2025). They are also largely language-independent and can therefore serve as an intermediate representation in NLP tasks.

Third, a component of many sign languages is **fingerspelling**. Fingerspelling involves the manual representation of written-language letters using specific handshapes, and is commonly used for proper names, technical terms, borrowings, or lexicalized items. From a computational perspective, fingerspelling poses unique challenges due to rapid articulation, coarticulation effects, and variation between fully spelled and lexicalized forms. Including fingerspelling in sign language research is crucial for building accessibility tools and applications that can accurately process names, specialized vocabulary, and other elements not covered by standard sign notation (Georg et al., 2025; Tanzer, 2025).

Across sign languages, the development of annotated large-scale datasets, particularly those containing continuous signing, has enabled progress in both linguistic research and NLP technologies (Power et al., 2025; Cihan Camgöz et al., 2020). These advances include improved automatic sign recognition, alignment between video and representations, and sign language translation systems (Koller, 2020; Núñez-Marcos et al., 2023). As resources continue to diversify in terms of modality and annotation scheme, careful consideration of these representational distinctions remains crucial for both corpus design and computational modeling.

## 2.2. Sign Language Research in Norway

NTS has received limited attention in terms of dataset creation. At the Western Norway University of Applied Sciences (HVL) in Bergen there is a research group focused on Sign Language, linguistics and interpreting.[5] Also, at the University of Bergen, there is a research group, TEIKN, researching both linguistics and the use of technology in processing Sign Languages.[6] Furthermore, at the Norwegian University of Science and Technology (NTNU) in Trondheim, there is a research project investigating how Deaf signers use NTS across linguistic and social contexts, while develop-

[3] https://www.signwriting.org/about/what/what02.html

[4] https://www.sign-lang.uni-hamburg.de/dgs-korpus/hamnosys-97.html

[5] https://www.hvl.no/en/research/group/TOLK/

[6] https://www4.uib.no/en/research/research-groups/teikn

ing a corpus and lexical database.[7] Researchers at the Department of Applied Data Science at Noroff University College in Kristiansand, investigated image classification models for NTS recognition (Svendsen and Kadry, 2023). Finally, SINTEF, an independent Norwegian research institute, is involved in a project aiming to develop an app for Sign Language translation for NTS.[8]

So far, existing efforts have primarily focused on linguistic documentation and research or small-scale educational resources, such as a publicly available sign dictionary, leaving a gap between available materials and the needs of computational research. Thus, the need for more NLP research and dataset creation for NTS is evident.

## 3. Norwegian Sign Language Datasets

This section reviews relevant resources for Norwegian Sign Language, starting with datasets that were explicitly created for NTS research or documentation, and continuing with alternative resources that primarily target the Deaf community but can be leveraged for computational experiments.

### 3.1. Curated Datasets

The *Norwegian Sign Language Corpus* encompasses four different datasets that were collected at NTNU between 2007 and 2024. These corpora were created for linguistic research and are in the process of being annotated. The resulting ELAN (EAF) files provide multi-tiered annotations for each signer, including right- and left-hand glosses, non-manual markers (body, face, head, gaze, mouthing), translations into Norwegian and English, and interactional features, allowing for detailed multimodal linguistic analysis. These datasets differ in the availability of ELAN files and their licensing:

- **Halvorsen (2012)**[9] – Collected in 2007, contains four deaf signers, two males and two females with varying age, which all have other deaf family members. The signers were asked to retell one children's book and answer one question. The dataset consists of 8 videos with a total duration of 18 minutes. The dataset is released under the CC BY-NC-SA 4.0 license. There are no annotations or translations available for this dataset. Additionally, all the signers lived in the central Eastern Norway and the videos may only capture one of the different dialects of NTS.
- **Pilot Corpus (Conversations)**[10] – The projectstarted in 2015 and aimed to document the language practices of elderly deaf signers and to introduce corpus research to the community. Seven elderly deaf signers from Oslo, Bergen, and Trondheim were recorded over two days at the OsloMet University. Participants engaged in dyadic and group conversations with other signers and deaf project team members. The total duration of this dataset is 5.5 hours. Parts of the dataset are available for non-commercial research with attribution, subject to approval of a research plan, compliance with privacy requirements, and restrictions against redistribution or creating derivative works. To the authors' knowledge, no translations are publicly available, though ELAN files exist for private recordings from this project (and for some Depicting Perspective videos) used in Ferrara et al. (2023).[11]
- **Depicting Perspective**[12] – Collected from 2017 to 2018, this dataset is a follow-up to previous L2 (sign language acquired as a second language) signing research. It captures how signers establish and maintain visual perspectives. Data were collected in Oslo, Bergen, and Trondheim with 21 deaf participants engaged in 13 naturalistic conversations (30–40 minutes each) with deaf or hearing native signers as interlocutors, totaling 7.5 hours of video. This dataset is licensed in the same manner as the Pilot Corpus. Some videos have additional restrictions, and their availability may be limited. Private recordings from this project and corresponding ELAN files[13] were also used in Ferrara et al. (2023).
- **Retellings and public events** – Collected from 2019 to 2024 as part of the **Language Ecology** project[14]. It contains retellings of three selected children's books and public presentations, recorded with 87 Norwegian Sign Language signers across multiple cities to study conversational interaction, multilingualism, and iconicity. The video duration is unknown. Retellings and Public Events is allow-

[7] https://www.ntnu.edu/isl/nts-language-ecology

[8] An interview about the SINTEF project can be found here.

[9] http://hdl.handle.net/11509/141

[10] http://hdl.handle.net/11509/147

[11] Some example ELAN files can be accessed via the OSF overview: https://osf.io/q8vf3/overview. Access to the full ELAN annotations may require contacting the corpus manager.

[12] http://hdl.handle.net/11509/144

[13] Some example ELAN files can be accessed via the OSF overview: https://osf.io/m4gh8/overview.

[14] http://hdl.handle.net/11509/155

| Dataset | Format | Size | Duration | Availability |
|---|---|---|---|---|
| Halvorsen | Video | 118 MB | 18 minutes | Non-commercial |
| Pilot Corpus | Video | 8,74 GB | 5,5 hours | Upon approval |
| Depicting Perspective | Video | 9,8 GB | 7,5 hours | Upon approval |
| Retellings and public events | Video | 13,8 GB | Unknown | Non-commercial |
| Conversations and private meetings/events | Video | 68 GB | Unknown | Upon approval |
| Alphabet Dataset | Images | 418 MB (24.300 images) | – | Available |
| Numbers Dataset | Video | Unknown | Unknown | Private |

Table 1: Overview of known Norwegian Sign Language datasets.

ing non-commercial use, adaptation, and sharing with attribution, while requiring derivative works to be released under the same license.

- **Conversations and private meetings/events** – Is also part of the **Language Ecology** project, and contains conversations, discussions of deaf-related issues, private meetings, and events recorded with 87 Norwegian Sign Language signers (aged 22–91) across multiple cities. Sessions, typically 2.5 hours in pairs, capture naturalistic conversational interaction to study multilingualism, reference, and iconicity. The total duration of these videos is unclear. Conversations and Private Meetings/Events is distributed under a CLARIN RES license, allowing non-commercial research use with attribution, requiring submission of a research plan and approval by the corpus manager, and prohibiting redistribution or sharing of derivatives. Some data have additional restrictions and must be requested separately.

In sum, the varying licenses and availability of ELAN files makes this corpus unsuitable for data-demanding processes such as sign language translation and recognition. In the context of the Norwegian Sign Language Corpus, a sign bank for NTS was also collected, but it is not publicly available.[15]

Two additional datasets have been reported in the recent literature. One consists of still images of hands representing fingerspelling of the Norwegian **alphabet** (Svendsen and Kadry, 2023), while another one (part of the SINTEF initiative) contains short video recordings of the **numbers** 0–10 in NTS (Uddin et al., 2025). These datasets are limited in size and scope. The data for the numbers is private, however, the article for alphabet recognition indicate the data is obtainable, and this webpage contains the data under the CC BY 4.0 licence. The characteristics of these datasets are summarized in Table 1.

[15] https://signbank.cls.ru.nl/datasets/NTS/

## 3.2. Alternative Data Sources

Beyond curated datasets, several digital resources intended for the Deaf community contain valuable NTS material. These include broadcast news, educational videos, and online sign language platforms. While not designed for research, such sources may be leveraged to construct datasets under appropriate legal and ethical frameworks.

One such resource for creation of a single-sign dataset is the online NTS **dictionary**.[16] This dictionary is part of a resource called *Tegnbanken*[17], which states that the material is available under the CC BY-NC-ND 4.0 license.

**SignPuddle**[18] is a community-driven initiative to make SignWriting accessible to Deaf communities worldwide. SignPuddle hosts language-specific SQLite databases containing SignWriting entries paired with glosses or written-language translations. The Norwegian SignPuddle database contains over 6000 entries.[19] Such databases can serve as valuable parallel data for experiments involving SignWriting transcription, gloss prediction, or cross-modal mapping. However, because they are often collaboratively developed and independently maintained, questions related to quality, documentation, version control, and long-term sustainability must be carefully considered when incorporating them into reproducible NLP research pipelines.

A recent master's thesis explored the construction of a continuous-signing dataset based on **news broadcasts** aligned with corresponding subtitles (Høgset, 2024), taken from programs of the national broadcaster of Norway, NRK. While this approach increases coverage, it also introduces challenges related to alignment accuracy, bias, and copyright

[16] https://www.minetegn.no/Tegnordbok-2016/tegnordbok.php

[17] https://www.minetegn.no/Tegnbanken-2016/index.php

[18] Info on the people behind SignPuddle can be found at https://steveslevinski.me/#infrastructure.

[19] https://signpuddle.com/client/#!/dictionary/nsl-NO-dictionary-public

restrictions.[20]

Besides news broadcast, there exists a large online sign language platform called *tegn.tv*[21], whose objective is to aggregate and broadcast sign language content produced by various contributors. Many of the videos are created with accessibility in mind and include spoken Norwegian and/or subtitles; however, the platform does not explicitly guarantee that all videos provide subtitles, and coverage may therefore vary across content. The multitude of contributors makes it difficult to determine the copyright situations and conditions of reuse. However, in contrast to the videos at NRK, this content is largely made by the Deaf for the Deaf community, leading to a more representative language content in the videos.

In sum, broadcast media offers continuous signing and a broad vocabulary but is typically subject to strict copyright restrictions, whereas other sources may yield good quality single sign content, with the risk of differing from the same sign in the "wild" or simply being the effort of individuals not fluent/educated in sign language. These trade-offs must be carefully considered when using such sources for research.

## 4. Going the Alternate Route

We present some preliminary experiments that rely on three alternative data sources: single-sign videos from the Tegnbanken online dictionary, the SignPuddle database for SignWriting, and a set of continuous-sign NRK news broadcasts. The goal of these experiments is to assess whether language independent annotation frameworks (such as SignWriting) and pretrained language independent annotation systems (such as those found in the Sign Language Processing open-source tools[22] by Amit Moryossef and Yoav Goldberg) can be leveraged for NTS in the absence of large manually annotated training data.

We evaluate how well automatic SignWriting transcription tools work for NTS. Prospectively, we aim to investigate if transcription to SignWriting (or to other language-independent sign representations) is useful for identifying single signs within continuous-sign videos, utilizing the SignWriting simply as a mapping to the dictionary gloss/translation.

### 4.1. Data Collection

The Tegnbanken dictionary is owned by Statped[23], and to obtain the single-sign videos, the authors programmatically harvested the website. An XML endpoint was queried to retrieve structured metadata, from which all video filename identifiers were extracted. For each identifier, a canonical URL pointing to the corresponding MP4 file was deterministically constructed. Videos were downloaded via parallelized HTTP requests with streaming transfer, file-integrity checks, and automatic handling of partial or failed downloads. Previously acquired files were skipped to ensure reproducibility and efficiency. The resulting dataset consists of 9119 locally stored lexical sign videos suitable for downstream tasks such as sign language recognition. The code can be found at https://github.com/elisaot/Code_NTS.git.

The NTS SignWriting reference is available as an individual SQLite database in the SignPuddle.[24] The database contained 6362 entries with SignWriting and corresponding gloss/translation to written Norwegian.

For the acquisition of NRK's news broadcasts it was decided to download one year of news, to maximize the vocabulary. We constructed the sign language news corpus by programmatically downloading episodes of NRK *Tegnspråknytt* from the NRK online media platform using the yt-dlp framework.[25] For each month URL, playlist metadata were extracted and automatically inspected for the availability of Norwegian subtitles. Only videos containing Norwegian subtitles were retained to ensure aligned written-language text. Selected episodes were downloaded in MP4 format together with their subtitle files, while previously acquired files were skipped to maintain reproducibility and efficiency. A structured download log was generated documenting episode identifiers and subtitle types. Due to copyright restrictions, we do not share the downloading code nor the videos themselves, but we release the processing log and the URLs used at https://github.com/elisaot/Code_NTS.git. The resulting dataset consists of 246 news broadcasts, with an average length of 4.2 minutes. Lastly, we automatically segmented the video material into subtitle-aligned clips by parsing time-stamped subtitle files and extracting corresponding video intervals. Subtitle times-

[20] NRK indicates that their products may be used for educational and research purposes: https://info.nrk.no/tv-spons-og-salg-av-nrks-innhold/arkiv/#innhold-til-undervisning-eller-forskning

[21] https://tegn.tv

[22] https://github.com/sign-language-processing

[23] Statped is the Norwegian national service for special education, providing support, resources, and expertise for individuals, including sign language users. https://www.statped.no

[24] The database for NTS (sgn69.db) can be found at http://signbank.org/swserver_data/puddle/.

[25] https://github.com/yt-dlp/yt-dlp?tab=readme-ov-file

| Dataset | Format | File size | Amount | Availability |
|---|---|---|---|---|
| Tegnbanken Dictionary | Video + Gloss | 2.72 GB | 5.29 hours | Non-commercial |
| SignPuddle database | Gloss + SignWriting | 1.8 MB | 6362 entries | Open source (MIT) |
| NRK News Broadcasts | Video + Text | 38.92 GB | 17.35 hours | Copyrighted* |

Table 2: Overview of the created/collected Norwegian Sign Language datasets.(* According to the NRK website, content for documentation and research purposes can be ordered via the National Library of Norway or through a local library.)

tamps were processed using pysrt, and video segments were extracted with moviepy. While this method enables efficient alignment of signed content with written-language subtitles, it relies on subtitle timing and does not guarantee precise linguistic segmentation. Additionally, the news broadcasts are interpreted from written Norwegian text and the person interpreting may be an L2 signer of NTS, introducing biases.

Table 2 shows the amount of data acquired from the three sources.

### 4.2. Inferring a Reference Dictionary With SignWriting Annotations

In order to evaluate automatic video-to-SignWriting tools, we need a reference dataset that associates single-sign videos with their SignWriting representations. The videos contains detailed gloss annotations (ID glosses) that distinguish between multiple sign variants (e.g., GLOSS-1, GLOSS-2). To enable alignment with the SignWriting data, these glosses are normalized into a simplified gloss set, where variant distinctions are collapsed (e.g., GLOSS-1, GLOSS-2 → GLOSS). The resulting simplified glosses serve as a pivot representation for cross-modal alignment between video and SignWriting.

As in written and spoken languages, multiple distinct signs in a sign language may share the same meaning, which is seen in our single-sign videos. Subtracting away their distinctions for alignment, performing no manual cleaning of these matches due to lack of SignWriting expertize, led to substantial lexical overlap, and consequently substantial multiplicity, in our reference dictionary.

Analysis of this revealed; In the SignPuddle database for NTS, 6362 entries were available, but only 1949 had a unique gloss. Of the 9119 single-sign videos examined, only 5598 of the pivot glosses occurred uniquely. When cross-referencing the two resources, only 2316 single-sign videos had a corresponding entry in SignPuddle. Thus, as a result of multiplicities in both the videos and the SignPuddle, the resulting dataset contained 5694 rows with only 519 unique glosses. For example, the gloss VANN ('water') showed the highest level of multiplicity across the datasets, yielding 66 entries in the reference dictionary.

Owing to limited expertise in both NTS[26] and SignWriting, manual verification to eliminate potential false matches between signs and SignWriting representations was not feasible. Consequently, while the dataset provides a valuable starting point for pose-to-SignWriting or gloss-based experiments, the presence of duplicate and potentially misaligned entries must be considered. For our project, this implies that any evaluation of automatic video-to-SignWriting tools using this dictionary should be regarded as preliminary and indicative rather than definitive, reflecting the quality of the underlying alignments rather than solely the performance of the models.

### 4.3. Automatic SignWriting Transcription

SignWriting transcription proceeds in three steps: first, the videos are transformed into **poses**, then the poses are segmented and lastly, the poses are transcribed into sequences of SignWriting signs.

For the first step, we make use of the Sign Language Processing tool Pose[27] by Amit Moryossef, Mathias Müller, and Rebecka Fahrni. An open-source toolkit that defines a standardized .pose format for storing temporally aligned body, hand, and facial keypoints extracted from video. The library provides utilities for reading, writing, preprocessing, and visualizing pose sequences, facilitating reproducible data preparation and consistent pose-based modeling in sign language recognition and translation tasks.

We applied the Linguistically Motivated Sign Language Segmentation (Moryossef et al., 2023) pipeline (utilizing the E1s model) to the .pose files. While not strictly necessary for single-sign videos, we applied the same tool to both continuous-sign and single-sign poses to maintain consistent processing and annotation.

The SignWriting transcription library[28], by Amit

[26] One of the authors is currently in the early stages of acquiring NTS as a second language (L2 signer).

[27] https://github.com/sign-language-processing/pose

[28] https://github.com/sign-language-processing/signwriting-transcription/tree/main

Moryossef, Rotem Zilberman, and Ohad Langer, provides a language-independent pre-trained model for pose-to-SignWriting transcription. The library defines a command-line interface that takes .pose files and corresponding ELAN files, normalizes and aligns the pose input, and runs a pre-trained neural transcription model to predict SignWriting symbol sequences.

Since the dictionary videos refer to single signs, we expect the transcription to mostly produce one SignWriting sign per video. However, we noted early on that dictionary videos were annotated with several SignWriting sequences, i.e. on average a dictionary video was annotated as two consecutive signs. This can be due to poor segmentation, the slower and larger signing, compared to natural conversation, due to the educational purpose of the videos, or a combination thereof. Additionally, some signs are multisyllabic, giving the appearance of two signs. In an attempt to counter the slow and large signing, we experimented with a pre-processing pipeline that generates speed- and/or trim-modified versions of the videos matched with entries from the SignPuddle. The transformations, including a 25% increase in playback speed and removal of 0.5 seconds from both the start and end of each video, were applied using moviepy.[29] This approach facilitates data augmentation and efficient preparation of large video datasets, while (optionally) preserving the original files.

### 4.4. Results and Analysis

For the evaluation of the SignWriting transcriptions, we relied on the work from Moryossef et al. (2024) on automatic evaluation metrics for SignWriting, relying on the data extracted from SignPuddle as reference transcriptions.

For manual evaluation SignWriting was visualized using the *signwriting_to_clip_image*[30] function from the SignWriting evaluation framework of Moryossef et al. (2024). To interpret the individual symbols, we consulted both the online Sutton SignWriting character reference[31] and the corresponding dictionary videos.

The metric results, found in 3, are very similar to the any-to-any comparison reported in Moryossef et al. (2024), where all sign pairs (of 1.000 randomly selected signs) are compared regardless of alignment. This suggests that the automatic annotations do not reliably align with the reference SignWriting: the metrics for supposed matches are comparable to the baseline of unrelated signs. The stability of these results across the experimental conditions (original, speed, cut, both) indicates that modification do not substantially affect this outcome. An important consideration regarding the metric scores; the misalignment of SignWriting in our results can be caused/exaggerated by the possible mismatches in our reference dictionary.

We conducted a manual evaluation of the similarity between predicted annotations for the dictionary videos and references for the alphabet, digits 0–10, and three randomly selected examples. We observed that mismatches between the video and the reference were particularly common in the alphabet. This is partly due to NTS including both one-handed and two-handed alphabet variants (Ferrara et al., 2023), which were not distinguished in our simplified gloss matching. Additionally, individual letters also had multiple similar variants, especially in the SignWriting references. Thus, in total, we had a manual evaluation dataset of 347 matches, where approximately half were either mismatches between videos and reference, videos without similar reference, or references without similar video.[32]

The following evaluation of predicted annotation quality considered videos and references that matched visually. For the alphabet, each of the four approaches (three modifications and the original) outperformed the others in approximately one-quarter of the cases. For digits 0–10, some handshape predictions were correct by chance, as the model tended to default to the most generic handshape (e.g., “5”). Finally, in the randomly selected examples, we observed handshapes and movements that were closer to the references, primarily in the original video predictions. However, there

[29] https://github.com/Zulko/moviepy

[30] https://github.com/sign-language-processing/signwriting-evaluation/blob/main/signwriting_evaluation/metrics/clip.py

[31] Which can be found here

[32] This manual verification was performed by one of the authors, who is currently acquiring NTS as a second language (L2 signer).

| Metric | Original | Speed | Cut | Both |
|---|---|---|---|---|
| CHRF | 0.2428 | 0.2427 | 0.2434 | 0.2431 |
| CLIPScore | 0.8708 | 0.8712 | 0.8694 | 0.8701 |
| SymbolsDistances | 0.3269 | 0.3265 | 0.3188 | 0.3197 |
| TokenizedBLEU | 0.0845 | 0.0857 | 0.0841 | 0.0843 |

Table 3: Mean metrics for the different experimental conditions on single-sign videos.

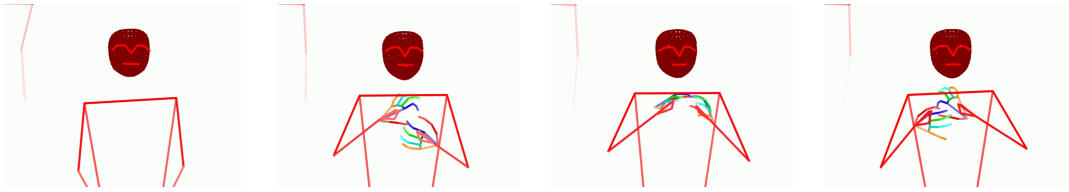


Table 4: Frames from the pose file for the NTS sign to gloss LENKE(girlander) ('chain(garlands)').

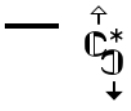
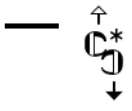
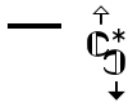
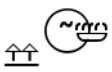
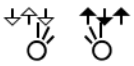
| Reference | Original | Speed | Cut | Both |
|---|---|---|---|---|
|  |  |  |  |  |
| |  |  | |  |

Table 5: SignWriting reference and predicted annotations for the NTS sign with ID-gloss LENKE(girlander) ('chain(garlands)'). From left to right: the reference SignWriting from SignPuddle; two consecutive predicted annotations for the original video; two consecutive predicted annotations for the video with increased playback speed; a single predicted annotation for the video trimmed by 0.5 seconds at the start and end; and two consecutive predicted annotations for the video that was both sped up and trimmed. (The horizontal line in the first predictions for original, speed and both modifications is the shoulders.)

are also inclusions of non-manual features that are not in the references from all video types.

Table 5 includes reference and predicted SignWriting for the NTS sign corresponding to gloss LENKE(girlander) ('chain(garlands)'), and table 4 shows the pose visualization of the sign. The automatic transcription produces multiple symbol sequences for a single sign, reflecting variations in signing speed, pose estimation noise, or poor segmentation. However, in this example, the sign is multi-syllabic and includes a hand-dominance switch; therefore, two separate predictions are reasonable.[33] The following is an evaluation of the different SignWritings (shown in Table 5) for the NTS sign corresponding to LENKE(girlander) ('chain(garlands)'):

- Reference (SignPuddle): The canonical SignWriting sequence captures the handshape, movement path, orientation, and spatial location consistently.
- Original video prediction (pose-to-SignWriting transcription): Handshapes are partially correct, almost forming a circle: the index and thumb should touch. The hands are forming the correct configuration in the first prediction. However, hand orientation is incorrect in both predictions, and the wrist rotation present in the sign is missing. One of the predictions introduces a non-manual feature (mouth movement) that is not present in the reference.
- Sped-up video predictions: The first prediction is identical to the first prediction of the original video. There is a slight improvement in that from the first to the second prediction the rotation of one of the hands seems to have been captured, but the second hand is missing, and the overall movement and placement are incorrect. A non-manual feature is again included that does not appear in the reference.
- Trimmed video predictions: Handshape is partially correct: the index and thumb form a circle, but the remaining fingers do not, as in the reference. Orientation appears correct, and the movement is predicted as up and down, whereas the actual sign involves wrist rotation and sideways movement. Additionally, the hands fail to interlock their circles.
- Sped-up and trimmed video predictions: The first prediction is identical to the original video's first prediction. The second prediction resembles the trimmed video prediction, but the movement is now forward and backward rather than the reference. The hands still do not interlock correctly.

Overall, performance remained limited. Moreover, the experiments did not improve segmentation quality, as all video types produced an average of two predicted annotations per sign. This highlights the need for more robust segmentation and transcription methods, as well as careful consideration of video quality and signer variability, to produce

[33] We thank the reviewer for this observation.

accurate SignWriting representations for NTS.

## 5. Discussion and Open Challenges

Regarding our collected data we encounter several common challenges regarding Sign Language Processing; Our video sources and alternative sources are primarily protected by **copyright**, limiting research. Our reliance on non-research-oriented sources, such as broadcast material, introduces variability in recording conditions, domain bias, and **legal complexity**. Source selection also introduces an additional layer of bias.

Our interpreted news broadcasts involve **translation** from written (or spoken Norwegian) into NTS, which may influence linguistic structure and register. Furthermore, the objectivity norms of journalism may limit the range of facial expressions, affective markers, and body language represented in the data, making such material less representative of everyday signing "in the wild". Lastly, the interpreter may introduce additional bias depending on L1 or L2 signer status.

The dictionary-based resources from Tegnbanken have limited **representativeness**, as they present isolated signs in controlled conditions. However, unlike typical single-sign datasets created primarily for research purposes (Svendsen and Kadry, 2023; Uddin et al., 2025), Tegnbanken is designed for both Deaf users and learners of NTS. It functions as a dictionary, providing standard video representations and ID-glosses, making it particularly useful for language learning and teaching. For some concepts (simplified glosses), Tegnbanken provides multiple video variants, reflecting natural variation in NTS and the kinds of signs learners encounter in authentic communication. While this variability can complicate direct comparisons and merging with other sources, it also enables evaluation of automatic SignWriting annotations under conditions closer to real-world use, even if it results in lower or more variable evaluation metrics.

Similarly, the community-driven platform SignPuddle rely on voluntary contributions, which may vary in linguistic accuracy, completeness, and consistency. This introduces uncertainty and potential bias depending on the contributors' expertise. Thus, our method of making a reference dictionary from the two sources may perpetuate the limitations of SignPuddle. Additionally, our combined data lacks manual cleaning, resulting in mismatches and multiplicities. However, if the reference dictionary is later revised by experts in NTS and SignWriting, and the glosses are rather the more linguistically rich, fine-grained ID glosses of the videos, this dataset could be a valuable asset to NTS in the single sign setting.

## 6. Conclusion and Future Directions

This paper has presented an overview of existing datasets and potential data sources for Norwegian Sign Language (NTS), highlighting key limitations related to quality, licensing, privacy, and reuse. Moreover, the paper reports on a pilot experiment to leverage a language-independent SignWriting transcription model to annotate single-sign NTS videos. Our analysis shows that the automatically produced transcriptions rarely correspond to the ground truth and that the model is sensitive to minor variations in speed and length. However, it may still be useful for identifying keywords in the news broadcast videos. To evaluate this, we plan to annotate the continuous-sign videos with SignWriting transcriptions in the same way and investigate (a) how many signs can be retrieved from the automatically annotated dictionary, and (b) whether the glosses referring to the dictionary entries correspond to the words in the subtitles. These experiments are currently in progress. Further work will also explore alternative approaches such as the integration of video LLMs to produce language independent representations of NTS signs, to be used as mappings between video and gloss.

Our survey and pilot experiments further indicate that legal constraints, limited public availability of continuous signing data, and challenges related to documentation and anonymization pose substantial obstacles to methodological development in NTS recognition and translation research. Future work should thus prioritize collaborative, community-informed approaches to resource development, ensuring that dataset creation and publication, annotation practices, and evaluation methods are not only aligned with the linguistic expertise and perspectives of the Deaf community, but also enable research and development of NLP tools for sign languages. Such cooperation will be crucial for building robust, representative, and ethically grounded resources that support long-term progress in NTS research.

Lastly, further works for NTS will need to address the re-identification concerns regarding signing videos, which inherently contain **personally identifiable information** ((Xia et al., 2024; Bigand et al., 2021)). This risk is amplified in the small signing community of NTS, where the limited number of fluent signers increases the likelihood of recognition. Good solutions of this challenge may also improve the community's outlook, and involvement on dataset and technology creation (Bragg et al., 2020).

## 7. Limitations

Our pilot project, in addition to previously mentioned limitations and challenges, has several inherent

limitations; Firstly, the pose transformation of our video data introduce its own set of limitations and bias from its reliance on MediaPipe Holistic[34] for keypoint extraction. Similarly, the segmentation model may introduce limitations and biases stemming from both the model itself and its training data. In addition, the segmentation model does not incorporate facial keypoints as inputs. Finally, the SignWriting annotation model may carry similar limitations, resulting in multiple stages of the pipeline being affected by methodological constraints and underexplored biases.

Our future use of the automatically annotated SignWriting of single signs is for identifying keywords in news broadcast videos. With this approach, we introduce an additional limitation: we consider intermediate representations (SignWriting) acceptable if they are predicted consistently. However, consistent predictions do not necessarily imply linguistic correctness, as our results in Section 4.4 demonstrate. Due to our limited expertise in NTS, we were unable to verify the linguistic accuracy of the generated representations for videos not included in our reference dictionary. Thus, we are unable to determine whether the model captures meaningful sign distinctions or merely produces stable but incorrect representations. Consequently, the utility of these annotations for downstream tasks such as keyword identification may rely more on internal consistency than on true linguistic validity, and the results should therefore be interpreted with caution.

## 8. Acknowledgements

We gratefully acknowledge *Tegnbanken* and the *SignPuddle* project for their valuable sign language resources, which were essential for our research on NTS.

## 9. Bibliographical References

Sobhan Asasi, Mohamed Ilyes Lakhal, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Beyond gloss: A hand-centric framework for gloss-free sign language translation. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA.

Félix Bigand, Elise Prigent, Bastien Berret, and Annelies Braffort. 2021. Machine learning of motion statistics reveals the kinematic signature of the identity of a person in sign language. *Frontiers in Bioengineering and Biotechnology*, 9:710132.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2020. Exploring collection of sign language datasets: Privacy, participation, and model performance. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA. Association for Computing Machinery.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10020–10030.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A large-scale multimodal dataset for continuous American Sign Language. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2734–2743.

Lindsay Nicole Ferrara, Benjamin Donald Anible, and Lena Mei Kalvenes Anda. 2023. Exploring sign-writing contact and multilingualism in the norwegian deaf community. In Ella Wehrmeyer, editor, *Advances in Sign Language Corpus Linguistics*, pages 66–89. John Benjamins Publishing Company.

Manfred Georg, Garrett Tanzer, Esha Uboweja, Saad Hassan, Maximus Shengelia, Sam Sepah, Sean Forbes, and Thad Starner. 2025. FSboard: Over 3 million characters of ASL fingerspelling collected via smartphones. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13897–13906.

Rolf Piene Halvorsen. 2012. *Tre diskursmarkører i norsk tegnspråk: en studie av blunk, blikkendring og nikk i åtte fortellinger*. Ph.D. thesis, University of Oslo.

Rolf Piene Halvorsen. 2025. Norsk tegnspråk. In *Store Norske Leksikon*. Accessed: 2026-02-04.

[34] https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/holistic.md

Thomas Hanke. 2004. HamNoSys – representing sign language data in language resources and language processing contexts. In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 1–6, Lisbon, Portugal. European Language Resources Association (ELRA).

Håkon Bygdås Høgset. 2024. Machine learning approaches to automated sign language translation. Master's thesis, NTNU, Trondheim. Accessed: 2026-02-04.

Tommi Jantunen, Johanna Mesch, and Lindsay Ferrara. 2024. Corpus research on signed languages in the nordic countries. *Nordic Journal of Linguistics*, page 1–14.

Oscar Koller. 2020. Quantitative survey of the state of the art in sign language recognition. arXiv: 2008.09918.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. The Sign Language Dataset Compendium: Creating an overview of digital linguistic resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association (ELRA).

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

Amit Moryossef, Rotem Zilberman, and Ohad Langer. 2024. signwriting-evaluation: Effective sign language evaluation via signwriting. arXiv: 2410.13668.

Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

Justin M. Power, David Quinto-Pozos, and Danny Law. 2025. Advances in the historical linguistics of signed languages. *Language and Linguistics Compass*, 19(6):e70026.

Valerie Sutton. 1990. *Lessons in Sign Writing*. SignWriting.

Benjamin Svendsen and Seifedine Kadry. 2023. Comparative analysis of image classification models for norwegian sign language recognition. *Technologies*, 11(4).

Garrett Tanzer. 2025. Fingerspelling within sign language translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 385–464, Albuquerque, New Mexico. Association for Computational Linguistics.

Md. Zia Uddin, Costas Boletsis, and Pål Rudshavn. 2025. Real-time norwegian sign language recognition using mediapipe and lstm. *Multimodal Technologies and Interaction*, 9(3).

Zhaoyang Xia, Yang Zhou, Ligong Han, Carol Neidle, and Dimitris N. Metaxas. 2024. Diffusion models for sign language video anonymization. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 395–407, Torino, Italia. ELRA and ICCL.

# Long-Term Sign Language Data Crowdsourcing Through Collaborative Lexicons

**Pierre Poitier, Jérôme Fink, Ariel Basso Madjoukeng, Adélaïde Couplet, Margaux Leleu, Benoît Frénay**
NADI/HuMaLearn at University of Namur
Rue Grangagnage 21, Namur, Belgium
{pierre.poitier, jerome.fink, ariel.bassomadjoukeng}@unamur.be

**Abstract**

While there exists a multitude of different sign languages (SLs) across the world, deaf communities often lack the digital tools required to document and process their languages. In this work, we introduce Mot-Signe (MOSI), an application designed in close collaboration with actors from the French Belgian deaf community. Our tool enables users to search for French Belgian Sign Language (LSFB) translations or to propose new ones by recording signs themselves. This crowdsourcing approach facilitates the collection of SL data in the wild, enriching the available documentation on LSFB and proposing an innovative response to the data scarcity issue inherent to sign language processing. To evaluate the sustainability of this community-driven data collection, a longitudinal user study was conducted. Following its public release, MOSI demonstrated significant real-world adoption, enabling the collection of over 3,000 distinct LSFB signs. Notably, MOSI captures highly valuable linguistic variations and specialized vocabulary often absent from traditional corpora.

**Keywords:** sign language, crowdsourcing, dictionary, data collection, data gathering, LSFB

## 1. Introduction

Sign Languages (SLs) serve as the primary mode of communication for deaf communities worldwide. Like all natural languages, SLs have evolved organically within their respective communities, resulting in a rich linguistic diversity and a multitude of distinct dialects. However, despite this prevalence and a recent surge of interest in the machine learning (ML) community (Rastgoo et al., 2021), SLs remain severely under-resourced. The deaf community still suffers from a lack of suitable digital tools, such as comprehensive dictionaries, large language models, or robust automatic translation systems.

Several structural barriers hinder the application of existing methods. First, SL data is sparse compared to textual data available on the web. The largest SL datasets contain approximately $200,000$ sign occurrences (Starner et al., 2023), whereas spoken language models benefit from datasets containing billions of word occurrences (Patel and Patel, 2020). Second, the sociopolitical status of SLs plays a role; according to the World Federation of the deaf, $59\%$ of countries have not yet officially recognized their national sign language, leading to stigmatization and a lack of funding for documentation. Third, the technical challenge is significant: SL data consists of high-dimensional video rather than text, making it computationally expensive to process. Furthermore, annotating this data requires skilled interpreters, making the process tedious, time-consuming, and costly. Consequently, SLs remain among the least documented and studied languages in the world (Fink et al., 2023a).

Beyond data scarcity, there is a fundamental linguistic challenge: there is no direct equivalence between SLs and spoken languages. Translating SL to a written form (e.g., French or English) is non-trivial due to differences in syntax and modality. Moreover, sign languages are dynamic; they evolve rapidly, with each generation introducing neologisms and variations that static dictionaries fail to capture. To effectively document the language, a system is required that can track these new signs and dialectal specificities in real time.

In this work, we introduce *Mot-Signe* [1] (MOSI), an application designed in close collaboration with actors from the French-Belgian deaf community. MOSI addresses the data sparsity issue by creating a sustainable, community-driven ecosystem. Unlike pure data collection initiatives, MOSI creates value for the user first, with data collection acting as a beneficial side effect. This aim gives MOSI a double-sided contribution. One of those is a social contribution that focuses on the needs of the community, and the other one is a technical one that is directed towards the data collection methodology. This paper focuses on the second one. In summary, the goals of the application are:

- to establish a long-term, sustainable data collection system to enrich existing dictionaries and corpora.
- additionally, to provide a well-integrated, free,

[1] https://www.mot-signe.be/



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 419–428
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

and public French to French Belgian Sign Language (LSFB) lexicon;

MOSI builds upon the static, studio-recorded corpus of the LSFB-Lab (Meurant, 2015; Fink et al., 2021) to deploy a dynamic crowdsourcing approach for "in-the-wild" data collection. This strategy captures realistic recordings from diverse signers, specifically targeting domain-specific vocabulary absent from academic datasets. We validate this approach through a user study on community retention and analyze data from the public release. Results are promising: among the $3,000$ new video entries contributed, over $80\%$ represent entirely new vocabulary not previously documented in the public lexicon.

## 2. Background and Related Work

This section reviews the background of Sign Language (SL) data collection and digital resources. We first present traditional data acquisition methods and their inherent limitations. Next, we examine existing online SL dictionaries and tools available to the deaf community. Finally, we explore crowdsourcing as a viable alternative, highlighting its benefits, prerequisites, and how previous works have applied it to build robust SL datasets.

### 2.1. Sign Language Data Acquisition

The creation of large representative SL corpora was initiated by linguists to document these languages and analyze their evolution. While these resources are now essential for leveraging machine learning (ML) algorithms, their primary design was for linguistic inquiry. The first notable corpus is the Auslan Corpus (Johnston, 2009) that contains $300$ hours of Australian SL recordings. Inspired by this approach, several teams started to collect their own SL corpus, e.g., the LSFB Corpus (Meurant, 2015) and the British SL (BSL-1K) Dataset (Cormier et al., 2012). To study how the language is used by several categories of the population, those corpora often involve a relatively large number of signers to be representative. The recordings are often performed in a studio with professional cameras and controlled lighting. Consequently, while these studio-recorded datasets offer high-quality documentation, a first limitation is that they are not well suited for ML algorithms. Models trained on this clean, controlled data often struggle to generalize to the visual noise and variable conditions of real-world webcam recordings. Another limitation is that the annotation process is time-consuming and requires skilled workers. Studies show that annotating one hour of video requires up to $100$ hours (Renz et al., 2021). Thus, only subsets of collected data are usually annotated and this acts as a major bottleneck for the linguistic analysis, even if the data exists. The format chosen for the annotation may also vary from one team to another, making them hard to leverage (Fink et al., 2023a).

Another source of sign language data are public lexicons developed by the deaf community. These lexicons contain videos of signs associated with their meaning in a spoken language. Such data are only relevant in the case of Isolated Sign Language Recognition (ISLR) (De Coster et al., 2023) that focuses on the recognition of sign videos containing only one sign. Usually, all signs in a lexicon are performed by the same signer and they are recorded in a controlled environment. Algorithms built upon those data are, therefore, unlikely to generalize to other signers and environments. In an effort to build larger and more diverse datasets, ML researchers started to seek data sources in the wild. The first data sources considered in the literature are TV broadcast SL translations. The Phoenix WTH dataset (Forster et al., 2012) is a collection of SL interpretations of German weather forecasts translated manually by interpreters. The BBC-Oxford British SL Dataset (BOBSL) (Albanie et al., 2021), building upon BSL-1K, introduces a method to align BBC broadcast transcription with its SL interpretation. Such datasets are larger and require less effort than collecting traditional SL corpus. Those solutions drastically increase the size of datasets, but the signers transcribing broadcasts are, usually, always the same and the recordings are still performed in a controlled environment. TV broadcasts also depict a specific vocabulary that differs from the usage of the language in the wild. The deaf community is skeptical about this approach, as interpreters are often non-native SL speakers and the translations of spoken languages are not genuine SL production (Fink et al., 2023a).

Social media constitutes another data source, e.g., Youtube-SL (Tanzer and Zhang, 2024), or video platforms where deaf people share vocabulary. The main issue with these methods lies in the fact that authors do not always give their consent for their data to be used in ML datasets or corpora (Saunders et al., 2021). Such recent datasets, e.g., ASL (Desai et al., 2024), tend to depict more realistic data captured in the wild.

Another limitation is that many existing SL datasets contain a number of biases (Atwell et al., 2024). For example, they are often produced under a controlled environment: uniform background, standard lighting conditions, etc. They may also only concern a lexical field specific to a particular theme, contain very few different signers or present a majority of non-native signers. All these biases make sign language recognition tools built upon aforementioned data sources difficult to transpose into real conditions.

## 2.2. Online Sign Language Tools

Several works have focused on the development of online tools or dictionaries dedicated to sign languages. Pioneer tools are Auslan and BSL SignBank, but many tools have been inspired by them, such as Elix [2], LSFB Dico [3], LifePrint [4], etc. Elix is a SL dictionary that maps the French language to the French SL (LSF). It consists of more than $27,000$ different signs and has approximately $70,000$ users per day. The LSFB Dico is a dictionary proposed for the translation from French to Belgian French SL (LSFB). It contains more than $5,900$ different signs and has the particularity of being very diversified, as many signs appear with multiple variants. These tools are specifically designed for a particular SL and do not account for multilingualism. Recently, multilingual dictionaries have emerged, covering several SLs. One of the most popular multilingual sign language tools is SpreadTheSign [5], a dictionary consisting of more than $200,000$ signs across $40$ sign languages. This dictionary has achieved considerable success in the context of multilingual sign languages. However, although these tools constitute a valuable resource for deaf communities, their primary objective is to be queried. They are not designed for sign language data collection, as they do not provide signers with the ability to add new signs. We propose a new sign language tool, MOSI, that acts as a collaborative sign language dictionary where data collection is embedded as a side effect. To the best of our knowledge, sign language dictionaries have never been paired with long-term crowdsourcing solutions.

## 2.3. Crowdsourcing

Crowdsourcing designates the process of outsourcing tasks to crowd (Hossain and Kauranen, 2015). There are three main reasons that explain why crowdsourcing might be useful in the context of SLs (Garcia-Molina et al., 2016): first, reducing **costs** to realize annotation tasks; second, reducing **latency** by parallelizing the tasks to multiple people; third, reducing **uncertainty** by comparing the outputs of multiple people. It has to be noted that none of those three objectives can be fully met at once. Reducing uncertainty often requires getting insights from a bigger crowd, increasing the cost and the latency. Members of the crowd also may need incentives to perform the tasks. The most common one is to give money in exchange for the task. But other incentives exist, such as philanthropy or entertainment. The Wikipedia project is a great example of philanthropic crowdsourcing. Users are not paid, but willingly participate in this project of general interest (Yuen et al., 2011). This is similar in MOSI, where the general interest is the documentation of the LSFB.

Crowdsourcing the collection and labeling of LSFB signs has the potential to collect a large amount of data recorded under various conditions. Also, the deaf community is used to share vocabulary on social media platforms to compensate for the lack of up-to-date lexicon. Riemer Kankkonen et al. (2018) use data posted by the members of the Swedish SL community to gather more vocabulary for their dictionary. This approach is interesting, as it leverages a well-known platform for the community and is not invasive. However, the various posts and videos shared on the platform must be collected manually, as this process is hard to automate.

To automate data collection, researchers have developed specialized crowdsourcing platforms for sign languages. For American SL (ASL), both Bragg et al. (2022) and Starner et al. (2023) present platforms to collect and validate SL data, demonstrating that it is possible to gather high-quality data. They also report that users are willing to contribute for a reasonable compensation. Dhruvo et al. (2023) built the Bangladeshi SL dataset and collected more than $21,000$ videos using crowdsourcing. Kapitanov et al. (2023) leveraged two crowdsourcing platforms (Yandex Toloka[6] and ABC Elementary[7]) for recording and validation, creating the Russian SL (RSL) dataset with $20,000$ videos across $1,000$ different signs. This demonstrates that crowdsourcing can be an effective way to build datasets, reducing costs by eliminating the need for financial incentives, although it slightly increases the time required for data acquisition.

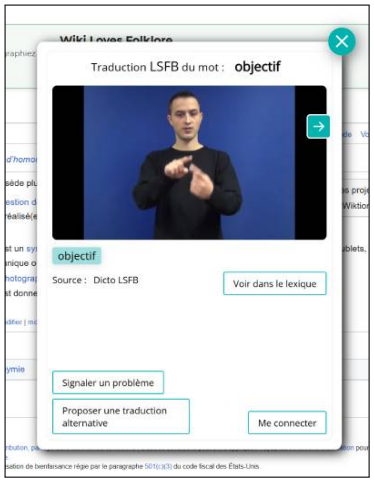
We observe that the primary motivation behind the existing platforms is data collection without starting from a need of the deaf community. This may result in a solution that is less relevant and not user-centered, reducing the likelihood of adoption and long-term engagement. In our study, we build on a need identified within the deaf community, using it as an incentive for application usage. Consequently, data collection and validation become secondary outcomes of the application. The expected result is improved user retention along with a long-term data collection process.

In summary, while traditional corpora and online dictionaries provide valuable linguistic resources, they remain static and do not leverage the community as an active contributor. The following section presents how MOSI bridges this gap.

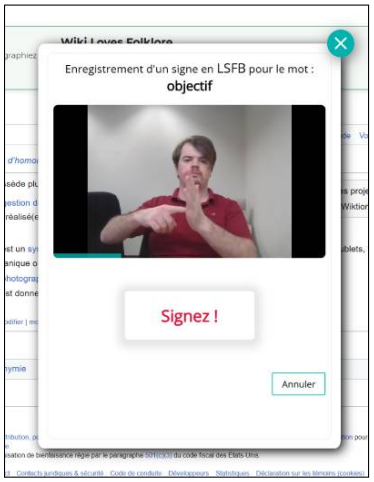
[2] https://dico.elix-lsf.fr/
[3] https://dico.lsfb.be/
[4] https://www.lifeprint.com/
[5] https://www.spreadthesign.com/
[6] https://platform.toloka.ai/
[7] https://elementary.center/




(a) Selection: The user highlights a word to trigger the extension icon.




(b) Translation: The overlay displays the LSFB video from the public or private lexicons.




(c) Recording: Users can record missing signs to add to their personal lexicons.

Figure 1: Typical usage scenario of MOSI. (a) Users select a word on any webpage, (b) consult available translations from the layered lexicons, and (c) add their own signs to their lexicons.

## 3. Our Tool: *Mot-Signe*

This section introduces *Mot-Signe* (MOSI), our platform designed to bridge French and French Belgian Sign Language (LSFB). We detail the motivations driving the tool's evolution from a simple lexicon to a collaborative documentation platform. We then discuss the design and key features, specifically focusing on how the architecture prioritizes user privacy through private lexicons, while enabling the enrichment of public LSFB corpora.

### 3.1. Goals and Motivations

The development of MOSI was prompted by a direct request from the deaf community in the French-speaking part of Belgium. Our research team was contacted by stakeholders who identified a severe lack of integrated tools for signers. Citing the utility of *Elix* for French SL (LSF) and *AuslanSignbank* [8] for the Australian SL, they expressed a strong interest in a similar solution that seamlessly provides LSFB equivalents for French words directly within a web browser. The primary goal is, therefore, to foster autonomous learning and digital inclusion, by reducing reliance on external human assistance for understanding daily web content.

This collaboration presented a secondary opportunity for the research community: the large-scale documentation of the language. Existing SL datasets, including LSFB, often lack specialized vocabulary or natural variations for given concepts. Drawing inspiration from collaborative initiatives like *signdict* [9], MOSI aims to fill this gap by empowering the community to document signs that are currently absent from static lexicons. This supports two research objectives: providing signers with a richer variety of expressions for a single word, and establishing a sustainable data collection system. This system is essential for enabling linguistic research into sign variations often missing from current corpora, as well as for improving future dictionary-based AI query systems that rely on comprehensive lexical coverage.

This study builds upon a preliminary prototype developed in a master's thesis (Ye and Sow, 2023) and focuses on the creation of a sustainable data collection system for LSFB, designed to be easily transposed to other SLs in the long term.

### 3.2. Features Overview

MOSI operates as a unified platform accessible via a web extension (compatible with Chrome and Firefox) and a dedicated website (`mot-signe.be`). The core functionality revolves around a privacy-first architecture that distinguishes between personal usage, trusted sharing with private lexicons, and public lexicons that every user can access.

#### Query Signs From a Selected Word

Users can query for LSFB translations, either directly on the website or via the web extension from any external web page. Upon selecting or typing a French word (see Figure 1a), the user triggers a query that aggregates results from three distinct sources (see Figure 1b):

1. **Public Lexicon:** Standardized signs managed by linguistic experts.

[8] https://auslan.org.au/dictionary/
[9] https://signdict.org/

2. **Shared Lexicons:** Signs contained in private groups (lexicons) the user has joined via invitation codes.

3. **Personal Lexicon:** Signs recorded by the user for their own usage.

This approach ensures that users can access authoritative vocabulary while simultaneously retrieving specialized or informal signs relevant to their specific context (e.g., a biology class or a family circle) without exposing those signs to the public.

#### Contribution and Private Sharing

If a translation is missing or if a user wishes to document a variation, they can propose an alternative translation directly through the extension interface (see Figure 1c). This process requires the user to be logged in. The tool opens a webcam stream, allowing the user to record, review, and confirm their sign.

By default, these recordings are **private**. To share them, users must create or join a **Lexicon**. A single sign can be associated with multiple lexicons. Each lexicon is managed by **administrators** who control membership via invitation codes and moderate the signs shared within that specific group. This ensures that signs are only visible to intended peers and that the group maintains its specific focus. The companion website serves as the management hub where users can organize their lexicons, manage their profile, and visualize their contributions. In addition, users can download the signs as GIF animations to use them in external tools (e.g., slideshows in teaching environments or interactive websites).

#### Expert Curation and Documentation

The platform facilitates a unique workflow for language documentation. Linguistic experts and administrators have access to the pool of recorded signs to identify missing vocabulary or emerging variations. Unlike traditional crowdsourcing where user videos are simply “approved”, MOSI uses user contributions as field data. Experts analyze these contributions and, when appropriate, produce high-quality, standardized studio recordings to enrich the **Public Lexicons**. This ensures that public resources maintain high production standards while being driven by community usage.

### 3.3. Data Governance and Corpus Enrichment

While the primary interface serves the immediate needs of the deaf community, the underlying architecture is designed to address the scarcity of digital LSFB resources. The data collection process is strictly governed by ethical protocols and GDPR compliance, ensuring that the goal of documentation never overrides user privacy.

#### Integration with Existing Corpora

Current LSFB datasets are predominantly static, studio-recorded corpora created by linguistic research groups: LSFB Corpus (Meurant, 2015) and LSFB Dictionary (Sonnemans, 2026). While high in quality, they often lack the "in-the-wild" diversity of natural signing and fail to capture neologisms or regional variations rapidly. MOSI acts as a dynamic extension to these existing corpora. It does not aim to replace them but to identify gaps (missing signs) and variations (accents, synonyms) through community usage.

#### Privacy and Anonymization Strategy

To ensure GDPR compliance and build trust with the community, the platform integrates strict privacy protocols. Upon registration, users are presented with clear guidelines regarding data usage to ensure informed consent. Users retain full sovereignty over their contributions, including the right to withdraw their data at any time. To guarantee anonymity while maintaining research utility, we employ a two-tier strategy.

**Curatorial Anonymization** Public exposure of user data is prevented by design. User recordings act solely as field data for linguistic experts. Validated signs are **re-recorded** by designated signers in a controlled studio environment for the public lexicon. This intrinsically anonymizes the source, as the public never views the original contributor’s video.

**Computational Anonymization** For downstream machine learning tasks and private research analysis, we mitigate re-identification risks by discarding pixel data in favor of vector representations. We utilize *MediaPipe* (Lugaresi et al., 2019) to extract pose landmarks (skeletons) from the videos. This removes sensitive visual identifiers (face, background) while preserving the linguistic motion required for analysis (Saunders et al., 2021).

## 4. User Study

Generally, the adoption of a tool depends on several criteria (e.g., ease of use, proposed features, etc.), the amount of data captured is correlated with its adoption by final users. To understand whether our tool meets these criteria and how it fits into users’ daily practices, a longitudinal user study was

conducted. This study had two objectives: (i) to collect qualitative data on the use of the application; and (ii) to assess how the application fits into users' daily practices.

### 4.1. Methodology

During this study, nine users, all proficient in LSFB, were asked to fill out diary entries over a period of time, allowing us to identify their habits and follow their journey while using the application. Participants were recruited in April 2024 during the launch event of the application at a bilingual (hearing and deaf) school. An account with access to all the features was created for all attendees, and the features were presented. The study was conducted over 10 opening days, with one entry per day. It began one week after the creation of the users' credentials. This allowed them to become more familiar with the application before starting to track their habits. The study was conducted remotely; therefore, instructions on how to fill out the diary were sent by email, and a reminder was sent every day at 4 p.m. Many questions of different types (features, accomplished tasks, difficulties faced, issues, etc.) were asked of the users. These questions allowed us to track the usage of the various features and the issues encountered by the users. After the study, a debriefing was conducted in person with each user to clarify some points in their diary and gather their opinion about the application.

### 4.2. Usage of the Application

The entries of the diary allow identifying which features are mainly used and how users adopted the application. Table 1 summarizes the features mentioned by the users in their diary entries along with the number of occurrences of each feature in the texts.

Table 1: List of the features mentioned in the diaries along with the number of times they are mentioned.

| Feature | Occurrences |
|---|---|
| Vocabulary search | 35 |
| Record vocabulary | 8 |
| Download sign GIF | 4 |
| Share signs in lexicons | 4 |

As expected, the *vocabulary search* (either using the website or the extension) is, by far, the most popular feature as it eases the search for vocabulary in texts. It was the initial feature requested by the stakeholders. Then, the users mentioned the *record sign* feature several times. It was often mentioned as teachers prepared independent reading exercises where pupils learned to read on their own. Two users explained that they selected the text the student would be reading and they made sure that all the vocabulary was already in the lexicon by recording the missing signs. They also prepared a list of synonyms as, currently, the search in the lexicon does not handle synonym search. During the reading class, deaf students could access a computer with MOSI available to search for the vocabulary. The teachers considered it as a way of increasing their students' autonomy.

Other users shared that they used the *record sign* feature to build their own vocabulary for exercises. Once recorded, they downloaded the corresponding GIF version of their recorded signs. It enabled them to embed the signs in their course material (e.g., flashcards or slides). This also explains why the *download sign GIF* feature is often mentioned in the diary study.

Finally, some users mentioned that they browsed the *lexicon pages* of the website to see the vocabulary shared by their colleagues and to validate some signs.

### 4.3. Users Feedback

User feedback is divided into two main categories: disincentive and incentive factors.

Disincentive factors refer to elements that negatively impact the user experience. In this category, the most frequent problems concern the vocabulary search. Indeed, the lexicon is too limited and some words too old or too specific are not available (7 reports). The current search engine is quite simple and could be improved. Therefore, when a user searches a conjugated verb, the application is not able to find the infinitive form (3 reports). Another consequence is the fact that the synonyms or homonyms are not detected by the search engine (4 reports). Sometimes, the application returns incorrect signs either because of similar spellings, annotation mistakes or because of the specific context of the sentence (3 reports).

On the other side, incentive factors are elements that make the application appealing to the user and would lead to a better retention. First, the fact that multiple public LSFB lexicons are merged and accessible directly via the extension is appreciated and saves time (reported 3 times). The ability to share vocabulary with other people is also praised (3 reports). The ability to download GIF images of the signs eases the work of teachers (3 reports). The integration of the extension in every web page also allows users to easily discover new signs (2 reports).

## 5. Community Adoption

Following the initial user study and subsequent iterative refinements, MOSI was released publicly to the broader deaf community in French-speaking Belgium, i.e., the whole LSFB community. This phase marks the transition from controlled testing to "in-the-wild" usage, allowing us to observe how the tool performs with real users in diverse, uncontrolled environments.

### 5.1. Quantitative Analysis

Since the public release, the platform has attracted 224 registered users, in addition to a broader base of unregistered users who can query the lexicons without an account but cannot record signs. As illustrated in Figure 2, the volume of contributions has shown a steady upward trend, indicating sustained user engagement rather than a momentary spike at launch. A promotional event held in April 2025 led to a notable surge in contributions, as visible in the figure.

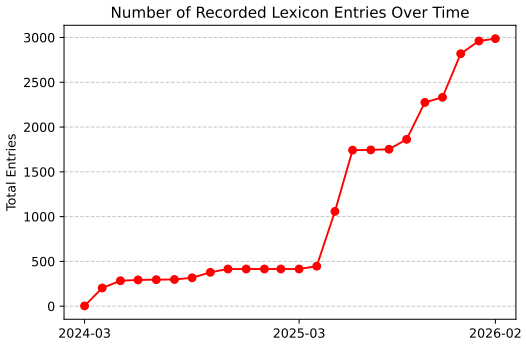



Figure 2: Cumulative evolution of the number of signs recorded on the platform since the public release. The sudden increase corresponds to a promotional event in April 2025.

**Data Collection Volume** To date, a total of $3,004$ distinct signs, identified through ID-glossing, have been recorded and added to private or shared lexicons. This collection volume is competitive with traditional studio-based recording sessions and guarantees a larger diversity in occurrences, validating the efficiency of the crowdsourcing approach.

**Novel Entries** Among the recordings, 2,501 ($83.26\%$) correspond to concepts that were previously absent from existing LSFB dictionaries. They primarily cover specialized vocabulary (e.g., technical or academic terms) and neologisms. Such entries would be difficult to collect with traditional approaches.

**Variations** The remaining 503 recordings ($16.74\%$) offer alternative signs for existing entries, capturing regional accents, generational differences, or context-specific variations that are crucial for a comprehensive understanding of the language. Additionally, users can associate multiple tags with each sign, providing metadata that helps contextualize the usage of a given variant (e.g., regional origin or thematic domain).

**Expert Curation and Taxonomy** Of the collected dataset, $22.7\%$ of the signs have been reviewed by linguistic experts in order to be included in the public LSFB lexical database. This proportion reflects the time-intensive nature of the curation process, as the experts involved are also the official maintainers and annotators of the LSFB Corpus. We expect this percentage to increase substantially as the curation effort continues.

### 5.2. Community Impact

Beyond the raw metrics, the qualitative reception from the LSFB community has been highly positive. Feedback collected through the platform's support channels indicates a strong sense of ownership among deaf users, who value the ability to directly influence the documentation of their language.

**Educational Usage** LSFB learners are also well targeted by MOSI. Usage logs suggest that those users go beyond passive lookups and use the tool as a verification mechanism, comparing their own signings with the Community contributions. This dual usage, as both a reference tool and a practice platform, highlights the role of MOSI in supporting continuous learning. Due to its diverse usage, we assume that other benefits remain unknown.

## 6. Conclusion

This work demonstrates the viability of sustainable, long-term Sign Language (SL) data collection by prioritizing the immediate needs of the Deaf community. By designing MOSI as a collaborative tool, we successfully flipped the traditional crowdsourcing paradigm: data collection became a natural byproduct of a tool built for daily utility. This collaborative approach proved highly effective and grounded in the principle of "nothing about us without us", a Community urge that is still not respected enough (De Meulder, 2025). It yielded thousands of new SL recordings while directly supporting autonomous learning and digital inclusion for its users.

From the community's perspective, the primary contribution is a living LSFB platform where users can both consult and enrich the available vocabulary, enabling the documentation and sharing of

a greater variety of signs, including regional and specialized variations.

From a computational perspective, this process yields new training data, particularly for signs and concepts that remain undocumented in existing corpora.

**Limitations** Despite these promising results, this study and the current iteration of the platform have notable limitations. First, the application focuses exclusively on isolated signs. While such data is valuable for dictionary-based search engines (Fink et al., 2023b) and training Isolated Sign Language Recognition (ISLR) models, it is insufficient for advancing more complex downstream tasks. Second, expanding future evaluations to include a larger proportion of native deaf users is necessary to draw more generalized conclusions about the community-wide impact of the tool.

**Future Works** Moving forward, several avenues of research and development will enhance the utility of the platform and the quality of the collected data. Linguistically, user feedback highlighted a strong need for semantic features, such as the ability to explicitly link signs (e.g., synonyms, homonyms) and attach rich metadata (e.g., regionalisms, thematic categories). Technically, we plan to extend the architecture to support the recording and processing of continuous SL sentences. Finally, to ensure long-term sustainability, we open source the codebase[10], and explore its transferability to other under-documented sign languages worldwide.

## 7. Acknowledgments

We would like to express special thanks to St. Marie School in Namur for their participation in the user study and their precious feedback.

This work was funded by both the Walloon region with a Ph.D. grant from FRIA (F.R.S.-FNRS) and by the SPWR under grant n°2010235 - ARIAC by DIGITALWALLONIA4.AI.

## 8. Bibliographical References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BBC-Oxford British Sign Language dataset.

Katherine Atwell, Danielle Bragg, and Malihe Alikhani. 2024. Studying and mitigating biases in sign language understanding models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 268–283.

Danielle Bragg, Abraham Glasser, Fyodor Minakov, Naomi Caselli, and William Thies. 2022. Exploring collection of sign language videos through crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24.

Kearsy Cormier, Jordan Fenlon, Trevor Johnston, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam, and Bencie Woll. 2012. From corpus to lexical database to online dictionary: Issues in annotation of the BSL corpus and the development of BSL SignBank. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 7–12, Istanbul, Turkey. European Language Resources Association (ELRA).

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1–27.

Maartje De Meulder. 2025. Deaf in AI: AI language technologies and the erosion of linguistic rights. *Language and Law/Linguagem e Direito*, 12(1).

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2024. ASL Citizen: A community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36.

Shahriar Elahi Dhruvo, Mohammad Akhlaqur Rahman, Manash Kumar Mandal, Md. Istiak Hossain Shihab, A. A. Noman Ansary, Kaneez Fatema Shithi, Sanjida Khanom, Rabeya Akter, Safaeid Hossain Arib, M. N. Ansary, Sazia Mehnaz, Rezwana Sultana, Sejuti Rahman, Sayma Sultana Chowdhury, Sabbir Ahmed Chowdhury, Farig Sadeque, and Asif Sushmit. 2023. Bornil: An open-source sign language data crowdsourcing platform for ai enabled dialect-agnostic communication.

Jerome Fink, Mathieu De Coster, Joni Dambre, and Benoît Frénay. 2023a. Trends and challenges for sign language recognition with machine learning. In *ESANN 2023: 31st European Symposium*

[10] https://github.com/LSFB-Team/mot-signe

*on Artificial Neural Networks, Computational Intelligence and Machine Learning 2023*, pages 561–570.

Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. 2021. LSFB-CONT and LSFB-ISOL: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Jérôme Fink, Pierre Poitier, Maxime André, Loup Meurice, Benoît Frénay, Anthony Cleve, Bruno Dumas, and Laurence Meurant. 2023b. Sign Language-to-Text dictionary with lightweight transformer models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, volume 9, pages 3785–3789, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hector Garcia-Molina, Manas Joglekar, Adam Marcus, Aditya Parameswaran, and Vasilis Verroios. 2016. Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 28(4):901–911.

Mokter Hossain and Ilkka Kauranen. 2015. Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal*, 8(1):2–22.

Trevor Johnston. 2009. Creating a corpus of AUSLAN within an Australian national corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*, pages 87–95. Cascadilla Proceedings Project.

Alexander Kapitanov, Kvanchiani Karina, Alexander Nagaev, and Petrova Elizaveta. 2023. Slovo: Russian Sign Language dataset. In *International Conference on Computer Vision Systems*, pages 63–73. Springer.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines.

Laurence Meurant. 2015. Corpus LSFB: Un corpus informatisé en libre accès de vidéos et d'annotations de la langue des signes de belgique francophone (LSFB).

Jay M Patel and Jay M Patel. 2020. Introduction to common crawl datasets. *Getting structured data from the internet: running web crawlers/scrapers on a big data production scale*, pages 277–324.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.

Katrin Renz, Nicolaj C Stache, Samuel Albanie, and Gül Varol. 2021. Sign language segmentation with temporal convolutional networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139. IEEE.

Nikolaus Riemer Kankkonen, Thomas Björkstrand, Johanna Mesch, and Carl Börstell. 2018. Crowdsourcing for the Swedish Sign Language dictionary. In *sign-lang@ LREC 2018*, pages 171–176, Miyazaki, Japan. European Language Resources Association (ELRA), European Language Resources Association (ELRA).

Benjamin Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. ANONYSIGN: Novel human appearance synthesis for sign language video anonymisation skeleton pose original image novel human appearances fig. 1: Novel appearance human synthesis examples generated by anonysign. In *IEEE International Conference on Automatic Face and Gesture Recognition 2021*. IEEE.

Bruno Sonnemans. 2026. Dictionnaire de LSFB en ligne. `https://dico.lsfb.be/`. Accessed: 2026-03-01.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, Daksh Sehgal, Saad Hassan, Bill Neubauer, Sofia Vempala, Alec Tan, Jocelyn Heath, Unnathi Kumar, Priyanka Mosur, Tavenner Hall, Rajandeep Singh, Christopher Cui, Glenn Cameron, Sohier Dane, and Garrett Tanzer. 2023. PopSign ASL v1.0: An Isolated American Sign Language dataset collected via smartphones. In *Advances in Neural Information Processing Systems*, volume 36, pages 184–196.

Garrett Tanzer and Biao Zhang. 2024. YouTube-SL-25: A large-scale, open-domain multilingual sign language parallel corpus.

Innocent Ye and Babacar Sow. 2023. Création d'un outil d'aide à l'apprentissage du français par la

lecture pour les élèves sourds et malentendants. Mémoire de master, Université à préciser, June. Master en informatique, spécialité Data Science.

Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2011. A survey of crowdsourcing systems. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE.

# Effect of Data Augmentation with Multi-View Perspectives of Signers on the DGS-Fabeln-1 Dataset

**Fabian Renner, Daksitha Withanage Don, Elisabeth André, Cristina Luna-Jiménez**
Chair for Human-Centered Artificial Intelligence,
University of Augsburg, Augsburg, Germany
{fabian.renner, daksitha.withanage.don, elisabeth.andre, cristina.luna.jimenez}@uni-a.de

**Abstract**

Sign languages constitute the principal form of communication for deaf communities across the globe. Nevertheless, the development of reliable Continuous Sign Language Translation (CSLT) systems is constrained by the lack of sufficient data and models able to handle spatio-temporal information. In this article, we explore the effect of adding multiview perspectives of the signer to the training set as data augmentation using the UniSign framework for the DGS-Fabeln-1 dataset. Our results reveal that increasing dataset size and using multiple camera perspectives significantly improve performance, with the best configurations achieving BLEU-4 scores of 4.20%. These results provide a competitive baseline for the DGS-Fabeln-1 dataset and guidance for further optimizations of CSLT systems.

**Keywords:** Sign Language Translation, Transformers, Transfer-Learning, Multi-view perspectives data augmentation

## 1. Introduction

By 2050, an estimated 2.5 billion people worldwide will experience some degree of hearing loss (Chadha et al., 2021). For many of these individuals, sign language serves as their primary mode of communication (Macht and Steinbach, 2019). However, due to fundamental differences in linguistic structure between sign languages and spoken languages, effective communication between user groups remains a persistent challenge (Macht and Steinbach, 2019), relying always on the availability of interpreters. Due to the existing research challenges and the motivation to develop technology that can foster independent communication, Sign Language Translation (SLT) is an active research topic.

Traditionally, SLT has followed the Isolated Sign Language Recognition (ISLR) approach, in which glosses are recognized from short fragments of videos. Although ISLR systems could be effective in reduced vocabulary scenarios (Luna-Jiménez et al., 2025) and useful for sign language spotting (Mercanoglu Sincan and Bowden, 2025); more recently, the SLT research community is driving their efforts to investigate Continuous Sign Language Translation (CSLT). This new paradigm is more natural to the translation task, receiving videos in Sign Language and translating them into Spoken Languages. Additionally, the absence of tedious gloss annotation saves human effort since only alignment between segments is required, making recording and data annotation more feasible for researching with large visual-language models (Li et al., 2024).

For this reason, the present article investigates the optimization of a Transformer-based model for SLT from DGS (German Sign Language) to German. Therefore, the contributions of this article are threefold:

- Considering the creative and varied nature of the narrative domain, we analyzed the impact of dataset size by comparing training configurations of six versus seven fairy tales. This assessment aimed to determine if incorporating additional, distinct story content improves the ability of the model to generalize and translate sentences across different narrative structures.
- Given the success of transfer learning in multiple tasks (Esteban-Romero et al., 2024; Luna-Jiménez et al., 2024), we investigated the transferability of knowledge within the UniSign transformer architecture to the DGS-Fabeln-1 dataset. This was conducted by comparing models initialized with random weights against models initialized with the weights learned from multilingual datasets.
- Additionally, since training large-scale models with more samples usually results in higher performance, we compared the effect of including only frontal camera perspectives against incorporating all eight synchronized perspectives available in the DGS-Fabeln-1 dataset.

To our knowledge, this is the first work proposing a competitive baseline for the DGS-Fabeln-1 for further investigations in SLT in a creative domain, as it is fairy tales.

The remaining part of the article is structured as follows: Section 2 provides foundational knowledge



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 429–437
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

on existing datasets annotated in DGS, and models employing Transformer-based architectures for performing SLT. Section 3 details the applied methodology, including the dataset preparation, feature extraction, and the experimental setup. Section 4 and 5 evaluate the experimental results by analyzing the influence of the distinct parameters. Finally, Section 6 outlines conclusions and future work.

## 2. Related Works

In this section, we introduce the datasets available for solving SLT in DGS. Next, we summarize existing models in the literature, focusing on transformers and Language Models-based architectures.

### 2.1. Datasets in German Sign Language

German Sign Language (Deutsche Gebärdensprache, DGS) is the main sign language employed by the Deaf community in Germany (Bross, 2020). DGS is expressed and perceived through the visual-gestural modality, encoding meaning in manual and non-manual articulators, including hand configurations and positions, facial expressions, head orientation, and body posture (Bross, 2020). Not only do manuals and non-manuals play a crucial role in Sign Language, but also the signing space.

Despite the relevance of the location of different events in the space, most of the available datasets for researching ISLR and CSLT only contain the frontal perspective of the signers. For example, RWTH-PHOENIX Weather 2014 T (Camgoz et al., 2018), one of the most used corpora by the machine learning community, only contains frontal perspectives of a signer interpreting broadcast news. Similarly, the AVASAG dataset (Bernhard et al., 2022) has only frontal-view videos of an interpreter signing sentences related to train station scenarios and wearing a motion tracking suit. From a linguistic point of view, one of the most popular datasets is the DGS-Korpus (Konrad et al., 2020), with more than 50 hours of videos of dialogues recorded with a frontal perspective and a profile perspective of the signers, although this profile perspective has the moderator plus the signers which would require further pre-processing to detect and extract the profile camera view of each individual. More recently, Avramidis et al. (2025) collected more than 900 hours of DGS in the multilingual TUB corpus covering up to 12 languages with 1.3 million subtitle segments containing 14 million tokens. The dataset covers topics related to politics, news, education, and social content. However, as in previous datasets, most videos are only available from the frontal perspective of the signer.

Finally, to support research in DGS and study the impact of having videos with different perspectives of the signer, the DGS-Fabeln-1 dataset (Nunnari et al., 2024) was introduced as a publicly available corpus for the linguistic and computer science community. It presents a native DGS signer retelling a set of seven German fairy tales, recorded from seven distinct camera angles as well as an additional backchannel perspective, to capture detailed spatial movements and facial expressions. The dataset has a total duration of 1 hour and 32 minutes, aligned with 1,428 written German sentences, making it a valuable resource for training and evaluating SLT systems.

### 2.2. Transformer Architectures for SLT

Since its introduction, Transformer architecture has been the leading model for sequence-to-sequence tasks, especially in natural language processing applications like machine translation. Its key innovation, the self-attention mechanism, allows the model to assess the relevance of input elements regardless of their position, making it highly effective at capturing long-term dependencies. Unlike recurrent neural networks, Transformers process all elements in parallel, which greatly improves training efficiency. The architecture typically follows an Encoder-Decoder structure: the Encoder transforms input sequences into contextual representations, which the Decoder uses to generate the output (Vaswani et al., 2017).

In SLT, this setup is particularly effective due to the sequential nature of both visual sign streams and textual output. The Encoder processes spatial and temporal patterns from sign language videos, such as hand positions and facial expressions, as well as their timing, learning rich, context-aware representations. The Decoder then uses these context-rich embeddings to produce fluent written or spoken translations in a differnt language (Vaswani et al., 2017; Duarte et al., 2021; Damdoo and Kumar, 2025; Cihan Camgöz et al., 2020).

From this family of models, UniSign (Li et al., 2024) is a Transformer-based architecture designed for sign language understanding that unifies several core tasks of this domain within a single framework: ISLR, which focuses on classifying individual signs; continuous sign language recognition (CSLR), which learns the alignment between sign sequences and their corresponding glosses; and SLT, which generates natural language descriptions from signed input. It leverages generative pretraining followed by task-specific fine-tuning, which enables effective knowledge transfer and optimization across tasks. UniSign is designed to be modular and extensible, allowing researchers to explore diverse input configurations and pretraining strategies across a range of sign language understanding scenarios.

Additionally, recent studies have further emphasized the potential of leveraging diverse camera angles; for instance, Ranum et al. (2024) demonstrated that incorporating multi-view data during training significantly enhances the model's ability to learn robust spatial representations, even when inference is restricted to a single frontal perspective, highlighting the advantages of this strategy as data augmentation for improving the performance of the model.

## 3. Methodology

This section describes the methodology used to adapt and evaluate the UniSign Transformer for DGS translation using the DGS-Fabeln-1 dataset. The process includes preparing the dataset by extracting pose landmarks and training multiple models under different configurations.

### 3.1. Dataset Preparation

The dataset employed in this article is the DGS-Fabeln-1 dataset. The corpus comprises 573 video segments of 7 fairy tales. Each sentence was recorded simultaneously from eight distinct camera perspectives. A central tablet provided a frontal view, while six additional tablets were positioned in a semi-circle at angles of 30°, 60°, and 90° on both sides of the signer. Finally, an eighth overhead camera captured the director's back-channeling, documenting the full interaction between the signer and the interlocutor.

For our experiments, we created three different configurations of the training and validation sets of the dataset according to the effect to explore in each of them:

1. The first dataset configuration contains six fairy tales and exclusively uses the front-view camera perspective. For the analysis, the samples are randomly split into an 80:10:10 train:validation:test ratios. This resulted in 338 videos for the training split, 42 videos for the validation split, and 43 videos for the test split. The last fairy tale ("Snow White") was left out of the test sets for evaluating data augmentation strategies and comparing the same 42 videos of the test set across experiments.

2. The second configuration also uses only the frontal-view videos, but includes the seventh fairy tales samples in the training and validation sets. In this case, the dataset resulted in 476 videos for the training set, 52 videos in the validation set and 43 in the test set.

3. Finally, the third configuration uses seven fairy tales and all perspectives, following the same allocation adopted in the second dataset configuration. In this way, if a sentence was in the train or validation split in the second configuration, it is also in the same split in this third configuration, with all available video perspectives for that sentence included. This ensures that all perspectives of a sentence always appear in the same subset. In these experiments, the train split contains 3,749, the validation split has 409 videos, and the test split has 43 videos.

As can be observed, the test split is identical across all the dataset configurations to ensure comparability of results in each experiment.

### 3.2. Landmark Extraction

For the feature extraction, we followed the UniSign pipeline (Li et al., 2024) to compute pose landmarks from the raw videos using OpenPose (Cao et al., 2021). OpenPose is an open-source tool that detects human keypoints from images and videos. This process produced frame-wise skeleton data capturing 134 keypoints of the hands, body and face.

### 3.3. UniSign Transformer

Figure 1 shows the architecture of the UniSign model. The UniSign model is capable of processing skeletal pose landmarks either independently or in combination with RGB video data, utilizing a multimodal fusion strategy to learn fine-grained spatio-temporal representations. The model contains several encoders that pass each of the features extracted from different body parts (i.e. hands, pose or face) to a projection layer and a three-layers Spatial Graph Convolutional Network (GCN) per modality. The outputs are then fed into the temporal encoders that are Spatio-Temporal Graph Convolutional Networks (ST-GCN) and the outputs are then combined with an average pooling to maintain the dimension of the embedding space of the multilingual language model mT5. Subsequently, the LM receives the resultant tokens to predict the most probable sentence represented in the video.

The vision encoder consists of an EfficientNet-B0 that receives the images of the hand for the experiments with RGB videos and fusion them through the Prior-Guided Fusion (PGF) module that combines the landmarks information from the hands with the visual features extracted from the hands of the original videos via an attention mechanism.

### 3.4. Evaluation Metrics

For comparing the model's performance across experiments, we used common metrics employed in

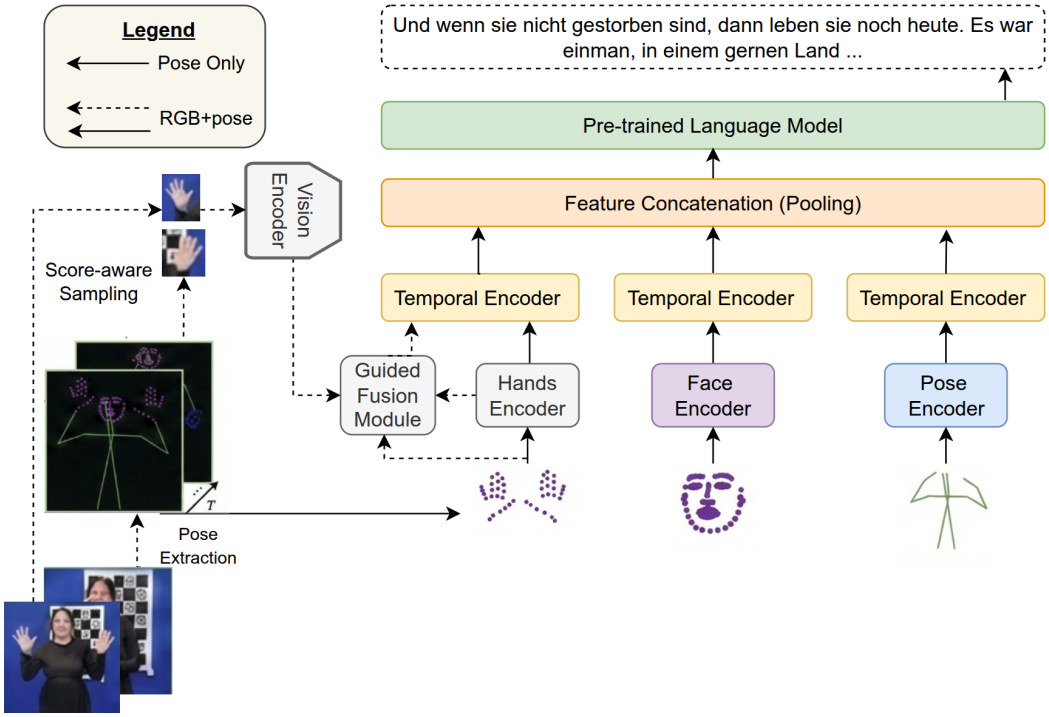



Figure 1: Overview of the UniSign architecture.

previous literature in SLT and MT (Li et al., 2024; Duarte et al., 2021; Camgoz et al., 2018; De Coster et al., 2023; Qin et al., 2025): BLEU-1 to BLEU-4 (Papineni et al., 2002), and ROUGE-L (ROUGE for short in the results tables) (Lin, 2004). These metrics are automatically computed within the UniSign framework by comparing the generated written German sentences against the corresponding ground-truth references. BLEU-1 assesses lexical accuracy through unigram precision, defined as the ratio of matching words to the total word count in the generated sentence, irrespective of their position. Conversely, BLEU-4 examines the alignment of four-word sequences (4-grams), providing an evaluation of the syntactic fluency. ROUGE-L quantifies structural overlap through the Longest Common Subsequence (LCS). This metric identifies the longest shared sequence of words that maintains the same relative order between the hypothesis and the reference. Unlike the fixed n-gram constraints in BLEU, the components of an LCS do not need to be contiguous. This characteristic allows the metric to accommodate variations in sentence structure while still rewarding correct word-order patterns. Additionally, to evaluate the efficiency in training times of each configuration, we included the training time. Finally, we added an auxiliary metric to measure the number of unique generated lemmas in inference time. Each lemma is the basic dictionary form of a word or sign, without grammatical endings or variations in form (e.g., “run” as the lemma for “running,” “ran,” and “runs”). Using lemmas in evaluation allows different word forms of the same meaning to be matched, making the metric less sensitive to inflectional variation. This metric also provides insight into the diversity of vocabulary. To operationalize this, we used the spaCy library (de_core_news_sm model) to process the predicted sentences and ground-truth texts. The process involved word-level tokenization, filtering for alphabetic characters and using the German language model to find the dictionary base form (lemma) of each word (e.g., converting "ging" to "gehen") in lowercase. The unique base forms were then aggregated to compare the overall vocabulary size of the translations and reference sentences. An illustrative application of these metrics on a sample sentence pair is provided in Table 1.

## 4. Experiments

In this section, we introduce the main experiments performed.

### 4.1. Experimental Set-Ups

To optimize our model, we conducted controlled experiments varying parameters while keeping other conditions constant. Specifically, we explored:

- *Effect of the Dataset size*: 6 vs. 7 fairy tales. As the domain is quite creative and fairy tales normally differ from each other except in certain sentences at the beginning and at the end,

| | | |
|---|---|---|
| **Reference (German):** "Der schlaue Fuchs bemerkte den Raben mit dem Käse auf dem Ast." **Reference (English Translation):** "The clever fox noticed the raven with the cheese on the branch." | | |
| **Prediction (German):** "Fuchs sehen Raben mit Käse auf Ast." **Prediction (English Translation):** "Fox see ravens with cheese on a branch." | | |
| **Metric** | **Value** | **Calculation Basis** |
| BLEU-1 | 85.71% | 6 matches / 7 words in prediction. Matches: {Fuchs, Raben, mit, Käse, auf, Ast}. |
| BLEU-4 | 0.00% | 0 matches. No sequence of four consecutive words matches the reference exactly. |
| ROUGE-L | 63.16% | F1-score derived from LCS length 6. Precision: 6/7 (prediction), Recall: 6/12 (reference). |
| Lemmas (Ref/Pred) | 9 / 7 | Unique lemmas Ref: {der, schlau, fuchs, bemerken, rabe, mit, käse, auf, ast}. Unique lemmas Pred: {fuchs, sehen, rabe, mit, käse, auf, ast}. |

Table 1: Illustrative metric calculation for a sample sentence pair.

we evaluated whether adding data from another fairy tale could improve the performance of the model for predicting sentences of the other fairy tales.

- *Effect of transfer learning*: Weights were either randomly initialized or initialized from the best pre-trained mT5 checkpoint provided by UniSign, which was derived from the multilingual mT5 Base model trained on large-scale multilingual text data.

- *Data Augmentation with additional Camera perspectives*: Front-only vs. All perspectives for training and validation. These experiments aim to evaluate whether adding data augmentation (i.e. training with more perspectives) could result in a final gain in the performance of the model.

For each experiment, we saved the best checkpoint based on the top BLEU-4 score in the test set, which was used for comparisons in the result section.

## 5. Results

In this section, we evaluate the results of the experiments described before. For this purpose, we first analyze the influence of each varied parameter individually, followed by a general overview. Specifically, we first conducted experiments using only the first configuration of the dataset with 6 fairy tales and the standard frame rate as a baseline. Subsequently, we kept the number of epochs constant and transferred the weights between models and then we evaluated the use of more camera perspectives as data augmentation to compare the performance of the model in each scenario.

### 5.1. Baseline on the Dataset size

In order to establish a first baseline, initial comparisons were conducted on the number of epochs from 50 to 100, which caused mostly no changes in the scores but increased the training times; for that reason, further experiments were performed only with 50 epochs. A complete set of all the performed experiments is added as supplementary material [1].

Regarding the experiments related to the dataset size, we employed a batch size of 4 with 50 epochs, random weights initialization and the regular number of frames of the original dataframe, corresponding with the first and second configurations of the datasets presented in Section 3.1. As can be observed in Table 2, the scores for experiments with 7 fairy tales were mostly significantly higher across all metrics than in the trials with only 6 fairy tales. For this specific configuration, an increment of 1.19 percentage points was obtained in BLEU-4 using the seven fairy tales in the training and validation sets, resulting in a generative model with a larger vocabulary, as can be observed in the lemmas. These results indicate that even in 'creative' topics in which sentences could differ from one fairy tale to another, adding more fairy tales to the training set results in higher performance of the models for predicting next sentences in other fairy tales. For this reason, in the next experiments, 7 fairy tales were kept for the training and validation.

| Tales | BLEU-1 | BLEU-4 | ROUGE | Lemmas |
|---|---|---|---|---|
| 6 | 13.66% | 0.97% | 13.73% | 9 |
| 7 | 18.17% | 2.16% | 15.80% | 16 |

Table 2: Results on the evaluation of the number of fairy tales used for training.

[1] https://cloud.hcai.eu/s/jJo2pdKEnZXqzLJ

### 5.2. Effect of Transfer Learning

As previous literature in deep learning has evidenced in different tasks, pre-training normally improves performance across similar tasks. Accordingly, we evaluated the previous top configuration trained from scratch with random weights versus applying Transfer Learning (TL) using pretrained weights, as shown in Table 3.

Overall, the training times were similar and pretrained weights delivered superior performance across the majority of experiments, with a particularly strong advantage for the Pose + RGB input type. This indicates that the pre-trained model, having learned foundational features from a large, diverse dataset, generalizes better and fine-tunes more effectively. Based on these findings, we concluded to use pre-trained weights for further model optimization.

| Weights | BLEU-1 | BLEU-4 | ROUGE | Lemmas |
|---|---|---|---|---|
| Scratch | 18.17% | 2.16% | 15.80% | 16 |
| TL | 17.88% | 4.15% | 16.09% | 28 |

Table 3: Results on Transfer Learning (TL).

### 5.3. Data Augmentation with additional Camera perspectives

Considering that different perspectives of the camera represent the same meaning when signing, another goal of the experiments was to explore whether adding other perspectives from the same sentences and person could help to improve the performance of the model when doing predictions on the frontal videos test set, following the third dataset configuration described in Section 3.1.

For the experiments with the Pose + RGB input type, no clear influence of the video perspective could be determined, especially when exploring the same model with different batch sizes. As shown in Table 4, results were generally similar but the training time increased significantly in trials using all perspectives by a factor of about three to four times. Given that in some cases, such as with the batch size of 8, the use of more perspectives increased BLEU-4 by +2.06 percentage points, for further optimizations, we continued using all the perspectives.

| Batch | Persp. | BLEU-1 | BLEU-4 | ROUGE | Lemmas |
|---|---|---|---|---|---|
| 2 | Frontal | 17.88% | 4.15% | 16.09% | - |
| 2 | All | 17.88% | 4.20% | **16.12%** | - |
| 4 | Frontal | 17.88% | 4.15% | 16.09% | 28 |
| 4 | All | 17.88% | 4.20% | 16.12% | 43 |
| 8 | Frontal | **18.32%** | 2.15% | 11.47% | 12 |
| 8 | All | 15.21% | **4.21%** | 11.43% | 18 |

Table 4: Results on Camera Perspectives as data augmentation.

## 6. Conclusion

Sign language is the means of communication for deaf communities around the world. However, Sign Language Machine Translation solutions are still under research. In this article, we explored optimization strategies as well as data augmentation approaches in the UniSign architecture to perform SLT from DGS to spoken German. The DGS-Fabeln-1 dataset was employed, given its flexibility and the recorded multi-view perspectives of the same signer. From the different explorations, we observed that employing pre-trained weights of the multi-language model improved the final performance of the model. Results from the data augmentation techniques suggest that for some configurations, employing all the camera perspectives resulted in an increment in the BLEU-4, however these additional number of training samples resulted in longer training times that should be considered. Similarly, adding more fairy tales to the training helped to predict fragments of other fairy tales, being in line with the idea that more data results in better performance. In summary, in this article, we propose a competitive baseline for the DGS-Fabeln-1 dataset that achieved a BLEU-1 of 17.88%, BLEU-4 score of 4.20% and ROUGE of 16.12% in the test set of 6 of the fairy tales of DGS-Fabeln-1. This model was trained with all the camera perspectives of the videos and a batch size of 2, achieving an increment of +4.22 percentage points in BLEU-1, and +3.23 in BLEU-4 with respect to the initial model. This baseline is in line with other proposals in the field for SLT employing transformers, as the solution proposed by Tarrés et al. (2023) for the How2Sign dataset that achieved a BLEU-4 of 8.03% having a considerable larger dataset. This study establishes a baseline for further improvements and explorations in the field.

As future work, further optimization steps can be taken by introducing and varying other parameters, such as the learning rate or warmup epochs. Additionally, exploring other libraries and datasets could provide insights about the limitations of the pose extraction modules, as OpenPose in this case, given that they play a critical role in the pose detections and hence in the sign predictions. Once higher scores are achieved, future work should also involve human evaluation to test the real-world application of the translations. Furthermore, changes to the model architecture, including more advanced LLMs could provide a boost in the final performance that could be explored as future work.

## 7. Limitations

From the experiments, the test loss indicates possible overfitting, a similar effect has been observed in similar works of SLT under limited resource scenarios (Ko et al., 2019). A plausible explanation for this is the small size of the dataset, which contains only about 4,000 videos, even when using all videos and perspectives. To address this issue, further optimization would need to increase the size of the dataset.

## 8. Acknowledgments

This contribution is funded by the German Ministry for Education and Research (BMBF) through the BIGEKO project, grant number 16SV9094. It has also received funding from the project FORSocialRobots, financed by the Bavarian Research Foundation (AZ1594-23).

## 9. Bibliographical References

Eleftherios Avramidis, Vera Czehmann, Fabian Deckert, Lorenz Hufe, Aljoscha Lipski, Yuni Amaloa Quintero Villalobos, Tae Kwon Rhee, Mengqian Shi, Lennart Stölting, Fabrizio Nunnari, and Sebastian Möller. 2025. The TUB Sign Language Corpus Collection. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, pages 1–7, Berlin Germany. ACM. doi:10.1145/3742886.3756709.

Lucas Bernhard, Fabrizio Nunnari, Amelie Unger, Judith Bauerdiek, Christian Dold, Marcel Hauck, Alexander Stricker, Tobias Baur, Alexander Heimerl, Elisabeth André, Melissa Reinecker, Cristina España Bonet, Yasser Hamidullah, Stephan Busemann, Patrick Gebhard, Corinna Jäger, Sonja Wecker, Yvonne Kossel, Henrik Müller, Kristoffer Waldow, Arnulph Fuhrmann, Martin Misiak, and Dieter Wallach. 2022. Towards automated sign language production: A pipeline for creating inclusive virtual humans. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '22, page 260–268, Corfu, Greece. Association for Computing Machinery. URL: https://doi.org/10.1145/3529190.3529202, doi:10.1145/3529190.3529202.

Fabian Bross. 2020. Object marking in German Sign Language (Deutsche Gebärdensprache): Differential Object Marking and Object Shift in the Visual Modality. *Glossa: a journal of general linguistics*, 5(1). doi:10.5334/gjgl.992.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, Salt Lake City, UT, USA. IEEE Computer Society. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Camgoz_Neural_Sign_Language_CVPR_2018_paper.html.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186. doi:10.1109/TPAMI.2019.2929257.

Shelly Chadha, Kaloyan Kamenov, and Alarcos Cieza. 2021. The World Report on Hearing, 2021. Global Report. World Health Organization Press. URL: https://www.who.int/publications/i/item/9789240020481.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10020–10030, Online. IEEE Computer Society. doi:10.1109/CVPR42600.2020.01004.

Rina Damdoo and Praveen Kumar. 2025. SignEdgeLVM transformer model for enhanced Sign Language Translation on Edge Devices. *Discover Computing*, 28(1). doi:10.1007/s10791-025-09509-1.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine Translation from Signed to Spoken Languages: State of the art and Challenges. *Universal Access in the Information Society*, 23(3):1305–1331. doi:10.1007/s10209-023-00992-1.

Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró i Nieto. 2021. How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language. In *CVPR*, pages 2735–2744, Online. Computer Vision Foundation / IEEE. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2021.html#DuartePVGDMTG21.

Sergio Esteban-Romero, Iván Martín-Fernández, Manuel Gil-Martín, David Griol-Barres, Zoraida Callejas-Carrión, and Fernando Fernández-Martínez. 2024. LLM-Driven Multimodal Fusion for Human Perception Analysis. In *Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor*, MuSe'24, page 45–51, Melbourne VIC, Australia. Association for Computing Machinery. doi:10.1145/3689062.3689084.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural Sign Language Translation Based on Human Keypoint Estimation. *Applied Sciences*, 9(13):2683. doi:10.3390/app9132683.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release. Universität Hamburg. Dataset. doi:10.25592/dgs.corpus-3.0.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2024. Uni-Sign: Toward Unified Sign Language Understanding at Scale. In *The Thirteenth International Conference on Learning Representations*, Singapore, Singapore. URL: https://iclr.cc/virtual/2025/poster/31250.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. URL: https://aclanthology.org/W04-1013/.

Cristina Luna-Jiménez, Lennart Eing, Annalena Bea Aicher, Fabrizio Nunnari, and Elisabeth André. 2025. Lightweight Transformers for Isolated Sign Language Recognition. In *Proceedings of the 27th International Conference on Multimodal Interaction*, ICMI '25, page 155–163, Canberra, Australia. Association for Computing Machinery. doi:10.1145/3716553.3750772.

Cristina Luna-Jiménez, David Griol, and Zoraida Callejas. 2024. Zero-shot ensemble of language models for fine-grain mental-health topic classification. In *Artificial Intelligence for Neuroscience and Emotional Systems: 10th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2024, Olhâo, Portugal, June 4–7, 2024, Proceedings, Part I*, page 88–97, Olhâo, Portugal. Springer-Verlag. doi:10.1007/978-3-031-61140-7_9.

Claudia Macht and Markus Steinbach. 2019. *33. Regionalsprachliche Merkmale in der Deutschen Gebärdensprache*, pages 914–935. De Gruyter Mouton, Berlin, Boston. URL: https://doi.org/10.1515/9783110261295-033 [cited 2026-01-26].

Ozge Mercanoglu Sincan and Richard Bowden. 2025. Spotter+GPT: Turning Sign Spottings into Sentences with LLMs. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, Berlin, Germany. Association for Computing Machinery. doi:10.1145/3742886.3756708.

Fabrizio Nunnari, Eleftherios Avramidis, Cristina España-Bonet, Marco González, Anna Hennes, and Patrick Gebhard. 2024. DGS-Fabeln-1: A Multi-Angle Parallel Corpus of Fairy Tales between German Sign Language and German Text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4847–4857, Torino, Italia. ELRA and ICCL. URL: https://aclanthology.org/2024.lrec-main.434/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. doi:10.3115/1073083.1073135.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A Survey of Multilingual Large Language Models. *Patterns*, 6(1):101118. doi:https://doi.org/10.1016/j.patter.2024.101118.

Oline Ranum, David Wessels, Gomèr Otterspeer, Erik J Bekkers, Floris Roelofsen, and Jari I. Andersen. 2024. The NGT200 dataset - geometric multi-view isolated sign recognition. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*. URL: https://openreview.net/forum?id=idkNzTC67X.

Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. 2023. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

*(CVPR): Workshops*, pages 5625–5635, Vancouver, Canada. IEEE Computer Society. URL: https://doi.ieeecomputersociety.org/10.1109/CVPRW59228.2023.00596.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Long Beach, California, USA. Association for Computing Machinery. URL: https://dl.acm.org/doi/10.5555/3295222.3295349.

# Lost in Expression: Diagnosing Systemic Challenges with Non-Manual Generalization in Sign Language Understanding Tasks

**Dmitriy Sazonov, Sevgi Z. Gurbuz, Evie Malaia**
University of Alabama, North Carolina State University, University of Alabama
Tuscaloosa, AL, USA, Raleigh, NC, USA, Tuscaloosa, AL, USA
desazonov@crimson.ua.edu, szgurbuz@ncsu.edu, eamalaia@ua.edu

**Abstract**

Incorporation of non-manual information is one of the most challenging aspects of Sign Language Understanding (SLU), as these features contribute to the semantic, syntactic, and pragmatic structure of signed communication as a critical feature of compositional meaning at sign, phrase and sentence level. Despite their key linguistic role, non-manuals are often an afterthought in SLU model and dataset design, with many recent models still neglecting to implement non-manual analysis or evaluate how articulators beyond the hands are contributing to the model prediction. In this work, we identify and analyze the challenges relating to recognition of non-manuals and generalization of their linguistic roles encountered by SLU models, offering new explanations for failures to properly model non-manual behavior. We perform a case study on the subtasks of Continuous Sign Language Recognition and Sign Language Translation by applying the Uni-Sign model to Isharah-1000, a Saudi Sign Language dataset. Using controlled partitioning and feature attribution, we further analyze model behavior and failure cases. With this work we hope to set the stage for the creation of diagnostic frameworks for generalization of non-manuals.

**Keywords:** Sign Language Translation, Continuous Sign Language Recognition, Non-manual Markers, Feature Attribution, Saudi Sign Language

## 1. Introduction

Historically, the field of Sign Language Understanding (SLU) has progressed from lower level tasks toward more complex and holistic tasks as technological capability has advanced. Early work was limited to lower-level tasks such as static sign recognition, where the objective is image classification focused on manual articulators(Ong and Ranganath, 2005). While this was a necessary step in the development of SLU systems, the field has largely remained focused on manual articulators since. The goal of these systems should ultimately be to model the full range of sign language articulation, including lexical, grammatical, and prosodic markers articulated via face, head, and body movement (Borneman et al., 2018; Malaia et al., 2018; Krebs et al., in press). Yet, non-manual markers (NMMs) such as facial expression or body movement remain underutilized in comparison to manual articulators (Kim et al., 2024). NMMs play an important role in sentence-level meaning by marking clause types, prosodic boundaries, and semantic distinctions, as part of a distributed system optimizing information transfer across articulators (Wilbur, 2021; Krebs et al., 2025; Malaia et al., 2018; Malaia and Wilbur, 2020).

Broadly, SLU includes (but is not limited to) Sign Language Recognition (SLR), which focuses on recognizing lexical units from sign language sequences, and Sign Language Translation (SLT), which is concerned with generating textual translations. SLR is typically further divided into Isolated (ISLR) and Continuous (CSLR) forms, addressing individual signs and sign sequence recognition, respectively. SLT can be further subdivided into gloss-based and gloss-free prediction, depending on whether glosses are predicted in addition to text (Alyami et al., 2026). Recent advances in machine learning have enabled substantial progress in ISLR, shifting attention toward CSLR and SLT. These tasks more directly demand integration of non-manual cues, motivating the development of multi-channel SLU models (Li et al., 2025; Ko et al., 2019; Camgoz et al., 2021).

Although NMM utilization is crucial for real-world deployment of SLU systems, it has received limited focus and is often not regarded as a primary challenge. This is due in part to common evaluation practices that treat performance gains from adding facial features as evidence of faithful NMM use. This issue underscores the need to amplify voices of native signers in SLU research, as they are ultimately the communities affected by this work. Without proper consideration for this issue, SLU research may optimize for benchmark performance while remaining unable to process the full range of grammatical and lexical meaning in native signing, including clause- and sign-modifying non-manuals (Tanzer et al., 2024). For example, in SSL, yes/no questions are marked by raised eyebrows and a forward head tilt, with no manual question particle (Altamimi and Alsager, 2023); similarly, the non-manuals "intense face" (combination of brow furrow, eye-squint, and downturned mouth) function as a non-manual adverb of degree modifying



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 438–449
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

verbs, adjectives, and classifier predicates (Morris and Schneider, 2012) without a manual component carrying the same meaning. Thus, a model that ignores these features cannot reliably distinguish a declarative from a question, or a neutral predicate from an intensified one.

This paper diagnoses systematic challenges in handling NMMs that arise throughout the SLU pipeline. Through a Saudi Sign Language (SSL) case study spanning CSLR and SLT, we show how comparative analysis using diagnostic measures and feature attribution can reveal failures to reliably integrate non-manual cues. We intentionally avoid restricting analysis to idealized conditions. Instead, we evaluate a representative end-to-end SLU pipeline and treat errors and occlusions as informative observations for future non-manual modeling practices.

## 2. Related Work

### 2.1. Challenges in CSLR and SLT

In order to create SLU models that are viable in real-world conditions, models must train and generalize on samples representative of real world signing conditions. Dataset structure can often be manipulated for evaluation purposes. Signer-independent evaluation tests whether models generalize beyond signer identity (Sincan et al., 2021). Unseen sentence evaluation is also used to test compositionality (Alyami et al., 2026). This is particularly critical for SLT, where the composition of lexical, semantic, and prosodic information from multiple articulators is a complex task. Both of these issues are closely linked to NMM utilization, where facial features may make identity obfuscation more difficult, and compositional generalization relies on faithful use of NMMs (Krebs et al., 2025).

Sample collection procedures also affect outcomes in real-world scenarios. Directed, controlled sample collection has been shown to degrade model prediction quality compared to samples collected in a natural dialogue setting (Kurtoğlu et al., 2024; DeHaan et al., 2025). Similarly, discourse context has been shown to be critical for SLT (Tanzer et al., 2024), yet most available datasets remain restricted to sentence-level clips. This directly translates into a limitation for non-manual modeling, as NMMs play a crucial role in shaping sentence-level context. Other qualities, such as environment and appearance variability, have been shown to limit model robustness when applied in different settings, although this can mostly be abstracted away for pose-based models.

Significantly, different sensors and modalities have varying efficacy in perceiving linguistic parameters. For example, while video and pose is advantageous for extracting spatial information (e.g. hand shape and NMMs), radio frequency (RF) sensors extract micro-Doppler signatures showing the velocity versus time of the backscatter during signing, providing a unique representation for characterizing kinematics (Gurbuz et al., 2020; Malaia et al., 2024), while also enabling higher resolution in the depth dimension. Thus, the modality through which linguistic information is extracted is inherent to the model's representation of sign language.

### 2.2. Multi-Channel Sign Language Understanding

To process the coarticulation of manual and non-manual markers, multi-channel SLU architectures use separate feature extractors for the hands, head, and sometimes upper body. This approach has been applied across ISLR (Pu et al., 2016; Ko et al., 2019), CSLR (Camgoz et al., 2021; Jiao et al., 2023; Mukushev et al., 2020), and SLT (Camgoz et al., 2020; Gueuwou et al., 2025; Li et al., 2025). However, the intended role of non-manual markers differs by subtask, changing the implicit role for NMMs towards the objective, particularly when lacking direct annotation (Mukushev et al., 2020). While in ISLR they might exclusively be used for lexical cues such as mouthing, in CSLR they may additionally function as prosodic signals for boundary cues between glosses. SLT has the strongest theoretical dependence on NMMs, as translation requires the composition of semantic, syntactic, and pragmatic information. The breadth of these challenges suggests that partial feature usage isn't enough for linguistically faithful modeling, motivating closer analysis of non-manual utilization.

### 2.3. Non-Manual Modeling

Multiple studies report modest gains when adding NMMs to manual features ( $\sim$1.5–13%) across sign languages (Aran et al., 2009; Luqman and El-Alfy, 2021; Brock et al., 2020; Mejía-Peréz et al., 2022; von Agris et al., 2008; Elons et al., 2014). However, some other studies report a drop in performance when using facial keypoints (Ko et al., 2019; Johnny et al., 2025). Ko et al. (2019) hypothesized that the performance drop was due to overparameterization caused by an excessive number of facial keypoints (70). However, Johnny et al. (2025) report a similar reduction despite using only 19 facial keypoints, suggesting that keypoint count alone may not be the sole cause of this degradation. Notably, these studies quantify the contribution of non-manuals primarily through overall task performance, rather than through metrics that faithfully reflect their linguistic function. As a result, it remains unclear whether the NMMs are truly being used for semantic and grammatical sentence-level information, as

opposed to cue-based memorization of samples.

### 2.4. Interpretability and Diagnostic Methods

Recent work has introduced interpretability tools, such as joint-wise feature importance (Holmes et al., 2024), temporal attention visualization (Dal Bianco et al., 2024; Zelezny et al., 2025), and saliency maps (Nam Pham and Avramidis, 2025). These methods provide insight into which features influence model predictions, allowing for further linguistic analysis of SLU model performance.

Despite this, the quantification of the functional contribution of non-manuals within SLU models has received limited attention. While Holmes et al. (2024) and Nam Pham and Avramidis (2025) examine contribution of non-manuals to performance, feature attribution alone does not prove faithful incorporation.

## 3. Dataset and Preprocessing

### 3.1. Dataset Overview

Isharah (Alyami et al., 2026) is a large multi-scene and multi-signer dataset for CSLR and SLT, divided into subsets Isharah-500, Isharah-1000, Isharah-2000 based on the number of unique sentences. The data consists of RGB videos of Saudi Sign Language (SSL) sentences with gloss annotation and transcription. The dataset was collected through video imitation of prompted sentences, so each target sentence is generally consistent across signers, with only slight variation in annotation when productions differ. As SSL remains underresourced in sign language technology, datasets such as Isharah provide an important setting for examining sentence-level SLU in a less-studied sign language. SSL, like other sign languages, uses non-manuals for grammatical marking (Morris and Schneider, 2012; Altamimi and Alsager, 2023), making their representation in Isharah a linguistic necessity. We use Isharah-1000, consisting of RGB videos of 18 signers performing up to 1,000 unique sentences for a total of 15,000 samples. With a sample count comparable to other popular CSLR and SLT datasets, such as Phoenix2014-T (Forster et al., 2014) (8,257 unique sentences across 8,257 samples) and CSL-Daily (Zhou et al., 2021) (6,598 unique sentences across 21,000 samples), the dataset provides a strong foundation for sentence-level CSLR and SLT experiments in an underexplored sign language. The dataset also features signer-independent (SI) and unseen sentence (US) partitions, allowing for controlled evaluation. The SI partition tests the model's accuracy on signers that were not seen in training, evaluating its ability to generalize across signing patterns. The US partition tests the model's ability to predict sentence combinations not seen in training, evaluating compositionality. This dataset was originally shared upon request by the authors, but has since been made publicly available.

A unique feature of Isharah is that several of the signers featured in the videos wear veils. The majority of these veils (and the only one from signers included in Isharah-1000) is the Niqab, which obscures all facial features aside from the eyes. This presents both a linguistic and logistical challenge for the model when predicting sign glosses from veiled signers, since many relevant non-manual markers (e.g. lip contour) are totally obscured by the veil, while other potentially informative ones (e.g., head motion) may not be effectively inferred.

#### 3.1.1. NMMs in Saudi Sign Language

NMMs perform both grammatical (clause-level) and morphological (sign-level) functions in SSL, making non-manual modeling a necessity for any SLU system for this language. At the clause level, non-manual cues are the primary means of distinguishing sentence types: yes/no questions are marked only by raised eyebrows and a forward head tilt, with no manual question particle, while wh-questions rely on facial expression whose form varies with context (Morris and Schneider, 2012). Thus, NMMs determine clause type independently of word order: a model that ignores facial features would be unable to distinguish a question from a statement.

At the sign and phrase level, SSL uses several non-manual morphemes as adverbs of degree. The "intense face" (a co-occurring combination of brow furrow, eye squint, and downturned mouth) functions as an obligatory modifier that can scope over individual adjectives, verbs, and classifier predicates (Morris and Schneider, 2012).

Tongue movements constitute another class of non-manual morphemes in SSL: forward tongue protrusion functions both as an intensifier and as an obligatory component of certain manual signs (e.g., the sign for red); side tongue protrusion appears as an intensifier restricted to negative affect and taboo contexts, and tongue flap marks small size (Morris and Schneider, 2012). These morphemes are obligatory: their omission changes the meaning of a signed sentence. The obligatory, multi-level linguistic functions of NMMs in SSL mean that they cannot be treated as auxiliary markers.

### 3.2. Pose Keypoint Extraction

We chose to use RTMPose (Jiang et al., 2023), a recent pose estimation model, to process Isharah-1000 into skeletal data. RTMPose is used in the original Uni-Sign model (as detailed in Section 4.1), making it an optimal choice for fine-tuning model




Figure 1: Uni-Sign Region Segmentation

weights pretrained on major datasets such as CSL-News (Li et al., 2025). Out of the 133 keypoints extracted by RTMPose, 69 are used by Uni-Sign: 42 for hands, 9 for body, and 18 for the face (representing facial orientation, jawline, lip contour, and nose). These keypoints and their corresponding regions can be seen in Figure 1. One linguistically critical articulator for NMMs (Wilbur, 2021) not included in this analysis is the eyebrow/eye area, being excluded from Uni-Sign's selected features.

## 4. Experimental Setup

### 4.1. Model Architecture

We use the pose-only Uni-Sign (Li et al., 2025) framework for both CSLR and SLT (gloss-free), fine-tuning publicly available weights pretrained on CSL-News (Li et al., 2025). This model introduced a pretraining framework that unifies downstream SLU tasks by treating them as variants of SLT, achieving state-of-the-art performance across ISLR, CSLR, and SLT datasets alike. This makes the model ideal for analytical comparison between tasks, minimizing any gaps between pretraining and downstream tasks that may bias the result. By sharing the same pose encoder and mT5-base decoder architecture between tasks (Xue et al., 2021), direct probing can reveal comparative differences in learned representations. This pose-only model reached nearly equivalent performance to the standard model, with additional allowance for joint-wise interpretability across regions of the body.

### 4.2. Tasks and Metrics

Several metrics are used to compare the two tasks, each serving a distinct purpose. We use Word Error Rate (WER) to evaluate CSLR, and BiLingual Evaluation Understudy (BLEU) to evaluate SLT, both of which are standard practice for the respective tasks.

| Partition | Train | Test |
|---|---|---|
| Control | Random | Random |
| SI | — | Signers not in Train |
| US | — | Sentences not in Train |

Table 1: Training-testing data partitions (US: Unseen Sentence, SI: Signer-Independent)

Additionally, we also evaluate exact-match accuracy, defined as the proportion of predictions that exactly match the ground truth. This was done to account for possible score inflation due to sentence memorization. We use the proportion of predicted samples within the sentences seen during training as an indirect measure for the model's ability to generalize. The difference between previously seen rate and exact-match accuracy indicates whether models learn surface-level pattern matching versus compositional semantics: a critical distinction for evaluating NMM learning. Finally, we use unigram overlap between the prediction and ground truth as a measure of sentence-level content overlap.

### 4.3. Dataset Partitions

We evaluate Uni-Sign in three data partitions (Table 1): Control, a standard random split; Signer-Independent, which evaluates generalization to unseen signers, distinguishing reliance on signer-specific versus linguistic cues; and Unseen Sentences, which evaluates on unseen sentences to test compositionality.

### 4.4. Integrated Gradients

To analyze how different body regions influenced model predictions, we used the Integrated Gradients (IG) method (Sundararajan et al., 2017) for feature attribution between the major keypoint regions used by Uni-Sign: hands, face, and body (including arms and head orientation). IG quantifies the contribution of each input feature by integrating the gradient of the model's output along a straight path from a neutral baseline pose to the actual input, providing a principled measure of how each region influences the prediction. The continuous form of Integrated Gradients is defined as:

$$\mathrm{IG}_i(x) = (x_i - \tilde{x}_i) \times \int_0^1 \frac{\partial F\big(\tilde{x} + \alpha(x - \tilde{x})\big)}{\partial x_i}\, d\alpha \quad (1)$$

where $F$ denotes the model's output, $x$ is the input sequence (skeleton keypoints), and $\tilde{x}$ is the baseline (zero-pose). The gradient $\frac{\partial F}{\partial x_i}$ measures the sensitivity of the model output to the $i$-th input feature, which is later aggregated over keypoints belonging to the same body region.

| Method | Partition | CSLR WER ↓ | SLT BLEU ↑ |
|---|---|---|---|
| Uni-Sign | Control | 17.78 % | 81.56 % |
| | SI | 49.56 % | 52.19 % |
| | US | 90.74 % | 2.13 % |
| Swin-MSTP | SI | 26.6 % | — |
| SMKD | US | 48.0 % | |
| GFSLT-VLP | SI | — | 39.4 % |
| MMTLB* | SI | — | 42.5 % |

Table 2: CSLR/SLT performance across test set of various partitions (* indicates gloss-based SLT model)

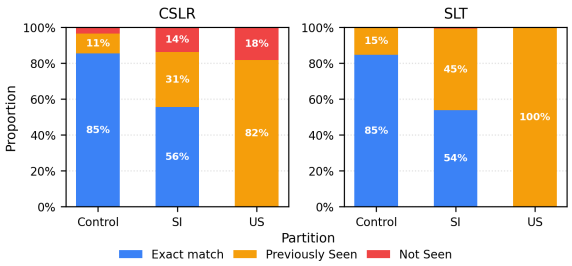



Figure 2: Proportion of predictions that are an exact match to the reference or a previously seen sentence, by partition for CSLR and SLT

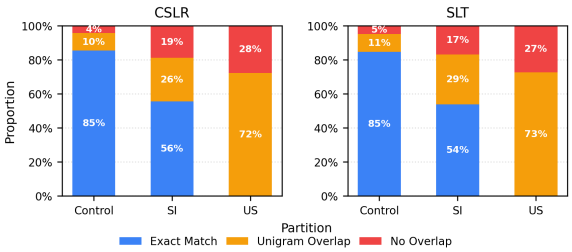



Figure 3: Proportion of predictions that are an exact match to the reference or have unigram overlap, by partition for CSLR and SLT

## 5. Results and Analysis

### 5.1. Overall Task Performance

Table 2 shows the performance of Uni-Sign across partitions for both CSLR and SLT, alongside the previous highest performing model benchmarks on Isharah-1000 (note that better performance is indicated by lower WER and higher BLEU). Across subtasks, Uni-Sign achieves similar results for each partition, with a moderate drop in performance for the signer-independent (SI) partition, and significant degradation for unseen sentences (US).

Comparison with previous benchmarks, such as Swin-MSTP (Alyami and Luqman, 2025), SMKD (Hao et al., 2021), GFSLT-VLP (Zhou et al., 2023), and MMTLB (Chen et al., 2022), shows that for CSLR Uni-Sign underperforms in both signer-independent and unseen sentence partitions. However, in SLT Uni-Sign performs higher than previous model benchmarks. Uni-Sign reaches 55.69 BLEU, significantly higher than 39.4 BLEU for GFSLT-VLP, or even 42.5 BLEU for MMTLB, a model that uses both gloss and text representations in training. In fact, the results for SLT would be regarded as highly competitive.

### 5.2. Sample Prediction Similarity

Figure 2 depicts the proportion of predictions across partitions that are an exact-match to the true sentence, along with the proportion of predictions that are an exact-match to any sentence seen in training. The figure shows much less incorrect non-exact-match sentences than expected, with CSLR having a small amount of generated sentences unseen previously, and SLT generating almost entirely previously seen sentences. For unseen sentences in particular, a large proportion of generated sentences were previously seen, yet none of the predicted sentences were exact matches. It's important to note that in both cases, predictions can often not match previously seen sentences but still be near accurate.

Figure 3 illustrates the presence of any non-zero unigram overlap between prediction and reference gloss annotation / sentences, representing shared sentence content. These values are almost identical between CSLR and SLT, as is exact-match accuracy.

### 5.3. Feature-Level Attribution

Tables 3, 4, and 5 reveal the most salient joints across each region contributing to prediction using the Integrated Gradients method. The figure reveals a pattern that holds across all partitions: contribution from the hands is the highest, closely followed by the body, while facial contribution is lowest.

In particular, Table 5 shows how the exact value of facial contribution differs between subtasks and partitions. Across partitions for SLT, the signer-independent partition achieves a significantly higher attribution value compared to unseen sentences and control. This reveals a connection between evaluation of signers not seen in training and facial feature contribution.

### 5.4. Observed Tracking Failures

As shown in Figure 4, face keypoint estimation often fails catastrophically for veiled signers. In some cases, the predicted facial keypoints are not merely imprecise, but spatially incoherent, failing to align



Figure 4: Keypoint Estimation Failure on Veiled Signer

| Joint | CSLR (%) | SLT (%) |
|---|---|---|
| 8:right_elbow | 10.38 | 7.50 |
| 6:right_shoulder | 7.13 | 6.70 |
| 7:left_elbow | 6.87 | 2.70 |
| 2:right_eye | 3.07 | 1.18 |
| 1:left_eye | 2.88 | 1.36 |
| 0:nose | 2.71 | 1.38 |
| 5:left_shoulder | 2.70 | 1.81 |
| 4:right_ear | 2.45 | 1.57 |
| 3:left_ear | 1.58 | 1.03 |
| **Total body share** | **39.76** | **25.23** |

Table 3: Top joints in Body by global attribution share (control partition).

even approximately with the visible eye region or a plausible facial configuration. By contrast, the hands are still being inferred with reasonable accuracy, despite the partial obstruction by surrounding objects.

## 6. Discussion

We identify three key sources of systemic error in non-manual modeling: model-level, dataset-level, and modality-level.

### 6.1. Model-Level Error

#### 6.1.1. Sentence Memorization

The results reveal several otherwise hidden failures in the model that couldn't be seen solely from the control partition's WER and BLEU metrics. Across the Isharah benchmarks, Uni-Sign provided reasonable performance for CSLR and state-of-the-art results for SLT. Despite this, the models were shown to have poor compositionality, instead preferring to memorize previously seen sentences, with SLT in particular making virtually no attempts to compose

| Joint | CSLR (%) | SLT (%) |
|---|---|---|
| R:124:right_middle_finger4 | 2.51 | 3.13 |
| R:129:right_pinky_finger1 | 1.89 | 3.00 |
| R:116:right_thumb4 | 1.92 | 2.98 |
| R:132:right_pinky_finger4 | 2.01 | 2.88 |
| R:120:right_forefinger4 | 2.48 | 2.87 |
| R:119:right_forefinger3 | 2.53 | 2.61 |
| L:98:left_forefinger3 | 1.50 | 2.56 |
| R:131:right_pinky_finger3 | 2.51 | 2.50 |
| L:99:left_forefinger4 | 1.30 | 2.40 |
| R:117:right_forefinger1 | 1.59 | 2.33 |
| **Total hands share** | **54.77** | **71.94** |

Table 4: Top joints in Hands by global attribution share (control partition).

new, unseen sentences in evaluation.

#### 6.1.2. Signer Identification

Feature-level attribution shows that the hands and body are most responsible for the prediction. The face share is small, yet consistent, indicating that face features are likely still being used as identification cues for the sake of memorization. Analyzing the region prediction contribution across partitions shows that face share is significantly higher for signer-independent in SLT, the partition where train/test is split based on signer identity. This suggests that facial features are used as peripheral signer identification cues rather than as sources of compositional information.

#### 6.1.3. Subtask Convergence

Though unified architectures can provide efficient and effective performance across downstream tasks, our results suggest that they may exhibit convergence of subtask behavior, limiting task-specific reasoning. In the case of CSLR and SLT, despite different reference text (gloss vs. sentence), the tasks exhibit highly similar memorization patterns to the point of having near identical exact-match accuracy. In addition, SLT does not demonstrate increased reliance on non-manual features compared to CSLR as expected from linguistic theory.

One possible explanation for this is that gloss and sentence targets occupy closely aligned regions of the decoder's embedding space, particularly when gloss-sentence mappings are nearly 1:1. This offers limited incentive to differentiate between differing linguistic demands of non-manuals, resulting in similar behavior. As a result, the model does not develop a functional distinction between recognition and translation.

| Joint | Control | | SI | | US | |
|---|---|---|---|---|---|---|
| | CSLR (%) | SLT (%) | CSLR (%) | SLT (%) | CSLR (%) | SLT (%) |
| 32:face-9 | 0.41 | 0.22 | 0.55 | 0.81 | 0.20 | 0.22 |
| 36:face-13 | 0.49 | 0.25 | 0.38 | 0.69 | 0.19 | 0.18 |
| 23:face-0 | 0.63 | 0.12 | 0.34 | 0.50 | 0.41 | 0.13 |
| 31:face-8 | 0.34 | 0.13 | 0.57 | 0.60 | 0.21 | 0.12 |
| 29:face-6 | 0.58 | 0.27 | 0.37 | 0.29 | 0.46 | 0.13 |
| 30:face-7 | 0.55 | 0.18 | 0.45 | 0.39 | 0.21 | 0.10 |
| 24:face-1 | 0.43 | 0.07 | 0.14 | 0.33 | 0.16 | 0.07 |
| 27:face-4 | 0.20 | 0.23 | 0.23 | 0.29 | 0.40 | 0.19 |
| 25:face-2 | 0.22 | 0.17 | 0.36 | 0.09 | 0.27 | 0.12 |
| 38:face-15 | 0.23 | 0.20 | 0.15 | 0.07 | 0.36 | 0.16 |
| **Total head share** | **5.47** | **2.83** | **4.95** | **5.80** | **4.39** | **2.24** |

Table 5: Top facial joints by global attribution share across partitions.

## 6.2. Dataset-Level Error

### 6.2.1. Gloss-Sentence Mapping

The dataset's high full-sentence similarity (1000 sentences with 1:1 gloss-sentence mapping) likely prevented the model from learning compositional semantics at all, including that contributed by NMMs. This highlights a critical limitation in dataset design for SLT/CSLR: models require high training data variability at the phrase and sentence level to learn compositional semantics.

This redundancy presents both benefits and limitations: it broadens coverage of articulatory variation within individual sentences, which was seen by the relative success of the signer-independent partition, yet also encourages memorization of a narrow set of reference sentences, undermining the model's ability to compose new sentences.

### 6.2.2. Low Non-Manual Diversity

The structure of this dataset also limits opportunities for non-manual learning by reducing instances in which syntactic or semantic disambiguation is required. Since each gloss sequence maps to a single reference translation, features vital to interpretation, such as non-manuals, may be underutilized. Although lexical content is recognized and utilized in this process, other linguistically critical features are not.

### 6.2.3. Directed Sample Collection

A broader issue is that of SLU data collection, where directed wording has been shown to produce movement patterns different from naturally occurring signing. In contrast, natural signing contexts allow greater linguistic variation, providing the depth and diversity necessary to capture non-manuals and other features of sign language communication. In this dataset, signers were prompted to sign by imitating a video, which may introduce such issues.

## 6.3. Modality-Level Error

### 6.3.1. Pose Estimation Error

As seen in Figure 4, the model encounters practical pose estimation errors, where facial content is obscured and facial keypoints are inaccurately estimated. In this context, aside from head orientation, minimal facial expression information can be inferred. The failure of the pose extraction model to even approximate the position of keypoints renders facial information largely unusable, further diminishing opportunities for non-manual learning. These cases require robustness from both pose estimation and SLU models to missing or degraded facial features, while still enabling the extraction of rich non-manual information when it is present. Challenges such as these are crucial to address as sign language understanding advances toward greater accessibility for all signers.

### 6.3.2. Incomplete Input Pose Data

Out of 68 available facial keypoints in COCO-WholeBody 2D, only 18 are used in Uni-Sign, meaning that eye and eyebrow movements are excluded, despite being known articulators of non-manuals (Wilbur, 2021). Such reductions are often implemented to decrease input size and improve training efficiency, but they may limit long-term and holistic learning of non-manual features. However, the omission does not preclude broader generalizability of the analysis of NMMs, although it limits its linguistic interpretability.

### 6.3.3. Pose-Only Data Limitations

Even when all available facial keypoints are included, facial movement remains highly complex. Subtle variations in facial morphology and prediction noise can cause keypoint trajectories to deviate from the true visual signal. As a result, skeletal

representations may fail to capture the full expressive nuance of facial articulation, highlighting the inherent limitations of the pose modality alone for non-manual analysis.

One might ask if evaluating veiled and non-veiled signers separately may have better isolated the effect of facial occlusion. However, framing occlusion as a condition to be controlled underestimates what a linguistically adequate CSLR or SLT model should be able to do. One meaningful comparison, or benchmarking bar, may be with a fluent human signer processing a sign language under similar conditions (e.g. when the signing partner is veiled). Research on neural bases of sign language comprehension shows that Deaf signers rely on predictive processing: rather than reconstructing meaning sign-by-sign from bottom-up visual input, fluent signers maintain hierarchical internal models of linguistic structure (unfolding syntactic structure, prior information about participants/arguments and likely event structure) that allow them to infer partially observable content from prior information contained in the visual signal (Malaia et al., 2021, 2023; Borneman et al., 2025). A signer who acquired SSL to native proficiency would have no difficulty communicating with veiled signers, because obligatory grammatical markers — including those carried by head movement — are recoverable by humans from the broader motion dynamics through predictive inference. It is, then, notable, that the current model appears to have an analogous capacity for the hands: as shown in Figure 4, even when a signer's hand passes behind a table, RTMPose infers plausible joint positions from the prior context. The limitation, then, is not that an occlusion occurs, but that no equivalent predictive capacity is acquired by the model for facial articulators. The more productive model benchmarking that we argue for should target human-comprehension-level performance, and evaluate if trained models can recover linguistically meaningful NMM information under naturalistic conditions including partial occlusion.

# 7. Conclusion

In this paper, we argued that modern SLU research remains systemically misaligned with the linguistic role of non-manual markers. Using Uni-Sign (Li et al., 2025) on SSL as a controlled comparison between CSLR and SLT, we evaluated three partitions (control, signer-independent, unseen sentences) and applied prediction-similarity analysis and integrated gradients to probe feature use.

We found that despite the fact that both models reached competitive prediction accuracy, they encountered several challenges in proper non-manual representation. Although non-manual markers were shown to contribute to the prediction, analysis suggested this was in the form of peripheral identification cues rather than semantic and syntactic compositional contributions. Additionally, tracking failures in the model pipeline highlighted broader challenges for reliable integration of non-manual markers.

This work highlights critical pitfalls in current approaches to non-manual modeling and compositional learning. Addressing these issues will require deliberate redesign across the SLU pipeline to explicitly account for the linguistic role of non-manuals and to frame NMM learning as a distinct training objective in SLU systems.

## 7.1. Future Work

Prior work has primarily evaluated the addition of non-manuals in terms of performance improvements, but this neglects to examine whether non-manual features are incorporated in a manner consistent with their linguistic function. Instead, benchmarks that explicitly evaluate non-manual learning across SLU subtasks should be developed. This should assess whether models meaningfully extract semantic and syntactic information from non-manual features. Design choices for such benchmarks must be deliberate to avoid the structural errors identified in this paper. This may be particularly challenging as many of these errors have opposing constraints. For example, the pose-based models are more effective at reducing reliance on identity cues compared to RGB video, but may sacrifice fine details of facial expression. Addressing these trade-offs will require broad and extensive experimentation.

In addition, this work has shown that there is no singular unimodal solution in sign language understanding. Multi-modal sensing has the potential to overcome the limitations of a single sensor, as shown by the 2025 ICCV Multi-modal Sign Language Recognition (MSLR) Workshop, which featured the first Italian Sign Language (LIS) dataset with both video and radar data (Luqman et al., 2025). While the work done at this workshop successfully achieved state-of-the-art performance (Islam et al., 2025; Juranek, 2025; Manjur et al., 2025; Sazonov et al., 2025), current approaches espouse a purely computational perspective to feature extraction that does not as of yet incorporate any linguistic perspectives in the data analysis, such as selecting the optimal sensing data for estimating task-cognizant linguistic parameters (Malaia et al., 2024). Consequently, multi-modal sensing for SLU remains an open area of research, with great potential to enrich the robustness and accuracy of SLU in real-world conditions.

## 8. Acknowledgements

We thank the AraSLP team for providing access to the Isharah-1000 dataset in video form.

## 9. Bibliographical References

Haya Altamimi and Haroon Alsager. 2023. Argument structure and word order in saudi sign language. *Journal of Language Teaching and Research*, 14:203–214.

Sarah Alyami and Hamzah Luqman. 2025. Swin-MSTP: Swin transformer with multi-scale temporal perception for continuous sign language recognition. *Neurocomputing*, 617:129015.

Sarah Alyami, Hamzah Luqman, Sadam Al-Azani, Maad Alowaifeer, Yazeed Alharbi, and Yaser Alonaizan. 2026. Isharah: A Large-Scale Multi-Scene Dataset for Continuous Sign Language Recognition. *IEEE Transactions on Multimedia*.

Oya Aran, Thomas Burger, Alice Caplier, and Lale Akarun. 2009. A belief-based sequential fusion approach for fusing manual signs and non-manual signals. *Pattern Recognition*, 42(5):812–822.

Joshua Borneman, Evguenia Malaia, and Ronnie Wilbur. 2018. Motion characterization using optical flow and fractal complexity. *Journal of Electronic Imaging*, 27:1.

Sean C. Borneman, Julia Krebs, Ronnie B. Wilbur, and Evie A. Malaia. 2025. Decoding predictive inference in visual language processing via spatiotemporal neural coherence. NeurIPS 2025 Workshop on Foundation Models for the Brain and Body.

Heike Brock, Iva Farag, and Kazuhiro Nakadai. 2020. Recognition of Non-Manual Content in Continuous Japanese Sign Language. *Sensors*, 20(19):5621.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10020–10030.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2021. Multi-channel transformers for multi-articulatory sign language translation. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, page 301–319, Berlin, Heidelberg. Springer-Verlag.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5110–5120.

Pedro Alejandro Dal Bianco, Oscar Agustín Stanchi, Facundo Manuel Quiroga, Franco Ronchetti, and Enzo Ferrante. 2024. SignAttention: On the Interpretability of Transformer Models for Sign Language Translation.

Kenneth DeHaan, Emre Kurtoğlu, Sabyasachi Biswas, Caroline Kobek Pezzarossi, Darrin J. Griffin, Chris Crawford, Ali C. Gurbuz, Evie A. Malaia, Abraham Glasser, Raja Kushalnagar, and Sevgi. Z Gurbuz. 2025. Rf -chesssign: Radar-enabled human-computer interaction in a real- time sign language-controlled game. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5000–5010.

A.S. Elons, Menna Ahmed, and Hwaidaa Shedid. 2014. Facial expressions recognition for Arabic Sign Language translation. In *2014 9th International Conference on Computer Engineering & Systems (ICCES)*, pages 330–335. IEEE.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025. SignMusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521, Vienna, Austria. Association for Computational Linguistics.

Sevgi Z. Gurbuz, Ali C. Gurbuz, Evie A. Malaia, Darrin J. Griffin, Chris Crawford, M. Mahbubur Rahman, Ridvan Aksu, Emre Kurtoglu, Robiulhossain Mdrafi, Ajaymehul Anbuselvam, Trevor Macks, and Engin Ozcelik. 2020. A linguistic perspective on radar micro-doppler analysis of american sign language. In *2020 IEEE International Radar Conference (RADAR)*, pages 232–237.

Aiming Hao, Yuecong Min, and Xilin Chen. 2021. Self-Mutual Distillation Learning for Continuous Sign Language Recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11283–11292.

Ruth M. Holmes, Ellen Rushe, and Anthony Ventresque. 2024. The key points: Using feature importance to identify shortcomings in sign language recognition models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15970–15975, Torino, Italia. ELRA and ICCL.

Md. Milon Islam, Md Rezwanul Haque, S M Taslim Uddin Raju, and Fakhri Karray. 2025. Fusionensemblenet: An attention-based ensemble of spatiotemporal networks for multimodal sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4983–4989.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose.

Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. 2023. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20619–20629.

Samuel Ebimobowei Johnny, Blessed Guda, Andrew Blayama Stephen, and Assane Gueye. 2025. AutoSign: Direct Pose-to-Text Translation for Continuous Sign Language Recognition . In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5071–5078, Los Alamitos, CA, USA. IEEE Computer Society.

Jakub F. Juranek. 2025. Multimodal italian sign language recognition with radar-video late fusion on the multimedalis dataset. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5079–5085.

Jung-Ho Kim, Changyong Ko, Mathew Huerta-Enochian, and Seung Yong Ko. 2024. Shedding Light on the Underexplored: Tackling the Minor Sign Language Research Topics. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 147–158, Torino, Italia. ELRA and ICCL.

Sang-Ki Ko, Changjo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9:2683.

Julia Krebs, Dietmar Roehm, Ronnie B. Wilbur, and Evie A. Malaia. in press. Visual-Linguistic Cues in Manual Communication: Neural and Behavioral Responses to Mouthings in Sign-Naïve Perception. *Brain and Language*. Special Issue: Gesture, Cognition and Language.

Julia Krebs, Ronnie B Wilbur, Dietmar Roehm, and Evie A Malaia. 2025. The interaction of syntax, non-manuals, and prosodic cues as potential topic markers in Austrian Sign Language. *Sign Language & Linguistics*, 28(1):1–48.

Emre Kurtoğlu, Kenneth DeHaan, Caroline Kobek Pezzarossi, Darrin J Griffin, Chris Crawford, and Sevgi Z Gurbuz. 2024. Interactive learning of natural sign language with radar. *IET Radar, Sonar Navigation*, 18(8).

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-sign: Toward unified sign language understanding at scale. In *The Thirteenth International Conference on Learning Representations*.

Hamzah Luqman and El-Sayed M. El-Alfy. 2021. Towards Hybrid Multimodal Manual and Non-Manual Arabic Sign Language Recognition: marsl Database and Pilot Study. *Electronics*, 10(14):1739.

Hamzah Luqman, Raffaele Mineo, Murtadha Aljubran, Ahmed Abul Hasanaath, Amelia Sorrenti, Sarah Alyami, Sadam Al-Azani, Maad Alowaifeer, JiHwan Moon, Vaclav Javorek, Tomas Zelezny, Marek Hruz, Gaia Caligiore, Silvio Giancola, Senya Polikovsky, Motaz Alfarraj, Sabina Fontana, Mufti Mahmud, Muhammad Haris Khan, Kamrul Islam, Sevgi Gurbuz, Egidio Ragonese, Giovanni Bellitto, Federica Proietto Salanitri, Concetto Spampinato, and Simone Palazzo. 2025. The SignEval 2025 Challenge at the ICCV Multimodal Sign Language Recognition Workshop: Results and Discussion . In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5086–5095, Los Alamitos, CA, USA. IEEE Computer Society.

Evie Malaia, Joshua D Borneman, and Ronnie B Wilbur. 2018. Information transfer capacity of articulators in American Sign Language. *Language and speech*, 61(1):97–112.

Evie A. Malaia, Joshua Borneman, and Sevgi Gurbuz. 2024. Capturing motion: Using radar to build better sign language corpora. In *Proceedings of the LREC-COLING 2024 11th Workshop*

*on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 384–389, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Evie A Malaia, Sean C Borneman, Joshua D Borneman, Julia Krebs, and Ronnie B Wilbur. 2023. Prediction underlying comprehension of human motion: an analysis of Deaf signer and non-signer EEG in response to visual stimuli. *Frontiers in Neuroscience*, 17:1218510.

Evie A Malaia, Sean C Borneman, Julia Krebs, and Ronnie B Wilbur. 2021. Low-frequency entrainment to visual motion underlies sign language comprehension. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2456–2463.

Evie A Malaia and Ronnie B Wilbur. 2020. Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(1):e1518.

Sultan Mohammad Manjur, Sabyasachi Biswas, and Ali C. Gurbuz. 2025. A multimodal video and radar fusion framework for high-accuracy isolated sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5061–5070.

Kenneth Mejía-Peréz, Diana-Margarita Córdova-Esparza, Juan Terven, Ana-Marcela Herrera-Navarro, Teresa García-Ramírez, and Alfonso Ramírez-Pedraza. 2022. Automatic Recognition of Mexican Sign Language Using a Depth Camera and Recurrent Neural Networks. *Applied Sciences*, 12(11):5523.

Carla Morris and Erin Schneider. 2012. On selected morphemes in saudi arabian sign language. *Sign Language Studies*, 13:103–121.

Medet Mukushev, Arman Sabyrov, Alfarabi Imashev, Kenessary Koishybay, Vadim Kimmelman, and Anara Sandygulova. 2020. Evaluation of manual and non-manual components for sign language recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6073–6078, Marseille, France. European Language Resources Association.

Dinh Nam Pham and Eleftherios Avramidis. 2025. The importance of facial features in vision-based sign language recognition: Eyes, mouth or full face? In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Sylvie Ong and Surendra Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE transactions on pattern analysis and machine intelligence*, 27:873–91.

Junfu Pu, Wengang Zhou, and Houqiang Li. 2016. Sign language recognition with multi-modal features. In *Advances in Multimedia Information Processing - PCM 2016*, pages 252–261, Cham. Springer International Publishing.

Dmitriy Sazonov, Kamrul Islam, Evie Malaia, and Sevgi Z. Gurbuz. 2025. Modality-specific benchmarks and radar range-doppler envelope classification for multimodal isolated sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5046–5053.

Ozge Mercanoglu Sincan, Julio C. S. Jacques Junior, Sergio Escalera, and Hacer Yalim Keles. 2021. ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge: Design, Results and Future Research.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, and David Uthus. 2024. Reconsidering sentence-level sign language translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6262–6287, Miami, Florida, USA. Association for Computational Linguistics.

Ulrich von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE.

Ronnie B Wilbur. 2021. Non-manual markers: Theoretical and experimental perspectives. In *The Routledge handbook of theoretical and experimental sign language research*, pages 530–565. Routledge.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference*

*of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tomas Zelezny, Jakub Straka, Vaclav Javorek, Ondrej Valach, Marek Hruz, and Ivan Gruber. 2025. Exploring pose-based sign language translation: Ablation studies and attention insights.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20814–20824.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

## 10. Language Resource References

Sarah Alyami, Hamzah Luqman, Sadam Al-Azani, Maad Alowaifeer, Yazeed Alharbi, and Yaser Alonaizan. 2026. *Isharah: A Large-Scale Multi-Scene Dataset for Continuous Sign Language Recognition*. PID https://doi.org/10.1109/TMM.2026.3664959.

# The SignBeach Dataset of Dutch Sign Language (NGT) signs

**Annika Schiefner**[1,2], **Gomèr Otterspeer**[1], **Beyza Sümer**[2], **Floris Roelofsen**[1]
[1]Institute for Logic, Language, and Computation, University of Amsterdam
[2]Amsterdam Center for Language and Communication, University of Amsterdam
{a.schiefner, g.otterspeer, b.sumer, f.roelofsen}@uva.nl

**Abstract**

This paper presents the SignBeach dataset, including 1401 lexical signs from Dutch Sign Language (NGT). The items in this dataset represent everyday vocabulary appropriate for primary school children and are part of a larger research project, investigating sign learning in a digital environment. Each sign is presented by four deaf signers in a controlled studio environment. For each item, high quality video recordings are available from five synchronised cameras, providing rich multi-view visual input suitable for linguistic analysis and the development of computer vision pipelines. In addition, we provide three types of computational derivatives: keypoint estimates using MediaPipe, handshape estimates using HaMeR, and 3D body reconstructions using SAM 3D Body. Signs are aligned with lexical entries in the NGT Signbank to provide interoperability of the database with other NGT resources. We outline the construction of the dataset and provide information on opportunities for reuse, for example in the context of psycholinguistic studies or in the context of sign language technology. All materials are available for non-commercial reuse under a CC BY-NC 4.0 license.

**Keywords:** Dutch Sign Language, NGT, sign language, lexical database, 3D-video recordings

## 1. Introduction

The primary contribution of this paper is a large-scale, multi-view, child-oriented lexical video database for NGT, enriched with aligned lexical identifies and multiple layers of computational derivatives, the *SignBeach* dataset. We outline the development of the dataset, describe its structure, and provide information on potential future scientific reuse[1]. The aim of the SignBeach project is to build a series of online learning games in which primary school-aged children throughout the Netherlands learn Dutch Sign Language (*Nederlandse Gebarentaal*, NGT), through simple translation games between NGT signs and Dutch words. The games are suitable for both deaf and hearing children, regardless of whether their preferred language is NGT or Dutch.

In order to select items that are suitable for this age group and purpose, we started from a book with signs that are relevant to children learning NGT, published by the *Nederlands Gebarencentrum* (engl. *Dutch Sign Language Centre*) (Nederlands Gebarencentrum, 2012). From there, we compared items to the NGT Signbank (Crasborn et al., 2020; Klomp et al., 2024) and created engaging and child-friendly recordings, linked to Signbank through alignment of ID glosses and item IDs. This process resulted in a dataset of 1401 individual lexical signs, each produced by four signers in a multi-camera setup. In addition, we provide derivative data and simple ways to link the dataset to other lexical resources of NGT to facilitate scientific reuse of the dataset.

[1]The dataset is available on Figshare at https://doi.org/10.21942/uva.c.8233876 under a CC BY-NC 4.0 license for non-commercial reuse.

## 2. Related work

Sign language databases differ in their purposes and thus in the types of information presented in each individual database. To clarify how the SignBeach dataset integrates with and complements existing resources, the following subsections present similar resources for NGT and other sign languages and highlight in particular how they address different research needs. We limit our review to lexical databases developed as resources for academic research that are similar to our projects in focus. Online dictionaries and materials developed in the context of language assessment tools are not covered here, as their purpose differs quite substantively from our target uses with important differences in design choices resulting from these different purposes.

### 2.1. NGT Signbank and related resources

Research on NGT has a long tradition within the context of sign language research as a scientific field of enquiry. NGT thus has a relatively well developed infrastructure of resources, building on the original development of the NGT Corpus (Crasborn and Zwitserlood, 2008), a collection of video materials of conversations in NGT including approximately 72 hours of recordings of which around 21% are annotated (Kopf et al., 2022). The NGT Signbank contributes a lexical database to that infrastructure,



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 450–458
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

providing standardised gloss-IDs for videos, alongside phonological and other linguistic information (Crasborn et al., 2020; Klomp et al., 2024). At the point of writing, the NGT Signbank includes 6863 individual public signs. By providing a set of standard gloss-IDs for individual signs with reference videos and phonological descriptions, the NGT Signbank serves as a useful connector, providing single identifiers for specific variants of signs. It is mainly designed as a resource for video annotation and can be imported as a controlled vocabulary in ELAN (2025; see also Crasborn et al., 2016). For this purpose, glosses are defined based primarily on manual forms, disregarding mouthings and other non-manual information. This, in turn, makes it a database that is suitable for looking up manual forms but currently provides very limited information on how these signs may look if produced in context with appropriate mouthings and other non-manuals. It does provide 'senses', alternative translations to Dutch and English, but these remain limited and additional information required for disambiguation is not available.

Besides typical academic licensing constraints around reuse, the videos are therefore not designed to be particularly engaging. By contrast, our project required signs targeted at children, presented in an engaging manner. In addition, in our dataset, signs are produced as translations for specific lexical items in Dutch, providing a dataset for a translation game between NGT signs and Dutch words. This requires the presence of mouthings, mouth gestures, and other non-manuals that contribute both semantic and pragmatic meaning appropriate to the game context. However, as mentioned above, the gloss-IDs of the NGT Signbank are a useful mechanism for aligning sign variants across datasets. In our dataset, we therefore identified all sign forms by their gloss-IDs, creating broad alignment with the NGT Signbank. In this way, we make sure that the variants recorded in our dataset can be easily identified and matched with signs used in other studies, opening up possibilities to bring together disparate pieces of information.

Another component of the existing infrastructure for the development of lexical datasets for NGT and other sign languages is the signCollect platform (Otterspeer et al., 2024). SignCollect provides a 'touchless' recording pipeline in which a signer can control the system through simple gestures recognized via computer vision (MediaPipe), enabling recordings of 60 to 120 signs per hour without the need for technical assistance staff. The platform integrates three key components: (i) signCollect Studio for multi-camera video recording with foot pedal or gesture-based control, automatic synchronization (using Tentacle Sync), and real-time review and approval of recordings; (ii) signCollect Dashboard for collaborative metadata management among team members; and (iii) signCollect Hub which connects the recording system to Global Signbank through automated synchronization of lexical entries. The present project makes use of the signCollect platform.

## 2.2. Lexical databases in other sign languages

Lexical databases are also available for a number of other sign languages. Some languages have Signbank lexicons that are similar to the NGT version and published under the database structure provided by the Global Signbank Project (Crasborn et al., 2018), for example the BSL Signbank (Fenlon et al., 2015), which streamlines annotations for the BSL Corpus (Schembri et al., 2017). Signbank databases thus primarily fulfil a lookup function for annotating video materials. In a similar vein, the German Sign Language (DGS) Corpus is linked with a number of lexical resources that serve similar purposes (Kopf et al., 2022; Otte et al., 2022; Schulder et al., 2024), as well as more public facing presentation modes that primarily serve as a language archive and resource for the deaf community (Konrad et al., 2020; Jahn et al., 2018).

Meanwhile, other lexical databases have been developed with the purpose of collecting and presenting detailed linguistic information and norms of a variety of psycholinguistic variables. These databases serve as a basis for developing psycholinguistic experimental stimuli and materials for standardised, educational or clinical tools. An early example is the lexical database for Spanish Sign Language (LSE) by Gutierrez-Sigut et al. (2016). This database includes highly detailed phonetic coding of signs and video materials that are visually optimised for neuroimaging studies and similar research that requires a high level of experimental control. Shortly after, ASL-LEX was created as a database that combines phonological information with different psycholinguistic variables, such as iconicity, familiarity and age of acquisition, measured through rating studies (Caselli et al., 2017; Sehyr et al., 2021). Initially, this resource was primarily created to allow for studying the effects of phonological neighbourhood density and creating stimulus sets for controlled psycholinguistic studies. In further iterations of the project, additional variables have been added and the database architecture has been made adaptable, creating a foundation for creating parallel databases in other sign languages (for examples, see Morgan et al., 2022; Trettenbrein et al., 2025). The database presented in this paper most closely resembles these types of databases in terms of structure and projected use. It is designed to present video materials that are

recorded in a highly controlled setting alongside useful derivate data, to maximise opportunities for scientific reuse. As such, it directly addresses the need for larger, public video databases articulated by Bragg et al. (2019).

## 3. Dataset design

The dataset for this project combines video data, metadata on the individual signs and information for linking the materials to NGT Signbank (Crasborn et al., 2020; Klomp et al., 2024), and derivative data (see Table 1).

### 3.1. Item selection

Items were initially selected from the book "Mijn eerste 1500 gebaren" (engl. *My first 1500 signs,* Nederlands Gebarencentrum, 2012) and referenced against the NGT-Dutch online dictionary (Nederlands Gebarencentrum, 2024). These items were then compared against the signs available on the NGT Signbank (Crasborn et al., 2020; Klomp et al., 2024) and if the variant in the book was not available on Signbank, a different variant was chosen from Signbank. If no corresponding item was available on Signbank, the sign as produced in the NGT-Dutch online dictionary (Nederlands Gebarencentrum, 2024) was recorded by the first author as a reference for the recording sessions. This strategy was also employed for NGT compounds, for which there are typically no unique items on Signbank. During the recording sessions, signers were able to provide additional input to ensure widely used variants were favoured in the recording process. Where signers disagreed with the proposed variant, we recorded both the initial variant and the proposed variant and checked with the other signers to select the final variant. Wherever this process resulted in an alternative variant being chosen, the corresponding item on Signbank was identified in post-processing. The items are thus aligned with Signbank glosses and item IDs, which can be resolved into the URL leading to each sign.[2] After collecting the dataset input list, we transferred the items into lists in signCollect (Otterspeer et al., 2024) and recorded them in individual recording sessions with the signers.

### 3.2. Linking the dataset to other lexical resources

As mentioned above, our dataset is aligned with the NGT Signbank through ID glosses and sign IDs. The sign IDs can be used to derive the URL leading to the relevant Signbank entry. In our dataset, we have therefore opted for directly including the URL, to facilitate access to Signbank. This alignment with Signbank allows for easy retrieval of linguistic information about the individual items, such as phonological annotations, as well as alignment with any other resource that is linked to Signbank IDs.

It should be noted that Signbank items are presented without non-manual information, meaning that the same manual form may be associated with multiple meanings. Multiple items in our dataset can thus be based on the same Signbank item, resulting in the same gloss and Signbank ID (1401 items:1297 distinct glosses, including compounds). We are therefore using the Dutch translations to identify signs in our dataset, as glosses and Signbank IDs are not unique to individual items.

## 4. Recording video dataset

### 4.1. Signers

Each sign is presented by four distinct, deaf signers, two men and two women. The signers were professionals with ample experience in sign language video recordings, who reported using NGT regularly in their everyday lives, including in professional contexts. Each of the initial four signers recorded all 1401 items for the dataset. However, due to technical issues in the items from one recording session that could not be re-recorded with the same signer, two additional signers joined the project to create new recordings for these items for the final dataset. In the final recordings, each sign is recorded by two male and two female signers. The signers were reimbursed for their participation and contributed to the project as professional contractors, rather than as research participants.

All signers were chosen for their experience recording materials for children. They are deaf professionals, well-versed in working in a recording context and produced engaging videos of the lexical items. They were asked to sign in a child-friendly manner and to produce mouthings and mouth gestures as they felt was appropriate for the concepts. Overall, the signs are recorded in a friendly and expressive manner, making them feel more natural than typical recordings in the NGT Signbank. However, it should be noted that affect is influenced by the semantic content of the signs recorded (e.g. the sign for BE CAREFUL is signed with a stern face across signers) but probably also by signers' fatigue over long recording sessions. We varied the order of items across signers and recorded over multiple sessions to alleviate issues of fatigue and the resulting recordings feel natural.

[2] Note that some items, including most compounds, are not available on Signbank. The Signbank gloss and ID are marked as NA for those items.

### 4.2. Recording setup

The recordings were made at the SignLab of the University of Amsterdam using a five-camera setup that extends the signCollect infrastructure described in Otterspeer et al. (2024). All cameras are Sony FX30 units with 35mm lenses, recording at 4K (3840×2160) and 60fps. The middle camera (M) faces the signer directly; cameras L and R are placed at 45° to the left and right; cameras A and B are positioned at approximately 80° to capture the depth of the hand that the other angles miss (see Figure 1). To sync frames across cameras, a screen in the recording space displays QR codes encoding the gloss ID, date, and timestamp. These codes are decoded in post-processing to align the five video streams. The embedded metadata also ensures the recordings remain identifiable if they outlive the signCollect system.

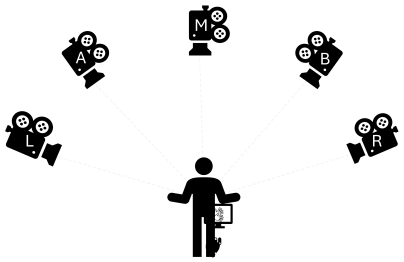



Figure 1: Schematic illustration of the recording setup.

For each individual sign, two signers produced the sign oriented straight towards the middle camera, while the other two signers were oriented at an angle, presenting the hands from the side. Which of the signers presented at which angle was rotated across signs. For the items in the game, that meant that each sign is available from both perspectives to somewhat mitigate the disadvantages of a 2-dimensional presentation of 3-dimensional materials (Watkins et al., 2024). In the present dataset, videos are available from all five cameras for both presentation angles.

Individual signs, timing information on the recording, and the review-screen for each sign were presented to the signer on a screen in front of them. The recording pipeline was set up for a default duration of four seconds, though signers were able to add time in one-second increments, if necessary. This means that videos may differ in length, particularly if they contain compounds. Navigating through the items, starting the recording, and adding additional time, was controlled by the signer through tapping a set of buttons on the foot. The first author of this paper served as a recording assistant, checking the quality of the recordings and alignment with the target variants along with the signer. In this way, we were able to efficiently record large numbers of items in a relatively short time frame.

### 4.3. Video processing pipeline

Raw footage is backed up in the cloud (UvA Research Drive), then videos of camera Left, Middle, and Right are automatically processed in DaVinci Resolve Headless mode to key out the green screen and replace it with a blue background. After that, they are cropped around the signer based on MediaPipe keypoints (Lugaresi et al., 2019), using the maximum extent of the detected hand and head landmarks while centered at waist height. The processed files are uploaded to the signCollect Hub for viewing and annotation. Then, derivative data are computed: MediaPipe pose estimates (`.pose`), HaMeR 3D hand meshes (Pavlakos et al., 2024a), and SAM 3D Body full-body reconstructions (Yang et al., 2025). Finally, all materials are archived on Figshare.

## 5. Computed derivatives

Our dataset provides both video files and computational derivatives from those videos for academic reuse (see Section 6 for information on licensing). Filenames are kept consistent between videos and computational derivatives for ease of alignment.

### 5.1. Keypoint estimates

For each video, we extract 2D and 3D pose estimates using MediaPipe Holistic (Lugaresi et al., 2019), which estimates body and hand keypoints. The output is stored in JSON format with the extension `.pose`, containing per-frame landmark coordinates for all detected keypoints.

Keypoint estimates can be used for pose estimation, to reconstruct three-dimensional information about the signs. These are particularly useful for extracting kinematic information about movement characteristics, such as the size of the signing space and velocity of movements throughout the sign. Similar representations have been used in sign language technology to drive sign detection and recognition and to analyse phonological information, such as hand location and movement (see, for example, Moryossef et al., 2020; Bragg et al., 2019).

### 5.2. Handshape estimates

3D hand mesh keypoints are extracted using HaMeR (Pavlakos et al., 2024a), a transformer-based model for hand mesh recovery from monocular video. For each frame, the model outputs 3D keypoint coordinates for both the left and right hand. Results are stored in JSON format with the extension `.hamer`, with keypoints ordered by frame.

HaMeR's 3D hand mesh keypoints create a reconstruction of the handshape through keypoints on the hand's surface. In this way, they allow reconstructing handshape estimations in occlusion situations (Pavlakos et al., 2024b). For research purposes, they can be used as a basis for categorising and comparing handshapes across signs and signers and may support the identification of the corresponding handshapes in fluent signing (Asasi et al., 2025; He et al., 2025).

### 5.3. 3D body reconstructions

Full-body 3D reconstructions are generated using SAM 3D Body (Yang et al., 2025), which estimates body, hand, and foot pose from a single image. The output is stored in NPZ format with the extension `.s3d` and can be converted to SMPL-X mesh representations for downstream use.

These 3D reconstructions can serve as a full-body representation of the signs. Providing a detailed representation of not only the hands but the full body, they allow the detailed tracking of bodily motions and poses over time. They can also serve as a basis for reconstruction of a sign in virtual avatars (Forte et al., 2023; Baltatzis et al., 2024).

## 6. The SignBeach dataset

### 6.1. Dataset structure

The SignBeach dataset contains recordings of 1401 lexical signs, each produced by four deaf signers, resulting in 5604 items. For each sign, video recordings are available from five camera angles (L, A, M, B, R) and in both orientations (front and side), resulting in 20 video files per lexical item (4 signers (2 front, 2 side) $\times$ 5 cameras).

The dataset is organized around Dutch target words (`wordNL`), which serve as the primary item identifiers. Each item is linked to the NGT Signbank through a gloss identifier (`glossID`) and a Signbank ID, from which the full Signbank URL can be reconstructed.[3] Table 1 summarizes the variables in the dataset.

Video files and their derivatives follow a consistent naming convention: `wordNL_signerID_orientation_cameraID_format.suffix` For example, `hond_theo_front_M_video.mp4` refers to the video recording (video) of the sign for 'hond' (dog), recorded by signer Theo, signing frontally (front) at the middle camera (M). This naming scheme allows immediate identification of the content without consulting external metadata, and enables straightforward parsing into dataset fields.

[3]Signs that are not available on Signbank have these fields marked as NA.

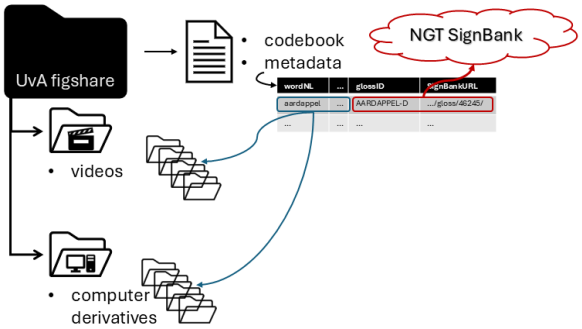



Figure 2: Illustration of the project architecture and links to the NGT Signbank as an external resource.

The derivative files follow the same base name with different extensions: `.pose` for MediaPipe estimates, `.hamer` for hand mesh data, and `.s3d` for full-body reconstructions.

### 6.2. Availability and license

The dataset is archived on Figshare at https://doi.org/10.21942/uva.c.8233876, organized into 50 sub-datasets grouped alphabetically by the first letter of the Dutch target word. This partitioning accommodates Figshare's limit of 500 files per dataset. Letters with many entries (e.g., 'S' with five parts, 'K' and 'V' with four parts each) are split across multiple datasets (see Figure 2). A Figshare Collection groups all datasets under a single citable DOI, with a total size of approximately 27 GB. All materials are released under a CC BY-NC 4.0 license for non-commercial reuse.

## 7. Conclusion and perspectives

This article presents a dataset of high quality videos for approx. 1400 NGT signs with derivational data suitable for scientific reuse, for example in the context of psycholinguistic studies or machine learning developments. High quality video recordings are available from five angles, providing 3D-information in the raw video data. In addition, we provide mediapipe, hammer, and SAM 3D reconstruction outputs for computational processing and analyses.

### 7.1. Links to other projects

The SignBeach dataset presented in this paper is the product of a larger project on sign learning in primary school-aged children. All resources and publications from this project will be be linked to the original video materials and derivational data provided in this dataset, permitting full academic reuse of these materials for future research.

In the SignBeach project, we are interested in investigating factors that contribute to sign learning. One factor that has been shown to impact sign

| Variable | Description | Possible values |
|---|---|---|
| wordNL | Dutch target word | *Dutch words or expressions* |
| wordEN | English translation | *English words or expressions* |
| signerID | Signer identifier | [theo, tobias, manouk, eva] |
| orientation | Orientation of the signer relative to the central camera | [front, side] |
| cameraID | Camera identifier | [L, A, M, B, R]<br>(left to right from signer's perspective) |
| glossID | ID gloss from the NGT Signbank | *e.g.*, HOND, SCHOOL;<br>compounds: GLOSS+GLOSS |
| SignBankURL | URL to the sign entry on the NGT Signbank | signbank.cls.ru.nl/dictionary/gloss/{id} |
| video | Video file | filenames*_video.mp4 |
| mediapipe | MediaPipe Holistic keypoint estimates | filenames*_mediapipe.pose |
| hamer | HaMeR hand mesh keypoint estimates | filenames*_hamer.hamer |
| s3d | SAM 3D Body full-body reconstruction | filenames*_s3d.s3d |
| *filenames | Naming system | *wordNL_signerID_orientation_cameraID* |

Table 1: Variables in the SignBeach dataset. File extensions indicate format: `.pose` for MediaPipe JSON, `.hamer` for HaMeR JSON, and `.s3d` for SAM 3D Body NPZ.

learning is iconicity (for a range of different learning contexts and effects, see for example Ortega, 2017; Caselli and Pyers, 2017, 2020; Karadöller et al., 2024). In another study within this project, we therefore collected iconicity ratings from Dutch and German non-signers and Dutch signers and computed transparency estimates from translation attempts by Dutch non-signers for all items in the dataset (Schiefner et al., 2026). Since the obtained iconicity and transparency data is aligned with the items provided in the present dataset by corresponding filenames, the two resources can easily be combined to answer additional research questions.

Data about childrens' learning trajectories, in turn, will also be made available, allowing us to link the visual information derived from video materials with children's learning trajectories over time or at specific time points, such as at first exposure. This creates an integrated set of resources in which lexical, visual, and behavioural data is brought together. This will facilitate work on NGT lexical processing and acquisition, as well as comparative studies that situate NGT within a broader cross-linguistic context, e.g., by comparing the NGT dataset with data from ASL-LEX (Caselli et al., 2017; Sehyr et al., 2021) and similar resources.

## 7.2. Limitations

While the SignBeach dataset is designed to maximise opportunities for scientific reuse, a number of limitations need to be highlighted that will restrict interpretation and opportunities for analysis. First, the lexical coverage of this dataset is constrained by the goals of the SignBeach project. Items were selected to be child-friendly and broadly known. As a result, the database focuses on everyday vocabulary and more specialised as well as later-learned vocabulary is likely to be underrepresented. Within the concepts represented, variant choice was driven by existing dictionaries and signers' input during the recording process, likely favouring widely used variants and falling short of reflecting the lexical variation present in NGT.

Second, signs are recorded in isolation in a highly controlled studio environment using the signCollect pipeline (Otterspeer et al., 2024). This setting ensures high recording quality but does not represent naturalistic, conversational signing styles. However, the instructions to envision a child audience for the signs should result in relatively lively renditions, that are likely to be more natural than what is often used in lexical references.

Finally, the dataset also only includes productions from a limited and fairly homogenous set of four signers, an issue highlighted for a number of similar databases by Bragg et al. (2019). This design supports the analysis of articulatory variation across signers but does not reflect regional, generational, or sociolinguistic diversity in the NGT signing community, nor does our sample of signers reflect the demographic diversity of the deaf community in the Netherlands.

## 7.3. Perspectives for future use

We anticipate that the dataset presented in this paper provides extensive opportunities for scientific reuse. The video materials and computational derivatives provide relevant information for phonological and kinematic analysis of the lexical signs provided, with the parallel recordings from four distinct signers providing opportunities to study articulatory variation. In parallel with sensor-based motion capture data, high-quality recordings and video-derived pose and motion estimation, such

as presented here, can significantly advance sign processing and analysis.

The video materials can also be reused as stimulus materials for future psycholinguistic studies, as is or in combination with the computational derivatives. For example, analysis of the computational derivatives may provide timing information for creating onset-controlled stimulus materials directly from the videos, or form the basis for creating synthetic representations of the signs by avatars that allow for targeted modification of the signs for even more controlled stimulus presentation.

Meanwhile, for computational research, the dataset provides a rich testbed for computer vision and sign language technology. This could include benchmarking pose estimation models, training and evaluating sign recognition systems, and exploring methods for realistic avatar animation grounded in naturalistic sign data.

By combining this rich, multi-angle video data with information from other studies in the SignBeach project, such as iconicity ratings, transparency scores, and information about the learnability of signs, analyses of visual information embedded in these lexical signs that contributes to these different factors will also be possible. Beyond the SignBeach project itself, the tight alignment with NGT Signbank gloss-IDs means that the SignBeach dataset easily integrates with the existing research infrastructure for NGT.

## 8. Acknowledgements

We thank Tobias de Ronde, Debby Musch, Machiel Ouwerkerk, Eva Prevaes, Theo Stoop and Manouk van Asperen who served as models for the video materials presented in this paper and gave valuable input in the sign selection process.

## 9. Bibliographical References

Sobhan Asasi, Mohamed Ilyas Lakhal, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Beyond gloss: A hand-centric framework for gloss-free sign language translation. ArXiv:2507.23575 [cs].

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024. Neural sign actors: A diffusion model for 3D sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1985–1995.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31, New York, NY, USA. Association for Computing Machinery.

Naomi K. Caselli and Jennie E. Pyers. 2017. The road to language learning is not entirely iconic: Iconicity, neighborhood density, and frequency facilitate acquisition of sign language. *Psychological Science*, 28(7):979–987.

Naomi K. Caselli and Jennie E. Pyers. 2020. Degree and not type of iconicity affects sign language vocabulary acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(1):127–139.

Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. 2017. ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods*, 49(2):784–801.

Onno Crasborn, Richard Bank, Inge Zwiterlood, Els Van Der Kooij, Ellen Ormel, Johan Ros, Anique Schüller, Anne De Meijer, Merel Van Zuilen, Ellen Nauta, Frouke Van Winsum, and Max Vonk. 2020. NGT Signbank. Nijmegen, NL. Radboud University, Centre for Language Studies. Version: 3.0.

Onno Crasborn, Richard Bank, Inge Zwitserlood, Els van der Kooij, Anique Schüller, Ellen Ormel, Ellen Yassine Nauta, Merel van Zuilen, Frouke van Winsum, and Johan Ros. 2016. Linking Lexical and Corpus Data for Sign Languages: NGT Signbank and the Corpus NGT. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 41–46, Portorož, Slovenia. European Language Resources Association (ELRA).

Onno Crasborn and Inge Zwitserlood. 2008. The Corpus NGT: an online corpus for professionals and laymen. In *Construction and Exploitation of Sign Language Corpora*, pages 44–49, Paris. ELDA.

Onno Crasborn, Inge Zwitserlood, Els Van Der Kooij, and Anique Schüller. 2018. Global SignBank manual.

J. Fenlon, K. Cormier, and A. Schembri. 2015. Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography*, 28(2):169–206.

Maria-Paola Forte, Peter Kulits, Chun-Hao Paul Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. 2023. Reconstructing signing avatars from video using linguistic priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12791–12801.

Eva Gutierrez-Sigut, Brendan Costello, Cristina Baus, and Manuel Carreiras. 2016. LSE-Sign: A lexical database for Spanish Sign Language. *Behavior Research Methods*, 48(1):123–137.

Low Jian He, Harry Walsh, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Hands-on: Segmenting individual signs from continuous sequences. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–5, Tampa/Clearwater, FL, USA. IEEE.

Elena Jahn, Reiner Konrad, Gabriele Langer, Sven Wagner, and Thomas Hanke. 2018. Publishing DGS corpus data: Different formats for different needs. In *Workshop Proceedings. 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, volume 12 May 2018. ELRA, pages 83–90, Miyazaki, Japan.

Dilay Z. Karadöller, David Peeters, Francie Manhardt, Aslı Özyürek, and Gerardo Ortega. 2024. Iconicity and gesture jointly facilitate learning of second language signs at first exposure in hearing nonsigners. *Language Learning*, 74(4):781–813.

Ulrika Klomp, Lisa Gierman, Pieter Manders, Ellen Nauta, Gomèr Otterspeer, Ray Pelupessy, Galya Stern, Dalene Venter, Casper Wubbolts, Marloes Oomen, and Floris Roelofsen. 2024. An extension of the NGT dataset in Global Signbank. In *Proceedings of the LREC-COLING 2024*, pages 178–183.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. *MY DGS – annotated. Public corpus of German Sign Language*, 3rd edition. Universität Hamburg.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association (ELRA).

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, and others. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*.

Max Planck Institute for Psycholinguistics, The Language Archive. 2025. ELAN (Version 7.0).

Hope Morgan, Wendy Sandler, Rose Stamp, and Rama Novogrodsky. 2022. ISL-LEX v.1: An online lexical resource of Israeli Sign Language. In *Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages*, pages 148–153, Marseille, FR.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020. Real-time sign language detection using human pose estimation. In *Computer Vision – ECCV 2020 Workshops*, pages 237–248, Cham. Springer International Publishing.

Nederlands Gebarencentrum. 2012. *Mijn eerste 1500 gebaren*, 1 edition. Nederlands Gebarencentrum, Bunnik, NL.

Nederlands Gebarencentrum. 2024. *Online Woordenboek NGT*. Nederlands Gebarencentrum.

Gerardo Ortega. 2017. Iconicity and sign lexical acquisition: A review. *Frontiers in Psychology*, 8:1280.

Felicitas Otte, Elena Jahn, Cornelia Loos, Julian Bleicken, and Annika Herrmann. 2022. The DGS corpus as a linguistic resource: student research and beyond. *Linguistische Berichte*, (272):447.

Gomer Otterspeer, Ulrika Klomp, and Floris Roelofsen. 2024. SignCollect: A 'touchless' pipeline for constructing large-scale sign language repositories. In *Proceedings of the LREC2024 11th Workshop on the representation and processing of sign languages: Evaluation of sign language resources*, pages 269–275, Turin. European Language Resources Association.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024a. HaMeR: Hand mesh recovery for the EgoExo4D hand pose challenge.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024b. Reconstructing hands

in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836.

Adam Schembri, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. 2017. *British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017*, 3 edition. University College London, London.

Annika Schiefner, Floris Roelofsen, and Beyza Sümer. 2026. Iconicity and transparency in Dutch Sign Language. *Behavior Research Methods*, 58(87).

Marc Schulder, Dolly Blanck, Thomas Hanke, Ilona Hofmann, Sung-Eun Hong, Olga Jeziorski, Lutz König, Susanne König, Reiner Konrad, Gabriele Langer, Rie Nishio, and Christian Rathmann. 2024. Data statement for the public DGS corpus. *DGS-Korpus Project Note AP06-2020-01*. Version: 3.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 project: A database of lexical and phonological properties for 2,723 signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277.

Patrick C. Trettenbrein, Jan Bümmerstede, Thomas A. Finkbeiner, Pia Gehlbach, Nina-Kristin Meister, Annika Schiefner, Petra B. Schumacher, Door Spruijt, Markus Steinbach, and Pamela Perniss. 2025. Towards a "DGS-LEX": A roadmap for the collaborative creation of a psycholinguistic database for German Sign Language (DGS). *Linguistische Berichte*, 281:91–100.

Freya Watkins, Diar Abdlkarim, Bodo Winter, and Robin L. Thompson. 2024. Viewing angle matters in British Sign Language processing. *Scientific Reports*, 14(1):1043.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. 2025. SAM 3D Body: Robust full-body human mesh recovery. *arXiv preprint*.

# Comparing Computer Vision Instruments for Eye Blink Analysis

**Margaux Susman, Carla Miquel Blasco, Jan Bulla**
University of Bergen
Bergen, Norway
margaux.susman@uib.no, carla.blasco@uib.no, jan.bulla@uib.no

**Abstract**

We compared four tools for analysing blink velocity and amplitude, examining how MediaPipe, OpenFace, InsightFace, and 3DDFA compare in terms of blink analysis. Building on previous findings that different tools yield different results (Kuznetsova and Kimmelman, 2024), we explored their fixed-effect estimates between linguistic versus non-linguistic blinks, within non-linguistic blinks (eye-watering blinks versus gaze-direction-change blinks), and within linguistic blinks (prosodic/turn-taking blinks, sign-aligned/list-marking blinks, and backchanneling blinks), while controlling for head pose (Pitch, Roll, Yaw). Using linear mixed-effects models on annotated French Sign Language data, we found tool-specific patterns: consistent negative effects for InsightFace and MediaPipe, but positive effects for 3DDFA. In addition, the influence of head pose varied across models (Pitch is strongly positive in MediaPipe but negative in InsightFace and some 3DDFA models; Roll and Yaw also vary in importance across tools). These discrepancies highlight methodological biases that can distort linguistic interpretations.

**Keywords:** blinks, computer vision, statistical analysis

## 1. Introduction

Computer vision tools are increasingly used in research on gestures and sign languages. In this context, they offer the possibility to track manual and non-manual movements, such as eye blinks, in a non-invasive and scalable way. Blinks constitute a particularly relevant object of study, as their timing and kinematic properties have been shown to relate to linguistic structure (Herrmann, 2010). Despite the growing use of computer vision techniques, little work has systematically evaluated how different tools compare when applied to fine-grained kinematic measures of non-manuals (Sargano et al., 2024; Kuznetsova and Kimmelman, 2024). In particular, potential tool-specific biases in the estimation of blink velocity and amplitude remain underexplored. The aim of this paper is to compare several computer vision tools: MediaPipe, OpenFace, InsightFace, and 3DDFA for the analysis of blink velocity and amplitude, using the same video data from the Dicta-Sign-LSF-v2 corpus (Belissen et al., 2020). The four tools examined in this study were selected to reflect established and more recent developments in computer vision. OpenFace and MediaPipe are widely used open-source frameworks for facial landmark detection and have been frequently adopted in linguistic and multimodal research, making them important reference points for comparison. InsightFace and 3DDFA, by contrast, represent more recent approaches that incorporate advances in deep-learning face analysis and 3D face alignment. We focus on blink velocity and amplitude across linguistically annotated blink functions. We contrast linguistic versus non-linguistic blinks, differences within non-linguistic blinks (eye-watering blinks versus gaze-direction-change blinks), and differences within linguistic blinks (prosodic/turn-taking, sign-aligned/list-marking, and backchanneling blinks) while controlling for head movements. Our evaluation reveals variation in measurements and statistical outcomes across tools. These include, e.g., higher estimated velocities for MediaPipe compared to OpenFace and reversed effect directions for 3DDFA in some models. We show that different tools exhibit different patterns and therefore highlight how measurement choices can shape the outcomes of blink kinematics analysis. We note that linguistic effects may partly reflect properties of the analysis pipeline rather than underlying behaviour.

## 2. Background

### 2.1. Blinks in Linguistic Research

Blinks are small-amplitude, high-frequency rapid closures (whether partial or complete) of the eyes. They are frequent in face-to-face interaction and have been studied in fields such as cognitive science, clinical research, and linguistics (Nyström et al., 2024; Herrmann, 2010). In research on gestures and sign languages in particular, blinks are of interest as non-manual events that may co-occur with discourse structure (e.g., marking turns and prosodic boundaries). In sign language research, it has been shown that blinks co-occur with lexical units as well. Because of their prominent role, blinks have been incorporated into sign language analyses since the mid-1990s. (Wilbur, 1994; Sze, 2004; Braffort and Chételat-Pelé, 2012; Herrmann,



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 459–467
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2010).

At the same time, blinks offer a methodological challenge for empirical work. They are short, involve small movements, and are susceptible to interference from head movements and gaze direction. These properties make blinks particularly vulnerable to annotation uncertainty and measurement error, whether blinks are annotated manually or extracted automatically. As corpus-based and experimental linguistic studies increasingly rely on computer vision tools to scale up the analysis of non-manual behaviours, it becomes necessary to understand how methodological choices affect the resulting blink measures.

Despite the growing use of automated facial analysis in linguistic research, blink measurements have not yet been examined in a conclusive manner. Consequently, different analytical pipelines may yield different estimates of blink velocity and amplitude, with direct consequences for statistical modelling and linguistic interpretation. Without explicit comparison, it remains unclear whether reported effects reflect properties of the linguistic data or echoe noise and artifacts of the measurement tool.

### 2.2. Computer Vision Tools

This study compares four widely used computer vision frameworks for facial analysis: MediaPipe, OpenFace, InsightFace, and 3DDFA. Although all four tools enable the extraction of facial landmarks and head pose information, they rely on different modelling assumptions, training data, and tracking strategies, which may affect their performance on subtle and fast facial movements such as eye blinks.

MediaPipe is an open-source framework developed by Google that provides real-time, cross-platform pipelines for perception tasks (Lugaresi et al., 2019). Its Face Mesh model estimates 468 three-dimensional facial landmarks using a lightweight neural network optimized for speed and robustness. MediaPipe relies on a combination of face detection and landmark regression models trained on large-scale datasets, enabling stable tracking under varying lighting and pose conditions. Due to its dense landmark representation around the eyes, MediaPipe has become popular for fine-grained facial motion analysis, including blink detection (e.g., Delgado Gómez et al., 2024; Rahman et al., 2025).

OpenFace is an open-source toolkit designed for facial behaviour analysis in research contexts (Amos et al., 2016). It provides facial landmark detection, head pose estimation, eye gaze tracking, and facial action unit recognition. OpenFace is based on a constrained local neural field (CLNF) landmark detector, which uses statistical models combining local visual cues and general facial shape information to track facial landmarks consistently over time. Compared to MediaPipe, OpenFace uses a smaller set of facial landmarks but integrates them within a well-established facial behaviour analysis framework. Its emphasis on interpretability and reproducibility has led to widespread adoption in linguistics and social signal processing research (Zadeh et al., 2018).

InsightFace is a deep-learning-based framework primarily developed for face recognition and analysis (Guo et al., 2021). It includes modules for face detection, alignment, and landmark localisation based on state-of-the-art convolutional neural networks. InsightFace relies on large-scale training data and modern backbone architectures, resulting in high accuracy in face alignment tasks. While it is less commonly used in linguistic research, its strong performance in landmark detection renders it a promising candidate for blink analysis.

3DDFA (3D Dense Face Alignment) is a model-based approach that fits a three-dimensional morphable face model to two-dimensional images (Guo et al., 2020). It estimates dense 3D facial geometry and head pose by optimizing the parameters of a statistical face model. Unlike purely landmark-based systems, 3DDFA explicitly models facial shape and pose in three dimensions, which may improve robustness to large head movements. However, because the model is fitted separately to each frame, the measurements can vary slightly over time and be influenced by image noise. This is especially problematic for small and rapid movements such as eyelid motion.

These tools differ in their facial representations, ranging from dense landmark meshes (MediaPipe) and sparse landmark models (OpenFace) to deep alignment networks (InsightFace) and parametric 3D face models (3DDFA). They also vary in their optimisation objectives and how they smooth measurements over time. As a result, they may respond differently to rapid, low-amplitude movements such as blinks. Understanding these methodological differences is essential for interpreting between-tool variation in blink kinematic measures and for assessing the robustness of computer-vision-based analyses of non-manual behaviour.

## 3. Methods

We used the Dicta-Sign-LSF-v2 dataset (Belissen et al., 2020), which contains recordings of discussions on European travel. We annotated the blinks from 8 videos using the ELAN (version 6.2). We annotated a total of 1078 blinks and considered their functions. Only 650 blinks were analysed further, because they had a single function, whereas the remaining blinks had two or

more concurrent functions.

For each frame in each video, we then extracted the eye coordinates and head pose (which are known to influence facial coordinates, see Kuznetsova and Kimmelman, 2024) using each tool under comparison (MediaPipe, OpenFace, InsightFace, and 3DDFA) The 3D coordinates for each video were saved to CSV files. This step was implemented in Python. After that, we calculated the Eye Aspect Ratio (EAR) for each blink using the extracted coordinates. This ratio was introduced by Soukupova and Cech (2016) and is a measure of the degree of eye openness. It was computed from a fixed set of six eyelid landmarks per eye; for MediaPipe, we selected the closest available Face Mesh landmarks corresponding to the standard dlib eye contour (MediaPipe landmarks: left eye indices: 249, 362, 373, 380, 385, 387 and right eye indices: 7, 133, 144, 153, 158, 160). Then we applied the same EAR formula as for the other tools.

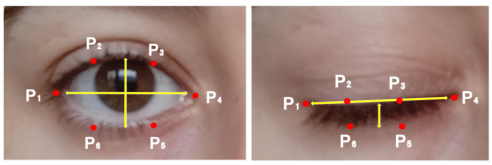



Figure 1: Eye landmark positions for the EAR calculation.

We merged the CSV files and removed rows without a blink and rows with multiple-function-blinks. For each blink, we retained only the first head-rotation measurement because of the short blink duration.

Subsequently, we derived blink kinematics (i.e. velocity and amplitude) from the smoothed EAR rather than from raw frame-by-frame EAR values in R (R Core Team, 2026; Bulla and Kimmelman, 2026). We selected smoothing splines as the smoothing method. More precisely, we relied on the function `ss` from the R package `npreg` (Helwig, 2020, 2024) to fit quintic smoothing splines via penalised least squares. The smoothing parameter $\lambda$ was chosen by generalised cross-validation (GCV). For further details on the implementation, see (Bulla and Kimmelman, 2026). The smoothing served to reduce measurement noise and for capturing the overall eyelid movement during a blink. Blink amplitude reflects how far the eyelids closed at the blink peak and is expressed in units of EAR. Average blink velocity reflects how quickly the eyelids moved during the blink, and it is expressed as the average rate of change of EAR over time.

After computing these values, we added an annotation column based on the blink name, containing the annotation label. The labels were divided into linguistic and non-linguistic functions, and further into subfunctions. Linguistic subfunctions include: i) backchanneling blinks (feedback to an interlocutor), ii) prosodic blinks (boundary marking), iii) turn-taking blinks (beginning and end of a turn), iv) list-marking blinks, and v) blinks occurring with specific lexical signs. Prosodic and turn-taking blinks were grouped into one larger category, because both mark boundaries. In addition, sign-marking blinks and list-marking blinks were grouped together, because of their co-occurrence with specific signs. Non-linguistic blinks included eye-watering blinks and blinks induced by a change in gaze direction. We created three different datasets: for i) all blinks, ii) only large linguistic blink categories, and iii) only non-linguistic blinks. After this preparation, we carried out the statistical analysis of velocity and amplitude in R. There were six different analyses per tool tested: more specifically, we examined differences in amplitude and velocity for each of the three above-described datasets. The modelling procedure followed this logic: we fitted a baseline linear mixed-effects model (no fixed effects, random intercept for signer, estimated via maximum likelihood) using the `lme()` function from the `nlme` package (Pinheiro et al., 2025). Then we performed stepwise model selection by extending the baseline model towards more complex models, adding one predictor at a time. At each step, we selected the preferred model using the likelihood ratio test (LRT) for nested models and the AIC for non-nested comparisons. To improve residual normality, we applied a log or square-root transformation when necessary. As the linguistic models involved multiple comparisons, we used the package `marginaleffects` (Arel-Bundock et al., 2024) to perform post-hoc pairwise comparisons with Holm correction, thereby controlling the family-wise error rate. The processed blink kinematic measures and analysis scripts are available at the Open Science Framework (OSF): https://osf.io/r9yfs.

## 4. Results

In the following, we present the results for blink kinematics across our three linguistic categories for the four computer vision tools. We first report amplitude, then velocity.

### 4.1. Amplitude

**Linguistic versus non-linguistic blinks** - Table 1 presents estimated fixed effects from our selected models for predicting blink amplitude across the

four facial analysis tools considered ($N = 631$). The outcome scales vary: amplitude values were square-root transformed for the first three tools, and we applied a log transformation for 3DDFA.

We found that non-linguistic blinks show consistently reduced amplitude compared to linguistic ones: OpenFace ($p = 0.002$), MediaPipe ($p = 0.044$), 3DDFA ($p < 0.001$). Model selection did not include blink category for InsightFace. Head movement effects diverge across tools: Pitch is strongly positive in MediaPipe ($p < 0.001$), but negative for InsightFace ($p = 0.008$) and 3DDFA ($p < 0.003$). Yaw and Roll are positive in OpenFace (both $p < 0.001$), and Pitch, Yaw, and Roll are negative in 3DDFA ($p = 0.003, 0.031, 0.006$).

**Non-linguistic blinks** - Table 2 summarises fixed effects for the selected models predicting non-linguistic blink amplitude, i.e., differences between eye-watering blinks and gaze-direction-change blinks (baseline) across our four facial analysis tools ($N = 120$). We applied a square-root transformation for the OpenFace, MediaPipe and InsightFace data, while we used a logarithmic transformation for 3DDFA.

Where included, the blink category shows strong negative effects indicating reduced amplitude: InsightFace ($p < 0.001$) and OpenFace ($p < 0.001$). MediaPipe did not include the blink category (Yaw-only best model). Head movement effects diverge by tool: Pitch is positive in InsightFace ($p = 0.033$) but negative in OpenFace ($p = 0.011$); Yaw is positive in MediaPipe ($p < 0.001$); in 3DDFA, Roll ($p = 0.041$) and Yaw ($p = 0.004$) are negative.

**Linguistic blinks** - Table 3 summarises fixed effects for the best models predicting linguistic blink amplitude, i.e., the differences between backchanneling blinks (baseline), prosodic/turn-taking blinks, and sign-aligned/list-marking blinks across our four facial analysis tools ($N = 511$ observations).

Linguistic categories show tool-specific divergence. MediaPipe reveals clear reductions (prosodic/turn-taking, $p = 0.029$; sign-aligned/list-marking, $p = 0.027$). InsightFace trends show smaller reductions (prosodic/turn-taking, $p = 0.043$, sign-aligned/list-marking, $p = 0.069$). OpenFace shows no detectable effects ($p > 0.55$). 3DDFA reverses the pattern with positive coefficients (prosodic/turn-taking, $p < 0.001$; sign-aligned/list-marking, $p < 0.001$).

In Table 4, MediaPipe confirms its statistically significant results: backchanneling ($3.72$) > prosodic/turn-taking ($2.65$, Holm $p = 0.005$) > sign-aligned/list-marking ($2.23$, Holm $p = 0.034$). Other tools show non-significant trends post-correction.

As we have seen, non-linguistic blinks are generally smaller than linguistic ones for most tools, but the size and direction of this effect depend on the tool. InsightFace and MediaPipe tend to show reduced amplitude for prosodic/turn-taking and sign-aligned/list-marking blinks relative to backchanneling, OpenFace shows little categorical contrast, and 3DDFA often reverses the pattern. Head pose (Pitch, Roll, Yaw) also affects amplitude, but its sign and importance vary strongly by tool.

## 4.2. Velocity

The results for velocity analysis across our four tools are presented in this section.

**Linguistic versus non-linguistic blinks** - Table 5 shows fixed effects from our best models predicting blink velocity across our four tools ($N = 631$). The outcome scales vary: untransformed velocity (InsightFace), square-root-transformed velocity (OpenFace and MediaPipe) and log-transformed velocity (3DDFA).

We found that non-linguistic blinks show consistently reduced velocity compared to linguistic blinks where included: OpenFace ($p = 0.014$), MediaPipe ($p = 0.043$), and 3DDFA ($p < 0.001$).

**Non-linguistic blinks** - Table 6 summarises fixed effects from our best linear mixed-effects models, predicting non-linguistic blink velocity across our four tools. The outcome scales were a square-root transformation for InsightFace, OpenFace, and MediaPipe, and log transformation for 3DDFA.

Non-linguistic blink velocity analysis exhibits consistent patterns across tools. InsightFace and OpenFace both included the blink category in their best model and show strong negative effects for eye-watering blinks ($p < 0.001$), indicating that eye-watering blinks are reliably slower than gaze-direction-change blinks. MediaPipe and 3DDFA highlight head movement effects with Roll significant in both tools ($p < 0.001$ and $p = 0.020$, respectively).

**Linguistic blinks** - Table 7 summarises fixed effects for the best models predicting linguistic blinks velocity, specifically the differences between backchanneling blinks (baseline), prosodic/turn-taking blinks, and sign-aligned/list-marking blinks across the four tools ($N = 511$). In Table 7, we see a consistent pattern of slowing across the blink categories compared to backchanneling blinks, our baseline, but the strength and the direction of the effect depend on both the tool the and outcome scale used. The outcome scale varied depending on the tool: no transformation was applied for InsightFace, a square-root transformation was applied for OpenFace and MediaPipe, and a log transformation for 3DDFA. Across the three different tools for which we modelled blink categories (InsightFace, OpenFace, and MediaPipe), prosodic/turn-taking and sign-aligned/list-marking blinks tend to have

| **Predictor** | InsightFace | OpenFace | MediaPipe | 3DDFA |
|---|---|---|---|---|
| | (P + cat) | (Y + R + cat) | (P + cat) | (cat + P + R + Y) |
| **Fixed effects** | | | | |
| (Intercept) | $0.267, p < 0.001$ | $0.175, p < 0.001$ | $0.497, p = 0.024$ | $-0.491, p = 0.013$ |
| Pitch | $-0.001, p = 0.008$ | | $17.657, p < 0.001$ | $-2477, p = 0.003$ |
| Yaw | | $0.045, p < 0.001$ | | $-1315, p = 0.031$ |
| Roll | | $0.052, p < 0.001$ | | $-954, p = 0.006$ |
| blink cat. (NOTLING) | $0.016, p = 0.016$ | $-0.014, p = 0.002$ | $-0.171, p = 0.044$ | $-0.405, p < 0.001$ |

Table 1: Statistical results of the four tools for the comparison of linguistic versus non-linguistic blink amplitude. (P = Pitch, R = Roll, Y = Yaw, cat = blink category). **Colour coding:** Green = significant positive (p<0.05), Red = significant negative (p<0.05), Gray = non-significant.

| **Predictor** | InsightFace | OpenFace | MediaPipe | 3DDFA |
|---|---|---|---|---|
| | (cat + P) | (cat + P) | (R) | (R + Y + cat) |
| **Fixed effects** | | | | |
| (Intercept) | $0.358, p < 0.001$ | $0.215, p < 0.001$ | $0.469, p < 0.001$ | $-1.176, p < 0.001$ |
| Pitch | $0.001, p = 0.033$ | $-0.056, p = 0.011$ | | |
| Roll | | | | $-3587, p = 0.041$ |
| Yaw | | | $206.290, p < 0.001$ | $-3099, p = 0.004$ |
| blink cat. (WAT) | $-0.088, p < 0.001$ | $-0.047, p < 0.001$ | | $-0.300, p = 0.111$ |

Table 2: Statistical results of the four tools for the comparison of non-linguistic blink categories (gaze-direction-change blinks (reference) and eye-watering blinks) on amplitude. (P = Pitch, R = Roll, Y = Yaw, cat = blink category). **Colour coding:** Green = significant positive (p<0.05), Red = significant negative (p<0.05), Gray = non-significant.

a lower velocity than backchanneling blinks (negative coefficients for sign-aligned/list-marking and prosodic/turn-taking blinks for InsightFace ($p = 0.003$ and $p = 0.050$ respectively), OpenFace (sign-aligned/list-marking $p = 0.016$) and MediaPipe (prosodic/turn-taking $p = 0.012$)). The InsightFace model suggests that backchanneling has the highest velocity, followed by prosodic/turn-taking blinks which are slightly slower and sign-aligned/list-marking blinks which are the slowest. Only the difference between backchanneling and sign-aligned/list-marking was statistically significant. OpenFace shows the same pattern as InsightFace but less pronounced: again, the difference between backchanneling and sign-aligned/list-marking is the only significant one. This suggests that OpenFace detects a similar slowing, but with weaker contrast. The MediaPipe model shows a clear separation between backchanneling and both prosodic/turn-taking and sign-aligned/list-marking categories: similar to InsightFace and OpenFace, backchanneling blinks have the highest velocity. The coefficients for prosodic/turn-taking (and to a lesser extent sign-aligned/list-marking) are negative and statistically significant. 3DDFA exhibits a completely different pattern. That is, prosodic/turn-taking and sign-aligned/list-marking show positive coefficients relative to backchanneling, implying higher velocity than baseline.

Overall, we note that non-linguistic blinks tend to have a slower velocity than linguistic ones (OpenFace/MediaPipe/3DDFA $p < 0.05$). Linguistic blinks show a backchanneling > prosodic/turn-taking > sign-aligned/list-marking gradient in InsightFace/OpenFace/MediaPipe, while we observe the opposite pattern with 3DDFA. For non-linguistic blinks, we note that eye-watering blinks are significantly slower than gaze-direction-change blinks (InsightFace/OpenFace $p < 0.001$).

## 5. Discussion

Our analyses show that the estimates of blink kinematics differ systematically across computer vision tools. In many cases, MediaPipe, OpenFace, and InsightFace agreed on effect direction, while 3DDFA tended to show weaker or opposite patterns, especially for linguistic blink categories. This shows that the linguistic effects on blink velocity and amplitude are not completely tool-independent, even when using identical video data and annotations.

The first implication is that we cannot assume that blink kinematic measures will behave identically across different analysis pipelines. In our data, non-linguistic blinks showed smaller amplitude and slower velocity than linguistic blinks in several tools. Linguistic categories showed the gradient: backchanneling > prosodic/turn-taking > sign-

| **Predictor** | InsightFace | OpenFace | MediaPipe | 3DDFA |
|---|---|---|---|---|
| | (P + cat) | (null) | (P + R + cat) | (cat + P + R) |
| **Fixed effects** | | | | |
| (Intercept) | $0.270, p < 0.001$ | $0.172, p = 0$ | $1.072, p = 0.256$ | $-0.748, p = 0.002$ |
| Pitch | $-0.001, p = 0.006$ | | $56.383, p < 0.001$ | $-2259, p = 0.008$ |
| Roll | | | $355.472, p = 0.007$ | $-459, p = 0.044$ |
| Yaw | | | | |
| blink cat. (PROS-TURN) | $-0.012, p = 0.043$ | | $-1.072, p = 0.029$ | $0.599, p < 0.001$ |
| blink cat. (SIGN-LIC) | $-0.014, p = 0.069$ | | $-1.496, p = 0.027$ | $0.641, p < 0.001$ |

Table 3: Statistical results of the four tools for the comparison of linguistic blink categories (backchanneling (reference), prosodic/turn-taking & sign-aligned/list-marking blinks) on amplitude. (P = Pitch, R = Roll, Y = Yaw, cat = blink category). **Colour coding:** Green = significant positive (p<0.05), Red = significant negative (p<0.05), Gray = non-significant.

| Tool | ADFE Mean (SE) | PROS-TURN Mean (SE) | SIGN-LIC Mean (SE) |
|---|---|---|---|
| InsightFace | $0.280\ (0.007)$ | $0.269\ (0.007)$ | $0.266\ (0.009)$ |
| OpenFace | $0.172\ (0.015)$ | $0.172\ (0.015)$ | $0.176\ (0.015)$ |
| MediaPipe | $3.720\ (0.927)$ | $2.650\ (0.951)^*$ | $2.230\ (1.052)$ |
| 3DDFA | $-0.888\ (0.235)$ | $-0.289\ (0.238)$ | $-0.247\ (0.252)$ |

Table 4: Post-hoc Pairwise Comparison (Response Scale, Holm Corrected) ($^*p < .01$ after correction; all others $p > .05$ except ADFE baselines), ADFE = backchanneling, PROS-TURN = prosodic/turn-taking, SIGN-LIC = sign-aligned/list-marking

aligned/list-marking blinks. However, the strength - and sometimes the direction - of the observed patterns varied by tool and, in some cases, by transformation (untransformed, square-root, log).

Second, the head-pose effects, which vary and are sometimes contradictory, show that controlling for pose is necessary but not sufficient: different tools estimate head pose in different ways. For example, some tools start from 2D landmarks, whereas 3DDFA fits a dense 3D morphable face model. Pitch contributed positively to amplitude and velocity in some MediaPipe models but negatively in others, and the relative importance of Roll and Yaw differed sharply between MediaPipe and 3DDFA. This suggests that head-pose predictors may compensate for tool-specific tracking behaviour rather than simply capturing signer movement.

The observed divergence across tools can also be attributed to their differences in modelling assumptions and facial representations. MediaPipe relies on a dense neural-network-based landmark mesh optimised for real-time robustness, OpenFace combines statistical shape constraints with landmark-based tracking, InsightFace employs deep alignment networks trained on large-scale datasets, and 3DDFA fits a parametric three-dimensional face model to each frame. Although we do not detail the exact causes, differences in tool design seem to affect blink estimates and statistical outcomes.

Note that our aim is to inform, not to endorse any specific tool as superior. We observed tool-dependent effects, so our study does not warrant strong recommendations about which tool should be favoured for blink analysis in sign language research. Instead, we encourage explicit reporting of the full analysis pipeline (tool versions, transformations, and model specifications).

Finally, this work has limitations that future research could address. Our sample is restricted to one sign language (LSF), a limited number of signers and recordings, and a particular set of kinematic measures derived from EAR-based tracking. Other annotation schemes and blink definitions might yield different results across tools. Extending the comparison to additional corpora could help determine the extent to which the tool dependence we observed generalises beyond the dataset considered here.

## 6. Author Contributions

The annotation and computer-vision processing and analysis were carried out by Margaux Susman, with input from Carla Miquel Blasco. The statistical analysis was performed by Margaux Susman under the supervision of Jan Bulla. The background on computer vision was drafted by Carla Miquel

| **Predictor** | InsightFace | OpenFace | MediaPipe | 3DDFA |
|---|---|---|---|---|
| | (P) | (P + cat) | (P + R + cat) | (cat + P + R + Y) |
| **Fixed effects** | | | | |
| (Intercept) | $0.020, p < 0.001$ | $0.094, p < 0.001$ | $0.254, p = 0.008$ | $-1.812, p < 0.001$ |
| Pitch | $-0.000, p = 0.041$ | $-0.011, p = 0.070$ | $6.457, p < 0.001$ | $-2809, p = 0.001$ |
| Yaw | | | | $-1048, p = 0.105$ |
| Roll | | | $31.639, p < 0.001$ | $-1036, p = 0.005$ |
| blink cat. (NOTLING) | | $-0.006, p = 0.014$ | $-0.078, p = 0.043$ | $-0.411, p < 0.001$ |

Table 5: Statistical results of the four tools for the comparison of linguistic versus non-linguistic blink velocity. (P = Pitch, R = Roll, Y = Yaw, cat = blink category). **Colour coding:** Green = significant positive (p<0.05), Red = significant negative (p<0.05), Grey = non-significant.

| **Predictor** | InsightFace | OpenFace | MediaPipe | 3DDFA |
|---|---|---|---|---|
| | (cat + P + R + Y) | (cat + P) | (R + P) | (R) |
| **Fixed effects** | | | | |
| (Intercept) | $0.171, p < 0.001$ | $0.105, p < 0.001$ | $0.232, p < 0.001$ | $-2.509, p < 0.001$ |
| Pitch | $0.001, p = 0.003$ | $-0.024, p = 0.027$ | $2.818, p = 0.016$ | |
| Roll | $0.000, p = 0.746$ | | $70.629, p < 0.001$ | $-1208, p = 0.020$ |
| Yaw | $-0.000, p = 0.996$ | | | |
| blink cat. (WAT) | $-0.024, p < 0.001$ | $-0.013, p < 0.001$ | | |

Table 6: Statistical results of the four tools for the comparison of non-linguistic blink categories (gaze direction change blinks (reference) and eye-watering blinks) on velocity. (P = Pitch, R = Roll, Y = Yaw, cat = blink category). **Colour coding:** Green = significant positive (p<0.05), Red = significant negative (p<0.05), Grey = non-significant.

Blasco, with the other sections written by Margaux Susman. All authors commented on and approved the final manuscript.

## 7. Acknowledgements

Funded by the European Union (ERC, NONMANUAL, project number 101039378). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## 8. Bibliographical References

Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20.

Vincent Arel-Bundock, Noah Greifer, and Andrew Heiss. 2024. How to interpret statistical models using marginaleffects for R and Python. *Journal of Statistical Software*, 111(9):1–32.

Valentin Belissen, Annelies Braffort, and Michèle Gouiffès. 2020. Dicta-Sign-LSF-v2: Remake of a continuous French Sign Language dialogue corpus and a first baseline for automatic sign language processing. In *12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 6040–6048, Marseille, France. European Language Resources Association (ELRA).

Annelies Braffort and Emilie Chételat-Pelé. 2012. Analysis and description of blinking in french sign language for automatic generation. In *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication: 9th International Gesture Workshop, GW 2011, Athens, Greece, May 25-27, 2011, Revised Selected Papers 9*, pages 173–182. Springer.

Jan Bulla and Vadim Kimmelman. 2026. Processing kinematics of nonmanual markers in r.

John E Delgado Gómez, Elena Muñoz España, Carlos F Rengifo Rodas, and Diego E Guzmán Villamarín. 2024. Proposal for a eye blink detection using mediapipe, eye aspect ratio and peak identification. In *International Conference on NeuroRehabilitation*, pages 345–349. Springer.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2021. Sample and compu-

| **Predictor** | InsightFace | OpenFace | MediaPipe | 3DDFA |
|---|---|---|---|---|
| | (cat + P) | (cat) | (P + cat + R) | (cat + P + R) |
| **Fixed effects** | | | | |
| (Intercept) | $0.021, p < 0.001$ | $0.089, p < 0.001$ | $0.282, p = 0.012$ | $-2.067, p < 0.001$ |
| Pitch | $-0.000, p = 0.065$ | | $7.154, p < 0.001$ | $-2707, p = 0.003$ |
| Roll | | | $22.166, p = 0.023$ | $-687, p = 0.005$ |
| blink cat. (PROS-TURN) | $-0.002, p = 0.050$ | $-0.003, p = 0.088$ | $-0.090, p = 0.012$ | $0.595, p < 0.001$ |
| blink cat. (SIGN-LIC) | $-0.003, p = 0.003$ | $-0.006, p = 0.016$ | $-0.091, p = 0.066$ | $0.615, p < 0.001$ |

Table 7: Statistical results of the four tools for the comparison of linguistic blink categories (backchanneling (reference), prosodic/turn-taking & sign-aligned/list-marking blinks) on velocity. (P = Pitch, R = Roll, Y = Yaw, cat = blink category). **Colour coding:** Green = significant positive (p<0.05), Red = significant negative (p<0.05), Grey = non-significant.

| Tool | ADFE Mean (SE) | PROS-TURN Mean (SE) | SIGN-LIC Mean (SE) |
|---|---|---|---|
| InsightFace | $0.0215\ (0.0013)$ | $0.0199\ (0.0013)$ | $0.0182\ (0.0015)$ |
| OpenFace | $0.0895\ (0.0067)$ | $0.0864\ (0.0067)$ | $0.0836\ (0.0069)$ |
| MediaPipe | $0.576\ (0.110)$ | $0.485\ (0.112)$ | $0.485\ (0.116)$ |
| 3DDFA | $-2.23\ (0.216)$ | $-1.63\ (0.220)$ | $-1.61\ (0.238)$ |

Table 8: Post-hoc pairwise comparison (response Scale, Holm Corrected, ($^*p < .01$ after correction; all others $p > .05$ except ADFE baselines), ADFE = backchanneling, PROS-TURN = prosodic/turn-taking, SIGN-LIC = sign-aligned/list-marking


tation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Nathaniel Helwig. 2020. Multiple and Generalized Nonparametric Regression. In *SAGE Research Methods Foundations*. SAGE.

Nathaniel E. Helwig. 2024. *npreg: Nonparametric Regression via Smoothing Splines*.

Annika Herrmann. 2010. The interaction of eye blinks and other prosodic cues in german sign language. *Sign Language & Linguistics*, 13(1):3–39.

Kuznetsova and Kimmelman. 2024. Testing mediapipe holistic for linguistic analysis of nonmanual markers in sign languages. *arXiv preprint arXiv:2403.10367*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Marcus Nyström, Richard Andersson, Diederick C Niehorster, Roy S Hessels, and Ignace TC Hooge. 2024. What is a blink? classifying and characterizing blinks in eye openness signals. *Behavior Research Methods*, 56(4):3280–3299.

José Pinheiro, Douglas Bates, and R Core Team. 2025. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-168.

R Core Team. 2026. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Anima Rahman, Roger Woodman, and Valentina Donzella. 2025. Advancing blink detection in driver monitoring with improved eye landmark analysis.

Allah Bux Sargano, Sébastien Vandenitte, Tommi Jantunen, and Vadim Kimmelman. 2024. Evaluation of head pose estimation algorithms for sign language analysis. In *2024 International Conference on IT and Industrial Technologies (ICIT)*, pages 1–6. IEEE.

Felix Sze. 2004. Blinks and intonational phrasing in hong kong sign language. In *Signs of the time: Selected papers from TISLR*, pages 83–107.

Ronnie Wilbur. 1994. Eyeblinks & ASL phrase structure. *Sign Language Studies*, 84:221–240.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

# Grounding Sign Language Representation Learning in Phonology

**Toon Vandendriessche, Mathieu De Coster, Joni Dambre**
Ghent University – imec – IDLab AIRO
Technologiepark-Zwijnaarde 126, 9052 Gent, Belgium
{toon.vandendriessche, mathieu.decoster, joni.dambre}@ugent.be

**Abstract**

Sign language recognition systems are commonly trained using gloss-level supervision, treating signs as holistic lexical units. While effective for classification, such approaches entangle sub-lexical structure and fail to capture the phonological parameters that govern sign formation, limiting interpretability, robustness, and cross-lingual transfer. In this work, we propose a phonologically informed representation learning architecture that explicitly structures the latent space according to linguistic principles. Grounded in the Dependency Model – a phonological model used to describe Flemish Sign Language (VGT) – our hierarchical architecture disentangles parameter-specific subspaces for handshape and location and is trained with multi-label phoneme supervision. To evaluate whether phonological information is directly encoded in the geometry of the embedding space, we introduce a non-parametric probing method that measures neighbourhood consistency across increasing scales. We show that conventional gloss-based networks achieve reasonable performance only for very small neighbourhoods, reflecting incidental visual similarity. In contrast, our disentangled representations maintain stable performance for larger neighbourhoods. This behaviour indicates that phonological structure is preserved across broader regions of the space, yielding more coherent and robust embeddings. Together, our results show that explicit phonological supervision – and crucially, disentangled representation learning – provides a principled foundation for interpretable and transferable sign language representations.

**Keywords:** Sign Language, Machine Learning

## 1. Introduction

In sign language processing, AI systems convert sign language video data into numerical representations to perform several tasks, such as recognition and translation. In the case of translation, these representations are often learned automatically from data, which allows the system to optimise directly for task performance, but results in representations of sign segments that are neither human-interpretable nor usable beyond the specific system they were trained in. Meanwhile, most recognition approaches rely on glosses, which serve as the predominant annotation format for sign language resources. However, the use of glosses for sign language processing is increasingly scrutinised within the field (Desai et al., 2024; De Coster et al., 2024). As glosses are merely lexical labels derived from spoken language, they impose a representational ceiling on what these systems can capture. Consequently, gloss-based approaches risk producing representations that are grounded in semantics alone, treating signs as holistic units and neglecting the distinct phonological and morphological properties of sign languages.

In this paper, we argue that sign language representation learning must move beyond glosses to capture the structure at a sub-lexical level. Representations should be grounded in the native modality of sign language and reflect the intrinsic building blocks of signs rather than their lexical annotations derived from spoken language. When learning from glosses, embeddings are implicitly conditioned on entire signs, meaning that any phonological information must be inferred from co-occurring combinations of phonemes rather than from the phonemes themselves. This introduces spurious dependencies between phonological parameters that are artefacts of lexical co-occurrence rather than linguistic structure. Instead, we propose to structure representations according to sign language phonology, targeting the phonemes directly as the primary units of meaning. While different phonological models exist (Battison, 1978; Brentari, 1998; Demey and van der Kooij, 2008), they all rely on a shared set of fundamental parameters, i.e. handshape, location, movement, orientation and non-manuals. In this paper, we focus on handshape and location as a starting point, given their central role in distinguishing signs and the relative maturity of annotation resources for these parameters.

Grounding representations in phonology rather than glosses carries several advantages for both sign language processing and the field more broadly. Even though phonological models are developed for specific languages, the cross-lingual overlap of phonological parameters is much larger than the cross-lingual overlap of gloss-level annotations. A phonological sign language representation thus facilitates cross-lingual knowledge transfer, which reduces the need for language- and task-specific data. Moreover, contrary to gloss-level representations, phonological representations in sign language recognition facilitate the recognition



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 468–476
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

of productive signs. Phonological representations allow systems to describe dynamic sign constructions in terms of their individual components instead of depending only on a predefined, fixed set of vocabulary items. While these are both longer-term goals, the investigation of phonologically grounded representations enables applications on the short-term (e.g. phonological dictionary search system, computer-assisted language learning) that lay the groundwork for long-term goals (Bragg et al., 2019).

To overcome the limitations of holistic gloss-based approaches, we propose an approach that aligns a network's learned representations with linguistic structure. We explicitly decompose these learned representations into distinct, parameter-specific components through phonological supervision, enabling a granular analysis of sub-lexical features of both handshape and location. Crucially, we do not separate these features arbitrarily; we ground our architectural choices in the Dependency Model (Demey and van der Kooij, 2008). By aligning our network's hierarchical structure with this phonological model, we provide a linguistically principled basis for feature separation. Furthermore, we posit that these phonological structures are robust enough to transcend specific languages. We test this by training our network on a large-scale American Sign Language (ASL) dataset to address data scarcity in Flemish Sign Language (VGT), empirically demonstrating that phonological representations enable cross-lingual generalisation.

In this paper, we begin by establishing the necessary background in Section 2. Section 3 then details our proposed phonologically disentangled representations. In Section 4, we validate the approach, presenting our evaluation methodology, empirical results, and discussion. Finally, Section 5 provides concluding remarks and outlines directions for future work.

## 2. Background

**Sign Language Representation Learning** is frequently formulated within the context of Sign Language Recognition (SLR). Broadly, SLR is categorised into two distinct streams: continuous (Camgoz et al., 2020; Zuo and Mak, 2022; Wei and Chen, 2023) and isolated (De Coster et al., 2020; Luqman, 2022; Laines et al., 2023). While continuous recognition could function as a prerequisite for full translation, we focus here on Isolated Sign Language Recognition (ISLR). This choice is driven by the immediate viability of robust ISLR applications, prioritising intermediary applications that provide bootstrapping benefits toward longer-term goals, such as full translation (Bragg et al., 2019).

In recent years, the use of keypoint estimators, such as MediaPipe Holistic (Grishchenko and Bazarevsky, 2020), has become a cornerstone of ISLR preprocessing. While multi-modal approaches – combining keypoints, video, and other inputs – also gained traction (Jiang et al., 2021; Chen et al., 2024; Renjith et al., 2026), the advantages of a strictly pose-based approach outweigh the occasionally flawed detections (Vandendriessche et al., 2026). Most importantly, keypoint representations serve as a critical countermeasure to data scarcity. By effectively abstracting away background noise and environmental artefacts, these representations reduce the high dimensionality of raw video. This ensures that the network focuses exclusively on the signer's motion rather than environmental variables. As a secondary benefit, keypoint representations provide inherent anonymisation of the signers, which is particularly advantageous for compliance within the European GDPR context.

While glosses are the predominant annotation standard, they impose semantics onto sign representations. To make representations more general, we turn to phonology. The integration of phonological information into ISLR is not uncommon (Tavella et al., 2022a; Kezar et al., 2023b; Gueuwou et al., 2025). Previous efforts have primarily relied on the Semlex Benchmark (Kezar et al., 2023a) or the WLASL-LEX dataset (Tavella et al., 2022b), utilising phonological labels from ASL-LEX (Sehyr et al., 2021). However, a critical limitation of these technical works lies in their treatment of sequentiality. Although ASL-LEX provides sequential annotations per morpheme[1], most existing approaches are restricted to the first morpheme of a sign. In contrast, we posit that a robust representation must capture the full temporal dynamics by predicting phonemes for *every* morpheme in a given sign.

One notable example of phonology-informed representation learning is the work of Kezar et al. (2023b), who encode isolated signs into a unified embedding using a Graph Convolutional Network (GCN). Although they employ separate classification heads to predict specific phonemes, these predictions are derived from a single, shared latent representation. Consequently, although the embedding implicitly captures phonological features, the specific components (such as handshape or location) remain entangled within the vector space. This entanglement poses a critical limitation for cross-lingual transfer and downstream applications. Because the structural information

[1]We adopt the terminology used by Sehyr et al. (2021). In this work, morphemes refer to sequential units of analysis annotated with phonemes. For instance, BIATHLON would contain two morphemes within this paradigm. While the authors acknowledge that 'morpheme' is linguistically imprecise in this context, we retain the term to maintain consistency with the dataset schema.

is locked behind language-specific classification heads, the embedding itself cannot be queried for specific phonological parameters when applied to a new sign language. The representation effectively remains a 'black box', preventing the granular extraction of features necessary for tasks such as error analysis or phonologically grounded retrieval. Addressing this lack of explicit structure is a primary objective of our work. To contextualise our proposed disentanglement strategy, we first outline several principles of sign language phonology.

**Sign language phonology** is commonly traced back to the work of Stokoe (1960), who proposed that signs are not holistic gestures but can be decomposed into a finite set of structural parameters. Stokoe identified three manual parameters, i.e. handshape, location, and movement. Battison (1978) later expanded this set with a fourth manual parameter: orientation. In sign language linguistics, these parameters are conceptualised as (manual) *phonemes*. They are distinct from *morphemes* – the smallest meaning-bearing units of language – in that phonemes do not carry semantic weight on their own. Instead, phonemes are contrastive: altering a single phoneme (e.g. the location in the VGT sings for MOEDER[2] (mother) and VADER (father)) is sufficient to distinguish one sign from another.

Different theoretical phonological models organise those phonological features in different ways. One influential example is the Prosodic Model by Brentari (1998), which was developed for ASL and later used for annotating several relevant sign language resources (Sehyr et al., 2021; Tavella et al., 2022b; Kezar et al., 2023a). This model distinguishes between inherent features and prosodic features of a sign. *Inherent* features do not change during the production of a sign segment (e.g. selected fingers), while *prosodic* features denote the dynamic properties (e.g. aperture of the hand). Brentari's model places these features in a hierarchical, prosody-informed structure to capture both static and temporal aspects of signs.

An alternative hierarchical organisation is the *Dependency Model* (Figure 1) by Demey and van der Kooij (2008). This model was originally developed based on a phonological study of Sign Language of the Netherlands (NGT) (van der Kooij, 2002) and was later further refined in a study of Flemish Sign Language (VGT) (Demey, 2005). We adopt this model, which posits the sign segment as the fundamental unit of analysis. A segment is structurally defined by the presence of a single or multiple simultaneous movements; consequently, while complex signs may consist of multiple segments, the majority of lexical signs map to a single segment.

[2]The URL to the corresponding video in the Flemish Sign Language dictionary is embedded within this text.

In this hierarchical model, a sign segment is represented as the root node assembled from underlying phonological parameters. The hierarchy characterises two types of relations between nodes as either *head* (vertical connections in Figure 1) or *dependent* (diagonal connections in Figure 1). A dependent specifies properties of its head. For instance, in the articulator node, the selected fingers function as the head, while finger configuration is modelled as a dependent, describing how the selected fingers are configured (dependent).

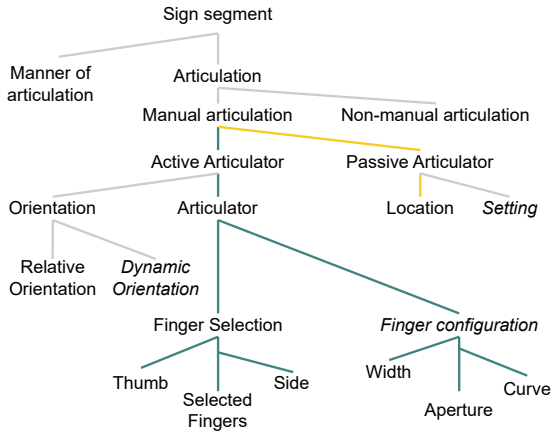



Figure 1: Overview of the Dependency Model (Demey and van der Kooij, 2008).

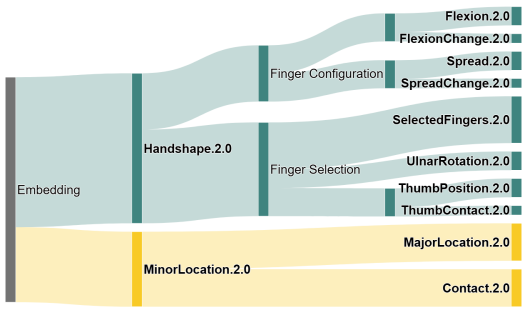



Figure 2: Hierarchical embedding partitioning following the structure of the Dependency Model (Demey and van der Kooij, 2008). The labels in bold are present in ASL-LEX, and the colours map to the coloured branches in Figure 1

In summary, the Dependency Model explicitly formalises the hierarchical organisation of sign phonemes – a structural characteristic that is highly advantageous for representation learning. While other theoretical models, such as the Prosodic Model (Brentari, 1998), have heavily influenced annotated resources (e.g., in ASL-LEX (Sehyr et al., 2021)), the Dependency Model offers a more explicit hierarchy between parameters. Moreover, as this model was empirically derived specifically for Flemish Sign Language (VGT), it provides the most

linguistically accurate basis for learning representations within this target language. Leveraging this theoretical foundation, the following section details how we translate this linguistic hierarchy into a computational architecture (Figure 2).

## 3. Phonologically shaping learnt representations

Although the Dependency Model is highly relevant for representation learning for VGT, no large-scale dataset provides annotations aligned with this structure. This absence makes it impractical to learn representations that directly mirror that model. Nevertheless, many alternative phonological models describe largely overlapping feature sets, albeit organised according to different theoretical perspectives. From a learning perspective, the use of ASL-LEX (Sehyr et al., 2021) – stemming from the Prosodic Model (Brentari, 1998) – in combination with several large-scale ASL datasets (Sehyr et al., 2021; Kezar et al., 2023a; Desai et al., 2023) is highly relevant, since it offers a rich source of phonological features in combination with a considerable amount of data.

We instantiate the hierarchical structure of the Dependency Model within our neural architecture by decomposing the latent space into modular sub-embeddings, each associated with specific ASL-LEX classification heads. Figure 2 illustrates the mapping between the theoretical dependency structure and our practical implementation. For instance, the embedding subspace designated for the *Handshape.2.0* feature is further partitioned into two sub-embeddings to reflect the dependency branching. Even though the intermediate nodes (i.e. finger configuration and finger selection) of the dependency tree are not explicitly annotated in ASL-LEX, the leaf nodes (the actual phonological components) are. By constraining specific classification heads to operate solely on their designated embedding partitions, we enforce a disentangled multi-task learning objective during pretraining.

While the handshape can be hierarchically deconstructed in great detail, the Dependency Model treats location at a broader level of granularity, as spatial features are specified more holistically within this model. ASL-LEX complements this by only distinguishing between *MajorLocation.2.0* (broad regions) and *MinorLocation.2.0* (specific contact points). Because *MinorLocation.2.0* provides a more precise spatial definition that implicitly encapsulates the broader Major Location, we prioritise this fine-grained parameter for our representation. Furthermore, while ASL-LEX includes annotations for additional parameters, these fall outside the scope of this initial exploration. This decision is strictly dictated by data availability in our target domain: the existing annotations provided in the VGT dictionary (Van Herreweghe et al., 2004; Vlaams Gebarentaalcentrum (VGTC), 2026) are currently focused on *Handshape* and *Location*.

## 4. Probing Phonological Structure

### 4.1. Learning Disentangled Representations

Pretraining is done on a combination of ASL-Citizen (Desai et al., 2023) and the Semlex Benchmark (Kezar et al., 2023a), leveraging phonological annotations from ASL-LEX (Sehyr et al., 2021) to align all (phonological) labels across datasets. Notably, whereas most existing works (Kezar et al., 2023a,b; Gueuwou et al., 2025) restrict evaluation to the first morpheme, our method predicts phonological attributes for every morpheme in a sign.

Our network architecture is based on the one used in (Vandendriessche et al., 2026), which was developed for dictionary lookup with similarity-based vector search. We adapted it to a dual-branch design: one branch processes hand keypoints and the other processes body keypoints. Intermediate lateral connections enable frame-level feature exchange between the branches. While previous approaches perform body-part fusion to improve recognition (De Coster et al., 2021; Laines et al., 2023), they typically yield entangled, gloss-specific features. Instead, we explicitly structure the embedding into phonological subspaces: the hand branch produces a handshape representation, whereas the pose branch yields a location representation.

Each subspace is routed to dedicated classification heads – illustrated in Figure 2 – and trained using a multi-label objective with sigmoid activations. We optimise the network using total focal loss (Lin et al., 2017), summed across heads, which allows independent prediction of concurrent phonological attributes without mutual suppression. For additional stability, the concatenated embedding ("Embedding" in Figure 2) is used to predict the sign gloss as an auxiliary supervision signal. Crucially, we posit that robust phonological representations should generalise across datasets and languages. To rigorously test this hypothesis, we deviate from standard in-domain testing and instead report results exclusively on a cross-lingual evaluation.

### 4.2. Assessing Structural Consistency

We perform cross-lingual phonological evaluation on the VGT Dictionary. At the time of writing, it contains 11,248 unique VGT signs, each with a single example video. It also provides phonological annotations for handshape and location. Of the 11,248

signs, 9,826 include phonological labels, with separate annotations for the start and end of each sign. This means that each sign has a minimum of one and a maximum of two annotations for both handshape and location. Consequently, a single sign may require multiple simultaneous predictions for a given parameter, rather than a single definitive label. This variable number of labels necessitates a flexible prediction method.

To accommodate this variability, we use these labels in a multi-label $K$-*nearest neighbours* (KNN) method (Figure 3) to investigate whether learned sign language embeddings capture the underlying phonological structure of signs. Our philosophy behind this method is simple: **signs that share the same physical properties (i.e. phonological descriptions), should have similar representations.**

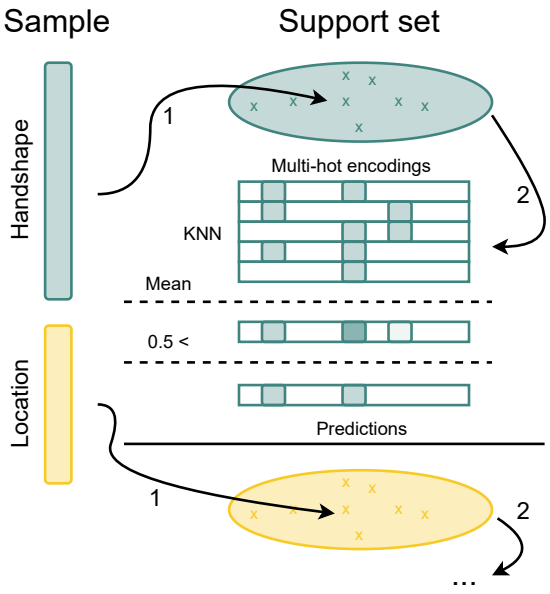



Figure 3: Overview of the multi-hot KNN phoneme classification procedure. Both support sets stem from the same collection of signs, but have distinct representations for the handshape (green) and location (yellow).

To validate this hypothesis, we divide the VGT dictionary into two parts: a set of unseen *test* signs and the rest of the dictionary, which we deem the *support set*. Both are collections of sign representations – existing out of a dedicated handshape section and a dedicated location section – and accompanying phoneme labels. When analysing a new, unseen sign from the *test set*, we calculate its similarity to every sign in the *support set* and identify a predefined number of closest matches – i.e. its "$K$ *nearest neighbours*." We then use majority voting within this *neighbourhood*. If a specific phoneme label (e.g. a specific handshape) appears in the majority of these *nearest neighbours*, we predict that the new sign contains that phoneme as well. In technical terms, we average the multi-hot encoded[3] labels across the closest neighbours; any phoneme for which this neighbourhood-based average exceeds 0.5 is assigned to the unseen sign. Using this approach, we effectively evaluate how well the similarity of the learned embeddings reflects linguistic similarity between signs.

We use this non-parametric probing method for two networks: a baseline and the proposed hierarchy architecture. The *baseline* is a gloss-based network that generates a single, holistic representation per sign. In this case, the entire embedding is used to predict both handshape and location labels. In contrast, the proposed *hierarchy architecture* produces separate embedding sections (Figure 2) specifically dedicated to handshape (green) and location (yellow). As shown in Figure 3, this allows us to retrieve relevant neighbours independently: we search the handshape space to predict handshapes (green), and the location space to predict locations (yellow).

We evaluate this using seven-fold cross-validation. In each round, roughly 86% of the data serves as the reference (support set) to predict labels for the remaining 14% (test signs). For both handshape and location phoneme classes, we report precision and recall as a function of neighbourhood size ($K$). Recall measures the proportion of ground-truth labels that are correctly retrieved, whereas precision measures the proportion of predicted labels that are correct. We compute these metrics per sample in the test set and average across folds to reflect overall performance.

Varying $K$ (the number of neighbours checked) allows us to test the consistency of the embedding space. We hypothesise that embeddings based only on general sign identity may quickly lose accuracy as $K$ increases, as the *neighbourhood* becomes "diluted" with phonologically distinct signs. Conversely, a specialised embedding that effectively captures phonological substructures should remain consistent, preserving high retrieval performance (recall) even as we look at a larger number of neighbours.

### 4.3. Results

Figure 4 illustrates the comparative performance of the baseline from (Vandendriessche et al., 2026) against our proposed hierarchical architecture. The evaluation relies on the multi-label KNN retrieval protocol described above. The figure is partitioned by phonological parameter: the left panel displays

[3]A multi-hot encoding is a binary vector in which 1 indicates the presence and 0 the absence of a class (in this case, a phoneme) in a given sample.

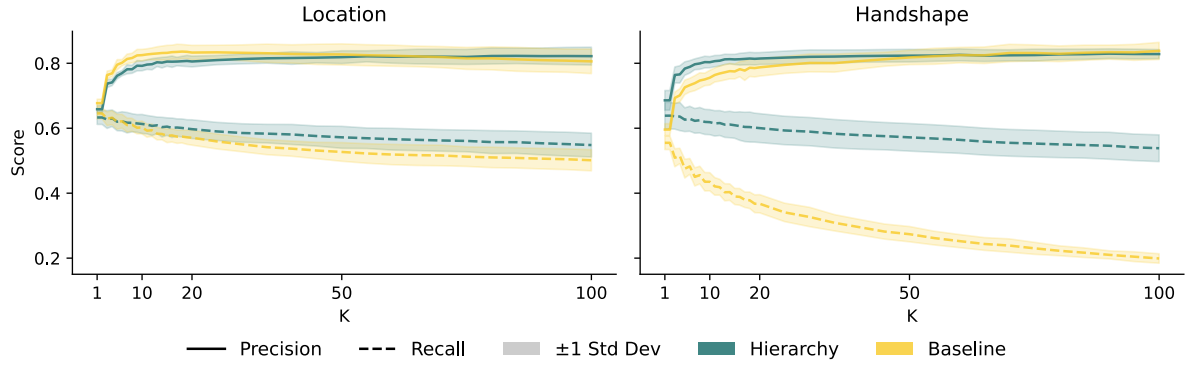



Figure 4: Cross validated precision and recall for multi-label KNN search of location and handshape on the VGT dictionary. The hierarchy network utilises dedicated embedding sections (Figure 2), while the baseline uses the same representation for predicting both handshape and location.

results for Location classification, and the right panel corresponds to Handshape. In both plots, Precision and Recall curves and their respective standard deviations are reported in function of the number of retrieved neighbours ($K$).

First, we observe the recall (dashed lines in Figure 4), as this reflects the proportion of correct targets retrieved. Focusing on handshape, we observe significantly superior stability with hierarchical phoneme modelling: while the baseline's recall degrades rapidly to a minimum of 0.221 (at $K = 100$), the hierarchical network maintains a robust recall of 0.594, even at its lowest point. Notably, this minimum is higher than the baseline's maximum recall of 0.588 (at $K = 1$). However, regarding location, this difference is far less apparent. While the baseline achieves a slightly higher peak recall (0.679 at $K = 2$) compared to the proposed method (0.668 at $K = 4$), the decline is less pronounced for the proposed method as $K$ increases (dropping to 0.534 and 0.590, respectively, at $K = 100$).

Analysing the precision (solid lines in Figure 4) provides further insight into prediction quality, as this reflects what proportion of all predictions is correct. All precision curves follow a similar trend: starting at a lower initial value, they rapidly rise and plateau at a high level. For instance, regarding handshape, the hierarchical network starts at its lowest point (0.716 at K=2) but increases to a maximum of 0.831 at $K = 90$. The baseline reaches a maximum of 0.838 (at $K = 100$). Similar observations hold for location classification, where all precisions plateau at values above 0.8.

Across both phonological parameters, the parameter-specific subspaces generally yield more stable performance than the baseline's global embedding approach, particularly as $K$ increases.

## 4.4. Discussion

The results support our hypothesis that robust phonological representations preserve structural information within the embedding space topology. First, we note that the evaluation method consistently yields high precision. This suggests that, regardless of the underlying pretraining strategy, the method produces reliable phoneme predictions for varying numbers of neighbours. This potentially stems from the thresholding technique, which requires a high presence of a given class within the neighbourhood.

However, analysing the recall provides deeper insight into the topology. In the case of the baseline's handshape classification, we observe a combination of high precision but rapidly degrading recall at higher $K$-values. This indicates that while the baseline correctly identifies the handshapes, it fails to retrieve the majority of relevant targets as the neighbourhood expands. This suggests that the baseline's clusters are fragmented or interspersed with varying handshapes. In contrast, the Hierarchical representation maintains high recall even at large $K$. This stability implies that its embedding space is significantly better structured, forming larger and more cohesive phonological clusters compared to gloss-based representations.

Although the hierarchical network substantially outperforms the baseline for handshape classification, the gains are less pronounced for location. Several factors may explain this difference. First, there is a considerable class imbalance within the location features of the VGT dictionary. The dominant class – neutral space – accounts for 6,079 out of 9,826 samples. Furthermore, the inherent ambiguity of the neutral space may blur the boundaries between distinct locations. Second, location may be inherently less amenable to hierarchical decomposition than handshape. Unlike handshape, whose internal structure is richly articulated in the

Dependency Model (Demey, 2005) – with multiple levels of nested dependencies between finger configurations – the passive articulator (which we operationalise as MinorLocation.2.0) is not assigned a comparably layered internal structure within the same phonological model. Location in sign language phonology is more categorical, defined primarily by spatial regions rather than by compositional subparts. As a result, the location branch of our hierarchy is shallower by design, which likely limits the gains that hierarchical supervision can provide for this parameter compared to handshape.

## 5. Conclusion

We introduced an approach for learning phonologically informed sign language representations that moves beyond conventional gloss-based supervision. By structuring the latent space according to a linguistic phonological model, the proposed approach captures the intrinsic sub-lexical properties of signs rather than relying on holistic sign representations. Grounded in the Dependency Model of Demey and van der Kooij (2008), our hierarchical architecture organises the embedding into parameter-specific subspaces, enabling interpretable and fine-grained phonological analysis through multi-label $K$-nearest neighbour retrieval.

Our experiments show that explicit phonological supervision is necessary to reliably recover sub-lexical features from unseen data. More importantly, the results demonstrate that supervision alone is insufficient: explicitly disentangling phonological parameters within the embedding space is critical for preserving phonological structure. The stability of our hierarchical network across increasing neighbourhood sizes indicates that the learned representations encode phonological information geometrically, yielding robust and interpretable neighbourhoods rather than incidental similarity. Furthermore, we find that the employed multi-label $K$ nearest neighbour retrieval inherently promotes highly reliable classification, yielding high-precision predictions across all networks.

Finally, successful transfer from ASL pretraining to evaluation on VGT highlights the broader applicability of phonological representations across languages. This cross-lingual generalisation suggests that modelling shared phonological structure provides a principled alternative to gloss-based systems, which remain constrained by fixed vocabularies and limited support for productive signing.

**Future work** will extend the hierarchical formulation to incorporate richer phonological distinctions. Additionally, incorporating detailed labels for movement, orientation, and non-manual features will further strengthen the representational capacity of the architecture. Overall, phonologically grounded embeddings offer a scalable and linguistically principled foundation for robust sign language recognition, retrieval, and transfer across languages.

## 6. Bibliographical References


Robbin Battison. 1978. Lexical borrowing in American Sign Language.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31.

Diane Brentari. 1998. *A prosodic model of sign language phonology*. MIT Press.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Hao Chen, Jiaze Wang, Ziyu Guo, Jinpeng Li, Donghao Zhou, Bian Wu, Chenyong Guan, Guangyong Chen, and Pheng-Ann Heng. 2024. SignVTCL: Multi-modal continuous sign language recognition enhanced by visual-textual contrastive learning. *arXiv preprint arXiv:2401.11847*.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2024. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, 23(3):1305–1331.

Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2020. Sign language recognition with transformer networks. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6018–6024.

Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2021. Isolated sign recognition from RGB video using pose flow and self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3441–3450.

Eline Demey. 2005. *Fonologie van de Vlaamse Gebarentaal: distinctiviteit en iconiciteit*. Ph.D. thesis, Ghent University.

Eline Demey and Els van der Kooij. 2008. Phonological patterns in a dependency model: Allophonic relations grounded in phonetic and iconic motivation. *Lingua*, 118(8):1109–1138.

Aashaka Desai, Maartje De Meulder, Julie A Hochgesang, Annemarie Kocab, and Alex X Lu. 2024. Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas. *arXiv preprint arXiv:2403.02563*.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic — simultaneous face, hand and pose prediction, on device. Posted by Research Engineers, Google Research.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H Liu. 2025. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3413–3423.

Lee Kezar, Jesse Thomason, and Zed Sehyr. 2023b. Improving sign recognition with phonology. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737.

David Laines, Miguel Gonzalez-Mendoza, Gilberto Ochoa-Ruiz, and Gissella Bejarano. 2023. Isolated sign language recognition based on tree structure skeleton images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 276–284.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Hamzah Luqman. 2022. An efficient two-stream network for isolated sign language recognition using accumulative video motion. *IEEE Access*, 10:93785–93798.

S Renjith, Aneesh Varghese, Manazhy Rashmi, and Poorna SS. 2026. Transformer-based motion-visual integrated fusion for isolated sign language recognition. *Computers and Electrical Engineering*, 130:110902.

William Stokoe. 1960. Sign language structure: An outline of the visual communication systems of the American deaf. *Studies in Linguistics, Occasional Papers*, 8.

Federico Tavella, Aphrodite Galata, and Angelo Cangelosi. 2022a. Phonology recognition in American Sign Language. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8452–8456. IEEE.

Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. 2022b. WLASL-LEX: a dataset for recognising phonological properties in American Sign Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 453–463, Dublin, Ireland. Association for Computational Linguistics.

Els van der Kooij. 2002. *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity*. Ph.D. thesis, LOT, Utrecht.

Toon Vandendriessche, Caro Brosens, Hannes De Durpel, Mathieu De Coster, and Joni Dambre. 2026. SignBuddy: From Sign Language Research to Scalable Co-Created Solutions. *Universal Access in the Information Society*.

Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621.

Ronglai Zuo and Brian Mak. 2022. C2SLR: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5131–5140.

## 7. Language Resource References

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2023. ASL Citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36:76893–76907.

Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sehyr, and Elana Pontecorvo. 2023a. The SemLex Benchmark: Modeling ASL signs and their phonemes. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–10.

Sehyr, Zed Sevcikova and Caselli, Naomi and Cohen-Goldberg, Ariel M and Emmorey, Karen. 2021. *The ASL-LEX 2.0 Project: A database of lexical and phonological properties for 2,723 signs in American Sign Language*. Oxford University Press.

Mieke Van Herreweghe, M Vermeerbergen, K De Weerdt, and Katrien Van Mulders. 2004. Woordenboek Nederlands–Vlaamse Gebarentaal/Vlaamse Gebarentaal–Nederlands online. https://woordenboek.vlaamsegebarentaal.be/.

Vlaams Gebarentaalcentrum (VGTC). 2026. Woordenboek Vlaamse Gebarentaal. https://woordenboek.vlaamsegebarentaal.be/search. Accessed: 2026-02-12.

# Towards Integrating Pose Estimation with Neuroimaging for the Analysis of Signed Language Video Stimuli

**Sébastien Vandenitte[1], Doris Hernández[2], Jarkko Keränen[2], Tommi Jantunen[2], Anna Puupponen[2]**

[1]University of Namur, LSFB-Lab; [2]University of Jyväskylä, Sign Language Centre

[1]Rue de Bruxelles 61, B-5000 Namur, Belgium, [2]P.O. Box 35, FI-40014 University of Jyväskylä, Finland

sebastien.vandenitte@unamur.be, {doris.m.hernandez-barros, jarkko.j.keranen, tommi.j.jantunen, anna.m.puupponen}@jyu.fi

**Abstract**

We present our project revisiting the video stimuli of an EEG study in Finnish Sign Language to ask whether kinematic properties of the videos impacted their processing by study participants. For each stimulus, an average measure of brain responses across participants is computed. To analyse movement properties in the video stimuli, we rely on MediaPipe for pose estimation. We subsequently report on our project to perform an exploratory analysis of the kinematic properties of the videos which may affect their processing. We focus on several landmarks: the signer's right and left wrists, nose, and upper torso. Our goal is to obtain a kinematic profile of each stimulus video using several average kinematic variables: velocity and acceleration for all selected landmarks, distance between the wrists, and surface covered by the triangular area defined by the left hand, the right hand, and the nose. We conclude by discussing the potential benefits and limitations of this methodological approach.

**Keywords:** pose estimation, EEG, kinematics

## 1. Introduction

Quantifying and analyzing motion characteristics in signing is crucial in signed language research, particularly in approaches that emphasize the importance of strong connections between linguistic theory and phonetic, psychological, or neurological knowledge (Bybee, 1999; Diessel, 2017). The necessity of conducting kinematic analyses also extends to materials used for studying the neural processing of signed languages. Using video stimuli in neuroimaging studies of signed language processing is crucial for achieving a naturalistic and ecologically valid approach (see, e.g., Hosemann et al., 2013; Hernandez et al., 2022b; 2024; 2025). However, this creates a trade-off, as videos also introduce challenges for brain analysis. Compared to still images, signed language video stimuli are inherently complex and create a less controllable experimental environment. As a result, most features of the video signal have traditionally been treated as noise and excluded from analysis. Yet doing so results in the loss of important information about the dynamics of a visual-kinesic language. Addressing these typically ignored features requires controlling human motion in ways that produce systematically distinct conditions—an inherently difficult task, given the natural variability of human movement (Hernandez et al., 2022b).

Quantifying human motion in video stimuli through motion-capture or pose-estimation methods may offer solutions for improving control over the conditions studied. Such approaches can support the analysis of neural variables and aid in developing methodologies for using video materials in future neuroimaging studies on signed languages (as well as in non-neuro experimental settings). One way to achieve this is by applying openly available pose-estimation tools to characterize the kinematic properties of signed video materials. The pose data can then be used to examine the interaction between the kinematic characteristics of the signed stimuli and the brain's processing of those stimuli.

To the best of our knowledge, there are few studies that combine neuro-based analyses of signed language processing with kinematic analysis of signed language materials used as stimuli. Rivolta et al. (2025) first recorded short signed narratives with a Kinect set-up, thereby acquiring videos as well as motion-capture measurements of the stories. They used this information to extract time series of velocity for five articulators, namely the right hand, the left hand, the head, the torso, and a combined variable of the aforementioned articulators ("whole signal"). Next, they used magnetoencephalography (MEG) recordings to analyze neuronal oscillations of participants exposed to the signed narratives as stimuli. Rivolta et al. then converted the MEG data to the sampling rate of the motion-tracking data, enabling an analysis of how MEG data and articulator speed were coupled.

In this paper, we describe ongoing work in which we use MediaPipe pose estimation to analyse signed stimuli after they were used to acquire brain imaging data for a processing study. We describe how pre-existing brain-imaging data can be revisited to determine the potential effects of kinematic variables into the brain responses already obtained. Rather than synchronizing the brain data and the pose data, our approach relies on the association between kinematic variable



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 477–483
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

averages, e.g. mean head velocity and average brain responses, e.g., mean amplitude of P3, for each stimulus video. We also discuss the benefits and limitations of this methodological integration beyond our research questions.

## 2. Methodological approach

The data and methods presented here build on our previous work on the processing of Finnish Sign Language (FinSL). This earlier work comprises video stimuli consisting of isolated 'signs', response-time indices, and EEG measurements collected in a detection task investigating signers' recognition of enactment in signs with varying degrees of this depictive strategy, as well as their behavioural responses to these stimuli (Hernandez et al., 2025). In the present paper, we describe how we conduct kinematic analysis on the existing FinSL video stimuli and set out to combine these analyses with the existing EEG measurements in order to investigate potential interactions between signers' processing of FinSL depiction and the kinematic features involved in that depiction.

### 2.1 Pose estimation

Pose-estimation technology has been a major methodological breakthrough for the study of multimodal communication, gesture research, and signed language linguistics. In addition to paving the way for new research questions, this technology also enables researchers to revisit previous research. Pouw et al. (2021) revisit the gestures filmed and manually annotated for a previous study by Motamedi et al. (2019) to produce a fine-grained analysis of gestures with pose estimation. Trettenbrein and Zaccarella (2021) point out that pose estimation provides a new way to control for kinematic differences in stimuli used to study the perception of gesture and signed languages:

> *OpenPoseR provides quantitative measures of motion based on velocity and acceleration of the actor in the video which can be used for controlling differences in these movement parameters, for example, between different conditions of an experiment. More precisely, the package makes it possible to straightforwardly compute the Euclidean norms of sums of all velocity or acceleration vectors* […] *and thereby provide a quantitative measure of motion for an entire video clip.*
>
> Trettenbrein and Zaccarella (2021, p. 2)

Properties of the visual signal have notably been used in studies on the processing of signed language articulation. Brookshire et al. (2017) quantify changes in the visual stimulus over time with the IVC (Instantaneous Visual Change), a "time-series of aggregated visual change between frames [...] computed as the sum of squared differences in each pixel across sequential frames of video" (2017, p. 6356). More recently, Rivolta et al. (2025) study the coherence between MEG data and motion-tracking information in signed videos, in particular the speed of the right- and left-hand, of the head, and of the whole upper body.

Similarly, we use pose estimation to revisit the video stimuli used in a neurocognitive study for a post-hoc analysis. This methodological approach is motivated by Hernández et al.'s (2025) study on the perception of Finnish Sign Language, which led the authors to posit that visual saliency and an increase in the number of enacting bodily articulators affect attentional resources allocated for the detection of enactment in the signed language signal. The video data comprised 120 video clips of signed expressions in FinSL. Figure 1 illustrates a stimulus set, each image still retrieved from a separate video. All four videos in a set refer to the same action (here, "typing on a keyboard") but they vary in their degree of enactment, the independent variable in Hernández et al. (2025).

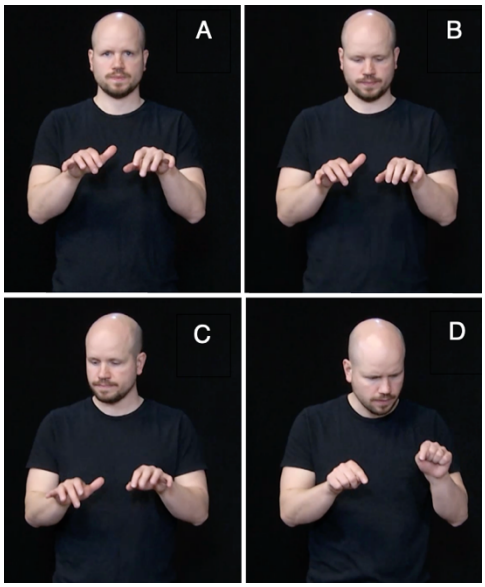



Figure 1: Stimulus set of 4 video stimuli with an increasing degree of enactment: no enactment (A), some enactment (B), substantial enactment (C), and full-blown enactment (D).

To analyse the kinematics of the stimuli, all 120 videos shown to participants were processed for pose estimation using MediaPipe Pose Landmarker (Lugaresi et al., 2019). Figure 2 illustrates the processing with original stills from a video stimulus (A) and stills from the same video processed for pose estimation (B). The dataframe generated after pose estimation contains three-dimensional coordinates for all landmarks. Given the recorded data is two-dimensional, the x and y coordinates are used whereas the z coordinates

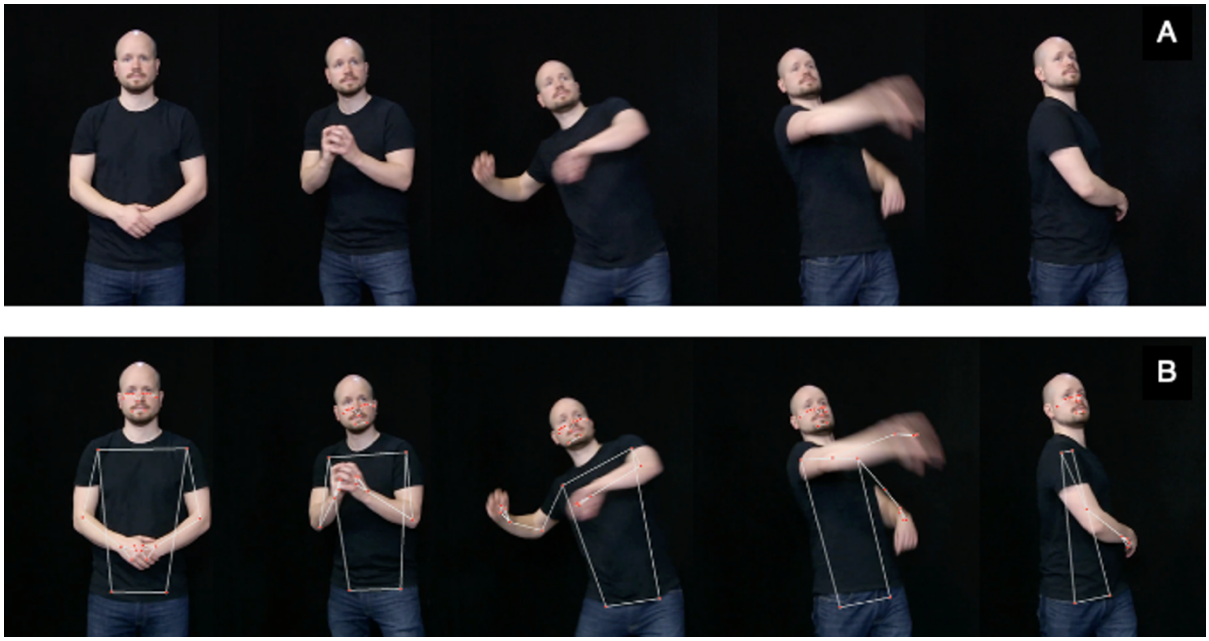



Figure 2: Original video stimulus (A) and video stimulus processed for pose estimation (B).

are discarded from the analysis, for reliability purposes (Ivashechkin et al., 2023). Next, landmarks corresponding to our articulators of interest are selected, namely the left and right wrist, and the nose. In addition, an upper-torso centroid landmark is generated as the mid-point between the left shoulder and the right shoulder. Our recording conditions provided good material for pose estimation: a single person is facing the camera in front of a dark background and stands at the same distance from the camera throughout the stimuli. The data processing proceeds with the generation of several kinematic variables based on the x and y coordinates.[1] These variables will be selected for their attested relevance in signed language phonetics while keeping a wide scope for the purpose of an exploratory approach:

- velocity (nose, upper torso, left wrist, right wrist)
- acceleration (nose, upper torso, left wrist, right wrist)
- distance between the left and right wrists
- surface of the area delimited by the nose, the left wrist, and the right wrist

The selection of these variables is motivated by previous studies. First, Jantunen et al. (2020) have shown significant differences in the velocity and acceleration of head and torso movements in motion-capture analyses of the phenomenon that we focus on. It remains to be seen whether these differences also affect processing. Second, we follow Rivolta et al. (2025) in analysing both the motion of individual articulators (head, left hand, right hand, torso) and the combined motion of varied articulators (here, the surface of the area within the perimeter surrounding the nose, the left wrist and the right wrist). Together, these variables provide us with complementary indices of the saliency of the motion signal, by measuring time derivatives and the size of the signing space. It is this saliency that Hernández et al. (2025) posited to affect participants' detection of enactment in the same video stimuli.

## 2.2 Processing of brain-based functional data

In neuroscience, several techniques can be used to examine either the structural (anatomical) or functional (operational) properties of the brain. Structural techniques, such as Magnetic Resonance Imaging (MRI), and functional techniques, such as Electroencephalography (EEG) and Magnetoencephalography (MEG), can both benefit from being combined with pose data. In this paper, we focus on EEG measurements of electrical activity, reflected as voltage changes arising from the synchronized post-synaptic potentials of large groups of neurons.

During the acquisition of EEG data, electrical activity is recorded from multiple electrodes placed on standardized scalp locations covering all the scalp. Following preprocessing (see Puce and Hämäläinen, 2017), the continuous EEG signal is divided into segments, or epochs, surrounding the events of interest (e.g. the onset of a signed language video). Each epoch usually includes a pre-stimulus baseline period to allow for the correction of voltage differences before the stimulus is presented. By aligning these epochs across multiple trials, the EEG signals for each stimulus can be averaged.

[1] To tackle errors in landmark detection and reduce noise in the generated kinematic variables, we apply Savitzky–Golay smoothing filters to the obtained coordinates as well as in the computation of time derivatives (Pouw, 2024).

| File name | Mean nose velocity | Mean RH acceleration | … | Mean LH-RH distance | Mean surfacre: LH-RH-nose area | Mean N2 amplitude | Mean P3 amplitude |
|---|---|---|---|---|---|---|---|
| video 1 | 0.0106 | 3.1114 | … | 0.1492 | 0.0090 | -0,4250 | 2,5329 |
| video 2 | 0.0252 | 4.2351 | … | 0.2369 | 0.0112 | -1,1591 | 3,0462 |
| video 3 | 0.0064 | 3.1789 | … | 0.2673 | 0.0087 | -1,4625 | 1,0295 |
| video 4 | 0.0367 | 5.1376 | … | 0.1907 | 0.0042 | -0,8466 | 0,5127 |

Table 1: Illustration of the data matrix combining the pose-estimation and EEG analysis.

This process enhances activity that is consistently synchronised with the stimulus, while reducing unrelated background activity. Instead of taking all epochs per person and averaging them, in the approach documented here we average the signal acquired for all subjects per epoch, in order to triangulate it with the pose-estimation analysis. In our experiment, as each stimulus was a video containing a single sign, the epochs were time-locked to the onset of the video. Therefore, an epoch represents a segment of the EEG signal that reflects the brain's response to a specific sign. Each epoch included a 200 ms pre-stimulus baseline period prior to video onset, followed by an 800 ms post-stimulus time window.

Epochs can be further used in several ways. One of these ways is to analyze the EEG data in the time domain by using event-related potentials (ERP). ERP is the term given to voltage fluctuations in the brain's electrical activity that occur in response to a particular sensory, cognitive or motor stimulus (like the presentation of a signed language video). ERPs are particularly useful as they enable researchers to examine the timing and sequence of neural events with millisecond precision (for a review of ERPs in research on signed languages, see Hernández et al., 2022a). Another way of further using epochs is by analyzing EEG data in the frequency domain by decomposing the signal into brain oscillations – rhythmic patterns in electrical brain activity. For one or several oscillations (i.e. theta, alpha, beta, gamma, etc.), several properties are commonly extracted (e.g., power, phase, bandwidth). They reflect the brain's response to the experimental conditions (for a review, see Drijvers and Mazzini, 2023).

Traditionally, EEG analyses have focused on temporal/frequency (ERPs, oscillations) patterns across trials. Individual epochs are often merely treated as inputs to these averages, with relatively little attention given to variability across trials or within single epochs (example Hernández et al., 2024; 2025). Instead of this, in the approach described here we focus on the information contained within each individual epoch rather than relying exclusively on their aggregated averages. In this approach, data from multiple subjects are combined at the group level, averaging across participants while maintaining a detailed representation of the neural responses in each epoch. This enables a more nuanced exploration of the neural processes associated with each stimulus presentation and captures subtle temporal (latency) or size (amplitude/power) variations that are usually overlooked in conventional EEG analyses.

### 2.3 Binding pose estimation and brain functional data

After pre-processing the pose-estimation and the EEG data, the database created for the analysis includes columns representing each extracted measure of kinematic and brain-derived variables (Table 1). Each row corresponds to one video presenting a signed language stimulus, rather than a participant, as is usually the case in EEG studies. In this dataframe, a column for each kinematic variable provides average measurements for each stimulus video, e.g., average nose velocity or average distance between the left and the right wrist. In addition, the database includes columns for the EEG variables, which provide average measurements of the neural responses to each stimulus across participants (per epoch), as described in Section 2.2. The specific variables in Table 1 (N2 and P3 ERPs) are motivated by our previous work and based on the processing and analysis of these components in that study (Hernández et al., 2025).

Once the potential associations have been identified, we plan to use linear mixed models (LMMs) to confirm their statistical robustness, because of their flexibility to include fixed and random effects and to control for covariates (West et al., 2022). The idea behind the confirmatory analysis is to know how kinematic variables explain part of the variance in the brain variables. Figure 3 demonstrates a typical workflow for the methodological integration presented here, modelled after our goal to study the processing of FinSL videos.

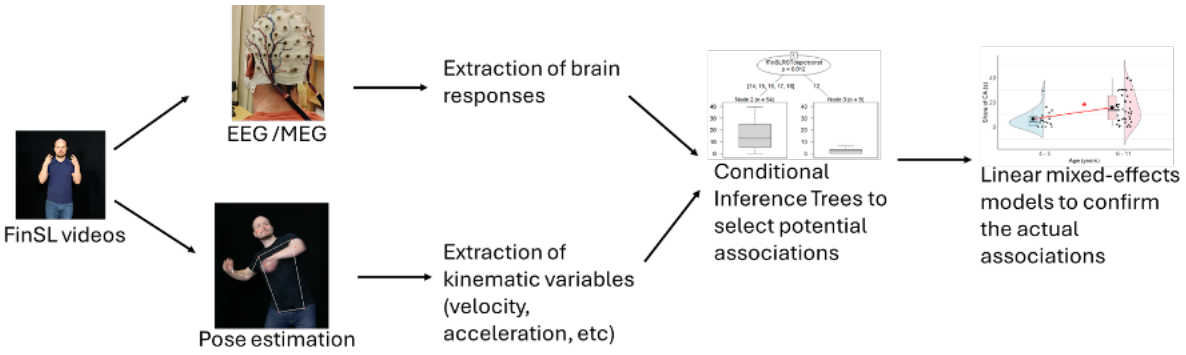



Figure 3: Diagram showing the study design and methodological approach suggested.

## 3. Discussion

In this paper, we have presented the combination of two methodological tools often used in different research traditions: motion tracking and brain imaging. Using fine-grained information from pose estimation enables joining brain analyses to visual information about the stimuli with an unprecedented level of detail. In what follows, we suggest some research avenues that could benefit from this methodological approach.

### 3.1 Benefits and application of the approach

Using kinematic analysis of signed language video stimuli is a methodological development that is beneficial for cognitive neuroscience. Pose-estimation methods can be used to characterize the videos (to be) used in neurocognitive studies. When used for post-hoc analysis, they provide detailed information about the movement properties of the stimuli. A direct benefit of this information is that it can be exploited to check for the effects of motion properties on the processing of the stimuli. In the long run, the variability in movement across stimuli could also be taken into account like other visible factors in video materials, such as background or luminance have been taken into consideration in previous studies (e.g. Hernandez et al., 2022b). This, in turn, can facilitate the building of video stimuli for future neurolinguistic studies on the processing of signed languages, enabling higher degrees of stimulus control and isolation of the target phenomena.

In addition, the increased explanatory power derived from pose estimation also facilitates pursuing several research questions that have emerged in signed language phonetics. It enables asking whether we find neural correlates to some of the phonetic phenomena studied or posited by signed language linguists. For instance, do differences in velocity and acceleration between signs and sign transitions lead to processing differences, notably at the level of perception (Jantunen, 2013)? Do we find neural evidence of a different processing of highly lexical signs as against depicting signs given their different kinematic signatures (Stamp et al., 2024)? If so, what is the nature of these differences?

Finally, another benefit of this methodological combination is its sustainability. Indeed, granted open access to such datasets, it enables conducting further research without having to use new resources for data collection. The maximal use of pre-existing materials is supported by current data protection policies (e.g., General Data Protection Regulation) as well as international guidelines for data management such as the FAIR principles (Wilkinson et al. 2016).

### 3.2 Limitations of the approach

Pose estimation and other forms of video-based motion-tracking offer clear advantages over methods that require specialized hardware, cameras and studio/laboratory environments, such as optoelectronic motion capture. These advantages include greater accessibility, lower cost, and the possibility of analysing pre-existing signed language video materials. However, these benefits come with a trade-off in accuracy, as such approaches do not achieve the same level of detail or precision as motion-capture methodologies in analyzing signed language kinematics (Bauer et al., 2024; Bux et al., 2024). This limitation affects the reliability of the resulting kinematic measurements when compared with those obtained through more specialized techniques. Nevertheless, the accuracy of pose estimation is expected to improve as the underlying technologies continue to develop.

In addition, the methodology presented here relies on average (kinematic and neural) measures associated with each stimulus, without synchronizing the EEG data to the estimated kinematic signals. In future work, the synchronization of the EEG data to the estimated kinematics signals is an important task (Gregori et al., 2023), though not a simple one. While EEG operates at a millisecond level, pose estimation is based on the framerate of videos. This can cause problems with latency measures, especially when working with ERPs or time-frequency

representations (TFRs) that need to be strictly locked to stimulus timing. At the moment, this can be done, for example, with Power Spectral Density (PSD) analysis (Rivolta et al., 2025) which looks at how the power of a signal is distributed across different frequencies disregarding time.

Finally, the exploratory study proposed in this paper is limited in scope. On the one hand, the dimensions of signed language motion suggested for analysis in this paper only capture part of the signed language motion signal. Other articulators in the stimuli, e.g., eye gaze or mouth actions, could also play a role in the processing of the videos. In the same vein, other kinematic features may also be of relevance. On the other hand, this study focuses on the potential influence of kinematic features on the neural processing of signed language video stimuli. However, further development of this approach should consider additional differences between the videos that may impact their processing, such as differences in sign meaning or degree of perceived iconicity.

## 4. Conclusion

In this paper, we have presented our experience combining EEG data with the pose data of the video stimuli to which study participants were exposed. After presenting pose estimation and the use of EEG for brain imaging, we exemplified one way of combining these datasets by using stimulus/epoch-specific averages of kinematic and neural activity variables. While pose estimation exhibits some limitations in terms of accuracy and relative schematicity, it provides a low-resource method to obtain kinematic information about the video stimuli. This methodological combination enables us to perform a post-hoc analysis of how motion properties of the video stimuli affected their processing, as measured by EEG.

## 5. Acknowledgments

This research was supported by the Research Council of Finland (ShowTell project, Grant No. 339268).

DH acknowledges funding from the Finnish Cultural Foundation - SKR (Grant No. 00250359).

SV is a Postdoctoral Researcher of the *Fonds de la Recherche Scientifique – FNRS*.

## 6. Bibliographical References

Bauer, A., Kuder, A., Schulder, M. and Schepens, J. (2024). Phonetic differences between affirmative and feedback head nods in German Sign Language (DGS): A pose estimation study. *Plos one,* 19(5), e0304040.

Brookshire, G., Lu, J., Nusbaum, H. C., Goldin-Meadow, S. and Casasanto, D. (2017). Visual cortex entrains to sign language. *Proceedings of the National Academy of Sciences,* 114(24), 6352-6357.

Bux Sargano, A., Vandenitte, S., Jantunen, T. and Kimmelman, V. (2024). Evaluation of Head Pose Estimation Algorithms for Sign Language Analysis. In *International Conference on IT and Industrial Technologies (ICIT)*, pages 1-6, Chiniot, Pakistan. https://doi.org/10.1109/ICIT63607.2024.10859486

Bybee, J. (1999). Usage-based phonology. *Functionalism and formalism in linguistics*, 1, 211-242.

Diessel, H. (2017). Usage-based linguistics. In Mark Aronoff (Ed.), *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, Oxford, United Kingdom. https://doi.org/10.1093/acrefore/9780199384655.013.363

Drijvers, L., and Mazzini, S. (2023). Neural oscillations in audiovisual language and communication. In *Oxford Research Encyclopedia of Neuroscience*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780190264086.013.455

Gregori, A. et al. (2023). A Roadmap for Technological Innovation in Multimodal Communication Research. In: Duffy, V.G. (eds) *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management.* HCII 2023. Lecture Notes in Computer Science, vol 14029, pages 402-438. Springer, Cham. https://doi.org/10.1007/978-3-031-35748-0_30

Hernández, D., Puupponen, A., Keränen, J., Vandenitte, S., Anible, B., Ortega, G. and Jantunen, T. (2025). Neuroelectrical and behavioral correlates of constructed action recognition in Finnish sign language. *Neuroscience,* 575, 140-149. https://doi.org/10.1016/j.neuroscience.2025.03.046

Hernández, D., Puupponen, A., Keränen, J., Ortega, G., and Jantunen, T. (2024). Between bodily action and conventionalized structure : The neural mechanisms of constructed action in sign language comprehension. *Brain and Language,* 252, Article 105413. https://doi.org/10.1016/j.bandl.2024.105413

Hernández, D., Puupponen, A., and Jantunen, T. (2022a). The contribution of event-related potentials to the understanding of sign language processing and production in the brain: Experimental evidence and future directions. *Frontiers in Communication,* 7, 750256. https://doi.org/10.3389/fcomm.2022.750256

Hernández, D., Puupponen, A., Keränen, J., Wainio, T., Pippuri, O., Ortega, G. and Jantunen, T. (2022b). Use of Sign Language

Videos in EEG and MEG Studies : Experiences from a Multidisciplinary Project Combining Linguistics and Cognitive Neuroscience. In J. H. Jantunen, J. Kalja-Voima, M. Laukkarinen, A. Puupponen, M. Salonen, T. Saresma, J. Tarvainen, and S. Ylönen (Eds.), *Diversity of Methods and Materials in Digital Human Sciences: Proceedings of the Digital Research Data and Human Sciences DRDHum Conference 2022*, December 1-3, pages 148-155. Jyväskylä, Finland. Jyväskylän yliopisto. http://urn.fi/URN:ISBN:978-951-39-9450-1

Hosemann, J., Herrmann, A., Steinbach, M., Bornkessel-Schlesewsky, I. and Schlesewsky, M. (2013). Lexical prediction via forward models: N400 evidence from German Sign Language. *Neuropsychologia,* 51(11), 2224-2237. https://doi.org/10.1016/j.neuropsychologia.2013.07.013

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics,* 15(3), 651-674.

Ivashechkin, M., Mendez, O. and Bowden, R. Improving 3D Pose Estimation For Sign Language. In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW);* 2023, pages 1–5. https://doi.org/10.1109/ICASSPW59220.2023.10193629

Jantunen, T. (2013). Signs and transitions: Do they differ phonetically and does it matter?. *Sign Language Studies,* 13(2), 211-237.

Jantunen, T., Weerdt, D., Burger, B. and Puupponen, A. (2020). The more you move, the more action you construct: A motion capture study on head and upper-torso movements in constructed action in Finnish Sign Language narratives. *Gesture,* 19(1), 72–96. https://doi.org/10.1075/gest.19042.jan

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C., Yong, M.G., Lee, J., Chang, W., Hua, W., Georg, M. and Grundmann, M. (2019). MediaPipe: A Framework for Building Perception Pipelines. *ArXiv, abs*/1906.08172.

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J. and Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition,* 192, 103964.

Pouw, W., Dingemanse, M., Motamedi, Y. and Özyürek, A. (2021). A Systematic Investigation of Gesture Kinematics in Evolving Manual Languages in the Lab. *Cognitive Science,* 45(7), e13014. https://doi.org/10.1111/cogs.13014

Pouw, W. (2024). Wim Pouw's EnvisionBOX modules for social signal processing (Version 1.0.0) [Computer software]. https://github.com/WimPouw/envisionBOX_modulesWP

Puce, A. and Hämäläinen, M. S. (2017). A review of issues related to data acquisition and analysis in EEG/MEG studies. *Brain sciences,* 7(6), 58.

Rivolta, C. L., Lizarazu, M., Costello, B., and Carreiras, M. (2025). Cortical tracking of sign language: The role of language knowledge in tracking of different articulators. *Proceedings of the National Academy of Sciences*, 122(51), e2512665122.

Stamp, R., Cohn, D., Hel-Or, H., and Sandler, W. (2024). Kinect-ing the Dots: Using Motion-Capture Technology to Distinguish Sign Language Linguistic From Gestural Expressions. *Language and Speech,* 67(1), 255–276. https://doi.org/10.1177/00238309231169502

Trettenbrein, P. C., and Zaccarella, E. (2021). Controlling video stimuli in sign language and gesture research: The OpenPoseR package for analyzing OpenPose motion-tracking data in R. *Frontiers in Psychology,* 12, 628728.

West, B. T., Welch, K. B., and Galecki, A. T. (2022). Linear mixed models: a practical guide using statistical software. Chapman and Hall/CRC. https://doi.org/10.1201/9781003181064

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 https://doi.org/10.1038/sdata.2016.18

# KWIC view on Constructed Action (CA) and its Collocates in German Sign Language (DGS) – Possibilities and Limitations

**Sabrina Wähl**
University Hamburg
Gorch-Fock-Wall 7, 20354 Hamburg, Germany
sabrina.waehl@uni-hamburg.de

**Abstract**

Constructed action (CA) is a phenomenon that is used in signed discourse to show the actions of a referent (cf. Cormier et al., 2015; for DGS cf. Fischer and Kollien, 2010). To achieve this, the signer adopts the role of the referent. Most studies use retellings as their data basis (e.g. Herrmann and Pendzich, 2018; Cormier et al., 2015). Consequently, there is less research on CA and its use in data that is not influenced by stimuli. Though there is a considerable number of studies on CA, the phenomenon is still not well understood. One possible way to understand this multifaceted phenomenon better is to analyse collocations in conversations. In spoken language lexicography concordance lines – also known as keyword in context (KWIC) – have proven to be a useful tool in the analysis of collocations. The data used in this study are *Free conversations* in the Public DGS Corpus.
This paper explores the possibilities and limitations of concordance lines as a tool to analyse collocational behaviour of CA. It also presents preliminary results regarding CA and its collocates which may be explored further in the future.

**Keywords:** keyword in context, constructed action, corpus

## 1. Introduction

There are different ways to express referents in sign languages by telling or showing. Constructed action (CA) is a way to express referents by showing. The term refers to constructions that depict the actions, behaviour, thoughts and utterances of a referent (e.g. Cormier et al., 2015; for DGS see Fischer and Kollien, 2010). This phenomenon was described from an early stage in sign language research (e.g. Friedman, 1975 or Mandel, 1977 for American Sign Language – ASL) though it was labeled very differently in various approaches. Different terms are among others: *personal transfer* (Cuxac, 2000), *body classifiers* (Suppalla, 1986) and *performatives* (Winston, 1991). However, the most frequently used terms are, *role shift* (e.g. Lillo-Martin, 2012; Herrmann and Steinbach, 2010) and *CA* (e.g. Metzger, 1995; Cormier et al., 2015). Quite often CA is regarded as gestural (cf. Liddell and Metzger, 1998; Emmorey, 1999; Ferrara and Johnston, 2014). It is expressed by non-manuals like eye gaze, facial expression or shifts of the torso, as well as the hands. Signers may also use *body partitioning* (Dudis, 2004), i.e. different articulators may represent different referents as shown in figure 1. In this example the signer articulates the one-handed DGS sign with the meaning 'to look at' with both hands. Thus, the signer maps one referent, i.e. the signer's partner, onto one hand and a second referent, i.e. the signer, onto the other hand. Both look with a frown (expressed by the signers face) at something on a website[1].

[1] Transcript, 00:12:22:12.



Figure 1: 'to look at sth. together'

There is a considerable amount of research on CA and the non-manuals marking CA, but it is also interesting to further investigate the manual activity during a CA and the preceding signs. One possible way to approach this is using concordance lines, also known as *keyword in context* (KWIC). These display signs or words that precede or follow a specific keyword. Collocational behaviour can reveal a lot about the use of a sign. This leads to the following research questions:

- What insights can be gained about CA by using a KWIC view for analysis?
- How useful is such a view of the data and where are its limitations?

## 2. Data

To answer these research questions, I used conversational data of German Sign Language (DGS;



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 484–490
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

*Deutsche Gebärdensprache*) collected by the DGS-Korpus project as a data basis. The DGS corpus contains approximately 560 hours of signed discourse (Konrad et al., 2024). Participants were filmed in couples and presented with different elicitation tasks such as *Free conversation*, *Discussion* or story retellings (e.g. *Frog Story*, *Sylvester and Tweety*). The data basis for the present research was the task *Free conversation* in the Public DGS Corpus (Konrad et al., 2020). The Public DGS Corpus is a publicly available selection[2] of annotated language material. In total, 34 transcripts of these *Free conversations* are available in the Public DGS Corpus, which is a total of 68 participants. For this task, the participants were asked to have a conversation about a topic of their choice. The session's moderator was not present during the *Free conversation*. Within the DGS corpus, this task is the one that is the least influenced by stimuli. As there already is a lot of research on CA in story retellings but little research on formats that are less influenced by stimuli, this task was chosen for the analysis.

The annotations in the DGS corpus include glosses, mouthing and translations (cf. Konrad et al., 2022). iLex – an integrated annotation tool and lexical database (Hanke and Storz, 2008) – was used to annotate the present data set. *Productive signs* are glossed as $PROD which is a collective gloss including various productive signs. The DGS Corpus project defines productive signs as follows:

> Unlike in vocal languages, signing discourse is not only a succession of lexical units, but an interplay between lexical signs (predominantly denoting something: *telling*) and productive signs (predominantly illustrating the intended meaning: *showing*). These productive signs – also known as classifier or polymorphemic signs – are fully iconically motivated. They have a general meaning to which each iconic value of its components contributes (Konrad et al., 2022, 12).

The glosses for gestures start with $GEST (cf. Konrad et al., 2022). They are defined as follows:

> Manual activity that is neither a conventional nor a productive sign is likely to be a gesture when this activity is also used in the hearing majority with same or similar functions (Konrad et al., 2022, 16).

Unlike other corpus projects, such as the Auslan (Australian Sign Language)[3] corpus, CA is not explicitly annotated in the gloss tier in the DGS corpus, it was annotated with glosses such as $PROD or $GEST instead. DGS corpus transcripts include, among others, separate tiers for glosses, mouthing, translations and comments. Annotators had the option to note supplementary information such as CA in a tier for comments. Some of the annotators also chose to make the distinction between CA and *constructed dialogue* (CD)[4]. As this was not obligatory, annotation of CA is therefore incomplete. Still, there is a considerable amount of annotations that can be used for preliminary analysis. In this paper, a KWIC view was used to conduct such an analysis which is described in the following sections.

## 3. KWIC

For the analysis I used an SQL[5] query that showed all annotations of CA combined with a KWIC (see figure 2). The query showed 486 occurrences of CA annotation from 47 signers across 27 transcripts. It showed:

1. which signs annotated in the gloss tier co-occur with the CA annotation and
2. which signs precede the CA annotations.

I chose to look at preceding and aligning signs only, as I was interested in seeing which signs are used to indicate a referent if it is expressed by a manual sign at all. Initially the query was designed so that it showed up to 5 signs preceding the CA annotations. I could not identify any relevant patterns within that range and narrowed it to 3 signs that were uttered before the CA. In the following sections, detectable patterns which appeared within the range of up to 2 preceding signs are discussed.

### 3.1. Top 5 signs aligned with CA comment

For the signs aligning with CA annotations, I analysed which signs occurred most often. In this context, aligned means that there are CA annotations in a comment tier and parallel to that gloss annotations in the gloss tier. The top 5 signs (i.e. the 5 signs or group of signs with the most occurrences) that are annotated aligned with CA are productive signs, gestures, referential signs, i.e. indexes and the signs for 'I' (I1 – index finger,I2 – flat

[2]For the selection criteria, see Salden and Konrad (2014).

[3]For the Auslan Annotation guidelines, see Johnston (2024).

[4]In this paper, CD is treated as subcategory of CA as it denotes CAs that construct communicative actions.

[5]SQL (Structured Query Language) is a standard query language designed for interacting with databases.




Figure 2: KWIC view: CA and preceding signs

| Sign aligned to CA comment | Number of Tokens |
|---|---|
| $GEST | 91 |
| $PROD | 71 |
| I1/ I2 | 39 |
| $INDEX | 39 |
| TO-LOOK | 34 |

Table 1: Top 5 signs aligning with CA annotations



Figure 3: 'to look at somebody with discomfort'

B-handshape) and signs that denote 'to look' (see table 1).

Of the the top 5, gestures – a group of types all starting with $GEST[6] – occur the most with 91 tokens. This finding could be considered in support of the claim that CA is gestural. Productive signs make up the group that is second most aligned with CA comments. The gloss $PROD contains a large number of classifier constructions. Dudis (2004) (among others) has described that classifier constructions and CA tend to co-occur whenever body partitioning takes place. Again, the results from the query could be seen as further evidence of patterns that have been observed before.

Third most common in *Free conversations* are referential expressions, i.e. I1/ I2 and $INDEX. Here a limitation of the KWIC view shows: The frequency of use is apparent but it is not possible in the KWIC to distinguish whether the signs for 'I' actually means me – the signer – or refers to a third person. I additionally analysed the use of I1/ I2 within CA further with regard to: **Is I1/ I2 referencing the signer or a third person?** 30 tokens were referring to the signer. Hence, these are first person uses of I1/ I2 within CA with the signer being the referent. This construction can be interpreted as: "I was like ..."[7]. The remaining 9 tokens referred to third person referents whose actions are represented. The frequency of I1/ I2 is probably an effect caused by the elicitation task which allowed for narratives of the signer's own experiences. Also, the use of I1 and I2 shows that there seems to be a need to clarify that the signer is the referent.

The last group of signs that aligned the most with the CA annotations is a group of signs that express 'to look at sth.' (for examples see, figures 1 or 3[8]). The group consists of signs using the V-handshape with different movement patterns (TO-LOOK-AT1,TO-LOOK-AT2, TO-LOOK-AT3, TO-LOOK-AWAY-AND-BACK1, TO-SEE-EACH-OTHER) and signs using the 5-handshape indicating a larger number of people (TO-LOOK-AT7, see figure 4[9]).

Hou (2022) described the usage of LOOK for ASL and stated: "While there is no unique marker or a

[6]This is a cluster of types. Accordingly, there is no DOI link whenever I refer to this cluster. For an overview of all types included in the Public DGS Corpus, see the types list.

[7]Lillo-Martin (2012) describes similar constructions and compares them to the use of "be like" in spoken English.

[8]Transcript, 00:02:35:08.

[9]Transcript, 00:00:11:49.

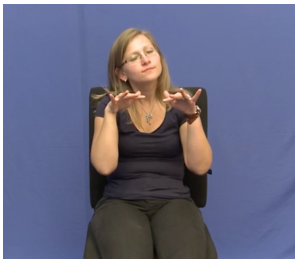


Figure 4: 'audience watches attentively a presentation'

unique group of markers that signals constructed action, LOOK seems to be a common trigger for signaling it in ASL and even other signed languages." The concordance shows that TO-LOOK indeed occurs with a certain frequency together with CA in the DGS data. At this point, it would be interesting to analyse more tokens that denote 'to look at sth.' in order to confirm that these are signs primarily used in CA. The query used for this paper would not be useful for this kind of analysis as annotation of CA is incomplete. Therefore, it would be necessary to annotate the given sign types for CA first. However, the result of the query points toward a similar use of TO-LOOK in DGS and ASL.

### 3.2. Top 5 signs immediately preceding CA comment

| Sign immediately preceding the CA comment | Number of Tokens |
|---|---|
| I1/ I2 | 75 |
| $GEST | 41 |
| $INDEX | 41 |
| $PROD | 37 |
| signs denoting persons | 29 |

Table 2: Top 5 signs immediately preceding CA annotations

The top 5 signs that precede CA again include indexes and I1 and I2, but the signs for 'I' occur the most (see table 2). Accordingly, DGS signers tend to use I1/ I2 as a preceding sign in *Free conversations*.
At this point the question arises: **Do the signs for 'I' introduce the following CA?** Again, the limitations of the KWIC view show here. It presents a string of glosses, but one cannot determine whether the preceding sign relates to the following CA or is part of another clause-like unit. I addressed this issue by analysing those 75 co-occurrences. Based on this analysis, I have determined that most occurrences – i.e. 67 tokens – serve as an introduction indicating that the signer is depicted in the subsequent CA. Another possibility to account for that shortcoming of the KWIC view is to have annotations for clause-like units that could be displayed together with the concordance lines indicating boundaries. The DGS corpus only provides translations that could be used as an approximation. This approach and its limitations have been discussed in the past. There is also a risk that the translation influences the analysis, i.e. its structures will be analyzed rather than the signed sequence itself which is the main reason why I decided against this approach.

$GEST and $PROD also belong to the top 5 though the number of tokens drops visibly. The number of tokens of $INDEX is relatively stable. Signs that denote a person also occur, which I grouped together into one category as there are 29 signs denoting all in all 20 different concepts such as 'deaf person', 'father', 'teacher' or 'boss'. The sign that was used the most was DEAF1A. This sign can be used with different senses, two of them being:

1. term for a person who is deaf
2. to be deaf (in a cultural or medical sense)[10].

The intended sense in the given context cannot be detected from the gloss alone. Thus *word-sense disambiguation* is necessary to determine the actual meaning of the token. Accordingly, this is also a limitation of the KWIC. It gives you an overview over types but the more fine grained differences in use are not necessarily reflected.
Taking the 67 tokens of I1/ I2 introducing CA and the signs denoting a person together, there are 96 tokens that introduce the referent of a following CA. That is 19,75% of the tokens preceding the CA annotations (based on the 486 occurrences – see section 3). This shows that DGS signers use manual referent introduction quite regularly in *Free conversations*. This aligns with research on referent shift strategies which have shown that "point-to-chest" (i.e. I1/ I2) and lexical signs are relevant strategies to mark a reference entity before the signer shifts (cf. Kocab et al., 2015; Stamp and Sandler, 2021).

### 3.3. Other patterns

So far we have looked at signs that align with the CA annotations and signs that immediately precede them. Interestingly, a pattern of two signs (on one occasion 3 signs) and a subsequent CA is also becoming apparent. This pattern is a sequence of a sign for a person followed by a sign of communication followed by a *constructed dialogue*.

[10] See Dictionary entry 8 in Digitales Wörterbuch DGS (DW-DGS) (Langer et al.), own translation.

As mentioned above, CD is the term for all communicative actions that are constructed in signed discourse. Examples for this pattern are:

- I1 TO-LET-KNOW1A 'I let someone know': Constructed dialogue
- BOSS1B SAY1 'The boss said': Constructed dialogue

This sequence of signs introduces the CD. This pattern occurs 14 times in the data set used for this paper. To further confirm this pattern, it would be necessary to expand the data set by adding another elicitation task. Alternatively, a further analysis that takes the communicative signs as a starting point could be conducted for further confirmation.

## 4. Discussion

To sum up the preceding sections, it can be said that KWIC provides an overview of a CA's neighbouring signs. There are several limitations of the approach presented in this paper which are:

- glosses are only an approximation to the actual meaning of the used signs,
- the concordance lines currently do not show boundaries of clause-like units.

These shortcomings can be dealt with by performing additional analyses, such as word-sense disambiguation, which would in any case be necessary for a lexicographic analysis.

So far, there is one issue that may affect the concordances described above: overall lexical frequency in a corpus. You can find a certain number of high frequency words or signs in any corpus. For the DGS corpus, the top 6 signs are[11]:

1. $INDEX (44,304 tokens)
2. I1 (37,181 tokens)
3. $GEST-OFF (29,442 tokens)[12]
4. I2 (12,788 tokens)
5. $GEST (12,027 tokens)
6. $PROD (10,467 tokens).

Hence, it is not surprising that these highly frequent signs appear as collocations in the KWIC. To address this issue one could calculate the mutual information score that measures the relation between two variables. Another possibility is to exclude the high frequency signs from the query which might help to identify patterns with less frequent signs, but also does not show the whole picture. I chose a mixed approach that combines quantitative and qualitative elements by describing the data and interpreting them. For the analysis, I revisited the concordance lines and analysed the range of uses of the signs discussed below. Thus the following ranges include signs from the third preceding sign up to aligned signs in the KWIC (for a schematic representation, see figure 5).




Figure 5: Schematic representation of the KWIC view with signs represented from 3rd preceding sign to signs aligned to CA annotations

The number of $INDEX tokens in the different positions (i.e. third preceding sign, second preceding sign, first preceding sign, aligned sign) is relatively steady and ranges between 37 and 43 occurrences. Because of this little variance of tokens, I conclude that this collocational pattern is likely due to the high frequency of $INDEX.
$GEST tokens preceding the CA annotations range between 29 and 41 tokens. The token number for $GEST and CA annotations aligned is 91 and doubles in comparison. This is similar for $PROD with preceding signs ranging between 27 and 31 tokens and 75 tokens that are aligned. Accordingly, a tendency that $GEST and $PROD are more often used in CA can be observed which is probably not only due to their high frequency in the corpus. I1 and I2 are used more often before CA than $GEST and $PROD. Token counts for I1 and I2 for preceding signs range from 58 to 75 – with 75 being the signs immediately preceding the CA annotations. The use of I1 and I2 drops (39 occurrences) when aligned with the CA annotations which indicates a tendency that signers may prefer constructions where signs denoting 'I' are preceding the CA.
The signs for the meaning 'to look' are also highly frequent (e.g. TO-LOOK-AT1 with 901 tokens throughout the DGS corpus) but are not as numerous as the signs discussed in the previous paragraph. Hence, high frequency effects might be less strong. Also it is interesting that the pattern is only relevant for the signs aligned to the CA annotations, which leads me to the conclusion that this indeed points to a preferred use of these signs within CA.

For the preceding signs in general, I found that 19.75% of the 486 occurrences are used to introduce the CA's referent. However, within that group I1/ I2 are the most frequent. Nevertheless it is interesting to observe that signers quite frequently intro-

[11] As of 18.02.2026.

[12] This gesture is often called *palm up open hand* (PUOH) in literature.

duce the following referent. This facilitates referent identification for the addressee. Non-manual markers or simply a prototypical representation could be used to mark the referent, but one could argue that prototypical representations do not work if the signer is the depicted referent as each person is unique. This might therefore be the reason for the more frequent use of I1 and I2 directly preceding a CA. However, in order to present valid results for DGS with regard to this, a consequent and complete annotation of CA would be needed throughout the DGS corpus or at least throughout a sub-corpus.

## 5. Conclusion

A KWIC view can be used as a means to explore collocational patterns of CA as it provides an overview and can point to patterns of interest for further analysis. Future studies could complete annotations for CA and clause-like units which may allow for a more detailed statistical analysis. Clearly, one has to keep in mind that certain signs appear with a higher frequency and take that into account. Concordance lines are a useful tool within a set of methods, but as a standalone method its limitations are reached fairly quickly.

## 6. Acknowledgements

I would like to thank all the participants of the DGS-Korpus project for taking part in the project and thus making such valuable signed data available for research. I also want to thank the DGS-Korpus project for the support and their work. This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

## 7. Bibliographical References

Kearsy Cormier, Sandra Smith, and Zed Sevcikova-Sehyr. 2015. Rethinking Constructed Action. *Sign Language & Linguistics*, 18(2):167–204.

Christian Cuxac. 2000. Iconicity of Sign Languages. In M. Martin Taylor, Françoise Néel, and Don Bouwhuis, editors, *The Structure of Multimodal Dialogue II*, pages 321–334. John Benjamins Publishing Company.

Paul G. Dudis. 2004. Body Partitioning and Real-Space Blends. *Cognitive Linguistics*, 15(2):223–238.

Karen Emmorey. 1999. Do Signers gesture? In Lynn Messing and Ruth Campbell, editors, *Gesture, Speech, and Sign*, pages 133–158. Oxford University Press.

Lindsay Ferrara and Trevor Johnston. 2014. Elaborating Who's What: A Study of Constructed Action and Clause Structure in Auslan (Australian Sign Language). *Australian Journal of Linguistics*, 34(2):193–215.

Renate Fischer and Simon Kollien. 2010. Gibt es Constructed Action in Deutscher Gebärdensprache und in Deutsch (in der Textsorte Bedeutungserklärung)? *DAS ZEICHEN*, 24(86):502–510.

Lynn A. Friedman. 1975. Space, Time, and Person Reference in American Sign Language. *Language*, 51(4):940–961.

Thomas Hanke and Jakob Storz. 2008. iLex – a Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 64–67, Marrakech, Morocco. European Language Resources Association (ELRA).

Annika Herrmann and Nina-Kristin Pendzich. 2018. Between Narrator and Protagonist in Fables of German Sign Language. In Annika Hübl and Markus Steinbach, editors, *Linguistik Aktuell/Linguistics Today*, volume 247, pages 275–308. John Benjamins Publishing Company.

Annika Herrmann and Markus Steinbach. 2010. Eine neue Perspektive auf Role Shift in Deutscher Gebärdensprache (DGS) – Perspektivwechsel als nichtmanuelles Kongruenzphänomen. *DAS ZEICHEN*, 24(84):112–118.

Lynn Hou. 2022. Looking for Multi-Word Expressions in American Sign Language. *Cognitive Linguistics*, 33:291–337.

Trevor Johnston. 2024. Auslan Corpus Annotation Guidelines.

Annemarie Kocab, Jennie Pyers, and Ann Senghas. 2015. Referential Shift in Nicaraguan Sign Language: A Transition from lexical to spatial Devices. *Frontiers in Psychology*, 5:1–13.

Reiner Konrad, Thomas Hanke, Amy Isard, Marc Schulder, Lutz König, Julian Bleicken, and Oliver

Böse. 2024. Corpus à la carte – improving Access to the Public DGS Corpus. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 390–399, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. Öffentliches DGS-Korpus: Annotationskonventionen / Public DGS Corpus: Annotation Conventions. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Gabriele Langer, Anke Müller, Sabrina Wähl, Felicitas Otte, Lea Sepke, and Thomas Hanke. Digitales Wörterbuch der Deutschen Gebärdensprache. Das korpusbasierte Wörterbuch DGS – Deutsch.

Scott K. Liddell and Melanie Metzger. 1998. Gesture in Sign Language Discourse. *Journal of Pragmatics*, 30(6):657–697.

Diane Lillo-Martin. 2012. Utterance Reports and Constructed Action. In Roland Pfau, Markus Steinbach, and Bencie Woll, editors, *Sign Language*, pages 365–387. De Gruyter Mouton.

Mark A. Mandel. 1977. Iconic Devices in American Sign Language. In Lynn A. Friedman, editor, *On the Other Hand: New Perspectives on American Sign Language*, Language, Thought, and Culture, pages 57–108. Academic Press.

Melanie Metzger. 1995. Constructed Dialogue and Constructed Action in American Sign Language. In Ceil Lucas, editor, *Sociolinguistics in Deaf Communities*, pages 255–271. Gallaudet University Press.

Uta Salden and Reiner Konrad. 2014. Auswahl von Aufnahmen für das Teilkorpus. Project Note AP06-2013-01v02, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Rose Stamp and Wendy Sandler. 2021. The Emergence of Referential Shift Devices in three young Sign Languages. *Lingua*, 257:103070.

Ted Suppalla. 1986. The Classifier System in American Sign Language. In Colette G. Craig, editor, *Typological Studies in Language*, volume 7, pages 181–214. John Benjamins Publishing Company.

Elizabeth A. Winston. 1991. Spatial Referencing and Cohesion in an American Sign Language Text. *Sign Language Studies*, (73):397–410.

# Beyond BLEU: Linguistic Invisibility and Interactional Repair Sequence in End-to-End Sign Language Translation

**Zirui Wang, Mayumi Bono**

SOKENDAI (The Graduate University for Advanced Studies), National Institute of Informatics
Shonan Village, Hayama, Kanagawa 240-0193 JAPAN, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 JAPAN
{wzr, bono}@nii.ac.jp

**Abstract**

Recent advances in end-to-end sign language translation (SLT) have achieved benchmark performance, yet little is known about whether these systems preserve the multi-channel linguistic structures that are essential for real-world communication. We argue that current optimization and evaluation practices create a form of linguistic invisibility, where interactionally decisive non-manual signals (NMS) are systematically underrepresented despite high translation scores.To empirically examine this issue, we analyze an interactional repair sequence from a Japanese Sign Language (JSL) conversational corpus as a diagnostic probe. Combining qualitative interactional analysis with kinematic measurements, we demonstrate a consistent manual–mouth decoupling pattern in which semantic resolution is carried primarily by mouthing while manual articulation remains largely constant. We show that such cross-channel contrast is unlikely to be preserved under current end-to-end training objectives that prioritize global motion similarity. Based on these findings, we argue that progress in SLT should be evaluated not only by sequence-level accuracy but also by the preservation of linguistically contrastive structures, motivating the development of diagnostic, multi-channel evaluation protocols for future SLT benchmarks. We therefore propose incorporating multi-channel diagnostic evaluation sets and decoupling-sensitive metrics into future SLT benchmarking frameworks, providing a pathway toward models that achieve both high performance and linguistic structural visibility.

**Keywords:** interactional repair sequence, non-manual signals, end-to-end model

## 1. Introduction

In the current era of Sign Language Processing (SLP), the dominance of end-to-end (E2E) neural architectures has led to a paradoxical state of "successful blindness." Driven by deep visual encoders and temporal alignment objectives, state-of-the-art models consistently achieve record-breaking BLEU scores on standardized benchmarks (Zhou et al., 2023; Wong et al., 2024; Hwang et al., 2025). However, we argue that this metric-driven progress masks a systematic "linguistic invisibility" where the fundamental multi-channel and contrastive structures of sign language are optimized away as noise to satisfy global sequence mapping.

The transition from modular, linguistics-heavy pipelines to data-driven E2E mapping has fundamentally redefined sign language within AI research (Camgoz et al., 2018). By treating signing as a continuous visual signal to be mapped onto linear text, the field has largely bypassed the "bottleneck" of linguistic annotation (e.g., glosses). While computationally efficient, this paradigm risks reducing a complex natural language to mere "gestures with labels," ignoring the simultaneous grammar such as non-manual signals (NMS) and spatial organization that distinguishes natural signing from pantomime (Bragg et al., 2019; De Meulder, 2021).

The core of this invisibility lies in the optimization target. Current models are trained to maximize surface-level similarity (BLEU/WER), which often rewards models for guessing context while missing the specific contrastive units that resolve ambiguity. This creates a dangerous gap: a model may achieve a high BLEU score by matching generic vocabulary, yet remain incapable of detecting the subtle interactional repair sequences such as a shift in mouthing or eye gaze that are decisive for real-world communication (Bono, 2017). Consequently, a "successful" translation on paper may fail entirely in a functional, human-centric context.

This paper challenges the E2E consensus by bridging a critical survey with micro-level kinematic evidence. We first analyze recurring design patterns in contemporary SLT pipelines that reinforce structural invisibility. We then present a pivot case study from a Japanese Sign Language (JSL) corpus, utilizing kinematic data to quantify how a signer resolves communicative "trouble" through manual-mouth decoupling. Finally, we provide empirical evidence that current E2E models tend to overlook these decisive pulses, and we outline a methodological roadmap toward maintaining linguistic structural visibility in future sign language resources.

Our contribution is threefold: (i) a diagnostic interactional case study revealing cross-channel repair mechanisms, (ii) quantitative kinematic evidence demonstrating manual–mouth decoupling that current E2E models are likely to miss, and (iii) a methodological proposal for evaluation frameworks that incorporate linguistic structural visibility.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 491–500
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

## 2. Scope and Method

We adopt a critical survey approach to analyze methodological trends in end-to-end (E2E) sign language processing, moving beyond performance comparison to examine recurring design patterns and their implicit linguistic assumptions, following the methodological perspective advocated by Yin et al. (2021). Our methodology integrates a longitudinal review of literature with a focused kinematic case study to provide empirical grounding for theoretical critiques.

### 2.1 Criteria for Inclusion

Our survey primarily covers Sign Language Translation (SLT), where E2E methods have progressed most rapidly and where "representational weakening" is especially visible. We concentrate on recent representative works, prioritizing models that have defined state-of-the-art benchmarks on datasets such as PHOENIX-2014T and How2Sign (Hwang et al.,2025; Forster et al., 2012; Duarte et al., 2021). We also include seminal papers that established the current evaluation protocols (e.g., BLEU, WER) to trace the origins of current metric-driven biases (Papineni et al., 2002; Snover et al., 2006).

### 2.2 Analytical Framework: A Dual-Track Approach

Rather than an exhaustive catalog, we summarize widely reused pipelines and extract recurring modeling choices specifically in feature fusion, temporal alignment (e.g., CTC-style objectives; Graves et al., 2006), and loss function design. We evaluate these choices against a "Visibility Criterion": Does the model architecture allow for the explicit detection and separation of simultaneous linguistic channels? Recent JSL-focused work has shown that mouth actions form linguistically motivated and structured categories rather than incidental articulatory noise (Regmi, 2025). A recent study explicitly classifies mouth actions into mouthing, mouth gesture, other mouth movement, and no mouth movement, and demonstrates that these categories can be reliably modeled even with limited data. The linguistic literature summarized in that work further reports that a large proportion of signs in languages such as Auslan, BSL, and JSL co-occur with mouth activity, and that such activity plays a crucial role in lexical disambiguation and grammatical marking. This supports our assumption that mouth actions constitute a contrastive non-manual channel whose suppression risks erasing linguistically relevant structure in SLT.

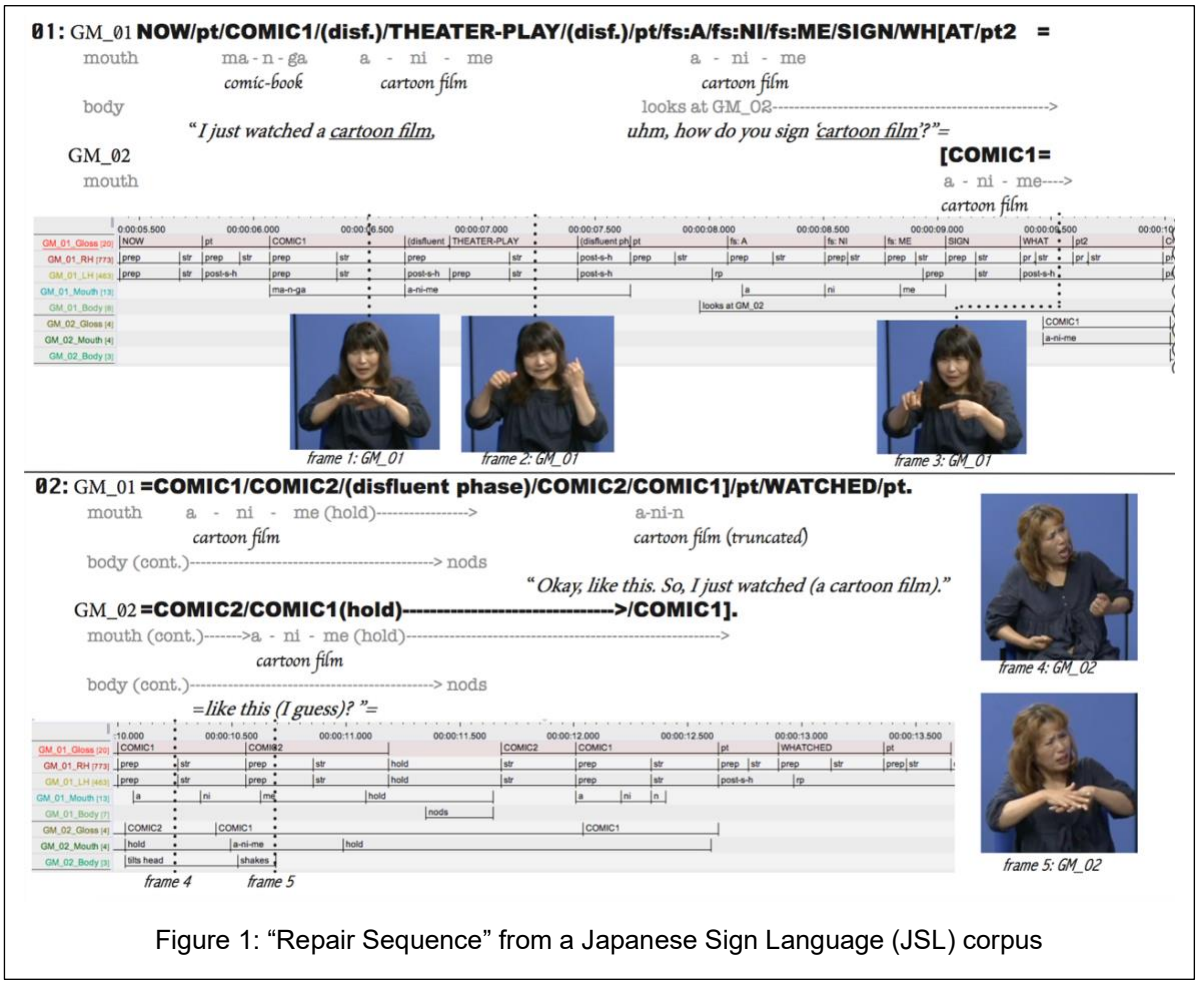



Figure 1: "Repair Sequence" from a Japanese Sign Language (JSL) corpus

To validate our findings, we supplement the literature survey with a Micro-level Kinematic Analysis. We utilize a specific interactional sequence from a Japanese Sign Language (JSL) conversational corpus (Bono, 2017), which serves as a "diagnostic probe." By calculating the decoupling between manual and mouth signals (Mouth Opening Magnitude vs. Hand Movement Energy), we quantify the precise linguistic structures that are systematically filtered out by current E2E optimization objectives.

### 2.3 Research Questions

Our analysis is guided by three central questions:

Representational Collapse: How do E2E visual encoders compress multi-channel signals (NMS vs. Manual) into flattened latent spaces?

Metric Blindness: To what extent do sequence-level metrics (BLEU/WER) mask catastrophic semantic errors in the repair sequence?

The Interactional Gap: Why is the "success" of current models hard to translate into usability in real-world, high-context Deaf communication?

## 3. Case Study: The Invisibility of Repair Sequence

To empirically examine the visibility criterion introduced in the previous section, we turn to a micro-level case study of an interactional repair sequence from a Japanese Sign Language (JSL) corpus involving pairs of Deaf adults from Gunma (Bono, 2017). Rather than serving as an isolated anecdotal example, the sequence functions as a diagnostic probe that allows us to observe how linguistic contrastivity is maintained in real interaction and to test whether such contrastive structures remain detectable under current end-to-end modeling assumptions.

### 3.1 The Interactional Challenge: Seven Trials of Repair Sequence

The session involves a lexical search for the term 'anime'. As the signer struggles to recall the specific lexical unit, they undergo seven distinct trials, shifting strategies between regional signs, similar phonological signs, and fingerspelling. Critically, the trouble is resolved in the final trial: the signer returns to a previous manual form but alters the mouthing from ma-n-ga to a-ni-me.

From an interactional linguistic perspective, this sequence is not merely an instance of lexical search difficulty but a structured process of collaborative meaning negotiation. In natural sign language interaction, repair sequences provide a privileged analytic window into how linguistic contrastivity is actively maintained in real time. What makes the present case analytically significant is that the decisive contrast does not occur in the manual channel, which remains largely constant across trials, but in the mouthing channel, which carries the final semantic disambiguation. This interactional configuration allows us to isolate a situation where linguistic success depends precisely on the model's ability to detect cross-channel decoupling rather than global motion similarity. Consequently, the repair sequence functions as a diagnostic probe: if a translation model fails to recognize the mouthing-based correction, it does not merely produce a lexical error but fails to capture the very mechanism through which meaning is stabilized in dialogue.

Table 1 provides a qualitative interactional summary of the repair trajectory across the seven trials, complementing the visual sequence shown in Figure 1. While the manual lexical form repeatedly converges on the regional sign COMIC1, the decisive semantic shift occurs in the mouthing channel, where the production changes from *ma-n-ga* to *a-ni-me*. This cross-channel asymmetry reveals that the communicative resolution of the repair sequence is not achieved through the introduction of a new manual sign, but through the stabilization of the mouth signal while the manual form remains largely constant. The figure therefore illustrates a key diagnostic property of repair sequence: meaning may be re-established through subtle cross-channel contrast rather than through large-scale manual movement, precisely the type of signal configuration that current end-to-end optimization pipelines are likely to underrepresent.

While the qualitative interactional analysis demonstrates where the decisive semantic contrast is located, a quantitative examination is required to understand why this contrast is likely to be suppressed in current optimization pipelines. We therefore turn to kinematic analysis to measure the degree of manual–mouth decoupling underlying the repair sequence.

| TRIAL | WHO | HAND | MOUTH |
|---|---|---|---|
| 1st | GM_01 | Lexical sign | ma-n-ga |
| 2nd | GM_01 | Lexical sign | a-ni-me |
| 3rd | GM_01 | Finger spelling | a-ni-me |
| 4th | GM_02 | Lexical sign | a-ni-me |
| 5th | GM_02 | Lexical sign | ↓ (hold) |
| 6th | GM_02 | Lexical sign: | a-ni-me |
| 7th | GM_02 | Lexical sign: | ↓ (hold) |

Table 1: Trials for sharing the semantic meaning

### 3.2 Kinematic Evidence: Manual-Mouth Decoupling

To quantify why this linguistic nuance is invisible to current models, we analyze the kinematic features of the sequence.

From a technical perspective, recent work has also shown that mouth movements exhibit temporal boundaries that do not necessarily align with manual gestures. Sandermann (2025) formulates mouth movement detection in JSL as a temporal segmentation problem and demonstrates that manual-focused pipelines often miss or misplace non-manual events. This work further reports difficulties in distinguishing interactionally relevant mouth movements from incidental facial activity, underscoring that mouth dynamics cannot be treated as a simple by-product of hand motion. This technical evidence closely aligns with our kinematic finding of manual–mouth decoupling and reinforces the claim that single-stream E2E encoders face a systematic blind spot for non-manual signals.

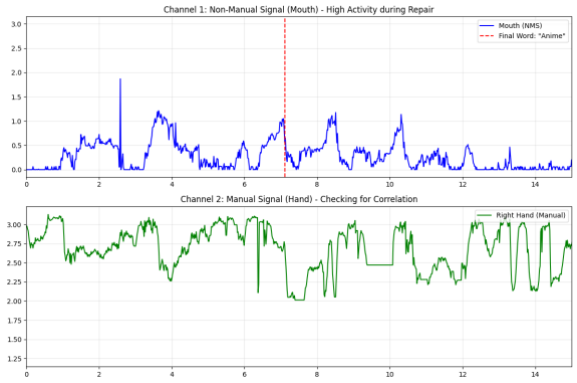



Figure 2: Kinematic decoupling between manual and non-manual signals
(A) Upper panel; (B) Lower panel

Figure 2 displays the kinematic contrast within the narrator's performance. In Panel A, we plot the Mouth Opening Magnitude (MAR) against the Right-Hand Movement Energy to observe a prominent “manual-mouth decoupling”. Between 0s and 7.11s, while the hand maintains a high-amplitude, repetitive oscillation, representing the ‘stagnant’ manual sign COMIC1, the mouth signal exhibits intense, pulsing activities. This suggests that the mouth independently carries the semantic load of the repair sequence. Under current E2E optimization schemes, which are biased toward globally dominant motion patterns (i.e., manual movement), such localized, high-frequency bursts of mouth activity are unlikely to be preserved and may instead be treated as temporal jitter or noise, despite being interactionally decisive.

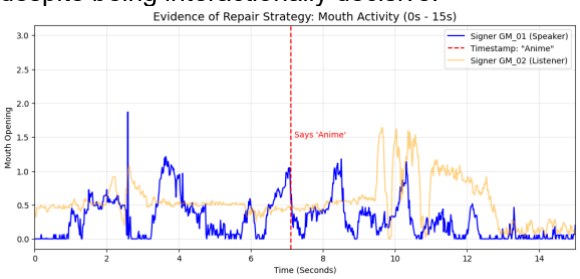



Figure 3: Interactional coordination and signal validation

Figure 3 illustrates the interactional relevance of the NMS signal through the listener’s response. Around the critical 7.11s mark, the narrator’s peak mouth activity (blue line) is followed by a temporally aligned change in the listener’s behavior (yellow line). While this temporal coupling does not by itself establish causality, it provides interactional evidence that the mouth signal constitutes a salient cue for the repair sequence, coinciding with the moment at which the communicative trouble is resolved in the dialogue. This cross-participant temporal alignment suggests that the NMS activity is not merely articulatory noise, but a socially consequential signal that is attended to by the interlocutor.

### 3.3 The Architectural Blind Spot: Visual Noise and Dyadic Interference

While the previous sections have shown that end-to-end models tend to compress multi-channel signals into flattened representations, the more critical question is what this implies for real signed interaction. Figure 4 provides a consolidated view of the architectural limitation in current end-to-end (E2E) frameworks by jointly displaying the manual and non-manual channels of both interlocutors. Based on this kinematic evidence, we identify three recurrent architectural blind spots through which standard visual encoders are likely to obscure interactionally decisive signals in the repair sequence.

Visual Noise Dominance (t = 2.5–4.0 s). At the initiation of the repair sequence, the linguistically relevant information is localized in the non-manual channel (the right signer’s mouth), whereas the manual channel exhibits high-frequency jitter with larger overall motion energy. Given the typical bias of visual encoders toward globally dominant motion patterns, the mouth-based cue is therefore likely to be attenuated in the learned representation in favor of the more salient manual signal, despite its central role in resolving the interactional trouble.

Channel Conflict and Attention Misalignment. This cross-channel imbalance creates a systematic challenge for attention-based mechanisms. Architectures that allocate attention weights primarily based on motion magnitude tend to prioritize the manual channel, which carries stronger but linguistically less informative variation in this segment, over the non-manual channel, which carries the decisive contrastive cue. As a consequence, the signal that is critical for the repair sequence is at risk of being treated as secondary or negligible within the model’s internal representation.

Diarization Failure ( t ≈ 5.0 s, interaction onset). A further source of error arises from the temporal overlap between the two participants, where the left signer’s activity decreases while the right signer initiates a response. In the absence of

explicit speaker diarization or spatially grounded attention, global encoders may conflate these two streams into a single feature trajectory. This conflation increases the risk of feature-level interference and, ultimately, of collapsing distinct interactional turns into an undifferentiated semantic representation.

Together, these three mechanisms illustrate how interactionally decisive non-manual signals in the repair sequence can be systematically overshadowed by globally stronger but linguistically less informative motion patterns in current E2E architectures.

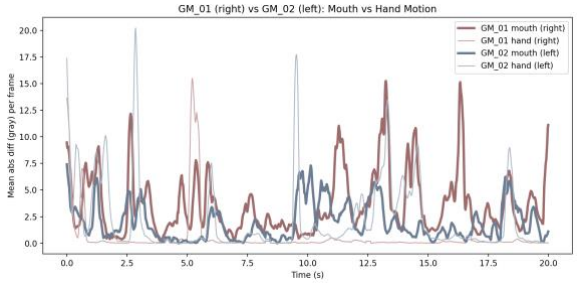



Figure 4: Distractor effect in dual-signer input In contrast to Figure 3, which isolates the mouth channel, this plot jointly displays all four channels: mouth (thick lines) and hand (thin lines) motion for both signers

### 3.4 Empirical Shortcomings of E2E Models

A quantitative analysis of the JSL repair sequence reveals a substantial kinematic divergence, raising questions about how current E2E models account for such discrepancies. We quantify the similarity between the kinematic time series using Pearson's correlation coefficient ($r$) computed over aligned temporal windows (±30 frames) around each target time point. When comparing the initial error (Trial 1, ma-n-ga) with the successful repair sequence (Trial 7, a-ni-me), we found that the Manual Signal Correlation is only 0.3448, while the Mouth Signal (MAR) Correlation drops to -0.1685.

This low correlation highlights a dual challenge for state-of-the-art architectures like GFSLT-VLP. First, the weak manual correlation indicates that the "stagnant" hand signal is not a static posture but a temporally jittered repetitive motion, which E2E encoders often struggle to align across different trials. Second, the near-zero mouth correlation confirms that the semantic shift is driven entirely by independent mouth activity.

Current models, which rely on global pooling or fixed-window attention, essentially "drown" these negative-correlation mouth signals in the noise of the physically more dominant manual channel. As a result, under current objectives, the models are biased toward overlooking the repair sequence, leading to a semantic collapse in which two distinct events are mapped to the same incorrect class, "Manga."

The empirical data in Table 2 provides quantitative evidence that helps explain the limitations of current E2E architectures. Semantically, the manual form remains constant (COMIC1), but kinematically it exhibits jittered repetition, which results in only a weak correlation ($r$ = 0.34). The mouth signal, by contrast, exhibits a near-zero negative correlation ($r$ = -0.1685). In a robust linguistic system, this lack of correlation in mouth activity should trigger a semantic shift (from "manga" to "anime"). However, although $r$ = 0.3448 is conventionally interpreted as a "weak correlation" in general statistics, under the global optimization objective of end-to-end models it can still become a dominant cue. This is because the physical motion energy of the manual channel is substantially larger than the subtle and localized mouth signal, leading the optimization process to implicitly assign greater weight to the manual channel in terms of representation learning and gradient contribution. As a result, models such as GFSLT-VLP (Zhou et al., 2023) are likely to

| Feature Channel | Signal Type | Correlation ($r$) | Linguistic Significance |
|---|---|---|---|
| Manual Channel | Right Hand Energy | 0.3448 | Weak correlation indicating temporal jitter in repetitive motion. |
| Non-Manual signal (NMS) | Mouth Opening (MAR) | -0.1685 | Strong negative/zero correlation confirming distinct mouthing patterns. |
| Expected Model Behavior | Latent Representation (conceptual) | - | Given global-motion-dominant optimization, E2E models are likely to preserve manual similarity while attenuating the contrastive mouth signal, leading to semantic collapse. |

Table 2: Kinematic correlation and semantic alignment analysis

prioritize the weak but high-energy manual correlation over the linguistically decisive mouth signal, and erroneously cluster both trials into the same semantic category, effectively rendering the model "blind" to the repair sequence shown in Figure 3.

### 3.5 Why This Invisibility Matters

Linguistic units are not merely "intermediate layers" to be optimized away for the sake of efficiency. When we flatten the simultaneity of signing into a single sequential mapping, we lose the contrastivity that defines meaning.

As demonstrated in the Gunma case, the resolution of communicative trouble relies on the entire multimodal repertoire. If an E2E pipeline filters out the NMS channel to improve the signal-to-noise ratio of manual signs, it effectively renders the most critical semantic corrections invisible. We argue that "progress" in SLT must be redefined: a model has not learned the language if it cannot perceive the very signals that human signers use to negotiate shared meaning. If a model can match 90% of a sentence's tokens but misses at the 10% that constitutes the repair sequence, it has not merely made a "minor error", it has fundamentally failed to participate in the communicative act.

## 4. The Linguistic Disconnect: What is Being Lost?

The shortcomings of E2E models to capture the JSL repair sequence is not an isolated technical glitch but a symptom of a systematic misalignment between machine learning objectives and sign language linguistics. In this section, we analyze the specific linguistic structures that become "invisible" when SLT is treated as a monolithic signal-to-text mapping task.

### 4.1 What Linguistics Demands: Contrastivity and Disentanglement

In linguistic theory, sign languages are defined by their simultaneous, multi-channel architecture. Meaning is not merely a sequence of symbols but a co-constructed product of manual signs and non-manual signals (NMS) (Bragg et al., 2019).

The principle of contrastivity is important. Linguistic units are defined by their differences. As demonstrated in the Gunma case, the "minimal pair" distinguishing ma-n-ga from a-ni-me resides entirely in mouthing, within the non-manual channel, rather than in the manual channel. When E2E encoders collapse these channels into a single latent vector, they violate the principle of contrastivity, effectively treating semantic shifts as intra-class variance (noise).

Interactional alignment implies Sign language is inherently social. The "repair sequence" proves that meaning is negotiated in real-time. The fact that the listener's response triggers exactly at the moment of the mouth pulse (Figure 2, Lower panel) underscores that NMS are not "auxiliary" but are the primary drivers of interactional resolution.

### 4.2 The Metric Mirage: Why BLEU and WER are "Linguistically Inadequate"

Current evaluation protocols systematically penalize linguistic depth by rewarding surface-level approximation.

This first influence is performance Inflation. Metrics like BLEU and ChrF are based on n-gram overlap (Papineni et al., 2002; Popović, 2015). In the JSL example, if a model correctly translates "I am reading a..." and "...it is interesting," it receives a high score even if it fails the core repair sequence (translating 'anime' as 'manga'). This creates a false sense of progress where models are "guessing" the context rather than "translating" the language.

Then sometimes it will cause the area-weight bias. In E2E loss functions (e.g., Cross-Entropy on pixels or keypoints), large-scale movements like arm swings carry more gradient weight than subtle facial or mouth activity. Consequently, the "linguistic signal" (the mouth) is drowned out by the "physical noise" (the stagnant hand). Current metrics provide no mechanism to penalize a model for missing a high-weight semantic correction if the global sequence looks "plausible."

The case study therefore establishes an empirical observation: interactionally decisive linguistic information may be concentrated in localized non-manual channels even when global motion patterns remain largely unchanged. This observation raises a broader methodological question: if such structures are systematically attenuated by current modeling and evaluation practices, which linguistic dimensions of sign language are being lost at scale? The next section addresses this question by examining the specific linguistic structures that disappear under current SLT paradigms.

## 5. Implications: Rethinking "Progress" in SLT

The "linguistic invisibility" observed in our JSL case study is not merely a technical limitation but a methodological consequence. As the field moves toward larger models and higher-order benchmarks, we must re-evaluate the criteria for successful translation. Representation design determines not only model performance but also which linguistic structures survive the pipeline.

Our analysis can be understood as a bridge between three strands of work: interactional analyses of repair sequence in sign language, linguistically motivated categorization of mouth actions, and technical efforts toward modeling the

temporal structure of non-manual signals. Recent JSL-focused studies have shown that mouth actions form structured linguistic categories, and technical work has demonstrated that their temporal boundaries often diverge from manual gestures. Yet, current end-to-end SLT pipelines still tend to absorb these signals into a single flattened visual stream.

Our L1–L3 framework is designed to make this tension explicit: it allows us to examine whether cross-channel contrast, which may be masked at isolated and local levels (L1/L2), becomes observable only at the interactional level (L3), where repair sequence is functionally resolved.

### 5.1 Linguistics as a Diagnostic Coordinate

End-to-end modeling does not render linguistics unnecessary. On the contrary, linguistic structure provides the essential coordinates for observing, diagnosing, and comparing model behavior (Bragg et al., 2019; Yin et al., 2021, see also Bono, 2026). Without structural visibility, it becomes impossible to judge whether a model has truly learned the language or has simply memorized a signal-to-text mapping. Our JSL analysis proves that if we do not know to look for manual-mouth decoupling, we cannot explain why a "high-performance" model is at risk of failing to resolve a basic communicative repair sequence.

### 5.2 Representation as a Survival Mechanism

In SLT, the choice of representation determines which structures are locatable and which are absorbed as noise. As emphasized in overviews of multimodal corpora (Bono, 2026), representational design directly constrains what kinds of linguistic structure can be observed, analyzed, and compared across modalities. Skobov and Bono (2023) propose a three-dimensional normalization of body movement that preserves analyzable structure while remaining compatible with E2E objectives. This trend is further exemplified by recent gloss-free architectures such as SpaMo (Hwang et al., 2025), which explicitly refines motion dynamics to integrate with large language models, avoiding the pitfalls of a single flattened visual stream. This illustrates that structural visibility is not incompatible with data-driven SLT; rather, it requires explicit representational interfaces. Ensuring identifiable "places" for key structures such as dedicated channels for NMS, enables the analysis of internal model logic.

### 5.3 Evaluation as the "Invisible Hand"

Evaluation protocols act as an invisible hand: what cannot be evaluated will eventually disappear from the research loop (De Meulder, 2021). Incorporating structural visibility into evaluation does not require abandoning BLEU or WER, but it does require recognizing their coverage limits. We therefore suggest a shift from single-score improvements toward diagnostic test sets (akin to NLP "CheckLists"), including repair-sequence benchmarks based on naturalistic dialogues (e.g., the Gunma JSL data with annotated "trouble" and "resolution"), as well as contrastive sensitivity tests that probe minimal pairs differing only in NMS (e.g., mouthing).

Taken together, these considerations point to a methodological shift: progress in SLT should be assessed not only by benchmark scores, but by whether models preserve, express, and support key linguistic structures (Yin et al., 2021). They also motivate diagnostic analyses that go beyond single-score evaluation and directly probe how models organize linguistic structure. To this end, we introduce a small-scale hierarchical context probe using a representative E2E model to examine whether successful and failed repair sequences exhibit different patterns of contextual integration in the model's internal representations.

### 5.4 A Hierarchical Context Probe: Computational Evidence for Context Alignment

To complement the interactional and kinematic analysis, we introduce a hierarchical context probe using a representative E2E model (GFSLT-VLP) to examine how successful and failed repair sequences are organized under different levels of contextual embedding. We consider three conditions: (L1) isolated segments (Trial 1 vs. Trial 7), (L2) local-context segments extracted around each trial (Trial 1_local vs. Trial 7_local), and (L3) the full 20-second interaction, which serves as a proxy for discourse-level "natural flow".

We analyze these conditions using two diagnostic measures derived from the last-layer self-attention: attention entropy (H), which reflects the degree of temporal information integration, and self-attention focus (F), which reflects the degree of frame-wise independence. In addition, we assess representation stability by comparing isolated and embedded representations, and compute the distance of each condition to the L3 baseline in the (H, F) space. These measures are used as indicators of contextual integration rather than as claims about human cognition.

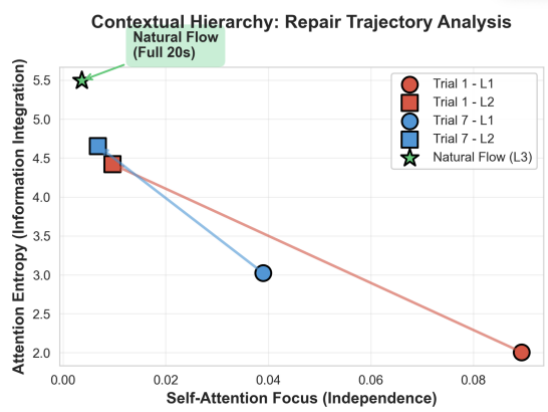



Figure 5: H–F trajectory across contextual levels

Trajectories of Trial 1 and Trial 7 in the attention entropy (H)–self-attention focus (F) space from isolated (L1) to local context (L2), with the full 20-second interaction (L3) shown as the natural-flow baseline

Figure 5 shows the trajectories of Trial 1 and Trial 7 in the (H, F) space from L1 to L2, relative to the L3 baseline, and Figure 6 reports the corresponding entropy and focus values. Two contrasts emerge. First, Trial 7 starts closer to the discourse-level baseline and undergoes a smoother transition from L1 to L2, whereas Trial 1 requires a substantially larger representational shift, suggesting that the successful repair sequence is already structurally closer to discourse-level organization. Second, across both L1 and L2, Trial 7 consistently exhibits higher entropy and lower self-attention focus than Trial 1, indicating a stronger reliance on contextual integration rather than frame-wise, citation-like independence.

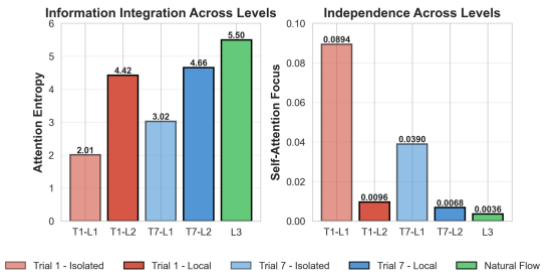



Figure 6: Information integration and independence across contextual levels

Attention entropy (H; left) and self-attention focus (F; right) for Trial 1 and Trial 7 across isolated (L1), local-context (L2), and full-sequence (L3) conditions

This contrast becomes especially clear in representation stability. When comparing isolated and embedded conditions, Trial 7 preserves 75% similarity between its representations, while Trial 1 drops to 18%. Likewise, the amplification of attention entropy from L1 to L2 reaches 120% for Trial 1 but only 54% for Trial 7. Together, these results indicate that the successful repair sequence is already pre-adapted to discourse-level processing, whereas the failed repair sequence relies on a more context-independent form that requires substantial reconfiguration when embedded into interaction.

Taken together, this hierarchical probe provides converging computational evidence for the context alignment hypothesis: successful repair sequence is achieved by producing forms that are structurally aligned with the ongoing discourse and therefore integrate smoothly into the interactional flow.

## 6. Conclusion

This paper has shown that linguistic invisibility in end-to-end sign language translation does not arise primarily from insufficient model capacity, but from optimization and evaluation regimes that privilege global signal similarity over cross-channel linguistic contrast. Through an interactional and kinematic analysis of a Japanese Sign Language repair sequence, we demonstrated that communicative resolution may depend on localized non-manual signals that remain largely undetected by current modeling pipelines despite high benchmark performance.

These findings suggest that translation accuracy alone is an incomplete indicator of linguistic competence in SLT systems. When interactionally decisive structures lack representational and evaluative visibility, models may appear successful while remaining functionally inadequate in real communicative settings. Ensuring structural visibility is therefore not a peripheral linguistic concern but a methodological requirement for reliable human-centered sign language technologies.

Future research should move beyond single-score benchmarking toward evaluation frameworks that explicitly test cross-channel contrastivity and interactional robustness. Developing diagnostic datasets, decoupling-sensitive metrics, and representations that preserve identifiable non-manual channels will be essential steps toward SLT systems that both perform well on benchmarks and faithfully capture the linguistic organization of natural signed interaction. Making these structures visible will be essential for ensuring that future SLT systems learn not only to translate sequences, but to participate reliably in human signed interaction.

## 7. Ethics Statement

This work is intended as a diagnostic and methodological analysis of current end-to-end (E2E) sign language translation (SLT) paradigms, rather than as a contribution toward deploying a production-ready translation system. Our goal is to make linguistically consequential structures, particularly non-manual signals (NMS) in interactional repair sequence visible to modeling and evaluation, and to highlight risks that arise when such structures are optimized away by metric-driven pipelines.

The empirical case study is based on an existing Japanese Sign Language (JSL) conversational corpus (Bono, 2017). Data collection and storage were conducted under the ethical protocols of the original project, including informed consent from participants (Bono et al., 2023). The present study performs secondary analysis only. We do not release any personally identifiable information, and all visualizations are presented in abstracted, de-identified kinematic form (e.g., motion energy and mouth opening measures), rather than raw video, to minimize risks of re-identification and to respect participant privacy.

We are mindful that SLT technologies can have direct social consequences for Deaf communities, especially if systems that achieve high benchmark scores nevertheless fail in real communicative settings. A central motivation of this work is therefore harm prevention: we argue that models optimized solely for surface-level metrics risk producing translations that are misleading or interactionally inappropriate. Our analysis does not advocate the deployment of such systems in high-stakes or user-facing scenarios; instead, it calls for linguistically informed, human-centered evaluation practices, and for the inclusion of Deaf perspectives in the design and assessment of SLT technologies, in line with prior ethical and participatory design discussions in the field. In short, this paper positions linguistics-based diagnostic evaluation as a means to improve the responsibility and reliability of future SLT research, rather than as a step toward replacing human interpreters or supporting unsupervised real-world deployment.

## 8. Limitations

This study has several important limitations that should be acknowledged.

First, the empirical analysis is based on a single interactional repair sequence from a Japanese Sign Language (JSL) corpus. The case is used as a diagnostic probe to demonstrate how linguistically decisive cross-channel contrasts can be obscured by current E2E optimization and evaluation practices. As such, our findings are not intended to provide statistical generalization about the frequency of such phenomena, but rather to establish the existence and methodological relevance of this failure mode.

Second, our analysis is language-specific in its empirical grounding. While the methodological concern, namely, the risk of collapsing multi-channel linguistic structure into global motion-dominant representations is likely to be relevant across sign languages, we do not claim that the exact same repair sequence dynamics or kinematic patterns will hold for other languages such as ASL, DGS, or CSL. Broader cross-linguistic validation remains an important direction for future work.

Third, the computational probe mainly focuses on representative contemporary E2E model (e.g., GFSLT-VLP) and related architectural assumptions. We do not claim that all possible future architectures will necessarily exhibit the same behavior. Rather, our analysis targets a dominant modeling paradigm in current SLT research, especially end-to-end models and highlights a structural risk inherent to this paradigm under common training objectives and evaluation metrics.

Finally, the kinematic measures used in this study (e.g., mouth opening magnitude and hand movement energy) should be understood as operational proxies for linguistically meaningful non-manual and manual activity. While these measures are sufficient for demonstrating manual–mouth decoupling and for constructing a diagnostic contrast, they do not exhaust the full richness of non-manual signals such as eye gaze, facial expression, or head movement. Future work should incorporate richer, linguistically grounded annotations and multi-channel measurements to refine and extend the present analysis.

## 9. Bibliographical References


Bono, M. 2017. Improvisational signing in sign language interaction through a lens of repair sequence. *The Japanese Journal of Language in Society*, 19(2), 20–31. DOI: 10.19024/jajls.19.2_59

Bono, M. 2026. Multimodal corpora. In H. Nesi and P. Milin (eds.), *International Encyclopedia of Language and Linguistics*, 3rd ed. Elsevier.

Bono, M., Okada, T., Kikuchi, K., Sakaida, R., Skobov, V., Miyao, Y., and Osugi, Y. 2023. Utterance unit annotation for the Japanese Sign Language Dialogue Corpus: Towards a method for detecting interactional boundaries in spontaneous sign language dialogue. In Ella Wehrmeyer (ed.), *Advances in Sign Language Corpus Linguistics*, SCL 108, pp. 353–382. John Benjamins. URL: https://cir.nii.ac.jp/crid/1360580230632859264

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N.,

Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Morris, M. R. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*, pages 16–31, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3308561.3353774

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

De Meulder, M. 2021. Is "Good Enough" Good Enough? Ethical and Responsible Development of Sign Language Technologies. In *Proceedings of Machine Translation Summit XVIII: Workshop on Automatic Translation for Signed and Spoken Languages* (AT4SSL). URL: https://aclanthology.org/2021.mtsummit-at4ssl.2/

Duarte, A., et al. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* DOI: 10.1109/CVPR46437.2021.00276

Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., and Ney, H. 2012. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Proceedings of LREC 2012*, Istanbul, Turkey. URL: https://aclanthology.org/L12-1503/

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of ICML 2006*.

Hwang, E. J., Cho, S., Lee, J., and Park, J. C. 2025. An Efficient Gloss-Free Sign Language Translation Using Spatial Configurations and Motion Dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

Regmi, B. S. 2025. Mouth Movement Classification in Japanese Sign Language. Internship Report, National Institute of Informatics, Tokyo, Japan.

Sandermann, T. H. 2025. Temporal Mouth Movement Segmentation in Japanese Sign Language. Internship Report, National Institute of Informatics, Tokyo, Japan.

Skobov, V. and Bono, M. 2023. Making Body Movement in Sign Language Corpus Accessible for Linguists and Machines with Three-Dimensional Normalization of MediaPipe. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1913–1926, Singapore. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. 2006. *A study of translation edit rate with targeted human annotation*. In *Proceedings of AMTA*.

Wong, R., Camgoz, N. C., and Bowden, R. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. In *Proceedings of the International Conference on Learning Representations*.

Yin, K., Kordjamshidi, P., and Roth, D. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 5774–5787.

Zhou, B., Zhang, S., Chen, X., Liu, M., and Wu, Y. 2023. *Gloss-Free Sign Language Translation: Improving from Visual-Language Pretraining*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

# Continuous Sign Language Recognition using Multimodal Input and Handshape-aware Boundary Detection

**Mingyu Zhao[1], Zhanfu Yang[1], Yang Zhou[1], Zhaoyang Xia[1], Can Jin[1], Xiaoxiao He[1], Shuhang Lin[1], Carol Neidle[2], Dimitris N. Metaxas[1]**

[1]Rutgers University, [2]Boston University

[1]110 Frelinghuysen Road, Piscataway, NJ 08854, USA

[2]Boston University Linguistics, 111 Cummington Mall, Boston, MA 02215, USA

{zhao.mingyu, zhanfu.yang, eta.yang, zx149, can.jin}@rutgers.edu

{xiaoxiao.he, shuhang.lin}@rutgers.edu, dnm@cs.rutgers.edu, carol@bu.edu

**Abstract**

This paper employs a multimodal approach for continuous sign recognition by first using ML for detecting the start and end frames of signs in videos of American Sign Language (ASL) sentences, and then by recognizing the segmented signs. For improved robustness, we use 3D skeletal features extracted from sign language videos to take into account the convergence of sign properties and their dynamics that tend to cluster at sign boundaries. Another focus of this paper is the incorporation of information from 3D hand configuration for boundary detection. To detect handshapes normally expected at the beginning and end of signs, we pretrain a handshape classifier for detection of 87 linguistically defined canonical handshape categories using a dataset that we created by integrating and normalizing several existing datasets. A multimodal fusion module is then used to unify the pretrained sign video segmentation framework and handshape classification models. Finally, the estimated boundaries are used for sign recognition, where the recognition model is trained on a large database containing both citation-form isolated signs and signs pre-segmented (based on manual annotations) from continuous signing—as such signs often differ a bit in certain respects. We evaluate our method on the ASLLRP corpus and demonstrate significant improvements over previous work.

**Keywords:** sign language recognition, boundary detection, handshape, multimodal fusion, skeletal features

## 1. Introduction

Signed languages are fully developed visual-gestural languages used by Deaf and hard-of-hearing communities worldwide. Unlike spoken languages, they convey meaning through hand shapes, orientations, movements, facial expressions, head gestures, and body posture (Brentari, 2019). American Sign Language (ASL) is the primary language used by more than 500,000 people in the United States (ASL-deafined, 2017), and it is used in other parts of North America, as well. It is also ranked as the third most studied non-native language studied in US colleges (Lusin et al., 2023). Despite advances in automatic speech and natural language processing, Sign Language Recognition (SLR) remains a complex problem for 2 main reasons: (1) the high-dimensional spatio-temporal dynamics of gestures, and (2) a general lack of large-scale annotated datasets (Huang et al., 2018). Much research to date has focused on Isolated Sign Language Recognition (ISLR), where each input clip contains exactly one sign. However, in real-world communication, signed languages are produced as continuous streams of signing, not isolated signs.

Compared with ISLR, Continuous Sign Language Recognition (CSLR) is more complex, in part because of significant differences in sign production in continuous signing as compared with citation forms (Neidle, 2023). Among many complicating factors in continuous signing are the potential independence of signing on the 2 hands, and the tendency for sign production to be influenced by the properties of adjacent segments. Note that in the current research, all of the signers are right-handed, and we are only attending to signs that include production by the dominant (i.e., right) hand, i.e., 1-handed signs produced on the dominant hand and 2-handed signs. We are not considering 1-handed signs produced on the non-dominant hand.

A major challenge in CSLR lies in the difficulty of locating precise temporal boundaries for each sign segment, i.e., identifying the start and end frames corresponding to each sign (Neidle, 2023). Another major difficulty arises from the fact that natural signing includes things that cannot just be looked up in a "dictionary" of isolated signs (even if they could be perfectly segmented)—including potentially: false starts, things like classifiers and gestures, not all of which belong to any finite classes of elements, as well as fingerspelled signs, potentially corresponding to any word or proper noun (name) from the spoken language (often with letters missing or transposed). This is why any ultimate strategy for recognition of signs from continuous signing needs to be able to distinguish among types of signs, as different recognition strategies need to be applied to different sign types (Yanovich et al., 2016). Note, however, that the research reported here does not



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 501–512
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

do this, so we have submitted our segmented signs for sign recognition only in cases where we have at least 6 examples of those signs in our training dataset. Thus, we are not currently carrying out end-to-end sign recognition, although that is the long-term goal. The sign recognition component of this research is intended to demonstrate the usefulness of segmentation for sign recognition.

Much of the prior work relies on relatively small datasets (Zhang and Duh, 2023; Zhang et al., 2026), which achieve better results compared to the use of larger datasets for experimentation (Gonzalez and Collet, 2011; Moryossef et al., 2023, e.g.). Other work (Rastgoo et al., 2024, e.g.) attempts to synthesize pseudo-continuous sign language datasets by concatenating isolated sign videos. However, relatively few of these strategies have been applied to real large-scale continuous sign language datasets for boundary detection (Moryossef et al., 2023). Most approaches to sign boundary detection follow a similar 2-stage pipeline. They typically first extract visual features from raw videos using tools such as OpenPose (skeletons) or 3D convolutional neural networks, and then feed these features into deep learning models such as LSTMs or Transformers for temporal modeling and boundary prediction (Rastgoo et al., 2024; Renz et al., 2021; Moryossef et al., 2023). These methods have shown some success, but they generally neglect 2 important perspectives: (1) multimodal insights from patterns involving handshape, motion, velocity, acceleration, facial expressions, upper body position, and head gestures, known to correlate with sign transitions, which have the potential to provide strong signals for boundary detection; and (2) the use of more specialized skeleton-based approaches for pose modeling, as some of methods rely on general-purpose deep models and overlook structured methods that may be better suited to use skeletal data.

Building on these observations, we propose a multimodal boundary detection framework, incorporating a pretrained handshape classification model with a sign boundary detection model. We use the estimated boundaries for sign recognition in a well-trained sign recognition model.

There is no standard written representation for signs. Researchers often use English-based ID glossing to identify signs, as we also do. However, a major challenge is ensuring consistency in gloss labeling, i.e., a 1-1 correspondence between signs and glosses, and many datasets fail to meet this requirement (Neidle et al., 2022a, e.g.). One major advantage of the datasets we are using is that we ensure consistency in gloss representations across all the datasets we use.

Our contributions are summarized as follows:

- We design a segmentation module based on a spatio-temporal convolutional network, incorporating velocity and acceleration features.
- We propose a novel method to enhance boundary detection by incorporating handshape information. A multimodal fusion module is introduced, which integrates features from a pretrained handshape classifier into the segmentation stream using a cross-attention mechanism with a gating mechanism.
- We present a new CSLR pipeline combining a segmentation model with a sign language recognition model. Experiments demonstrate the effectiveness of our segmentation framework.

## 2. Related Work

### 2.1. Isolated Sign Language Recognition

Recent approaches to Isolated Sign Language Recognition (ISLR) have leveraged both visual and skeletal modalities to recognize signs from trimmed video clips. Early methods mostly relied on handcrafted features and Hidden Markov Models (HMMs) (Starner and Pentland, 1998). With the rise of deep learning, convolutional neural networks (CNNs) (Pigou et al., 2015) and hierarchical attention networks (HANs) (Huang et al., 2018) became widely used to model spatial and temporal dependencies in isolated signs.

More recent approaches adopt BiLSTM (Camgöz et al., 2017), Transformer-based architectures (Rastgoo et al., 2024), and graph-based models (e.g., ST-GCN (Zhou et al., 2024)) to capture spatio-temporal dynamics. Some work also explores multimodal fusion (Zhou et al., 2020), combining RGB, depth, optical flow, and skeletal data to improve recognition accuracy. In particular, skeletal features extracted from pose estimation models offer a compact and robust representation of sign gestures, leading to improved performance in data-limited scenarios (Li et al., 2020).

Despite impressive progress, ISLR methods typically assume well-segmented input clips, limiting applicability to real-world CSL scenarios. This motivates the development of methods that can extend to continuous sign recognition.

### 2.2. Continuous Sign Language Recognition

Continuous Sign Language Recognition (CSLR) aims to segment and recognize signs from natural sign language sentences. This has important potential applications to real-world scenarios, as it could provide useful input to facilitate such things as live interpretation or video captioning.

Early approaches relied more on alignment-based strategies, which offered a good way to model temporal sequences. Vogler and Metaxas

(1999; 2001) proposed parallel HMM frameworks that modeled the simultaneous movements of multiple articulators (e.g., hands and facial expressions), enabling recognition of continuous signing without requiring manual segmentation. Such work laid the foundation for structured temporal modeling in CSLR. Deep learning methods introduced alignment-free techniques such as Connectionist Temporal Classification (CTC) (Graves et al., 2006), which allowed sequence-level supervision without the need for frame-level annotations.

Later research introduced iterative alignment schemes using RNNs or TCNs to refine temporal segmentation. More recently, to address the intrinsic ambiguity in gloss-to-frame alignment, several studies propose incorporating linguistic priors and contextual constraints. For example, Zhang et al. (2023) introduce a gloss-conditioned decoder integrating prior gloss context into the decoding process. Guo et al. (2023) enhance temporal modeling via cross-temporal knowledge distillation. Reseachers have also introduced gloss-level embeddings as guidance for visual feature learning (Guo et al., 2024; Zhang et al., 2025).

### 2.3. Sign Boundary Detection

For CSLR, it is essential to identify the precise temporal boundaries of individual signs within continuous video streams. Early approaches treated boundary detection as a frame-wise classification problem or relied on dynamic programming to align gloss-level annotations with video frames (Gonzalez and Collet, 2011). Still, these methods overlook the movement dynamics of continuous signing and sign boundaries.

Subsequent methods leveraged temporal convolutional networks, such as TCNs (Renz et al., 2021), or recurrent neural networks, such as LSTM models (Jui et al., 2022), to better capture dynamic dependencies in signing sequences for estimation of sign boundaries. Moryossef et al. (2023) later encoded pose features, which they used as input into an LSTM; they applied a greedy segmentation decoder to identify sign segments.

Recently, Transformer-based architectures have been proposed for sign boundary detection. Varol et al. (2021) use transformers for boundary detection in CSL. Bull et al. (2021) utilize BERT (Devlin et al., 2019) for boundary detection. Rastgoo et al. (2024) propose a Transformer-based framework for boundary detection, but simplify the problem by using artificially constructed CSL sentences rather than natural sign language sentences. Rastgoo et al. (2025) employ a GCN-based model to extract skeletal features that are then embedded into a Transformer-based encoder for boundary detection. These transformer-based approaches exhibit strong performance in modeling multiframe dependencies, but still fail to integrate richer information, such as motion patterns and acceleration, which could provide additional signals for more accurate boundary detection and recognition.

## 3. Method

### 3.1. Problem Formulation

In this work, we formulate the sign language segmentation task as a frame-wise "BIO" annotation problem over continuous sign language video sequences. That is: given a sequence of video frames $X = \{x_1, x_2, ..., x_N\}$, the goal is to predict a corresponding label sequence $L = \{l_1, l_2, ..., l_N\} \in \{B, I, O\}^N$, with the label "B" indicating the boundary frame of a sign segment, "I" indicating that the frame lies within a sign, specifically between the start frame and the end frame, and "O" indicating a frame that is outside any sign segment. This BIO annotation is better suited for modeling the continuous sign language segmentation task than traditional IO tagging, as sign segments in continuous sign language video often follow one another without any O frames in between, which results in these two adjacent segments being merged and counted as one segment (Moryossef et al., 2023).

### 3.2. Spatio-Temporal Convolutional Segmentation

We define S(x) as the skeleton extracted model (Fang et al., 2023). From a video frame sequence $X = \{x_1, x_2, ..., x_N\}$, we extract the corresponding frame-wise 27 skeleton joint representations $Y = \{Y_1, Y_2, ..., Y_N\}$, where $Y_i = S(x_i)$. To meet input requirements of the STGCN-based model, we pad all extracted skeleton sequences along the temporal dimension to a fixed length of T frames. Ground-truth labels are extended to include 4 classes: B, I, O, and P (padding), corresponding to label indices 3, 2, 1, and 0, respectively.

The extracted skeleton sequences $Y = \{Y_1, Y_2, ..., Y_T\}$ are then fed into a Spatio-Temporal Convolutional Segmentation Network. As shown in Figure 1, the front part of the network is constructed based on the ST-GCN framework proposed by Han et al. (2020), which effectively models the spatio-temporal dynamics of human motion. Specifically, we stack 10 ST-GCN layers to progressively capture high-level spatio-temporal patterns across the skeleton sequence.

The output features from the final ST-GCN layer are then passed into a segmentation module designed for frame-wise prediction. To aggregate information across joints, we first apply a spatial pooling operation over the 27 joints at each time step. The pooled features are then processed through 2 temporal convolutional layers, followed by a linear

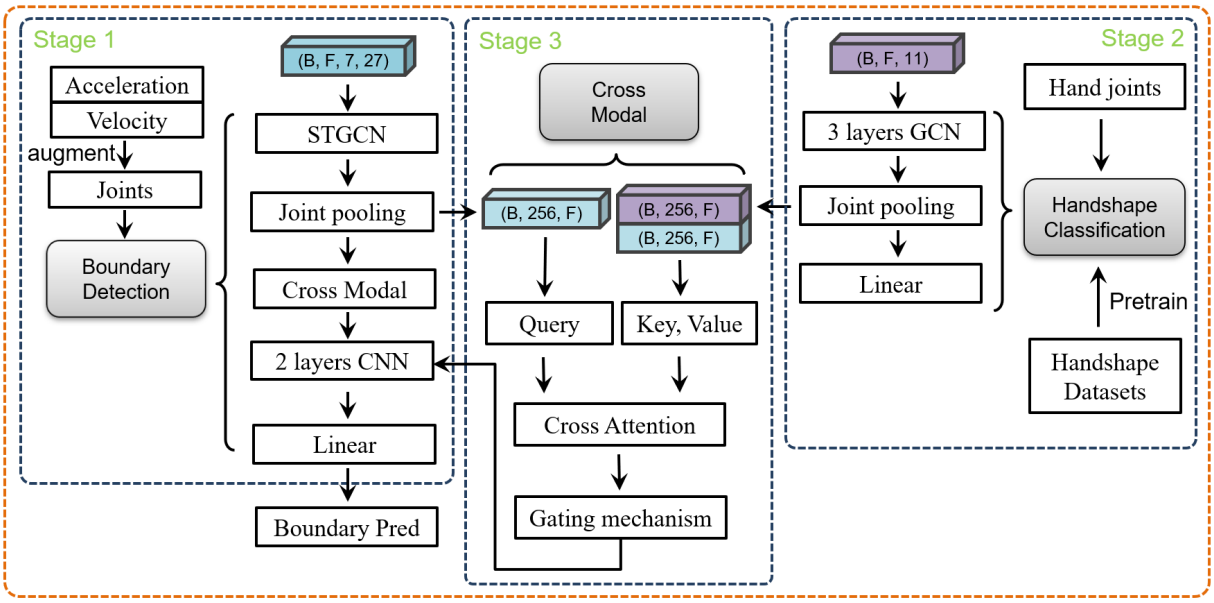



Figure 1: Overview of our proposed multimodal boundary detection. The left branch is the segmentation network based on ST-GCN and CNN modules, which processes all skeletal joints augmented with their velocity and acceleration. We pretrain this branch without the Cross-Modal layer in stage 1. The right branch shows the handshape classification module, which is pretrained on curated handshape datasets using a 3-layer GCN and later used to extract frame-wise handshape features. We pretrain this branch in stage 2. In the middle, a cross-modal attention module with a gating mechanism fuses the handshape and segmentation features to enhance temporal boundary prediction. We fine-tune the right branch with the cross-modal attention module in stage 3. The final segmentation output is further used to extract sign clips and then used for sign recognition.

projection layer that produces frame-level predictions over the segmentation labels. This design allows the model to first learn rich spatial and temporal dependencies using the ST-GCN backbone and then map the encoded motion features to fine-grained frame-wise segmentation outputs.

We adopt a combined loss function consisting of a frame-wise prediction loss and a boundary loss.

The frame-wise prediction loss is defined as a weighted cross-entropy loss. Specifically, we assign a higher class weight to boundary frames (class 3) and a lower weight to background frames (class 0), in order to emphasize the importance of accurate boundary detection and reduce the influence of dominant background frames. The loss is computed as:

$$\mathcal{L}_{\text{frame}} = -\sum_{t=1}^{T} w_{y_t} \cdot \log\left(\hat{y}_t[y_t]\right),$$

where $\hat{y}_t \in \mathbb{R}^C$ is the predicted class probability at time $t$, $y_t \in \{0, 1, 2, 3\}$ is the ground-truth label, and $w_{y_t}$ is the class-specific weight.

The boundary loss is defined as the absolute difference between the number of predicted boundaries and ground-truth boundaries, averaged over the utterance:

$$\mathcal{L}_{\text{boundary}} = \frac{1}{U}\sum_{u=1}^{U}\left|N_{\text{pred}}^{(u)} - N_{\text{true}}^{(u)}\right|,$$

where $N_{\text{pred}}^{(u)}$ and $N_{\text{true}}^{(u)}$ represent the number of predicted and true boundaries in the $u$-th utterance, and $U$ is the number of all utterances.

The final loss $L_{seg}$ is a weighted combination of the 2 components:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{frame}} + \lambda \cdot \mathcal{L}_{\text{boundary}},$$

where $\lambda$ is a hyperparameter controlling the contribution of the boundary loss.

### 3.3. Handshape Classification

We use the set of 87 handshapes from the NCSLGR (National Center for Sign Language and Gesture Resources, 2012) that can be linguistically significant. We formulate the handshape recognition task as a frame-wise classification problem over hand regions in sign video sequences. Given a video frame sequence $X = \{x_1, x_2, ..., x_N\}$, we extract the corresponding frame-wise hand joint representations $X_H = \{x_1^h, x_2^h, ..., x_N^h\}$. We use the established set of 87 canonical handshape categories, indexed from 0 to 86. Rather than assigning an absolute label to each frame, the goal is to estimate the degree of similarity between the observed handshape and these canonical categories. Thus, we want to find a handshape classifier denoted as $H(x_H) \in \{0, 1, ..., 86\}^N$. However, in practice, some frames exhibit blurred or

occluded hand configurations. To address this, we manually sample a subset of clear frames to construct a high-quality training dataset denoted as $X_H^S = \{x_{h_1^s}, x_{h_2^s}, ..., x_{h_k^s}\}$, where $h_1^s, h_2^s, ..., h_k^s$ are selected k frames. The handshape classifier $H(x^H)$ is then pretrained on this dataset and integrated into our multimodal segmentation framework. See Section 4.1.1 for details of the handshape dataset and its construction.

For the handshape recognition model, since we operate directly on skeletal data, we adopt a straightforward 3-layer GCN architecture. This design is intuitive and intentionally kept shallow to mitigate overfitting, while still being expressive enough to capture the discriminative skeletal patterns of different handshapes, as shown in Figure 1. We use standard cross-entropy loss for training.

It is worth noting that our model is not intended to assign a definitive handshape label to every frame in a video (nor would it be realistic to attempt to do so). This is because almost all frames contain hand configurations that deviate in some way from the canonical handshapes. These canonical handshapes are idealized abstractions, whereas in practice, handshapes often fall somewhere in between. Instead, the purpose of our classifier is to extract handshape features by estimating the degree of similarity between the observed handshape and handshapes in the canonical set.

## 3.4. Multimodal Module

The goal of the multimodal module is to incorporate the additional per-frame information from the handshape classification module into the segmentation network. To do this, we introduce a unified cross-attention-based fusion module into the segmentation network; see Figure 1.

### 3.4.1. Self-Attention

For context, we begin by reviewing the self-attention mechanism (Vaswani et al., 2017). Given an input sequence $X \in \mathbb{R}^{T\times d}$, the attention is computed by projecting it into query, key, and value spaces via learned matrices $W_Q, W_K, W_V$, and computing:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad (1)$$

where $Q = W_Q X$, $K = W_K X$, and $V = W_V X$, $\sqrt{d}$ is the scaling factor computed from Q, K and V.

### 3.4.2. Cross-Attention Module

We treat the segmentation features $X^{\text{seg}} \in \mathbb{R}^{T\times d}$ as the source of the query, and concatenate the keys and values from both the segmentation and handshape streams to form the joint attention pool.

Let $X^{\text{hand}} \in \mathbb{R}^{T\times d}$ be the pooled handshape features. We compute:

$$Q_c = W_Q X^{\text{seg}},$$
$$K_c = \begin{bmatrix} W_K X^{\text{seg}} \\ W_K X^{\text{hand}} \end{bmatrix}, \quad V_c = \begin{bmatrix} W_V X^{\text{seg}} \\ W_V X^{\text{hand}} \end{bmatrix}. \quad (2)$$

The resulting cross-attention is then used to update the segmentation representation as follows:

$$\hat{X}^{\text{seg}} = X^{\text{seg}} + \sigma(g) \cdot \text{softmax}\left(\frac{Q_c K_c^\top}{\sqrt{d_c}}\right)V_c. \quad (3)$$

Here, $\sigma(g)$ is a learnable scalar function with parameter g that modulates the fused output, and $\sqrt{d_c}$ is a scaling factor (Vaswani et al., 2017).

This formulation enables the model to leverage both internal temporal dynamics and relevant external cues from the handshape stream, while the gating mechanism ensures that cross-modal information is selectively integrated in a learnable and data-dependent way.

## 3.5. Sign Recognition Using Boundaries from Segmentation

The predicted boundaries are used to segment the video into sign clips, among which only those segments that have been successfully associated with a ground truth sign and for which we have at least 6 examples in our training dataset are passed to a sign recognition model. Specifically, we integrate into our framework a sign classification model (Zhou et al., 2024), which is used to predict the gloss label for each segment correctly identified by our boundary detection model. Given the output of the boundary detection model, we first identify the corresponding segments from the input sequence. These segments are then passed to the sign language classification model for recognition, laying the groundwork for a future continuous sign language recognition pipeline (Zhou et al., 2024). We use this to show how good boundary segmentation is for sign recognition.

# 4. Experiments

## 4.1. Dataset Construction

### 4.1.1. Handshape Recognition Dataset

We integrated 3 different ASLLRP ASL datasets (Neidle et al., 2022b)—two collections with isolated signs, ASLLVD (Neidle and Metaxas, 2024a; Neidle et al., 2012) and DSP (Neidle and Metaxas, 2024b), and one collection with continuous signing, ASLLRP Sentences (Neidle and Metaxas, 2025a) (henceforth ASLLRP-S)—to construct a unified dataset for handshape recognition. Note that the ASLLVD and ASLLRP-S datasets are shared on the Web from https://dai.cs.rutgers.edu/dai/s/signbank. Statistical information about

these datasets is available from https://dai.cs.rutgers.edu/dai/s/runningstats.

We also used a set of videos collected specifically to illustrate the 87 handshapes in motion from different camera angles and made available through the Boston University NCSLGR (National Center for Sign Language and Gesture Resources, 2012). For each handshape category, the NCSLGR videos provide synchronized multi-view video recordings for the 87 linguistically defined handshapes. We sampled at least 20 frames per handshape from different camera angles, enabling the model to learn cross-view representations of each handshape. In addition, we augmented the dataset by sampling 30 instances per handshape from the 3 other sign language datasets. For rare handshapes that do not meet the 30-sample threshold, we collected as many representative samples as possible. Through this process, we constructed a comprehensive handshape dataset that captures rich visual diversity across categories and views.

#### 4.1.2. Sign Recognition Dataset

For continuous sign data, we adopt the method used in Zhou et al. (2024) We constructed a unified sign classification dataset by integrating video clips from 6 different sources. We use 4 collections with citation-form ASL signs—WLASL (Li et al., 2020) (with modified glossing, as explained below), RIT (Neidle and Metaxas, 2024c), ASLLVD (Neidle and Metaxas, 2024a; Neidle et al., 2012), and DSP (Neidle and Metaxas, 2024b)—as well as 2 datasets in which signs had been extracted from ASL sentences based on manual annotations carried out using SignStream® (see http://www.bu.edu/asllrp/SignStream/3)—ASLLRP-S (Neidle and Metaxas, 2025a) and DSP Sentences (Neidle and Metaxas, 2025b) (hereafter DSP_S).

As previously mentioned, it is critically important that the datasets for this research all make use of consistent glossing conventions. To ensure this, we have modified some of the gloss ID labels for the WLASL dataset to conform to the conventions used for the other datasets (Neidle et al., 2022a; Dafnis et al., 2022; Neidle and Ballard, 2022b,a). Note also that we are relying on "class label" ID glosses to represent signs, which group together minor variants in sign production (Neidle and Opoku, 2022; Neidle and Ballard, 2022a).

We exclude from the sentence datasets signs designated as "hidden". This label identified false starts and signs that otherwise seriously deviate from their canonical articulation. For further details, see https://dai.cs.rutgers.edu/asllvd/signbank/Additional-information-6-29-25.pdf. We used the hidden signs for purposes of segmentation, but excluded them from the sign recognition research. We further restrict the recognition research to signs that were correctly segmented and for which our training set includes at least 6 examples.

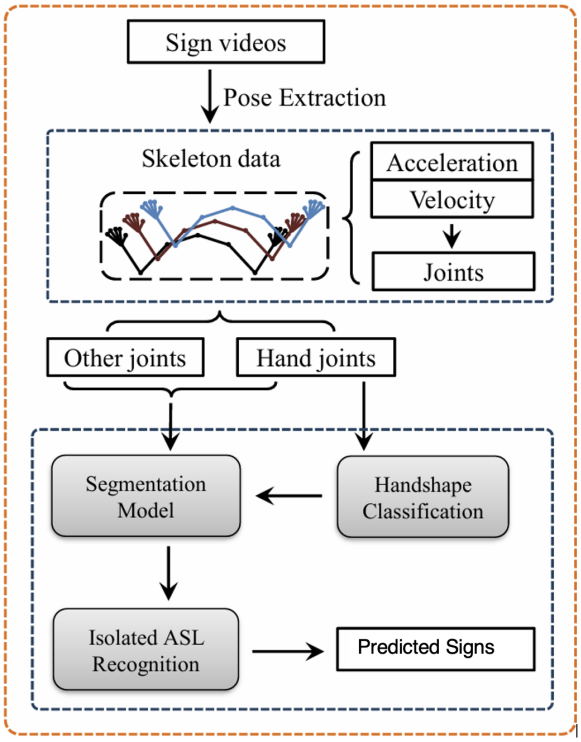



Figure 2: Overview of the experimental procedure. We begin by extracting 2D skeletal keypoints from sign language videos using a pose estimation tool. We use velocity and acceleration information to augment the joint feature for each frame. These are then separated into two branches: hand joints are fed into a pretrained handshape classification model, while all joints are used as input to a segmentation model. The outputs of both branches are fused to produce improved segmentation boundaries. For sign recognition, the segmented sign clips are further passed into a sign recognition model trained on both citation-form signs and signs segmented from continuous signing, to produce final sign predictions.

#### 4.1.3. Boundary Detection Dataset

For boundary detection, we use all of the pre-segmented clips from the ASLLRP-S (Neidle and Metaxas, 2025a) dataset, including the "hidden signs," as described in the previous section. Training for segmentation was performed on the basis of the annotations for the start and end frames of each sign.

### 4.2. Experiment Setup

#### 4.2.1. Data Preprocessing

For our constructed datasets, we first extract skeletal features, including keypoints of the torso, arms,

and hands, from RGB frames using the AlphaPose model (Fang et al., 2023). We also incorporate velocity and acceleration information derived from the keypoint sequences to enrich the temporal dynamics. For both the boundary detection and continuous sign language recognition tasks, we employ a skeletal graph consisting of 27 nodes. In contrast, for the pretrained handshape recognition model, we utilize a skeletal graph constructed solely from hand joints.

In addition, we apply normalization and data augmentation techniques inspired by Jiang et al. (2024), Boháček and Hrúz (2022), Zhou et al. (2022), and Jin et al. (2025). Specifically, we normalize the skeleton keypoints to the range $[-1, 1]$ to enhance cross-dataset transferability. To further improve model robustness, we also apply random horizontal flipping as a data augmentation strategy during training.

#### 4.2.2. Experiment Details

To ensure a fair comparison with the baseline (Renz et al., 2021), we adopt the same 4:1 train-test split ratio and randomly partition the ASLLRP-S dataset. As shown in Figure 2, our overall pipeline consists of skeletal feature extraction, a pretrained handshape classification model, and segmentation and recognition modules. We fine-tune the model with LoRA (Jin et al., 2025) for parameter-efficient adaptation. For the boundary detection task, the model is trained only up to the segmentation module to predict temporal boundaries. For the sign language recognition task, the recognition results of each correctly predicted segment are used to further optimize the boundary detection module. All models are trained on 8 NVIDIA RTX 8000 GPUs.

#### 4.2.3. Evaluation Metrics

We define different evaluation metrics for different tasks. For boundary detection, to facilitate comparison with prior related research, we adopt mF1B (mean F1 score of boundary) and mF1S (mean F1 score of segment) proposed by Renz et al. (2021). The mF1B metric measures the mean F1 score for boundary detection, where a predicted boundary is considered correct if its distance to a ground-truth boundary is less than a predefined threshold. The mF1S metric measures the mean F1 score for sign segmentation, where a predicted sign segment is considered correct if its intersection-over-union (IoU) with a ground-truth segment exceeds a predefined threshold. We adopt the same threshold settings as the baseline paper (Renz et al., 2021), to ensure fair comparison.

We also propose a more intuitive evaluation metric. Specifically, we introduce a tolerance-based matching criterion to evaluate the alignment between predicted and ground-truth sign segments, which we will refer to as Boundary Tolerance. A predicted segment is considered correct if both its start frame and end frame are within a predefined tolerance of the ground-truth start and end frames, respectively. The tolerance is adaptive to the duration of the sign: ±2 frames for signs with duration of 1-5 frames, ±3 frames for signs with 6–10 frames, ±4 frames for signs with 11–23 frames, and ±5 frames for longer signs. This adaptive metric ensures that both boundaries are accurately captured within a reasonable margin. Allowing for some divergence takes account of the fact that neighboring frames may be very similar in appearance; there may also be a little variability in manual annotations for the same reason. Thus, this tolerance allows for reasonable evaluation.

For the CSLR task, we use our tolerance-based matching criterion, Boundary Tolerance, to determine whether we have achieved a successful segmentation. We consider a predicted segment to be aligned with a sign in the ground truth if both the start and end frames of the two signs are within the tolerance range. We then compute the top-1 recognition accuracy across all predicted sign segments that were successfully matched based on this criterion.

### 4.3. Evaluation Benchmark

We separately evaluate our model on 3 tasks: boundary detection, sign recognition from continuous signing, and handshape classification.

#### 4.3.1. Segmentation Comparison

We compare our segmentation results with Renz et al. (2021), although our experiments are carried out on the ASLLRP-S dataset, which is more complex and challenging than the PHOENIX14 dataset (Forster et al., 2014). Specifically, ASLLRP-S contains approximately 2,000 distinct sign classes and fewer than 20,000 annotated sign clips, whereas PHOENIX14 includes around 1,200 sign classes with approximately 23,000 annotated sign clips.

Despite the fact that our dataset is significantly more complex, we adopt the same training-testing split ratio used in prior work and directly compare segmentation performance. As shown in the comparison with other methods in Table 1, Schembri et al. (2013) report a segmentation performance of 61.12 (mF1B) on the BSL-1K test dataset. For the PHOENIX14 dataset, Renz et al. (2021) achieve an mF1B of 71.50.

In contrast, our model achieves substantially stronger segmentation performance on the more challenging ASLLRP-S dataset, with an mF1B of 79.40. Given that ASLLRP-S presents greater complexity than PHOENIX14 in terms of vocabulary

| Method | Dataset | mF1B | mF1S |
|---|---|---|---|
| $RF_{GEO}$ | BSLCORPUS | 68.68 | 41.71 |
| $I3D, TCN$ | PHOENIX14 | 71.50 | 52.78 |
| $Our\ MHB$ | ASLLRP-S | **79.40** | **58.26** |

Table 1: Comparison with other methods. We compare to the geometric features proposed by Farag and Brock (2019) on BSLCORPUS and the I3D + MS-TCN method proposed by Renz et al. (2021) on Phoenix14. Our method significantly outperforms the others on the different metrics.

size and class distribution, these results demonstrate the robustness and effectiveness of our proposed segmentation approach.

#### 4.3.2. Sign Recognition

In the sign recognition task, we adopt a different setting than was used for the standalone boundary detection task, as described earlier. To the best of our knowledge, this is the first work to evaluate sign recognition on continuous sign language datasets using segments derived from predicted boundaries. Therefore, during evaluation we assess performance from two perspectives: boundary detection and sign recognition.

For boundary detection, we adopt our Boundary Tolerance matching criterion. Specifically, we report 2 metrics to evaluate segmentation performance: (i) the proportion of ground-truth signs that have predicted segments that match under our Boundary Tolerance criterion; and (ii) the proportion of predicted segments that match a ground-truth sign under our Boundary Tolerance criterion. This provides a clear assessment of segmentation accuracy, as shown in Table 2.

| Segment type | Match | Total | Prop |
|---|---|---|---|
| Ground-truth signs | 3783 | 6595 | 57.4% |
| Predicted sign segmts | 3783 | 6477 | 58.5% |

Table 2: Segmentation accuracy evaluated based on our tolerance-based matching criterion. For each segment type, **Match** denotes the number of sign segments that have corresponding sign segments under our matching criterion, **Total** is the total number of segments in the ground-truth or predicted sign segments, and **Prop** is the ratio of matched to total (First row is the Recall; Second Row is the Prediction).

We evaluate the improvement in sign recognition based on the predicted segmentation boundaries by conducting experiments under varying thresholds k for the minimum number of training samples per sign class. Evaluation is restricted to predicted sign segments that can be aligned to ground-truth sign segments under our Boundary Tolerance criterion. For varying values of k, we report the top-1 sign recognition accuracy in Table 3, which shows that top-1 recognition accuracy on boundary-matched segments increases consistently as the minimum sample threshold k rises, ranging from 80.2% to 83.3%. These results demonstrate that our predicted segmentation boundaries provide reliable alignment with ground truth, and that recognition performance is strong and improves with stricter data thresholds, confirming the robustness of our approach for downstream sign recognition tasks.

| Threshold $k$ | Top-1 Accuracy |
|---|---|
| 6 | 80.23% |
| 10 | 80.86% |
| 15 | 81.67% |
| 20 | 82.17% |
| 30 | **83.30%** |

Table 3: Sign recognition results on the ASLLRP-S dataset under different thresholds $k$, where $k$ denotes the minimum number of training samples required per class. Top-1 Accuracy is the recognition accuracy based on the Top-1 prediction for each segment.

### 4.4. Ablation Study

We also conduct an ablation study; see Table 4. When both the handshape classification module and the multimodal fusion module are removed, leaving only the spatio-temporal convolutional segmentation model, we observe a small but significant performance drop. The model still outperforms existing methods, demonstrating the strength of our base segmentation framework.

| Method | mF1B | mF1S |
|---|---|---|
| $I3D, TCN$ | 71.50 | 52.78 |
| $MSTCN, HaMeR$ | 76.22 | 50.18 |
| $Ours\ w/o\ handshape\ module$ | 77.29 | 56.98 |
| $Ours\ with\ handshape\ module$ | **79.40** | **58.26** |

Table 4: Ablation study for the boundary detection task. We compare to the I3D + MS-TCN method proposed by Renz et al. (2021) and the MS-TCN + HaMeR method proposed by He et al. (2025). The 2nd row shows our method without the handshape module, and the 3rd presents the full model. Best performance is achieved when the handshape module is incorporated.

## 5. Conclusion

Here we address the task of boundary detection for CSLR and propose a novel pipeline for continuous sign language recognition. We construct a

merged ASL handshape classification dataset and evaluate our method on the ASLLRP-S dataset, measuring boundary detection performance and the accuracy of recognition of successfully segmented signs. The results demonstrate that our approach significantly outperforms existing baselines in boundary detection, as shown in table 4. The ablation studies and other experiments confirm that accurate boundary modeling contributes to the overall recognition performance.

## 6. Acknowledgments

We are grateful to the many people who have helped with the collection, linguistic annotation, and sharing of the ASL data upon which we have relied for this research. In particular, we are endebted to the many ASL signers who have contributed to our database; to Matt Huenerfauth and his team for data collection at RIT; to the many who have helped with linguistic annotations; and to our invaluable Web designer, Augustine Opoku. Thanks also to anonymous reviewers for their suggestions. This work was supported in part by NSF grants 2212302, 2212301, and 2212303, but any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## 7. Bibliographical References

ASL-deafined. 2017. Deaf and Hard of Hearing Population in the United States (expanded overview). https://blog.asldeafined.com/2025/04/deaf-and-hard-of-hearing-population-in-the-united-states-expanded-overview/.

Matyáš Boháček and Marek Hrúz. 2022. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 182–191.

Diane Brentari. 2019. *Sign Language Phonology*. University of Chicago Press.

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. 2021. Aligning subtitles in sign language videos. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11552–11561.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084.

Konstantinos M. Dafnis, Evgenia Chroni, Carol Neidle, and Dimitri Metaxas. 2022. Isolated sign recognition using ASL datasets with consistent text-based gloss labeling and curriculum learning. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 13–20, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2023. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173.

Iva Farag and Heike Brock. 2019. Learning motion disfluencies for automatic sign language segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7360–7364.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matilde Gonzalez and Christophe Collet. 2011. Sign segmentation using dynamics and hand configuration for semi-automatic annotation of sign language corpora. In *International Gesture Workshop*, pages 204–215. Springer.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376.

Leming Guo, Wanli Xue, Qing Guo, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Shengyong Chen. 2023. Distilling cross-temporal contexts for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10771–10780.

Leming Guo, Wanli Xue, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Dimitris Metaxas. 2024. Gloss prior guided visual feature learning for continuous sign language recognition. *IEEE Transactions on Image Processing*, 33:3486–3495.

Haoyu Han, Mengdi Zhang, Min Hou, Fuzheng Zhang, Zhongyuan Wang, Enhong Chen, Hongwei Wang, Jianhui Ma, and Qi Liu. 2020. Stgcn: A spatial-temporal aware graph learning method for poi recommendation. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1052–1057.

Low Jian He, Harry Walsh, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Hands-on: Segmenting individual signs from continuous sequences. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–5. IEEE.

Jiani Huang, Wengang Zhou, Houqiang Zhang, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 2257–2264. Article No. 275.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193, Miami, Florida, USA. Association for Computational Linguistics.

Can Jin, Ying Li, Mingyu Zhao, Shiyu Zhao, Zhenting Wang, Xiaoxiao He, Ligong Han, Tong Che, and Dimitris N. Metaxas. 2025. Lor-VP: Low-rank visual prompting for efficient vision model adaptation. In *The Thirteenth International Conference on Learning Representations*, pages 9359–9376.

Tonni Das Jui, Gissella Bejarano, and Pablo Rivas. 2022. A machine learning-based segmentation approach for measuring similarity between sign languages. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 94–101.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on Applications of Computer Vision*, pages 1459–1469.

Natalia Lusin, Terri Peterson, Christine Sulewski, and Rizwana Zafer. 2023. Enrollments in languages other than English in US institutions of higher education, fall 2021 - Modern Language Association of America.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

Boston University (BU) National Center for Sign Language and Gesture Resources. 2012. NCSLGR Handshapes. Videos available from `https://www.bu.edu/asllrp/cslgr/pages/ncslgr-handshapes.html`.

Carol Neidle. 2023. Challenges for linguistically-driven computer-based sign recognition from continuous signing for American Sign Language. *arXiv preprint arXiv:2311.00762*.

Carol Neidle and Carey Ballard. 2022a. Revised Gloss Labels for Signs from the WLASL Dataset - Boston University. `http://www.bu.edu/asllrp/wlasl-alt-glosses.pdf`.

Carol Neidle and Carey Ballard. 2022b. Why alternative gloss labels will increase the value of the WLASL dataset - ASLLRP Project Report no. 21, Boston University. Project reports available from `https://www.bu.edu/asllrp/reports.html`.

Carol Neidle and Dimitris Metaxas. 2024a. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank: Boston University American Sign Language Lexicon Video Dataset (ASLLVD) version 8 - June. Videos and annotations available from `https://dai.cs.rutgers.edu/dai/s/signbank`.

Carol Neidle and Dimitris Metaxas. 2024b. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank: DawnSignPress (DSP) ASL Sign Collection, version 6. Videos provided by DawnSignPress.

Carol Neidle and Dimitris Metaxas. 2024c. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank: Rochester Institute of Technology (RIT) collection, version 6. Videos

and annotations available from https://dai.cs.rutgers.edu/dai/s/signbank.

Carol Neidle and Dimitris Metaxas. 2025a. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank: ASLLRP Continuous Signing Corpora - version 3 - June 2025. Videos and annotations available from https://dai.cs.rutgers.edu/dai/s/signbank.

Carol Neidle and Dimitris Metaxas. 2025b. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank: DawnSignPress (DSP) Sentences Collection, version 4. Videos provided by DawnSignPress.

Carol Neidle and Augustine Opoku. 2022. Addendum to: Documentation for Download of ASLLRP Sign Bank Citation-form Sign Datasets - ASLLRP Project Report no. 20, Boston University. Project reports available from https://www.bu.edu/asllrp/reports.html.

Carol Neidle, Augustine Opoku, Carey Ballard, Konstantinos M Dafnis, Evgenia Chroni, and Dimitris Metaxas. 2022a. Resources for computer-based sign recognition from video, and the criticality of consistency of gloss labeling across multiple large ASL video corpora. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, Marseille, France, 25 June 2022*, pages 165–172.

Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022b. ASL Video Corpora & a Sign Bank: Resources available through the American Sign Language Linguistic Research Project (ASLLRP). *arXiv preprint arXiv:2201.07899*.

Carol Neidle, Ashwin Thangali, and Stan Sclaroff. 2012. Challenges in development of the American Sign Language Lexicon Video Dataset (ASLLVD) corpus. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 143–150, Istanbul, Turkey. European Language Resources Association (ELRA).

Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. 2015. Sign language recognition using convolutional neural networks. In *Computer Vision – ECCV 2014 Workshops*, pages 572–578. Springer.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2024. A transformer model for boundary detection in continuous sign language. *Multimedia Tools and Applications*, 83(42):89931–89948.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2025. A non-anatomical graph structure for boundary detection in continuous sign language. *Scientific Reports*, 15(1):25683.

Katrin Renz, Nicolaj C Stache, Samuel Albanie, and Gül Varol. 2021. Sign language segmentation with temporal convolutional networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139.

Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. 2013. Building the British Sign Language corpus. *Language Documentation & Conservation*, 7:136–154.

Thad Starner and Alex Pentland. 1998. Real-time American Sign Language recognition using desk and wearable computer-based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Christian Vogler and Dimitris Metaxas. 1999. Parallel Hidden Markov Models for American Sign Language Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 116–122.

Christian Vogler and Dimitris Metaxas. 2001. A framework for recognizing the simultaneous aspects of American Sign Language. *Computer Vision and Image Understanding*, 81(3):358–384.

Polina Yanovich, Carol Neidle, and Dimitris Metaxas. 2016. Detection of major ASL sign types in continuous signing for ASL recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3067–3073.

Huaiwen Zhang, Zihang Guo, Yang Yang, Xin Liu, and De Hu. 2023. C2st: Cross-modal contextualized sequence transduction for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21053–21062.

Sen Zhang, Xiaoxiao He, Di Liu, Zhaoyang Xia, Mingyu Zhao, Chaowei Tan, Vivian Li, Bo Liu, Dimitris N. Metaxas, and Mubbasir Kapadia. 2026. Large sign language models: Toward 3D American Sign Language translation. In *Proceedings of the IEEE/CVF Winter Conference on*

*Applications of Computer Vision (WACV)*, pages 3109–3119.

Xuan Zhang and Kevin Duh. 2023. Handshape-aware sign language recognition: Extended datasets and exploration of handshape-inclusive methods. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2993–3002.

Yan Zhang, Wanli Xue, Yuxi Zhou, Tiantian Yuan, and Shengyong Chen. 2025. Core: multi-link graph attention network with inter-regional collaboration for continuous sign language recognition. *Pattern Recognition*, 167:111716.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2020. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13009–13016.

Yang Zhou, Zhaoyang Xia, Yuxiao Chen, Carol Neidle, and Dimitris N. Metaxas. 2024. A multimodal spatio-temporal GCN model with enhancements for isolated sign recognition. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 408–419, Torino, Italia. ELRA and ICCL.

# Author Index


Alikhani, Malihe, 226
Alizadeh, Tohid, 212
Andre, Elisabeth, 297, 306, 429
Antzakas, Klimis, 361
Atkinson, Joanna, 113
Avramidis, Eleftherios, 80

Báez-Montero, Inmaculada C., 140
Bao-Fente, María C., 140
Barberà, Gemma, 1
Barbieri, Isabela, 31
Basso Madjoukeng, Ariel, 10, 419
Battisti, Alessia, 17
Bayanov, Daniyal, 398
Benetti, Ângelo, 31
Benoit, Frenay, 10
Beskow, Jonas, 256
Beuls, Katrien, 100
Blat, Josep, 361
Bleicken, Julian, 155, 361
Blekos, Kostas, 361
Boddu, Raviteja, 31
Bono, Mayumi, 491
Börstell, Carl, 42
Bowden, Richard, 51
Braffort, Annelies, 174, 203
Broto Clemente, Inés, 1
Brown, Matt, 51
Brown, Phillip, 120
Bulla, Jan, 61, 459

Cassany Viladomat, Roger, 1
Chan, Frederick, 71
Cheng, Qi, 71
Cormier, Kearsy, 51
Couplet, Adelaide, 10, 419
Curiel, Arturo, 174
Czehmann, Vera, 80

Dai, Zixuan, 93
Dambre, Joni, 468
Dang, Anh, 341
De Coster, Mathieu, 468
De Vos, Liesbet, 100
Dimou, Athanasia-Lida, 113
Duppen, Yves A., 120

Ebling, Sarah, 17, 371
Efthimiou, Eleni, 113
Eing, Lennart, 297, 306
Elsendoorn, Ben, 361
Esteban Romero, Sergio, 297

Fabre, Diandra, 129, 203
Fernández Soneira, Ana, 140
Ferrara, Lindsay, 148
Fiedler, Anike, 155
Filhol, Michael, 164
Fink, Jérôme, 419
Fish, Edward, 51
Fotinea, Stavroula-Evita, 113
Frénay, Benoît, 419

de la Garza, Lorena, 174
Gebhard, Patrick, 297, 306
Georges, Munir, 31
Gibet, Sylvie, 184
González-Montesino, Rayco H., 140
González, Marco, 306
Gouiffès, Michèle, 174, 203
Goulas, Theodore, 113
Gren, Gustaf, 193
Gurbuz, Sevgi Z., 438

Halbout, Julie, 129, 174, 203
Hamidullah, Yasser, 80
Harbour, Eric, 324
He, Xiaoxiao, 501
Henter, Gustav Eje, 256
Hernández, Doris, 477
Herrmann, Annika, 155, 361
Hoffmann-Lamplmair, Doris, 113

Imai, Saki, 226
Imashev, Alfarabi, 212
Inan, Mert, 226
Isard, Amy, 341

Jantunen, Tommi, 477

Jin, Can, 501

Kaltay, Aruzhan, 398
Kankkonen, Nikolaus Riemer, 193
Karampidis, Konstantinos, 361
Karel, Tessa, 226
Kenmogne, Belise Edith, 10
Keränen, Jarkko, 477
Khan, Sarmad, 239
Khristoforova, Evgeniia, 248
Kimmelman, Vadim, 61
Klezovich, Anna, 256
König, Lutz, 262
Konrad, Reiner, 262
Konstantakopoulos, Athanasios, 361
Kopf, Maria, 262
Kosmopoulos, Dimitrios, 361
Krebs, Julia, 324
Kuder, Anna, 351

Langer, Gabriele, 262
Lascar, Julie, 129, 174, 203
Leleu, Margaux, 10, 419
Lepp, Lisa, 120, 275
Levow, Gina-Anne, 71
Lin, Shuhang, 501
Lopes da Silva, Joed, 31
Loy, Lisa, 287
Luna Jimenez, Cristina, 297, 306, 429

Maina, Ezekiel, 315
Malaia, Evguenia A., 324
Malaia, Evie A., 438
Marques, Ricardo, 361
Marrocu, Maria Grazia, 333
Marshall, Anna, 226
Marteau, Pierre-François, 184
Martetschläger, Julia, 324
Martinez, Niels, 174
De Martino, José Mario, 31
Martinod, Emmanuella, 164
Mavridou, Ifigeneia, 120
McLoughlin, Simon, 239
de Melo Afonso, Natália, 31
Mesch, Johanna, 256
Metaxas, Dimitri, 501
Meurant, Laurence, 100
Miquel Blasco, Carla, 459
Morgan, Hope E., 287, 341
Mostowski, Piotr, 351
Müller, Mathias, 371
Murtagh, Irene, 239, 361

Neidle, Carol, 501
Ntova, Tarsita, 113
Nunnari, Fabrizio, 80, 297, 306

O'Brien, Catherine, 371
Obuhuma, James, 315
Okrouhlíková, Lenka, 387
Orazumbekov, Batyrbek, 398
Othamar, Elisabeth, 408
Otterspeer, Gomèr, 450

Paulus, Liona, 361
Pedrini, Hélio, 31
Perniss, Pamela, 306
Poitier, Pierre, 10, 419
Poryadin, Roman, 248
Proctor, Heidi, 51
Puupponen, Anna, 477

Ranum, Oline, 51
Renner, Fabian, 429
Reverdy, Clément, 184
Roehm, Dietmar, 324
Roelofsen, Floris, 450

Sako, Shinji, 93
Sandygulova, Anara, 398
Sant, Gerard, 371
Santos, Thyago, 31
Sazonov, Dmitriy, 438
Scherrer, Yves, 408
Schiefner, Annika, 450
Schneeberger, Tanja, 297
Schulder, Marc, 155, 262, 361
Schwameder, Hermann, 324
Senel Withanage Don, Daksitha, 306
Shterionov, Dimitar, 120, 275
Sidler-Miserez, Sandra, 17
De Sisto, Mirella, 120, 275
Sümer, Beyza, 450
Susman, Margaux, 459

Teichmann, Birgit, 113
Tissi, Katja, 17
Tsatali, Marianna, 113
Tsolaki, Magda, 113

Valls, Eva, 361
Van Eecke, Paul, 100
Vandendriessche, Toon, 468
Vandenitte, Sébastien, 477
Vartampetian, Markarit, 129
Venâncio Barbosa, Felipe, 31

Vieira Leite, Guilherme, 31
Vinaixa Roselló, Xavier, 1

Wähl, Sabrina, 484
Wang, Zirui, 491
Wanzare, Lilian, 315
Wilbur, Ronnie B., 324
Withanage Don, Daksitha, 429
Wójcicka, Joanna, 351
Woll, Bencie, 51, 113

Xia, Zhaoyang, 501

Yang, Zhanfu, 501
Yazdani, Shakib, 80

Zhao, Mingyu, 501
Zhou, Yang, 501
Zimmer, Alessandro, 31