Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC 2022), pages 88–93
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

MC-TRISLAN: A Large 3D Motion Capture Sign Language Data-set

Pavel Jedlička, Zdeněk Krňoul, Miloš Železný, and Luděk Müller
NTIS - New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, 306 14 Pilsen, Czech Republic.
{jedlicka, zdkrnoul, zelezny, muller}@ntis.zcu.cz

Abstract

The new 3D motion capture data corpus expands the portfolio of existing language resources by a corpus of 18 hours of Czech sign language. This helps to alleviate the current problem, which is a critical lack of high quality data necessary for research and subsequent deployment of machine learning techniques in this area. We currently provide the largest collection of annotated sign language recordings acquired by state-of-the-art 3D human body recording technology for the successful future deployment in communication technologies, especially machine translation and sign language synthesis.

Keywords: Motion Capture, Sign Language, Human Body Pose

1. Introduction

Sign languages (SLs) are natural means of communication for deaf people. Hundreds of sign languages are used around the world today. Czech sign language (CSE) is one of sign languages in Europe and, in general, each country has one or more native sign languages (Timmermans, 2005).

Although significant progress has been made in recent years in the field of spoken language machine learning techniques, the field of SL processing struggles with a critical lack of high quality data needed for the successful application of these techniques. For comparison, WaveNet-based speech synthesis method has been trained on data set contained 10,000 utterances (about 14 hours of speech) of one professional male speaker (Vít et al., 2018). SL resources are scarce – they consist of small SL corpora usually designed for a specific domain such as linguistics or computer science. There are some motion capture data-sets for American Sign Language (ASL) and French Sign Language (Lu and Huenerfauth, 2010; Naert et al., 2017). The total recorded time of motion is up to 60 minutes in those data-sets. The situation is even worse for “small” languages like CSE.

There are techniques for the 3D reconstruction of human body pose/motion from RGB and depth images and this is a common approach for capturing human body movement (MMPose Contributors, 2020; Cao et al., 2017). Current SL data-sets are mostly video-based (Joze and Koller, 2019; Zelinka and Kanis, 2020). Although video is a natural way of capturing sign languages, these readily available data sources are ambiguous in the sense that they do not contain spatial (3D) information. For comparison SIGNUM, one of the largest video-based SL data-sets, contains approximately 55 hours of SL recordings (Koller et al., 2015), other example of large data-set is DGS-Corpus with more than 47 hours of SL recordings, see (Wolfe et al., 2022). On the contrary, one of the largest 3D motion capture data-sets contain only 60 minutes of SL recordings (Naert et al., 2017; Naert et al., 2020).

Motion capture technologies guarantee high precision recording of the signer’s movements in 3D space at the cost of a more complex preparation phase compared to standard video recording. Optical marker-based motion capture has become the industry standard for capturing movement of the human body. One of the first publicly available scientific motion capture SL data-set was recorded in 2016, see (Benchiheub et al., 2016).

In (Jedlička et al., 2020), we collected the first 3D motion capture data-set for CSE, which covers the weather forecast domain. This data-set is rather limited in size and contains recordings of one signer only. 18 mocap simultaneously recording cameras were used to capture SL, which was our first step in the research towards a new concept of sign language capture in 3D. The total length of recordings was 42 minutes.

A large number of cameras eliminates the frequency of marker occlusions and thus the loss of measurements. However, this method turned out to be very time consuming and not suitable for large data and records from multiple markers. The negative aspect is its high complexity both in the recording and the post-processing phase. In principle, this approach does not allow us to get more data for a given price. Additionally, the occlusion issue is not resolved in this case. Occlusion has been shown to be a significant problem for hand markers in general, where hand poses in sign language are often in contact with each other or with the face.

In this work, We deliver a new recording protocol and a large 3D motion data-set collected using high precision optical marker-based motion capture system in order to extend the existing portfolio of language resources with Czech sign language (CSE) data. The contribution of our work can be summarized as follows:

- Proof of concept for large-scale motion capture recording by splitting hand-configuration and body recording;

- 3D motion capture protocol to cover wider domains, grammatical context and more signers. We assume proper data post-processing, annotation, and tools for data extraction from the collected data;

- The largest SL motion capture data-set of sign language consisting of recordings of continuous CSE phrases and a vocabulary of six native SL speakers from carefully selected domains, in total more than 18 hours. The data-set is available at *https://live.european-language-grid.eu/catalogue/corpus/18324*

2. Related Work

Recently, spoken language research is directed to machine learning algorithms, deep neural learning in particular. In the field of sign languages, common tasks are translation, speech recognition and synthesis (Zelinka and Kanis, 2020; Stoll et al., 2020; Gruber et al., 2021). The goal of sign language synthesis is to generate natural, natural, and intelligible video-utterances of SL based on methods capable of mimicking human SL's performance.

There are techniques developed for the pose estimation from the image or video, e.g. OpenPose (Cao et al., 2017) or MMpose (MMPose Contributors, 2020). There methods are marker-less with no restriction on the freedom of movement of the hands but the 3D precision is in principle lower than the actual 3D pose measuring provided by MoCap systems.

Some data-sets using different motion capture techniques were created in recent years (Lu and Huenerfauth, 2010; Naert et al., 2020; Jedlička et al., 2020). (Lu and Huenerfauth, 2010) recorded American SL using magnetic-based motion capture for hand and finger tracking. The evolution of motion capture data-sets collected in French SL is described in (Gibet, 2018). They recorded several MoCap data-sets in the last 15 years. All of them contain manual and non-manual components of SL.

The project HuGEx (2005) used the Vicon MoCap system in combination with Cybergloves for recording finger movements and for the body and the facial movements. The total recording time was 50 minutes. The Sign3D project ((Lefebvre-Albaret et al., 2013)) recorded the position of the body and hand with the same system in combination with the eye gaze recorded with a head-mounted oculomotor (MocapLab MLab 50-W). However, it contains 10 minutes of recorded data only. More recently in (Naert et al., 2020), the authors collected the LSF-ANIMAL corpus that composed of captured isolated signs and full sentences that can be used both to study LSF features and to generate new signs and utterances.

In contrast, we assume that we can reconstruct the hand pose and other SL components with only one technique and with minimal restrictions on signers' body movement. We rather follow SignCom ((Gibet et al., 2011)) and we use the Vicon MoCap system to record 3D pose with limited markers per hand and face. We provide a protocol suitable for acquiring large volumes of SL data using the motion capture system.

There is a continual need for a large amount of data to utilize machine learning techniques. Although the quality and size of data-sets are increasing, there is still a lack of such data. The usual size of those data-sets is between 10 and 60 minutes of recording time.

3. Objective

The aim of this research is to create a new large data-set of sign language suitable for sign language synthesis based on machine learning techniques. 3D data are essential for synthesizing new, natural, and realistic utterances of a data driven avatar (Naert et al., 2020). The problem of synthesis lies in modifying and connecting captured movements. One of the main problems is how to capture a shape and motion of human body in 3D space with sufficient precision.

By fulfilling this goal we gain the opportunity to work on large scale data. In particular, data-set contains occurrences of signs and grammar structures in natural context. This is beneficial for analysis of movement, linguistics and other phenomena in SLs. We will use the data as a ground truth for design, observations, and evaluation for new algorithms for SL synthesis.

Movement of human body during sign language utterances is very specific and complex. The movements of hands, and body, as well as facial expressions are made simultaneously in SL utterances. We assume that continuous speech is most natural manifestation of sign language. So we are solving a problem where complex movements demands specific and elaborate setup and on the other hand large volume of data is needed. Our approach to data-set acquisition attempts to meet both demands.

4. Data-set MC-TRISLAN

We have done experiments with different setups and protocols in order to find one suitable for recording large scale of motion capture data of sign language. As a result we have made the MC-TRISLAN data-set.

4.1. Methods and Experiment

The recording is divided into two separate parts, according to our protocol: One is the recording of a body movement with a simplified hand and face model, and the other one being highly detailed hand pose and movement. This division allows us to use different motion capture system settings for both recordings, each using simpler settings and therefore a reduced number of cameras. It also allows faster motion capture system preparation and fine tuning for a new SL speaker.

4.1.1. Recording Setup

We used our VICON motion capture system based on infrared high-speed cameras T-20. This system uses passive retro-reflexive markers placed on special suit or directly on body. The choice of marker sizes and their exact locations on the body of the SL speaker is crucial for precise measurement of the movement during recording. The resulting movement is modeled as a trajectories of a skeleton, which is composed of bones representing measured rigid body parts. The recording was made using 8 cameras set to frame-rate of 100 fps. This frame-rate is a compromise between need for capture SL dynamics with enough precision and increasing noise levels. Recording setup was extended with standard RGB camera for reference video at 25 fps, see Figure 3.

The cameras placement and speaker preparation depends on the type of recording. For whole body movement, we used standard VICON 3-finger body setup. The markers are located on the poles of the axis of rotation of the joints of the skeleton. Each body part is defined by at least 4 markers except the fingertips, see Fig. 1. Total number of markers tracked in full body recording is 59. Tracked fingertips are the thumb, index and pinky. The fingertips are not well defined and, in general, lost tracking can not be corrected or traced from another marker. We used 7 markers for the recording of non-manual component. This setup is used for the whole body recording of continuous speech and dictionary items, particularly for capturing all three parts at once: hand configuration (HC), position and non-manual component (Sandler and Lillo-Martin, 2006).

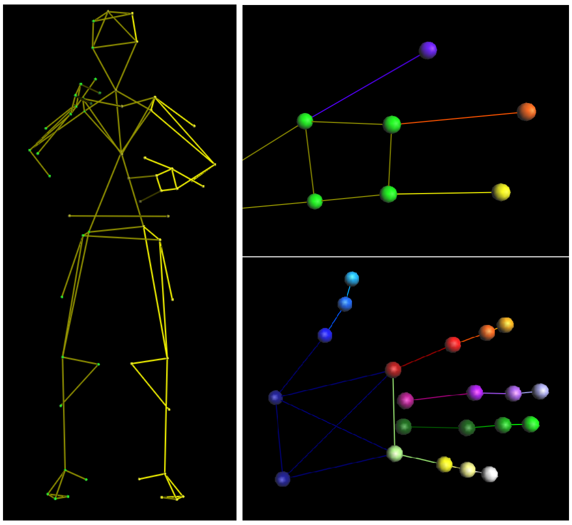

Figure 1: The marker-setups for body on left and hand pose on right. The body pose marker-setup integrates simple hand pose by 7 markers (right up). High detailed hand poses are reconstructed from hand marker-setup data (right down).

We used 21 marker setup developed specifically for hand configuration (HC) recording. There is a set of HC used in a vocabulary extended with some common HCs, see Fig. 2. HCs are recorded separately with limited arm movement. One HC is recorded at one time and the movement of the recording hand is from a relaxed position to the given HC and then back to a relaxed pose. Data are recorded for dominant hand only, we use those data also for non-dominant hand. The exact position of markers on the hand is very important for the 3D reconstruction of the hand skeleton. From the point of view of capturing all degrees of freedom, the location of the markers is not unified (Hoyet et al., 2012).

Our hand marker setup is based on our previous research (Jedlička et al., 2020). We are newly proposing a small marker for each finger joint, which is placed on the top of the hand to prevent the hand from moving as much as possible during signing. Additionally, we put one marker for each fingertip and two markers on the wrist. We used a total of 21 hemispherical markers with 4 mm diameter. The markers were attached to a skin with a double-sided adhesive tape. Note, that 7 of these markers are at the same positions as markers in the whole body setup, and therefore, can be used as reference for data composition. Motion capture cameras are placed closer to the speaker, 2 from above and 2 from below the speaker's hand, the other 4 surround the hand from the sides.

4.1.2. Data-set Design and Data Acquisition

To select suitable domains and to estimate the amount of SL recordings to cover them, we cooperated with CSE linguists, translators and native speakers. The data-set design was done so that it contains sufficient informational data, and including multiple instances of the same signs in different grammar contexts. All recordings were made twice. The possibility to choose between instances of the same movement segment is beneficial for the fine setup of synthesis (Gibet, 2018). We limited the linguistic domain to two specific fields to reduce the number of unique signs. Weather forecasts and animal descriptions from the zoological garden domain were selected by CSE linguists (Dictio Contributors, 2022), see Table 1. Linguists have provided us a list of all HCs that occur in these domains, see Figure 2. The data-set is collected from 6 native CSE speakers, who differ in body size, age, and gender.

The data-set was collected during 37 recording sessions, the recording team of each session consisted of our recording staff, an SL speaker and one or more SL quality control expert(s). The sessions were divided into two separate tasks. The first task was to record a complete set of hand configurations (HCs) that are used in the selected topics. The signer is obligated to perform each HC separately. The movement starts from the relaxed HC, then changes to the given HC and back to the relaxed HC.

The second task was to record whole body movement

Topic	Weather forecast	ZOO tour
Structure	36 individual forecasts (one forecast ∼30 sec continuous speech)	20 different animals - Structured description
Vocabulary Type	diversity - cover forecast topic (3 forecasts per month)	Large sample of letters (Latin names - finger-spelling) (extensible)
Vocabulary Size	Limited (> 300 signs), Large sample of numbers	Limited (> 800 signs)
Data Characteristic	Multiple instances of the same sign in different context (frequent signs more than 20 repetitions)	Repetition of similar sign groups (biotope, food, lifespan, ...)

Table 1: Topics, vocabulary and data characteristic of MC-TRISLAN data-set.

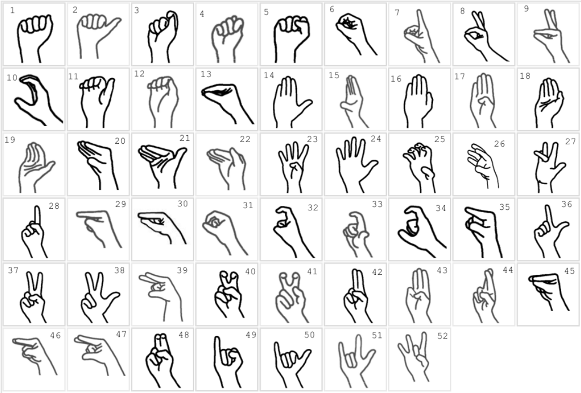

Figure 2: The list of all hand configurations.

using the 3-finger setup. This task consisted of isolated signs for vocabulary and continuous SL utterances. For each topic, the first signer was carefully selected so that his entries would serve as a template for the other signers' entries. These signers were informed of the required recorded content in advance by watching a reference video. Thus the content was the same for all signers. Instructions for each task were displayed on a large screen in front of the signer. Signers could choose whether they wanted a text template, a video or their combination. But signers had always been instructed to make the most natural and realistic sign language production possible. The signer was obliged to perform the given utterance in such a way that he started from the T-position, shifted to the rest position, performed the given utterance and finally returned through the rest position to the T-position.

4.1.3. Data Annotation

An essential step is the annotation of captured SL utterances. We used a reference video, that is time-synchronised manually and the ELAN tool, see Figure 3. The annotation of a sign is done by SL experts giving the information of the sign's meaning (gloss), and the right and the left HC. If the sign consists of more than one defined HC, the HC are annotated as a set of HC. Both the activities are very laborious and time-consuming. To successfully complete this task, we are involving several trained annotators who worked in parallel.

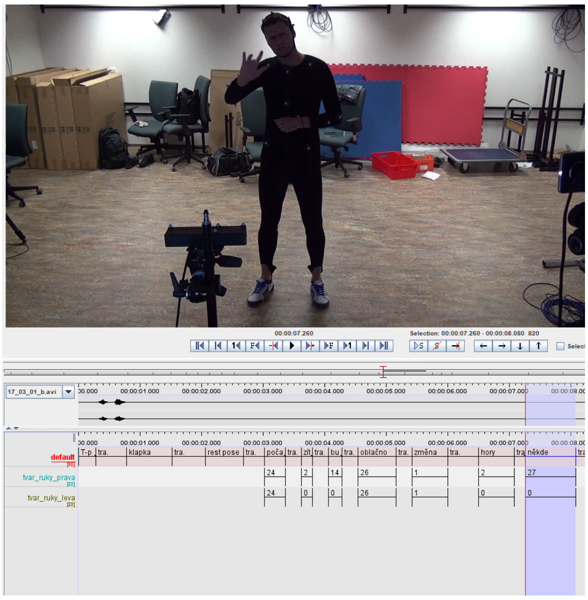

Figure 3: Example of annotation work in ELAN, we use reference video annotated by SL experts.

4.1.4. Data Post-processing

Post-processing consists of data-cleaning, whole-body motion reconstruction, and data-solving. Data-cleaning removes noise and fills gaps in the raw 3D data caused by frequent mutual occlusions of markers during signing, and other noise caused by the environment. Motion reconstruction recalculates the position of the marks on the motion of the skeletal model using a data solver.

The data of both setups was post-processed. For HC setup, we reconstructed small gaps by the interpolation standard technique as long as the trajectory was simple enough. Note, that the recording speed is 100 fps, which is fast enough to contain minimal changes in trajectory between frames. We used semi-automatic 3D reconstruction of marker trajectories and labeling, and manual cleaning of swaps and gaps. For the body parts defined by at least four markers, filling in the trajectories of the marker is well automatised because at least three points are enough to define the missing position.

The body marker setup uses only one marker per fingertip and some larger gaps caused by more complex self-occlusions of body parts can obscure three or more markers in one rigid segment. Post-processing in those cases is more complicated and gaps must be filled in manually.

The full SL body movement is achieved as a composition of the body movement and corresponding data of the HCs setup. For this purpose, the annotation of HCs provides us with temporal segmentation of the recordings, see Figure 3. Thus the 3-fingertip motion in the segment provides information about dynamic changes during the performance of the HC in a particular data frame.

The middle part of the segment is always completed from reference data according to the HC(s) assigned by the annotation for each hand. We captured full fingers motion only for the transition of the given HC from and to the neutral HC. Thus, for the reconstruction of the other frames of the segment, the nearest hand pose with the smallest reconstruction error (1) were used. We consider only those frames that contain the trajectory of fingertips and where the error is below a given threshold τ. The remaining frames will have gaps in the final trajectories of high detailed hand pose.

We solved the above problem as point-set alignment via Procrustes analysis that arises especially in tasks like 3D point cloud data registration. The rigid transformation of two sets of points on top of each other minimises the total distance in 3D between the corresponding markers (Arun et al., 1987). Since the data is noisy, it minimises the least-squares error:

$$err_f = \sum_{i=1}^{N} ||R_f M_f^i + t_f - M_{rf}^i||, \qquad (1)$$

where M_f and M_{rf} are current and reference frame(s) respectively as a set of N 3D points with known correspondences, R_f is the rotation matrix and t_f the translation vector for given frame f. We assume $3 \leq N \leq 7$ points of the 3-fingers setup, see Figure 1. We aligned only the rotation and translation because the 3D the transformation preserves the shape and size (same HC and SL speaker). For the non-dominant hand, we mirrored the reference frame(s) recorded for dominant hand only.

The last step of the post-processing is motion datasolving. It is a process of reconstruction of the 3D motion of the skeleton from the 3D marker trajectories. For this purpose, we use the VICON software. The skeleton is well defined to directly control the SL avatar animation or handle animation retargeting.

4.2. Discussion

The key factor for optical motion capture is the correct identification of each trajectory in 3D, so-called labeling. We chose the marker setup, that reduces the amount of occlusions and marker swaps. A lower marker count placed on the hands and a reduced number of facial markers reduce significantly the labeling complexity. This is crucial for processing large volume of data. In order to complement the data, detailed HC recording is done. This detailed HC recording uses a high number of markers that provide precise information at the cost of demanding manual post-processing. HC recording is done only once for each SL speaker and does not increase up with the number of recorded signs.

5. Conclusion

SLs are not sufficiently supported through technologies and have only fragmented, weak, or no support at all. We propose a new protocol that solves the problem of complex data-set creation and provides a procedure for obtaining sufficient diversity of SL speakers, grammar and character contexts. In contrast to the all-in-one recording setup, the body movement is recorded separately from the highly detailed recording of hand poses. This separation reduces the complexity of camera setup and data during post-processing, making SL recording more flexible and making adjustments for new SL speakers or large data easier. Data processing procedures are an integral part of the experiment. The protocol therefore provides complete instructions for the necessary post-processing and annotation.

As result, a professionally created SL data-set via state-of-the-art 3D motion capture technology is introduced. The data-set provides data for the wider research community. We have recorded 18 hours of sign language and recorded six different speakers for two different domains. This makes this data-set more versatile and useful in many different areas of research, such as other linguistic and SL analysis.

We assume our results will be beneficial for other applications such as next generation SL synthesis that uses a 3D animated avatar for natural human movement reproduction or SL analysis or gesture recognition and classification in general.

6. Acknowledgements

We want to thank staff of Institute of Special Education Studies, Palacký University Olomouc, mainly J. Langer, P. Kučera and J. Grohmannová for their cooperation in signers selection, data elicitation and linguistics support. The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects and the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ. The ELG project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825627.

7. Bibliographical references

Arun, K. S., Huang, T. S., and Blostein, S. D. (1987). Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700.

Benchiheub, M., Berret, B., and Braffort, A. (2016). Collecting and Analysing a Motion-Capture Corpus of French Sign Language. In *Workshop on the Representation and Processing of Sign Languages*, pages 7–12, Portoroz, Slovenia, January.

Cao, Z., Simon, T., Wei, S., and Sheikh, Y. (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.

Dictio Contributors. (2022). Online monolingual dictionary. `https://www.dictio.info/about?lang=en`.

Gibet, S., Courty, N., Duarte, K., and Le Naour, T. (2011). The SignCom System for Data-Driven Animation of Interactive Virtual Signers : Methodology and Evaluation. *ACM Transactions on Interactive Intelligent Systems* , 1(1):6.

Gibet, S. (2018). Building French Sign Language Motion Capture Corpora for Signing Avatars. In *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC 2018*, pages 53–58, Miyazaki, Japan, May.

Gruber, I., Krňoul, Z., Hrúz, M., Kanis, J., and Boháček, M. (2021). Mutual support of data modalities in the task of sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3424–3433, June.

Hoyet, L., Ryall, K., McDonnell, R., and O'Sullivan, C. (2012). Sleight of hand: Perception of finger motion from reduced marker sets. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D '12, page 79–86, New York, NY, USA. Association for Computing Machinery.

Jedlička, P., Krňoul, Z., Kanis, J., and Železný, M. (2020). Sign language motion capture dataset for data-driven synthesis. In E. Efthimiou, et al., editors, *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages*, pages 101–106, Marseille, France, May. European Language Resources Association (ELRA).

Joze, H. R. V. and Koller, O. (2019). Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 100. BMVA Press.

Koller, O., Forster, J., and Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Lefebvre-Albaret, F., Gibet, S., Turki, A., Hamon, L., and Brun, R. (2013). Overview of the Sign3D Project High-fidelity 3D recording, indexing and editing of French Sign Language content. In *Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT) 2013*, Chicago, United States, October.

Lu, P. and Huenerfauth, M. (2010). Collecting a motion-capture corpus of american sign language for data-driven generation research. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pages 89–97. Association for Computational Linguistics.

MMPose Contributors. (2020). OpenMMLab Pose Estimation Toolbox and Benchmark. `https://github.com/open-mmlab/mmpose`.

Naert, L., Larboulette, C., and Gibet, S. (2017). Coarticulation analysis for sign language synthesis. In Margherita Antona et al., editors, *Universal Access in Human–Computer Interaction. Designing Novel Interactions*, pages 55–75, Cham. Springer International Publishing.

Naert, L., Larboulette, C., and Gibet, S. (2020). LSF-ANIMAL: A Motion Capture Corpus in French Sign Language Designed for the Animation of Signing Avatars. In *LREC 2020*, Marseille, France, May.

Sandler, W. and Lillo-Martin, D. (2006). *Sign Language and Linguistic Universals*. Cambridge University Press (CUP), The Edinburgh Building, Cambridge CB2 2RU, UK, 02.

Stoll, S., Camgöz, N. C., Hadfield, S., and Bowden, R. (2020). Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *Int. J. Comput. Vis.*, 128(4):891–908.

Timmermans, N. (2005). *The Status of Sign Language in Europe (Integration of People with Disabilities*. Council of Europe Publications.

Vít, J., Hanzlíček, Z., and Matoušek, J. (2018). On the analysis of training data for wavenet-based speech synthesis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5684–5688.

Wolfe, R., McDonald, J. C., Hanke, T., Ebling, S., Van Landuyt, D., Picron, F., Krausneker, V., Efthimiou, E., Fotinea, E., and Braffort, A. (2022). Sign language avatars: A question of representation. *Information*, 13, April.

Zelinka, J. and Kanis, J. (2020). Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March.