Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC 2022), pages 144–147
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Facilitating the Spread of New Sign Language Technologies across Europe

Hope E. Morgan[1], Onno Crasborn[1] Maria Kopf[2] Marc Schulder[2] Thomas Hanke[2]
[1] Centre for Language Studies, Radboud University;
[2] Institute of German Sign Language and Communication of the Deaf, University of Hamburg

[1] {hope.morgan, onno.crasborn}@ru.nl
[2] {maria.kopf,marc.schulder,thomas.hanke}@uni-hamburg.de

Abstract

For developing sign language technologies like automatic translation, huge amounts of training data are required. Even the larger corpora available for some sign languages are tiny compared to the amounts of data used for corresponding spoken language technologies. The overarching goal of the European project EASIER is to develop a framework for bidirectional automatic translation between sign and spoken languages and between sign languages. One part of this multi-dimensional project is that it will pool available language resources from European sign languages into a larger dataset to address the data scarcity problem. This approach promises to open the floor for lower-resourced sign languages in Europe. This article focusses on efforts in the EASIER project to allow for new languages to make use of such technologies in the future. What are the characteristics of sign language resources needed to train recognition, translation, and synthesis algorithms, and how can other countries including those without any sign resources follow along with these developments? The efforts undertaken in EASIER include creating workflow documents and organizing training sessions in online workshops. They reflect the current state of the art, and will likely need to be updated in the coming decade.

Keywords: sign language resources, sign language corpora, sign language lexicons, training

1. Introduction

Various inputs are needed to develop functional workflows for language technologies. These technologies are varied, including visual recognition of signed utterances, cross-lingual transfer and naturalistic avatars. Annotated corpora associated to rich lexical databases have an important role to play. In the case of sign languages, these corpora have to be annotated manually, as there is no way of doing so automatically yet. Unlike video data with interpreters or videos from signers of various skill levels on social media (Bragg et al. 2019, De Meulder 2021, Leeson 2021), high quality linguistic corpora more often include the natural language use of fluent deaf signers and conversational rather than monologic discourse. Even more importantly, they contain detailed time-aligned linguistic data rather than merely translations. Yet, there are many well-known challenges with using these corpora, including the fact that they are rather small compared to what language engineers for spoken languages work with, and that their content is very diverse, leading to low type-token ratios. This leads to challenges for many language technologies that rely on significant quantities of training data. The problem we focus on here is that for many European countries there are still very few annotated corpora at all.

Two current projects, EASIER[1] and SignON[2] (each running from 2021-2023), both endeavor to advance the automatic translation of sign languages. These two projects have some overlapping and some complementary goals. One of the notable aspects of the EASIER project is a concerted effort to focus on language documentation datasets; specifically, how to integrate them into the translation workflow and how to make sure that datasets from under-resourced languages are not overlooked presently and in the future. In this paper, we describe the steps that EASIER has taken and will take to do this.

Even within EASIER's core sign languages, Sign Language of the Netherlands (NGT), German Sign Language (DGS), British Sign Language (BSL), French Sign Language (LSF), Greek Sign Language (GSL), Italian Sign Language (LIS), and Swiss German Sign Language (DSGS), there is substantial variation with respect to the size and nature of language resources available. These languages were pragmatically selected because of expertise in the languages or use of the datasets among the project partners. Other European sign languages for which sizeable corpora and lexicons are available include Swedish Sign Language (SSL), Finnish Sign Language (FinSL), and Polish Sign Language (PJM).

Plans to address the inclusion of large datasets, as well as partial or new datasets from various languages are addressed below. EASIER will direct special attention on how to support linguists and deaf communities in countries with partial datasets or new projects to create sign resources that are commensurate with emerging European standards. The following sections sketch how we aim to achieve this.

2. Overview of Datasets for Sign Languages in Europe

At the onset of the EASIER project, it was recognized that preparing datasets in other languages would be important, not only for potential benefit to the current project (as data inputs), but well into the future. This desire to include more sign languages also dovetails with the ethical consideration to not leave out smaller or less-resourced languages in Europe from participating in technological advances.

However, at the beginning of the project, there was no comprehensive or current survey of these datasets. Therefore, the first step was to gather information about all the known sign language resources in the EU that would meet the needs of the EASIER pipeline. This was accomplished in the report *Overview of the Datasets for the Sign Languages of Europe* (Kopf, Schulder, and Hanke 2021) which identifies and describes 26 corpora and 41 lexical resources covering 24 sign languages.

[1] https://www.project-easier.eu

[2] https://signon-project.eu

By clarifying the existing resources of these languages, it will be possible to build a bridge for them to participate in at least some parts of a machine translation pipeline, giving these languages a head start when it comes to further developing or integrating resources to ultimately enable full two-way translation.

One of the findings in the Kopf et al. (2021) report was that high-quality training data for language technologies does not yet exist for the majority of European SLs. Fragmented and small datasets can be found for approximately half of the European SLs; for the rest no suitable resources could be identified.

3. Harmonizing Existing Datasets

Having identified quality datasets, the next challenge is to make sure their contents are machine-readable. Over the decades, as new language corpus projects were implemented, they borrowed some methods and annotation conventions from previous documentation projects starting with Johnston's seminal work on Auslan (Johnston 2010), but each team also developed their own conventions and notations along the way. The EASIER project recognized that each of these idiosyncratic systems would need to be translated into a common interchange format in order to be usable for language technology pipelines.

In order to understand exactly how the datasets differed, the report *Specification for the Harmonization of Sign Language Annotations* (Kopf et al. 2022) analyzed each available set of annotation conventions and the associated annotations of the available corpora for over 20 aspects, including segmentation, compounds, repetition, name signs, directional verbs, etc. This report also summarizes the notation of non-manuals and compares handshape coding across corpora.

With a much clearer picture of how the corpus resources of European sign languages both align and differ in their notation, the report then proposes a basic single unified interchange format that would be able to encode the information relevant to the EASIER translation pipeline. Because this format must be easily and unambiguously parsable by software, we propose using a JSON container structure to encode signs and other linguistic units (buoys, fingerspelling, etc.).

This interchange format will continue to develop as converters for individual corpora are written. The initial effort and most work by project partners within EASIER will be given to converting corpus data from the six core project sign languages. However, the EASIER project would also like to be able to incorporate data from other sign languages. This would allow the inclusion of more languages in the translation system as well as providing additional training data. Even outside of the EASIER system, use of the interchange format could support and speed up the integration of datasets into technology pipelines and the use of multiple datasets in quantitative linguistic studies.

With the detailed picture of relevant sign language resources in Europe and the basic interchange format established, the next issue is how to facilitate the entry of this data into the EASIER pipeline for resource managers. These managers include language documentation teams, institutions with national corpora, and possibly individual researchers. There are three broad audiences among them: (1) those that already have relatively large-scale resources that are richly coded, (2) those with partial language resources (e.g., a good online dictionary, but no corpus), and (3) those who have just recently or will soon start language documentation projects. For each of these audiences, it should be determined what they need to know to be able to integrate their data with the EASIER pipeline. There are only a few examples of the first type of audience that are not already in the EASIER or SignON project. Among them are the datasets for SSL, FinSL, and PJM mentioned above. Given that expertise was developed in these countries to create large-scale annotated corpora, significant capacity-building has already taken place. This makes it likely that resource managers for these languages will be able to use our published documentation to develop their own converters for the interchange format with minimal input from the EASIER project partners. However, the other two audiences may need further support. The EASIER project therefore designed a specific work package to reach out to these groups, described next.

4. Extending to Other Sign Languages

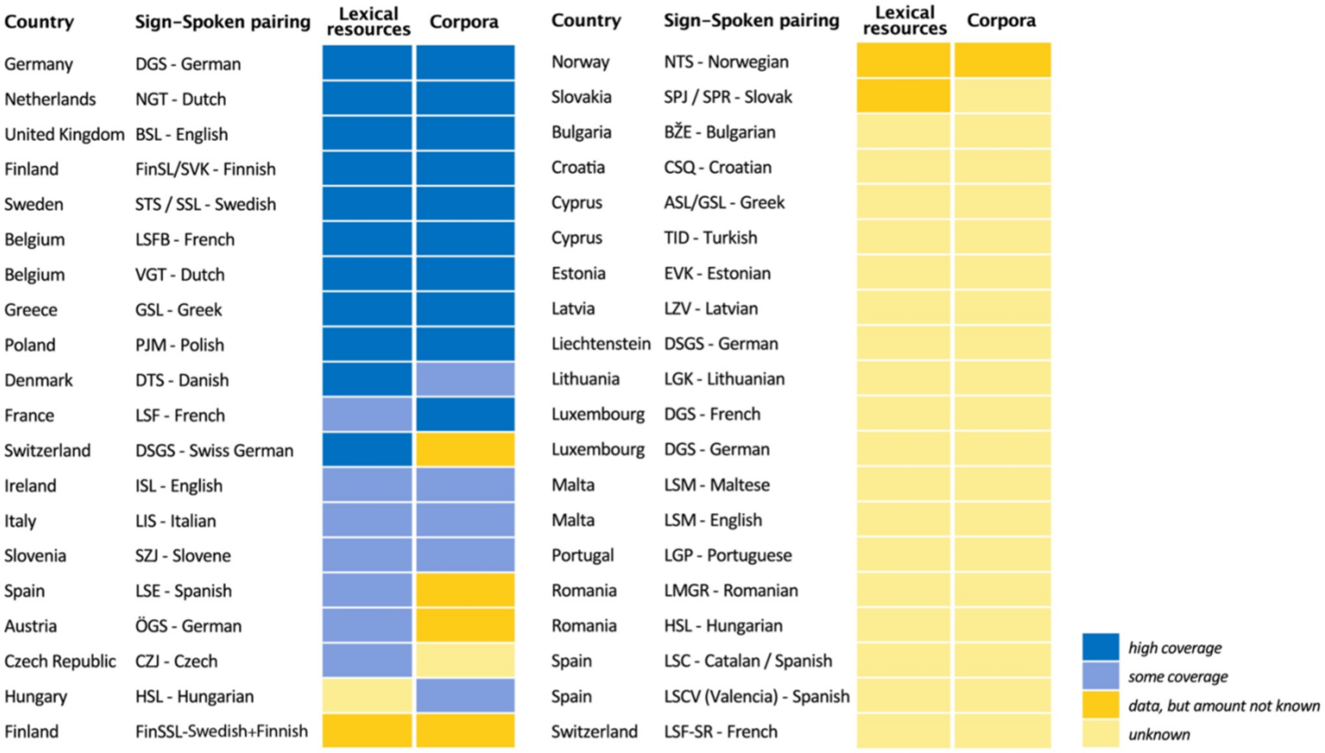
In this section, we describe the steps to extend the fruits of the EASIER project to reach more sign languages. This is a long-term endeavor that will not be realized within the short timeframe of this project, but we hope will prepare sign language resources to be ready for the next steps in machine translation in the future.

4.1 Defining the "Minimal Contents" for a New Language Dataset

For the two audiences who do not already have relatively rich corpora and/or lexical resources – that is, those with partial language resources and those who have just recently or will soon start language documentation projects – it is important to provide guidance on what it would entail to create, modify, or update resources to be ready for inclusion into the machine translation pipeline based on what we currently know. One important question to address is how large datasets should be in order to lead to translations that match the quality of those for the seven project languages.

This question remains difficult to answer in terms of exact quantities, but an indication of the size can promote resource development throughout Europe, in the sense that grant applications and lobbying efforts would have something they can refer to, and new documentation projects can work with tangible benchmarks in the near term, even if these continue to evolve in the future.

Therefore, a report is planned to provide an overview of what would be minimally necessary based on current standards and best practices: what are the ranges for size in terms of hours of annotated and non-annotated interaction, and associated lexical resources? This report will thus provide recommendations for both the creation and coding of *corpora* (i.e., linguistic, technical, and ethical criteria) and *lexical resources* (e.g., software, quantity, ID-glossing, phonological coding, etc.). The report is currently in production and will be published on the EASIER website in 2022.



Country	Sign-Spoken pairing	Lexical resources	Corpora
Germany	DGS - German	🟦	🟦
Netherlands	NGT - Dutch	🟦	🟦
United Kingdom	BSL - English	🟦	🟦
Finland	FinSL/SVK - Finnish	🟦	🟦
Sweden	STS / SSL - Swedish	🟦	🟦
Belgium	LSFB - French	🟦	🟦
Belgium	VGT - Dutch	🟦	🟦
Greece	GSL - Greek	🟦	🟦
Poland	PJM - Polish	🟦	🟦
Denmark	DTS - Danish	🟦	▨
France	LSF - French	▨	🟦
Switzerland	DSGS - Swiss German	🟦	🟨
Ireland	ISL - English	▨	▨
Italy	LIS - Italian	▨	▨
Slovenia	SZJ - Slovene	▨	▨
Spain	LSE - Spanish	▨	🟨
Austria	ÖGS - German	▨	🟨
Czech Republic	CZJ - Czech	▨	□
Hungary	HSL - Hungarian	□	▨
Finland	FinSSL-Swedish+Finnish	🟨	🟨

Country	Sign-Spoken pairing	Lexical resources	Corpora
Norway	NTS - Norwegian	🟨	🟨
Slovakia	SPJ / SPR - Slovak	🟨	□
Bulgaria	BŽE - Bulgarian	□	□
Croatia	CSQ - Croatian	□	□
Cyprus	ASL/GSL - Greek	□	□
Cyprus	TID - Turkish	□	□
Estonia	EVK - Estonian	□	□
Latvia	LZV - Latvian	□	□
Liechtenstein	DSGS - German	□	□
Lithuania	LGK - Lithuanian	□	□
Luxembourg	DGS - French	□	□
Luxembourg	DGS - German	□	□
Malta	LSM - Maltese	□	□
Malta	LSM - English	□	□
Portugal	LGP - Portuguese	□	□
Romania	LMGR - Romanian	□	□
Romania	HSL - Hungarian	□	□
Spain	LSC - Catalan / Spanish	□	□
Spain	LSCV (Valencia) - Spanish	□	□
Switzerland	LSF-SR - French	□	□

🟦 *high coverage*
▨ *some coverage*
🟨 *data, but amount not known*
□ *unknown*

Figure 1: Chart of European sign language resources shown in sign-spoken language pairs; data is based on the selection criteria and findings in Kopf et al. (2021).

4.2 Locating New Language Documentation Projects

The immediate scope of EASIER is all sign languages in the EU (plus the UK, which at the time of submitting the project proposal was still an EU member). Ultimately, these technologies will become available as open source tools for any sign language. Those countries who stand to benefit most are the few who already have existing datasets while countries with fewer resources and who have not invested in sign language documentation projects are at a disadvantage.

In order to determine which countries and sign languages may need specific support, the survey report by Kopf et al (2021) described in §2 was used to create a list of all sign languages in Europe, categorizing the availability of lexical resources and corpora that meet the criteria set up for their possible integration into the EASIER pipeline in terms of quantity and – roughly – quality. This is illustrated in Figure 1, showing four levels of resources: high coverage resources (dark blue), resources with some coverage (light blue), resources that exist but the extent is not known (dark yellow), and no resources found (light yellow).

What we can observe in this overview is that most languages with high or medium coverage are already participating in machine translation projects, in either SignON or EASIER, while most with partial resources and whose extent is not known are not involved in these projects. These 'partial resource' languages will be able to take advantage of the relevant portions of the definition of minimal contents for datasets in §4.1 and the workflow documents in §4.3.

In addition, there are a striking number of European sign languages with no language resources at all. Therefore, one current task in EASIER is to discover whether any new language documentation projects are underway or planned in the future for those languages colored yellow in the chart. This involves a two-pronged approach, reaching out to (i) researchers in those countries to find out about possible projects within academic institutions and (ii) contacting representative members of the European Union of the Deaf to connect with potential projects led by deaf community and other social institutions outside of academia. This also involves an online media effort to request help from the public on identifying projects. To the extent that this uncovers sign language resources not currently in the Kopf et al. (2021) report, we will make updates in a new version. Any new or in-progress documentation projects can take advantage of our report on minimal contents for language datasets, the workflow documents, and training sessions for new documentation, discussed next.

4.3 Workflow Documents for New and Existing Datasets

The LREC workshop series *Representation and Processing of Sign Languages* along with a series of other European workshops (e.g., Crasborn 2010, Cormier et al. 2016) has resulted in a substantial body of knowledge regarding sign language resource creation. Written output of those events has been collected in the 'sign-lang@LREC Anthology'.[3] The many hundreds of papers there constitute a valuable source of information for universities and deaf associations starting the creation of new sign language resources. However, this collection is bewilderingly diverse, and it can be difficult for language resource managers to extract key information. For that reason, another aim in EASIER is to compile the most essential information on how to

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

create valuable SL resources into a set of workflow documents that can serve as a starting guide. These documents will cover linguistic questions (e. g. granularity of annotation) as well as technological questions (e.g., studio setup).

4.4 Training Sessions for New Documentation

The workflow documents mentioned above will be accompanied by online training sessions, where linguistic and technical aspects, tools and open issues can be discussed and researchers can provide support to each other.

One of the workshops will specifically focus on how to deal with the translation of neologisms. As the pipeline developed in EASIER will include a post-editing environment for humans it will be possible to provide high-quality translations that even take into account the use of new terms in either the spoken or the signed language. Sign language interpreters come across neologisms and challenging vocabulary on a day-to-day basis, and the aim is to bring them together, discuss existing solutions across European SL productions and see how they can enrich the machine translation output.

4.5 Infrastructure to Automatically Analyze Other Datasets

Lastly, a hurdle for the creation of automatic analyses may be a lack of technological infrastructures within smaller projects. Therefore, EASIER will support data creators with video processing services in the form of an infrastructure running on high-performance clusters. In this way, less-resourced research projects can use state-of-the-art 2D pose estimation techniques which then again can be used to feed sign language translation pipelines and other sign language technologies, e.g., classifiers for the verification of manual annotation.

5. Conclusion

Language technologies for signed languages are in an emerging state, where initial application areas are explored and served with the latest of technical advances in computer vision, machine translation, and animation. These developments are foreseen to increase in speed over the coming decade. It is our responsibility as developers to look beyond the ‘test languages’ that we currently can work with, and that have benefited from major investments in language resources over the last ten to twenty years. The present efforts within the EASIER project to increase the scope to all of Europe’s sign languages that we described in this paper will hopefully contribute to best practices in this field when it comes to extending the use of technologies to less-resourced languages. Although the focus of EASIER lies within Europe, modern practices in sharing both software and research data will hopefully further broaden its impact throughout the world.

6. Acknowledgements

This work is supported in part by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Union’s Horizon 2020 research and innovation programme, grant agreement n° 101016982.

This work has in part 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

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C. and Ringel Morris, M. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st international ACM SIGACCESS conference on computers and accessibility*, pp. 16–31. DOI: 10.1145/3308561.3353774

Cormier, K., Crasborn, O. and Bank, R. (2016). Digging into Signs: Emerging Annotation Standards for Sign Language Corpora. In *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining.* Paris: European Language Resources Association (ELRA), pp. 35–40. https://www.sign-lang.uni-hamburg.de/lrec/pub/16015.html

Crasborn, O. (2010). The Sign Linguistics Corpora Network: Towards standards for signed language resources. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias (Eds.). Valletta, Malta: European Language Resources Association (ELRA), pp. 457–460. https://aclanthology.org/L10-1009

De Coster, M., Shterionov, D., Van Herreweghe, M. and Dambre, J. (2022). Machine Translation from Signed to Spoken Languages: State of the Art and Challenges. arXiv preprint. DOI: 10.48550/arXiv.2202.03086

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)*, pp. 12–22. https://aclanthology.org/2021.mtsummit-at4ssl.2

Jantunen, T., Rousi, R., Rainò, P., Turunen, M., Moeen Valipoor, M. and García, N. (2021). Is There Any Hope for Developing Automated Translation Technology for Sign Languages? In *Multilingual Facilitation*, pp. 61–73. DOI: 10.31885/9789515150257

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. In *International Journal of Corpus Linguistics*, 15(1), pp. 106–131. DOI: 10.1075/ijcl.15.1.05joh

Kopf, M., Schulder, M. and Hanke, T. (2021). D6.1 Overview of Datasets for the Sign Languages of Europe. Technical report, Universität Hamburg. DOI: 10.25592/uhhfdm.9560

Kopf, M., Schulder, M., Hanke, T. and Bigeard, S. (2022). D6.2 Specification for the Harmonization of Sign Language Annotations. Technical report, Universität Hamburg. DOI: 10.25592/uhhfdm.9841

Leeson, L. (2021). Deliverable 1.1: Case Studies and Evidence Analysis. Report, Trinity College Dublin. https://signon-project.eu/publications/public-deliverables/