Challenges with Sign Language Datasets for Sign Language Recognition and Translation

Mirella De Sisto*, Vincent Vandeghinste†, Santiago Egea Gómez‡
Mathieu De Coster§, Dimitar Shterionov*, Horacio Saggion‡

*Tilburg University, the Netherlands
m.desisto, d.shterionov@tilburguniversity.edu

†Dutch Language Institute, Leiden, the Netherlands
Centre for Computational Linguistics, Leuven.AI, KU Leuven, Belgium
vincent.vandeghinste@ivdnt.org

‡Universitat Pompeu Fabra, Barcelona, Spain
santiago.egea, horacio.saggion@upf.edu

§IDLab-AIRO – Ghent University – imec, Ghent, Belgium
mathieu.decoster@ugent.be

Abstract

Sign Languages (SLs) are the primary means of communication for at least half a million people in Europe alone. However, the development of SL recognition and translation tools is slowed down by a series of obstacles concerning resource scarcity and standardization issues in the available data. The former challenge relates to the volume of data available for machine learning as well as the time required to collect and process new data. The latter obstacle is linked to the variety of the data, i.e., annotation formats are not unified and vary amongst different resources. The available data formats are often not suitable for machine learning, obstructing the provision of automatic tools based on neural models. In the present paper, we give an overview of these challenges by comparing various SL corpora and SL machine learning datasets. Furthermore, we propose a framework to address the lack of standardization at format level, unify the available resources and facilitate SL research for different languages. Our framework takes ELAN files as inputs and returns textual and visual data ready to train SL recognition and translation models. We present a proof of concept, training neural translation models on the data produced by the proposed framework.

Keywords: sign language translation, sign language recognition, sign language corpora, unified data format, machine learning

1. Introduction

More than 70 million people worldwide are deaf.[1] In Europe alone, for approximately half a million of deaf and hard of hearing (DHH) people, Sign Languages (SLs) are the main or preferred means of communication (Pasikowska-Schnass, 2018). SLs are natural languages which exploit the visual-gestural channel. Ethnologue lists 150 sign languages (Eberhard et al., 2021); however, this list is far from exhaustive, and many SLs remain undocumented. Even in the case of recognised and described SLs, the limited amount of data available relegates them into the category of low resource languages.

In recent years, there has been a growing interest in applying natural language processing techniques to SLs (Camgoz et al., 2018; Camgoz et al., 2020; De Coster et al., 2021; Zhou et al., 2021; Yin et al., 2021; Moryossef et al., 2021) and, in particular, in developing tools for automatic translation between SLs and spoken languages. Namely, the progress in deep learning has drawn an inspiring context to develop tools to automatically translate SLs. The two main tasks involved to automatically translate signs to spoken language words using neural networks are: (1) SL recognition, in which each signer interaction should be recognized and matched with the correct sign meaning; and (2) SL translation, which focuses on generating translations into spoken or other sign languages. Unfortunately, there are several issues that hamper the progress in these research areas.

In 2021, two large-scale European projects started working on developing communication services for sign and spoken languages, SignON[2] (Shterionov et al., 2021; Saggion et al., 2021) and EASIER[3]. The research efforts by these and other researchers world-wide have pushed the state-of-the-art to a new level, while at the same time they have stumbled upon various challenges.

A severe obstacle to the advances of SL recognition and translation is data sparseness. As stated by Bragg et al. (2019), SL corpora are generally extremely smaller than speech recognition corpora. SL corpora usually

[1] Estimation from the World Federation of the Deaf (WFD) https://wfdeaf.org/

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

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

contain less than 100,000 articulated signs in contrast to millions or even billions of words in a typical speech corpus. In the context of neural machine translation (NMT), typically one would require a couple of millions of parallel sentences to achieve decent quality.[4] However, the German SL (DGS) corpus (Prillwitz et al., 2008) — one of the largest annotated SL corpora — encompasses 50 hours of publicly available data (video material and transcriptions) which correspond to around 60 thousand parallel sentences.

In addition to that, annotation is mostly a manual process, which makes it very time-consuming. Consequently, the completion of the annotation task is often delayed in comparison to the other tasks within a project; even sometimes it is still in progress at the time of data release. Accordingly, the amount of data collected does not always correspond to the amount of data that is actually annotated. For instance, the Corpus VGT (Van Herreweghe et al., 2015) contains 140 hours of conversations in VGT (Flemish SL); however, only 10% of it is currently annotated. Similarly, only 25% of the videos in NGT (SL of The Netherlands) of the Corpus NGT (Crasborn et al., 2020) have annotations.

For other datasets, not dubbed as *sign language corpora*, the situation is worse, as they often consist of television broadcasts augmented with an SL interpreter. The quality of the SL in such cases is often debatable, depending on the concrete setting in which the interpretation was made. Most often, the interpreter is a hearing, non-native signer, trying to convey the meaning of the spoken message, as good and as quickly as possible. Another limit of similarly-obtained data is that the SL is the target language, hence, the SL stream will always be somehow affected by the source spoken language from which it was interpreted. This results in *translationese* SL.

Using non-native and/or interpreted data to train translation systems that should be able to recognise, and generate, spontaneous authentic/native language will lead to less natural and very likely inaccurate translations (or transcriptions).[5]

Besides the aforementioned problems, things do not get easy even once SL corpora with authentic signers as informants and SL as the source language, have been identified. Projects on SL recognition and translation still face a number of challenges, such as: (1) the difficulty in acquiring data as downloadable datasets, (2) the lack of a common annotation format, and (3) the limited usability of the available data formats for machine learning.

[4]One of the most commonly used corpora for MT, Europarl (Koehn, 2005), contains approximately 2 million parallel sentences for the high-resource language pairs and around 500 thousands for the low-resource ones. On the extreme side, consider the work of Hassan et al. (2018) which claims achieving human parity and presents a model trained on more than 25 million parallel sentences.

[5]See also (De Meulder, 2021).

In the present paper we discuss these challenges and propose a methodology to facilitate the interoperability across corpora and their usability in automatic SL recognition and translation. We focus on processing ELAN (Sloetjes and Wittenburg, 2008) files from various SL corpora into a unified data format which is suited for machine translation.

We will make the code and the data which do not have copyright nor GDPR limitations available.

2. Issues when Acquiring Sign Language Datasets

2.1. Access to Sign Language Corpora

Traditionally, SL corpus collection was targeted towards SL linguists and not towards the language technology community. EASIER's deliverable 6.1 (Kopf et al., 2021) provides a very useful overview of existing datasets for SL in Europe, including a short description of how the data were collected and a table with relevant metadata such as licence, project URL, etc.

That the datasets are not geared towards the language technology community is apparent in the fact that these corpora are often accessible only through a web interface. Whereas individual files (video and corresponding annotations) can be downloaded, the corpus as a whole, e.g., in a single file or archive, is often not available. In some cases this limitation could be resolved by personally contacting the website owners and asking whether such a single file download could be foreseen. Thankfully, large parts of these corpora are unrestricted. However, accessing them in pieces hinders their usability for the machine learning community.

The ECHO (European Cultural Heritage Online) corpus is a freely available multilingual corpus including video materials and annotations in three SLs (Nonhebel et al., 2004). In order to download it one needs to use the dataset persistent identifier[6] and download separate files one by one.

The British SL (BSL) Corpus (Schembri et al., 2011) is a collection of videos of people using BSL, together with background information about the signers and written descriptions of the signing in ELAN. In order to download the unrestricted data in the BSL corpus, we first had to contact the website owners as the provided link was outdated. Then, we could browse the archive and download 454 separate files.

The Corpus VGT (Van Herreweghe et al., 2015) is a collection of 140 hours of videos in Flemish SL. 120 deaf people contributed to the Corpus VGT as informants. Age, region and gender were taken into account when selecting the informants. The corpus was only available through its online search interface.[7] While it probably would have been possible to scrape the website, after contacting the authors, the complete corpus

[6]https://hdl.handle.net/1839/00-0000-0000-0001-4892-C

[7]https://www.corpusvgt.be/

has been made available and will soon be downloadable for research purposes through the CLARIN infrastructure.[8]

The Corpus Nederlandse Gebarentaal (NGT) (Crasborn et al., 2020) was available for download at the Language Archive,[9] but again not as a single file. In collaboration with the corpus authors and their system administrators, the corpus will also be made available for download as a single file through the CLARIN infrastructure.[10]

iSignos (Cabeza and García-Miguel, 2018) was created within the RADIS (Actantial Relations and Signed Discourse) project of the University of Vigo with the purpose of facilitating the investigation of grammatical processes used in Spanish SL (LSE) (Pérez et al., 2018). The annotated corpus contains 20 recordings of eight signers mostly telling a story or giving elicited examples.[11] The source LSE videos have glosses and a Spanish translation. The corpus can be consulted on the project website[12] but cannot be downloaded.

A parallel Spanish-LSE corpus was compiled by (Porta, 2014) using material from a psycholinguistic study (Rodríguez Ortiz, 2005). The corpus contains Spanish texts of various topics which were translated by an interpreter into LSE. The resulting LSE videos were transcribed into glosses and then re-translated into Spanish by a CODA (Child Of Deaf Adults) interpreter; this was done in order to verify the accuracy of the LSE translations (Porta, 2014, p. 37). The data provided by Porta (2014) contain the LSE glosses and both source and translated Spanish texts. The corpus is available in the appendix of Porta's thesis (Porta, 2014).

The Fundación CNSE (State Conferedation of Deaf)[13] has produced various material in LSE, such as an online driving license manual platform.[14] The (probably interpreted) signed videos are accessible online but the source texts do not appear to be available.

A corpus of Catalan SL (LSC) was created for a thesis project at the University Pompeu Fabra in Barcelona (Sanabre, 2012). Weather forecast texts in Catalan were extracted from the website of the Meteorological Service of Catalonia,[15] and then translated into LSC by two native signers, two LSC professors from the Deaf Cultural Recreation Centre of Barcelona and the Barcelona Deaf House. The recordings have been annotated with iLex[16] (Hanke, 2002) in terms of manual articulator glosses, non-manual articulator glosses, morphemes associated with movement, place of articulation and orientation, and Catalan text (Sanabre, 2012, p. 109). Unfortunately, the corpus and its annotations are not publicly available.

The Signs of Ireland corpus (Leeson and Nolan, 2008) is one of the largest, most richly annotated SL corpora with Irish SL (ISL) in natural use by male and female adult signers across a range of ages. It contains annotated signed data of ISL, but is not (yet) available for online search nor for download. Through personal contacts with the authors we were able to get hold of this corpus and its annotations.

It is clear that the authors of the above mentioned corpora have not prepared the corpora for ease-of-use by the machine learning community.

2.2. Machine Learning Datasets

Some SL datasets are not geared towards SL linguistics but are explicitly aimed at SL machine translation or SL recognition. A downside of these datasets is that they often consist of *non-authentic* SL: TV broadcasts that have been augmented with a SL interpretation, possibly with additional parallel information streams, such as OpenPose (Cao et al., 2019) coordinates providing information on body part positions in the source videos and a written representation in the form of autocues or closed caption subtitles. In these cases we cannot assume that the signers belong to the respective SL community of the SL they sign, as most often they are hearing SL interpreters.

The Content4All dataset (Camgöz et al., 2021) [17] (not present in (Kopf et al., 2021)) contains such TV broadcast material with signing and its associated subtitles. A login and password had to be requested and at a first attempt there were some broken links, but after contacting the webmaster, this was quickly fixed.

The BBC-Oxford British Sign Language Dataset (BOBSL) (Albanie et al., 2021) is not usable as the licence states that each researcher has to request access individually, which is rather unpractical, and that any cooperation with commercial partners is not allowed. For consortia like SignON or EASIER or other groups involving industry partners, the use of this dataset is prohibited.

The RWTH-PHOENIX-Weather 2014T dataset (Camgoz et al., 2018) contains TV broadcast material, more specifically weather broadcasts, interpreted from spoken German to DGS by hearing interpreters. It is available for download on the website of the research group that first published it.[18] It is designed for research into machine translation for SLs and is available as a single archive file, with the file format (CSV) directly usable for training translation models. It could serve as an example of openness and ease of use, as it poses no issue

[8] http://hdl.handle.net/10032/tm-a2-u4
[9] https://archive.mpi.nl/tla/
[10] http://hdl.handle.net/10032/tm-a2-u5
[11] http://isignos.uvigo.es/information
[12] http://isignos.uvigo.es
[13] https://www.fundacioncnse.org
[14] https://www.fundacioncnse.org/dgt/
[15] http://www.meteo.cat
[16] http://www.sign-lang.uni-hamburg.de/ilex/
[17] https://www.cvssp.org/data/c4a-news-corpus/
[18] https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/

whatsoever in its usage in a machine learning context.
The How2Sign Dataset (Duarte et al., 2021) contains 83 hours of instructional videos from How2[19] translated from English into American SL (ASL). The videos are recorded from multiple viewpoints and the annotations contain, besides glosses, OpenPose and 3D pose estimation. The parallel corpus is designed for research and is downloadable from the website,[20] however, currently only the video modalities and the English texts can be downloaded.

3. Different Annotation Formats and Content

3.1. Different Annotation Formats

Given the multimodal characteristics of SL data, most of the available corpora are annotated using ELAN[21] (Sloetjes and Wittenburg, 2008), an annotation tool that supports multiple layers of annotations, called *tiers*, which are synchronized with the audio or video timeline. The fact that sign language corpora are stored using the ELAN format is of great aid for research. Furthermore, the fact that ELAN files are in fact in XML format — containing annotations and timestamps — makes it suitable to use any XML processing library. While using ELAN is indeed helpful in sharing data, challenges arise due to the lack of standardised annotation format.
For instance, the Signs of Ireland corpus contains tiers for describing the activity of various non-manual articulators (i.e., eyebrows, eye aperture, eyegaze, etc.). The ECHO dataset for BSL, besides non-manual articulator features, also includes direction and spatial location of the two hands (Nonhebel et al., 2004). The corpus VGT has annotations for some non-manual articulators, i.e., head movement and eye blinking; more non-manual articulators are mentioned in the guidelines (Van Herreweghe et al., 2013, pp. 9-11), however, they do not appear to be implemented in some annotation files. The BSL and NGT corpora, instead, mainly focus on manual articulators (Cormier et al., 2017; Crasborn et al., 2020). The only non-manual articulator tiers in the corpus NGT have information about head shakes and mouthing (Crasborn et al., 2020, p. 41).
In addition, information might be organised differently across tiers. For instance, the NGT, VGT and BSL corpora, and the ECHO dataset for BSL have two different tiers for glosses, depending on whether the sign is produced by the right or the left hand; the Signs of Ireland corpus, instead, has a unique tier for *Lexical Gloss*.
Other corpora examined in this paper are not annotated using ELAN, which adds a large spectrum of variation concerning data format as well as its content.
The iSignos corpus (Cabeza and García-Miguel, 2018) uses a CSV format, which contains values for left hand and right hand glosses, a Spanish translation, and start and end timestamps in milliseconds of the annotated event.
In Porta's Spanish-LSE parallel corpus, each sentence is displayed in three versions, namely the Spanish source, the LSE translation, in the form of glosses, and the Spanish re-translation (Porta, 2014, p. 37).
The Content4All dataset uses a JSON format, which contains OpenPose, time and frame information, and the *annotation*, which consists of the Belgian Dutch or Swiss German sentences that were translated into VGT or DSGS (Swiss German SL). As the dataset is aimed at end-to-end machine translation, no further annotations, such as glosses, are included.
As mentioned above, The RWTH-PHOENIX-Weather 2014T dataset is in a format suitable for machine translation. Its CSV files have a field for annotation; this contains the transcription into DGS glosses of the original German spoken weather broadcasts. The glosses do not exhibit supplementary information such as pointing signs or classifiers,[22] however, this might be due to the specific and limited semantic field of the material.

3.2. Differences in Content Organisation and Glossing Conventions

When turning to content-motivated organisation of annotations, another layer of discrepancy which constitutes an issue for employing SL corpora data in automatic translation is the lack of common glossing conventions. Consequently, the type and format of information included in the glosses does not always coincide in the various corpora, and, in some cases might not be fully compatible.
The first major difference is constituted by the spoken language used for glosses: glosses are generally based on lexicalized representations borrowed from a language spoken in the country where the SL is used. Consequently, while the NGT and VGT corpora have Dutch-based glosses, glosses in the BSL and ISL corpora are based on English, those in the RWTH-PHOENIX-Weather 2014T dataset on German, and those in iSignos and Porta's parallel corpus on Spanish.
As could be inferred from the naming of the gloss tier in the Signs of Ireland corpus, i.e., Lexical Gloss, the glosses annotated in it only concern lexical units, or lemmas, which coincide with lexical entries of a dictionary, with no other supplementary semantic information. In the BSL dataset from the ECHO project, the gloss tiers for right and left hand (*Gloss RH* and *Gloss LH*) either indicate the lemma, e.g., "MOUSE", or a short periphrasis of the action taking place, e.g.,

[19] https://github.com/srvk/how2-dataset
[20] https://how2sign.github.io/
[21] https://archive.mpi.nl/tla/elan

[22] Classifiers in SL are used to describe properties, such as movement, location and shape, of an entity or an object. Constructions including classifiers are part of the productive lexicon, hence, are not-conventionalised nor predefined. Therefore, they cannot be encoded in a dictionary nor linked to a lexical entry. See also (Zwitserlood, 2012).

"mouse move around trapped lion".[23] Additional information about how many hands articulate the sign (i.e., 1h, 2h), pointing signs (i.e., IND) and many-meaning components (i.e., p-, which stands for *poly*) —usually involving classifiers constructions —are also included in the annotations (Nonhebel et al., 2004). The glossing annotations of the BSL, NGT, VGT, iSignos and Porta's Spanish-LSE parallel corpora include supplementary information besides the lemma; nevertheless, the conventions used in those corpora are not identical. For instance, the BSL and NGT corpora annotation guidelines follow similar conventions for annotating classifier constructions. Both guidelines distinguish four types of classifier predicates based on their type of movement, namely (1) MOVE, i.e., from one location to another, (2) PIVOT, i.e., change of position of a referent, (3) AT, i.e., localization of a referent, and (4) BE, i.e., none of the previous mentioned (Crasborn et al., 2020; Cormier et al., 2017). However, in BSL annotations, this classification is preceded by another distinction made according to function of depicting signs, namely if the classifier construction refers to (1) a whole entity, (2) part of an entity, (3) the handling of an object or (4) size and shape (Cormier et al., 2017, p.10). In iSignos annotation conventions, a classifier construction is introduced with an initial *cl.* followed by an abbreviation which defines its type, namely (1) entity, *e*, (2) manual interaction, *m*, (3) descriptive, *d*, and (4) corporal, *c*; meaning follows the abbreviations and is structured into three elements: the meaning related to the hand configuration, the predicate and the referent of the non-dominant hand (Pérez et al., 2018). Classifiers are annotated in the Spanish-LSE corpus by (Porta, 2014) based on a six type distinction: descriptive constructions (CLD), locative constructions (CLL), pronominal constructions (CLS), body-as-subject related actions (CLC), body-part related actions (CLCP) and instrumental constructions (Porta, 2014, p.38). Nevertheless, marking does not seem to have been applied uniformly and consistently (Porta, 2014, p.38). The corpus VGT guidelines do not refer to classifier constructions.

The five corpora use different strategies for glossing pointing signs to signer and interlocutor. For example, if we focus on the pointing sign to refer to the signer, BSL annotations use PT:PRO1SG (Cormier et al., 2017, p. 8), in which 'PT' indicates that it is a pointing sign, and 'PRO1SG' encodes the first person singular pronoun; in the Corpus NGT the annotation is IK 'I' (Crasborn et al., 2020, p. 30), and in the Corpus VGT the annotation is WG-1, in which 'WG' stands for *wijsgebaren*, 'pointing sign'. Finally, in iSignos this is represented as INDX.PRO:1sg (Pérez et al., 2018) and in Porta's Spanish-LSE parallel corpus as YO 'I'.

In addition to that, not all corpora have guidelines which are easily accessible; consequently, the conventions used during data annotation are not always fully transparent. This poses significant limits to determining the compatibility of annotations across corpora.

The corpora discussed in this section all have a tier, a key-value pair or an entry with text in a spoken language. However, it is important to stress the difference between the cases in which the text into a spoken language is the source from which SL was interpreted, and those in which this is a translation of the originally signed utterances.

The Signs of Ireland corpus, BSL, iSignos, VGT and NGT corpora and the BSL dataset of the ECHO corpus are part of the second group. Whereas most of these corpora have one translation tier, the Corpus NGT has both a *TranslationNarrow* and a *TranslationFree* tier, hence, distinguishing between a text as close as possible to the source language, and a text more adapted to the target language requirements.

As mentioned above, Porta's Spanish-LSE parallel corpus contains two types of transcriptions in Spanish: one that constitutes the source language, and one which is the re-translation from LSE.

In Content4All, the source Belgian Dutch or Swiss German text is in the key/value *annotation*. In the RWTH-PHOENIX-Weather dataset, the German source text is displayed in the entry *translation*.[24]

Even if at a first look the texts in spoken languages in the various datasets could appear similar, they deeply differ in relation to the signed stream and in terms of their function in the dataset.

These discrepancies in what information is annotated and how, and in the types of glossing used in the corpora, constitute a serious obstacle to the building of a multilingual corpus. In particular, the glossing styles and their granularity vary significantly; hence, they might be difficult to combine if not incompatible. A common standardized glossing format is needed in order to facilitate the interoperability between corpora.

Another challenge in terms of glossing is added when we consider that the data might consist of either monologues or dialogues. In the case of the latter, one or two extra gloss tiers will be included in the annotation. The NGT, VGT and BSL corpora contain dialogues, which means two gloss tiers for each informant, one for each hand, for a total of four gloss tiers in each file. The data of the Signs of Ireland corpus, iSignos, Porta's Spanish-LSE parallel corpus and of the ECHO corpus BSL dataset, instead, consist of monologues in which informants tell a story, one signer at a time; hence, they only have gloss tiers or values referring to one informant.

Table 1 summarizes the properties and differences of the available SL corpora and SL machine learning datasets that were outlined in this and the previous Section (2).

[23] Examples from the file BSL_CN_fab3.eaf, ANNOTATION_ID: a200 and a199

[24] This might seem counter-intuitive, but German is the source from which DGS utterances were interpreted.

	Publicly available	Format	Source Lang.	Signers per file	R & L hand glosses	Non-manual features	Available annotation guidelines	OpenPose
ECHO Corpus	separate files	ELAN	BSL	1	2	✓	✓	✗
BSL Corpus	separate files	ELAN	BSL	2	2	✗	✓	✗
Corpus VGT	separate files	ELAN	VGT	2	2	✓	✓	✗
Corpus NGT	separate files	ELAN	NGT	2	2	✓	✓	✗
iSignos	on website	CSV	LSE	1	2	✗	✓	✗
Porta's corpus	thesis	-	ES	1	1	✗	✓	✗
Signs of Ireland	private	ELAN	ISL	1	1	✓	✗	✗
Content4All	with account	JSON	NL/DE	1	n/a	n/a	n/a	✓
RWTH-PHOENIX	yes	CSV	DE	1	1	✗	✗	✗
How2Sign	not fully	CSV	EN	1	1	-	✗	✓
CNSE's corpus	only video	-	ES	-	-	-	-	-
LSC corpus	✗	iLex	CAT	1	2	✓	✓	✗
BOBSL corpus	restricted	-	EN	-	n/a	✗	n/a	✗

Table 1: Summary of the properties of the SL corpora and SL machine learning datasets

4. A Framework to Unify SL Corpora

In order to alleviate the standardisation issues at format level in SL corpora, we provide a framework to process raw ELAN files and generate data ready to be fed to machine learning algorithms. As ELAN is a widely used tool to annotate SL corpora, we focus on this particular format here; although the framework can be easily extended to other input formats.

As Figure 1 shows, the proposed output data are organized in a folder structure that allows direct access to annotation segments and media data contained in ELAN files. Additionally, parallel text files ready to train NMT systems (Item 2 in Figure 1) are generated by aligning and merging annotations previously stored in separate files (Item 1 in Figure 1). Text data are stored in raw files using UTF-8 encoding, and video frames are saved as JPG files easily identifiable by timestamps (Item 4 in Figure 1).

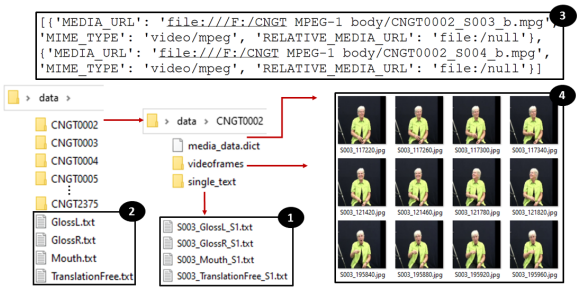


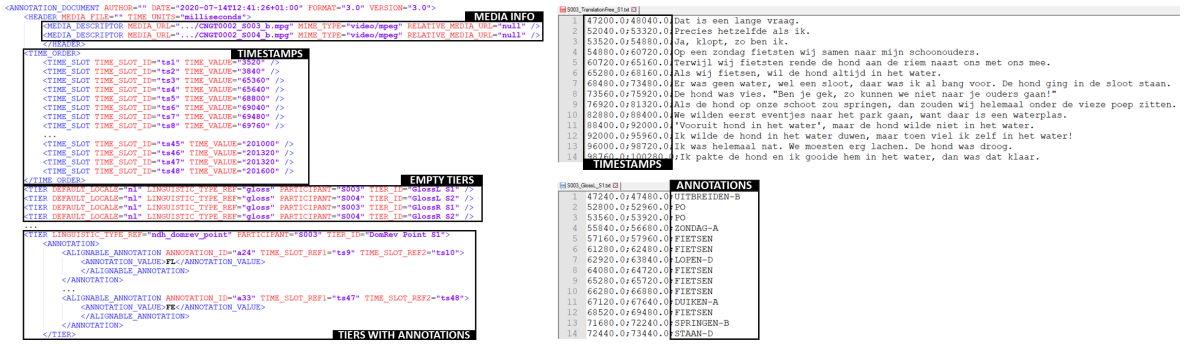
Figure 1: Folder structure output by our framework.

The proposed framework comprises the following steps:

(1) ELAN file parsing. It is very common to find that all corpus annotations are in different ELAN files contained in a top folder (this is the case for NGT, VGT, ECHO and other SL corpora). The proposed method reads all these files and extracts the target data from the different data containers (or tiers). As can be observed in Figure 2a, this step involves skipping irrelevant information and/or empty annotation containers, and recognising the different participants (if more than one). To fulfil the ML tasks at hand, we extracted the media information (paths or URLs to videos), all annotations and their timestamps. This information is stored in different directories creating a hierarchical folder structure for the sake of a proper data organization. Namely, the annotations along with their timestamps are stored in the *single_text* folders as CVS-like format as it is shown in 2b, whereas media data are stored in files (Item 3 in Figure 1) in the folders created for each ELAN file.

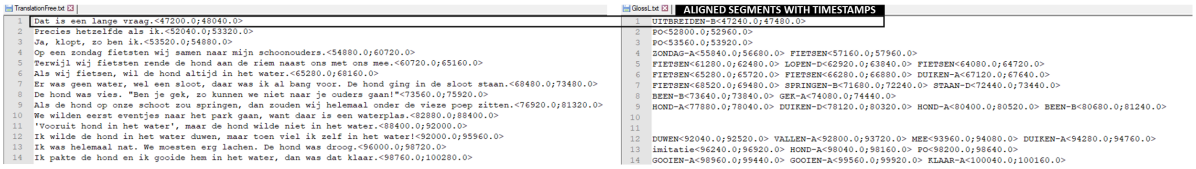
(2) Aligning and merging annotations. In order to train NMT models, we need to generate parallel text files with the source and target annotations (or data modalities). To this aim, the user must define one leading text modality and the required modalities to process. Then, the timestamps presented in the leading modality files will be used to align the required modalities, and create aligned sequence of text segments. The recommended choice is using the sign transcriptions or translations as the leading modality, since these annotations usually have the longest timeslots; and setting other annotations (such as right/left hand annotations and mouthing annotations) as the required modalities. This process is recursively applied to each subfolder producing one text file per modality (both the leading and required) which are stored in the top output folder as Figure 1 describes (Item 2). The ELAN files in which leading and required modalities were not annotated are skipped in this step. Finally, annotation timestamps are included in aligned text files using the special markers `<timestamp1;timestamp2>` to avoid information loss (Figure 2c).

(3) Video frames extraction. Finally, to train neural models with visual information, the video frames in which the participants are producing signs must be accessible. Obviously, this task can only be success-



(a) Input ELAN format

(b) Annotations with Timestamps created in Step 1



(c) Aligned Text created in Step 2

Figure 2: Data Input and Outputs in our Framework

fully performed when videos are available and correctly match participants. In this regard, we found difficulties to match videos and corresponding signers in some of the examined corpora. Thus, automatic video-participant matching is planned for future releases of this framework. In the cases in which videos are clearly identified with the corresponding participants, the files generated in (1) are used to check the timestamps in which signers are producing information, and those video frames are extracted, resized (to 224 by 224 pixels) and stored in a subfolder called *videoframes* (Figure 1, Item 4). As high resolution images will be needed in future research stages, we plan to leverage this step to meet this requirement.

Through the presented framework, we are able to process SL corpora in ELAN format to extract text and visual data to train neural networks. For the former, ELAN annotations are parsed (1), merged and aligned to produce parallel text files to experiment with NMT models (2). For the latter, video frames with participants signing are extracted and labelled with timestamps (3). The data outputted in (1) and (3) can be used to incorporate visual information in SL translation and recognition models.

Through the preliminary results presented in the next section we aim to assess the effectiveness of our framework.

5. Empirical Evaluation

To showcase the effectiveness of our framework, i.e. reducing the manual efforts to extract datasets ready to train NLP models with, we train and evaluate two neural models. Namely, we use the parallel data generated by our framework to train Transformer models (Vaswani et al., 2017) for gloss to text translation.

5.1. Experimental Settings

We process the NGT corpus using our framework selecting the *Free Translation* tier as leading modality; and *GlossL*, *GlossR* and *Mouth* as required modalities. As results, we obtained four text files containing 8344 aligned utterances for these data tiers. Afterwards, the resulting parallel segments were split in training, validation and test subsets using 20% and 10% of the data for validation and test respectively. In this experiment, we consider *GlossL*, *GlossR* and *Mouth* as inputs to the transformer and *Free translation* as output. We trained two NMT models, one including all inputs and one excluding mouthing.

The goal of our experiment is not to obtain the highest performance from the translation models, but rather to create a proof of concept with the data generated by our framework. Therefore, we used a simple transformer architecture without an extensive hyperparameter tuning. The encoder and decoder have the same architecture consisting of 3 transformer layers with 4 attention heads, and a hidden dimension of 1024; the embedding vector size used is 512. We employed Huggingface's[25] implementations for our model. To tokenize the input sequences, we trained a Sentence Piece model (Kudo and Richardson, 2018) on the training partitions with a joint vocabulary size of 5000 tokens.

Finally, we use a batch size of 64 sentences and train the model for 250 epochs using the Adam optimizer (Kingma and Ba, 2015) with 10^{-5} as learning rate; we apply beam search decoding with 5 beams during generation.

[25] https://huggingface.co

5.2. Results

To evaluate the translation quality, we use the BLEU metric (Papineni et al., 2002) as implemented in the SacreBLEU package (Post, 2018). We also analyze the loss curves for the training and validation partitions.

Figure 3 shows the loss curves for the model including mouthing data. We can observe that training and validation losses decrease between 1 and 50 training epochs, showing that the employed model is able to learn from the annotations generated by our framework. After this point, the curves follow the typical behaviour when a model is overfitted on training data: the training loss continues to drop, whereas the validation loss increases. Further hyperparameter tuning could be used to reduce overfitting, but this is out of scope for this paper. The loss curves for the model with mouthing exhibited a similar behaviour to the plotted in Figure 3.

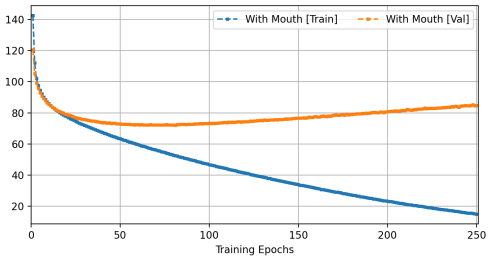


Figure 3: Train and validation losses for model with mouthing

Figure 4 plots the BLEU scores per training epoch for the two models. As can be observed, the models can learn from the different annotation modalities outputted by our framework: the model including mouthing improves consistently between 1 and 160 epochs reaching its highest values around 3.35 BLEU and converging. The model without mouthing follows a similar trend in the beginning of the training, but the curve for this model is generally less steep and the convergence is not as clear as for the previous model. Interestingly, we can note that including mouthing modality in the model boosted the performance metric. The BLEU scores on the test partition at the end of the training process for both models are: 3.04 BLEU for the model including mouthing and 2.57 for the model without mouthing.

While neither the transformer with mouthing nor the one without mouthing reach significant quality levels, it is important to note that these models can in fact learn (judging by the progressively increasing performance curves as well as the decreasing loss curves) from the data generated by the proposed framework. Furthermore, our framework generates parallel data, i.e. data aligned along all tiers, which is directly suitable for sequence to sequence tasks, as shown above with our NMT experiments.

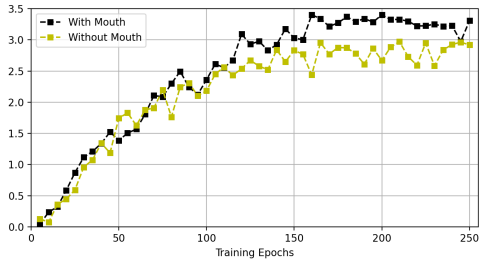


Figure 4: BLEU scores on the validation partition for the models with and without mouthing.

6. Conclusion and Future Work

By comparing the availability and properties of various SL corpora and SL datasets, we outlined the challenges which are currently limiting the advances of the fields of SL recognition and translation. We described the difficulty in acquiring data and the divergences amongst the various resources in terms of data and annotation formats. The type, granularity, and amount of information annotated in SL corpora and datasets varies extremely. In addition, often the same or similar information is encoded following different conventions. This fact poses limits to the compatibility of data from different datasets and to the creation of multilingual datasets. The only solution in this case is introducing standardized annotation and glossing conventions.

In order to approach the challenges concerning the lack of a common data format, we proposed a framework that adapts ELAN files into a unified format which is suitable for SL recognition and translation models. We employed the annotation data outputted by our framework to train an NMT model; and our preliminary results prove that neural models can indeed learn from data in the proposed format.

Future work is needed in order to expand the potential of these preliminary findings. Firstly, the proposed framework may be improved to deal with video-participant matching and higher resolution video frame extraction. Besides, our framework only processes ELAN format, and can be extended to other formats as iLex [26]. Regarding the experimental side, the data generated by our framework for different corpora will be used to train multilingual models for SLs.

Acknowledgements

Work in this paper is part of the SignON project.[27] This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101017255.

Mathieu De Coster's research is funded by the Research Foundation Flanders (FWO Vlaanderen): file number 77410.

[26] http://www.sign-lang.uni-hamburg.de/ilex/

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

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 Morris, M. R. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *ASSETS 2019 – 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Camgoz, N. C., Koller, O., Hadfield, S., and 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*, pages 10023–10033.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cormier, K., Fenlon, J., Gulamani, S., and Smith, S. (2017). BSL Corpus Annotation Conventions. Version 3.0. Technical report, Deafness Cognition and Language (DCAL) Research Centre, University College London.

Crasborn, O., Zwitserlood, I., Van der Kooij, E., and Bank, R. (2020). Annotation conventions for the Corpus NGT. Technical report, Radboud University Nijmegen, Centre for Language Studies and Department of Linguistics.

De Coster, M., D'Oosterlinck, K., Pizurica, M., Rabaey, P., Verlinden, S., Van Herreweghe, M., and Dambre, J. (2021). Frozen pretrained transformers for neural sign language translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 88–97, Virtual, August. Association for Machine Translation in the Americas.

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)*, pages 12–22, Virtual, August. Association for Machine Translation in the Americas.

David M. Eberhard, et al., editors. (2021). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, USA, twenty-fourth edition.

Hanke, T. (2002). ilex – a tool for sign language lexicography and corpus analysis. In Manuel González Rodríguez et al., editors, *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 923–926. París: ELRA.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pages 79–86.

Kopf, M., Schulder, M., and Hanke, T. (2021). Overview of datasets for the sign languages of europe. Deliverable 6.1, Easier project.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Leeson, L. and Nolan, B. (2008). Digital deployment of the signs of Ireland corpus in elearning. In Onno Crasborn, et al., editors, *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 112–122, Marrakech, Morocco, June. European Language Resources Association (ELRA).

Moryossef, A., Yin, K., Neubig, G., and Goldberg, Y. (2021). Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual, August. Association for Machine Translation in the Americas.

Nonhebel, A., Crasborn, O., and van der Kooij, E. (2004). Sign language transcription conventions for the echo project. annotation convention. version 9. Technical report, University of Nijmegen.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pasikowska-Schnass, M. (2018). Sign languages in the EU. Technical report, European Parliamentary Research Service.

Porta, J. (2014). *Towards a rule-based Spanish to Spanish sign language translation: from written forms to phonological representations*. Ph.D. thesis, Universidad Autónoma de Madrid. Departamento de Tecnología Electrónica y de las Comunicaciones.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Pérez, A., García-Miguel, J. M., and Cabeza, C. (2018). Convenciones de Anotación. Corpus RADIS de la LSE. Version 2. Technical report, Universidade de Vigo.

Rodríguez Ortiz, I. d. l. R. (2005). *Comunicar a través del silencio: las posibilidades de la Lengua de Signos Española*. Universidad de Sevilla. Secretariado de Publicaciones.

Saggion, H., Shterionov, D., Labaka, G., de Cruys, T. V., Vandeghinste, V., and Blat, J. (2021). SignON: Bridging the gap between Sign and Spoken Languages. In *XXXVII Spanish Society for Natural Language Processing conference (SEPLN2021)*.

Sanabre, G. M. (2012). *Desenvolupament d'un sistema de traducció automàtica estadística cap a la llengua de signes catalana*. Ph.D. thesis, Universitat Pompeu Fabra. Departament de Traducció i Ciències del Llenguatge.

Shterionov, D., Vandeghinste, V., Saggion, H., Blat, J., De Coster, M., Dambre, J., van den Heuvel, H., Murtagh, I., Leeson, L., and Schuurman, I. (2021). The SignON project: a Sign Language Translation Framework. In *the 31st Meeting of Computational Linguistics in the Netherlands*.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Van Herreweghe, M., Vermeerbergen, M., Demey, E., De Durpel, H., Nyffels, H., and Verstraete, S. (2013). Corpus Vlaamse Gebarentaal: annotatierichtlijnen. Technical report, Universiteit Gent and KU Leuven.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., and Alikhani, M. (2021). Including Signed Languages in Natural Language Processing. *arXiv:2105.05222 [cs]*, May. arXiv: 2105.05222.

Zhou, H., Zhou, W., Qi, W., Pu, J., and Li, H. (2021). Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.

Zwitserlood, I., (2012). *Classifiers*, pages 158–186. De Gruyter Mouton.

8. Language Resource References

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

Carmen Cabeza and José M. García-Miguel. (2018). *iSignos: Interfaz de datos de Lengua de Signos Española (versión 1.0)*.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., and Bowden, R. (2021). Content4all open research sign language translation datasets. *CoRR*, abs/2105.02351.

Crasborn, O. and Zwitserlood, I. and Ros, J. and van Zuilen, M. (2020). *Corpus NGT, 4e editie*.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Prillwitz, Siegmund and Hanke, Thomas and König, Susanne and Konrad, Reiner and Langer, Gabriele and Schwarz, Arvid. (2008). *DGS corpus project–development of a corpus based electronic dictionary German Sign Language/German*.

Schembri, A. and Fenlon, J. and Rentelis R. and Stamp, R. and Cormier, K. (2011). *British Sign Language Corpus Project*.

Van Herreweghe, M. and Vermeerbergen, M. and Demey, E. and De Durpel, H. and Verstraete, S. (2015). *Het Corpus VGT. Een digitaal open access corpus van videos and annotaties van Vlaamse Gebarentaal, ontwikkeld aan de Universiteit Gent i.s.m. KU Leuven*.