Bootstrapping Pre-trained Word Embedding Models for Sign Language Gloss Translation

Euan McGill
Universitat Pompeu Fabra
Barcelona, Spain
euan.mcgill@upf.edu

Luis Chiruzzo
Universidad de la República
Montevideo, Uruguay
luischir@fing.edu.uy

Horacio Saggion
Universitat Pompeu Fabra
Barcelona, Spain
horacio.saggion@upf.edu

Abstract

This paper explores a novel method to modify existing pre-trained word embedding models of spoken languages for Sign Language glosses. These newly-generated embeddings are described, visualised, and then used in the encoder and/or decoder of models for the Text2Gloss and Gloss2Text task of machine translation. In two translation settings (one including data augmentation-based pre-training and a baseline), we find that bootstrapped word embeddings for glosses improve translation across four Signed/spoken language pairs. Many improvements are statistically significant, including those where the bootstrapped gloss embedding models are used.

Languages included: American Sign Language, Finnish Sign Language, Spanish Sign Language, Sign Language of The Netherlands.

1 Introduction

There has been a surge in research interest on Sign Language machine translation (SLMT) in recent years, but the data scarcity problem (De Sisto et al., 2022) and lack of standardised annotated data (Cormier et al., 2016) remain substantial obstacles to overcome.

At the heart of the labelling problem is the fact that although writing and transcription systems exist for SLs (Grushkin, 2017), none are used day-to-day by signers. Glosses are a semantic labelling tool for signs. They typically use lexemes from the *ambient* spoken language of the hearing community where the SL is used in order to convey the semantic sense of a given sign. However, glosses cannot be considered an orthographic system for SLs as they often differ between datasets, are not used by signers to write their languages (Müller et al., 2023), and may not include linguistic phenomena which are crucial to understand an utterance (Yin and Read, 2020).

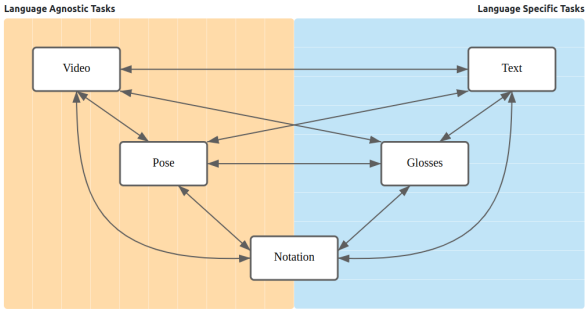

Figure 1: Intermediate, or subtasks of SLMT (Moryossef and Goldberg, 2021). This work focuses on translation between text and glosses.

SLMT is inherently multimodal (Bragg et al., 2019), and it is helpful to conceptualise it as a constellation of sub-tasks at the interface of NLP and computer vision. End-to-end SLMT between SL video and text in a spoken language exists, but performs poorly compared to translation broken down into intermediate steps where signs are represented by some orthographic form (*e.g.* in glosses (De Coster et al., 2023), or SL notation system (Walsh et al., 2022; Jiang et al., 2023)) - except for restricted domains and datasets (*e.g.* (Camgöz et al., 2020; Albanie et al., 2021; Zhang et al., 2023)). These subtasks are neatly shown in a diagram from Moryossef and Goldberg's (2021) overview of the field in Figure 1.

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 116–132
June 24-27, 2024 ©2024 European Association for Machine Translation

Even though both text-to-SL gloss (Text2Gloss) and SL gloss-to-text (Gloss2Text) are sequence-to-sequence tasks using machine-readable text, the amount of parallel data available for any SL is orders of magnitude smaller than equivalent pairs of spoken languages. According to Duarte and colleague's survey (2021), the largest parallel corpus between SL glosses and text available to researchers[1] contains 21,000 parallel utterances (Zhou et al., 2021a). It is reasonable to refer to all SLs as *extremely* low resource languages (Moryossef et al., 2021), and therefore data augmentation approaches must be adopted in order to improve the performance of translation models which include them.

In this paper, a novel method to generate semantic representations for Sign Language (SL) glosses is described. They are created by bootstrapping pre-trained word embedding models from spoken languages which already exist and their use is demonstrated in multilingual Text2Gloss and Gloss2Text machine translation experimental settings.

This paper is structured as follows: In Section 2, previous work where linguistic information is used to supplement gloss representations and its impact on SLMT is described, as well as work to create computational semantic resources for SLs in general. Section 3 sets out the process to generate SL gloss embeddings from pre-trained word embeddings, before Section 4 demonstrates their use in translation experiments. Findings from these experiments are described in Section 5 and discussed in Sections 6 and 7, along with potential future research directions using these embedding representations.

2 Background

One way of mitigating the semantic bottleneck created by gloss representation of signs is to explore techniques for low-resource neural machine translation (Sennrich and Haddow, 2016). These include data augmentation methods involving linguistic features (Armengol Estapé and Ruiz Costa-Jussà, 2021) as well as techniques specifically designed for Text2Gloss translation (Moryossef et al., 2021; Zhou et al., 2021b).

Zhu and colleagues' (2023) comprehensive study of these methods found, for DGS[2] corpora, that: (1) a combination of data augmentation strategies, and (2) transfer learning[3] are viable methods to improve translation performance for Text2Gloss. They also highlight that it is important to ensure that these findings are generalisable to other SLs so further investigation such as the present study is required.

Other studies focused on injecting linguistic features into the embedding table for Text2Gloss and Gloss2Text (Egea Gómez et al., 2022; Chiruzzo et al., 2022), and found that transfer learning was again beneficial, as well as using linguistic features such as part-of-speech (PoS) and syntactic dependency tags - including PoS tags for SL glosses (McGill et al., 2023).

It may also be beneficial to use semantic information about signs into translation models, instead of (or as well as) using syntactic or grammatical information. No previous study with a parallel methodology was found, but other studies do use embeddings as part of SLMT models. Walsh and colleagues (2022) use sentence-level word2vec (Mikolov et al., 2013) or BERT (Devlin et al., 2019) embeddings to support Text2Notation (in HamNoSys (Hanke, 2004)) translation. Other studies use visual embeddings to support joint Sign2Gloss2Text (De Coster et al., 2023), SL recognition (Wong et al., 2023), or to encode phonological information for isolated SL recognition (Kezar et al., 2023).

This paper investigates using semantic information about words and glosses as a transfer learning strategy, and also its performance in combination with the syntax-based data augmentation methods seen in previous works.

2.1 Semantic data sources

Despite the fact that there are many word embedding collections for a great number of spoken languages, the same cannot be said about SLs. This problem is accentuated because the size of current SL corpora is not large enough to create high quality word embedding sets. Schuurman and colleagues (2023) propose SignNets, a database containing rich information about signs in a given SL, indexed by either gloss or an equivalent lexeme in a spoken language. This type of representation would be ideal to map meaning between signs and

[1] How2Sign intended to include 35,000 parallel English/ASL text/glosses, but annotation was suspended indefinitely.

[2] German Sign Language (Deutsche Gebärdensprache)

[3] Such as pre-training on larger, language-agnostic models

between SLs and spoken language senses. However this research is in its early stages, and not ready for use in applications such as SLMT yet.

In contrast, Signbanks (Cassidy et al., 2018) are a well-established and extensible lexicon resource. Signbanks typically store information like ID-glosses, definitions or equivalent senses in a spoken language, phonological specification, images, and video for a given sign.

Semantic resources which allow the understanding of meaning in context, or calculating similarity of a given lexeme to another, are known as pre-trained word embedding models. There are no extant models of this type for SLs, which means that novel ones must be created. However, training models like word2vec or GloVe (Pennington et al., 2014) requires a large quantity of written utterances and it has been established that written SL data does not exist anywhere in large quantities.

Fortunately, it is possible to leverage data from the ambient spoken language in which glosses are written: For example, English for Auslan glosses or Dutch for glosses in Flemish Sign Language. One possible approach could be to just use pre-trained word embeddings without any modification for SL data - *e.g.* a Spanish word2vec model for Spanish Sign Language (LSE)[4] tasks. However, in previous studies this approach has been shown to degrade the performance of Gloss2Text (Chiruzzo et al., 2022) and PoS-tagging (McGill et al., 2023) tasks.

Moreover, in studies in spoken languages, it has been shown that using high-quality English pre-trained embeddings as "anchors" to train bilingual word embedding models for low-resource languages is a promising strategy (Eder et al., 2021). Another study shows that English-lower resource languages bilingual lexica can be used to bootstrap the development of NLP-based tools in under-resourced languages (Wang et al., 2022).

3 Sign Language gloss embeddings

The motivation behind the present methodology is the proposition that, by mapping ID-glosses and their equivalent senses from a Signbank, it is possible to alter the weights of a pre-trained word embedding model from a spoken language in order to simulate the semantic interactions between signs in a given SL.

[4]Lengua de Signos Española

Spoken Language			Sign Language		
ID	**Dims.**	**#Embs**	**ID**	**#Signs**	**#Embs**
English	300	3.00M	ASL	5079	+2605
Finnish	100	247k	FinSL	3120	+1178
Spanish	300	1.00M	LSE	1221	+316
Dutch	320	627k	NGT	4144	+2938

Table 1: Left hand side: For each spoken language, the dimensionality and total number of word embeddings in its word2vec model. Right hand side: For each SL, number of signs in its Signbank(s) and the number of additional word vectors added to the new, bootstrapped word2vec model

As such, each gloss in a given SL is mapped to pre-trained embedding weights in one of three ways, along with some examples from the SLs in this study:

1. If the mapping between ID gloss and spoken language senses in a given Signbank is **one-to-one**, use the embedding weights from that sense (usually this is the same lexical item *e.g.* "TIME-D"[5] = "time" in NGT[6])

2. If there is a **one-to-many** relationship between ID gloss and spoken language senses, take the mean embedding weight from those senses (*e.g.* "WATER" $\in$ {"water", "to drink"} in LSE, "RAT" $\in$ {"rat", "rodent", "mouse", "freshman", "rookie"} in ASL)

3. If there are no senses which match any existing sense in the spoken language Signbank, use the embedding weight from the word embedding model's 'unknown' token (*e.g.* "GALLAUDET"[7] = "UNK" in ASL)

The rest of the weights in the original pre-trained word embedding models remain the same if there is no gloss with the same label and are retained in the model to allow for the mapping of out-of-vocabulary token mapping.

3.1 Datasets

In order to create these bootstrapped pre-trained word embedding models for glosses, a given SL/spoken language pair must have all of the following dataset types available: (1) a Signbank with ID-glosses and translations in the ambient spoken language, (2) a pre-trained word embedding model for the spoken language, (3) parallel corpora of

[5]Glosses derived from other languages than English are translated here

[6]Nederlandse Gebarentaal

[7]The name of a well-known University for DHH students

continuous signing utterances, with both text in the spoken languages and glosses as annotations.

As seen in SL resource surveys (Duarte et al., 2021; Moryossef and Goldberg, 2021), SL-spoken language pairs reaching all these criteria are few in number. Therefore, the bootstrapping of word embedding models and translation experiments are performed on the following language pairs: Spanish Sign Language (LSE)-Spanish; American Sign Language (ASL)-English; Sign Language of the Netherlands (NGT)-Dutch; and Finnish Sign Language (FinSL)[8]-Finnish.

In all language pairs, a word2vec model was chosen, then all unique gloss-definition pairs from a given Signbank were processed following the technique outlined at the beginning of Section 3. The Signbanks and pre-trained word embedding models chosen for each language pair are shown in Appendix B. Table 1 shows the resources used to generate gloss embeddings along with some statistics. The parallel corpora used for translation experiments and a description of data preprocessing is described in Section 4.

3.2 Examples

This section demonstrates the operation of the embedding creation methodology, and shows the potential effectiveness of more accurately representing semantic relations between SL glosses. What follows are examples of gloss embeddings, and then the embedding space is shown visually.

Using cosine similarity to obtain the most similar word vectors, it is possible to compare representations of the same gloss/word in NGT and Dutch respectively. For example, the meaning mapping for "STAGE[9]" $\in$ {"theatre", "the stage", "stage acting"}" results in a slightly different semantic field for NGT and Dutch respectively. In Dutch, the most similar words include "*theatre, folk theatre, play, Bolshoi*". In NGT, the most similar (cosine similarity) glosses include "*ACT-A, ACT-B, VIOLIN, play*", incorporating the verb senses of the gloss. Note that similar NGT glosses contain lexemes which only exist in NGT like "ACT-B". This is a positive sign, as it shows it is possible to map semantic relations to novel lexemes.

Figure 2 shows a visual example, for "Africa" in English, and "AFRICA $\in$ {Africa, continent, geography}" in ASL, as a 2D representation of the vector space (t-SNE (van der Maaten and Hinton, 2008)) and the twelve most similar lexemes for each, as well as for the "RAT" example from Section 3. In English, similar words tend to be the names of nations, whereas similar terms for ASL are more terms related to geographical features. For "rat", the dominant sense seems to be related to the "rookie" definition in ASL, as opposed to the animal in English.

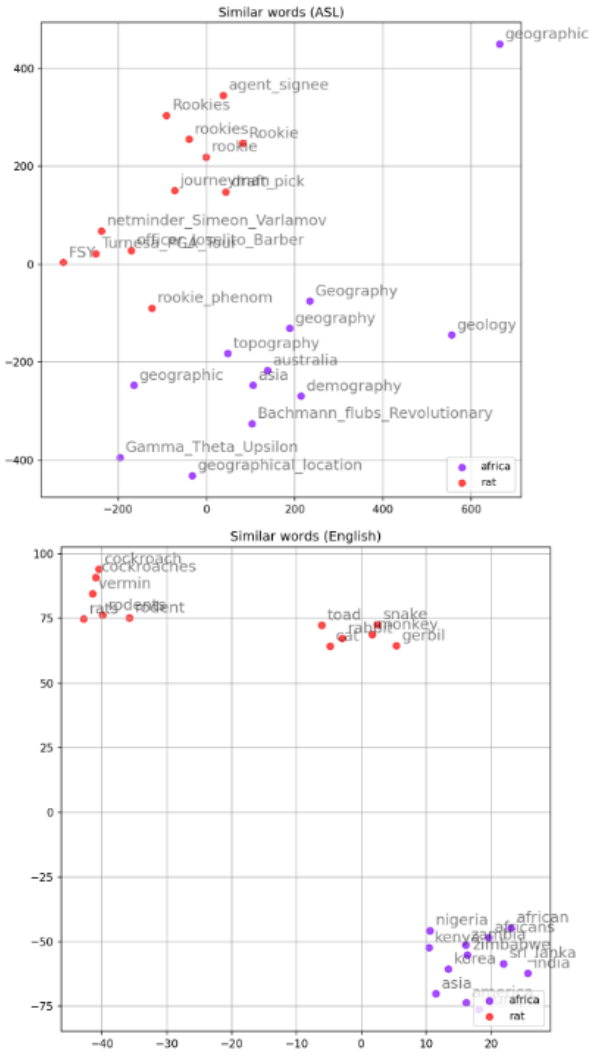

Figure 2: Top N similar words plot for "Africa" and "rat" in ASL (top) and English (bottom)

As seen in Table 1, some glosses introduced to the bootstrapped embedding models do not exist in the original spoken language models. An interesting example of this are three LSE glosses derived from the Spanish lexeme "blood": "SANGRE1" $\in$ {"passion", "to carry sth. in the blood"}; "SANGRE2" = "blood"; "SANGRE3" $\in$ {"glass", "blood"}. A plot for this example is

[8]Suomalainen viitomakieli
[9]TONEEL, in NGT gloss

shown in Appendix C.

3.2.1 Vector space

Turning to an overview of the semantic space overall, Figure 3 was created by plotting these 300-dimensional vectors in joint 2D space (also with t-SNE): (1) all unique glosses from the LSE Signbank, which (2) have an entry in both the original Spanish, and bootstrapped LSE word2vec models. This plot shows that the vector space is altered by the transformations made by the present methodology, and hopefully means that the bootstrapped word2vec model can better simulate SL semantics.

However, it is important to note that the total of 1221 LSE gloss vectors plotted here are the only ones whose weights may have been altered, while the rest of the 1M vectors in the original Spanish word2vec remain the same. This is hopefully not a large concern, as one would expect glosses in a parallel corpus to largely overlap with the ones used to create the modified word2vec model (see also, Table 3 for statistics on overlap).

4 Translation experiments

For each language pair, we perform both Text2Gloss and Gloss2Text experiments in two settings. Firstly, a baseline (Section 4.1) with each parallel corpus for each language. Then, following previous similar experimental setups (Moryossef et al., 2021; Chiruzzo et al., 2022; Zhu et al., 2023), a *warm start* transfer learning approach (Section 4.2) is executed. In other words, first a translation model is pretrained with a larger silver corpus and shared silver and gold vocabulary, and then finetuned on the same parallel data as the baseline.

Translation experiments are performed using OpenNMT-py 3.4.2 (Klein et al., 2017). OpenNMT is an open source translation toolkit which is based on LSTM encoder-decoder model with attention. All other running parameters are set to default, unless stated in Appendix A.

4.1 Baseline experiments

The baseline experiments involve Text2Gloss and Gloss2Text translation between the four spoken language-SL pairs. The specific parallel (or 'gold') datasets are described in Section 4.3. In order to evaluate the utility of these novel word embedding representations in real translation settings, the encoder and decoder (or both) embedding spaces, that start in a random state by default in OpenNMT, are replaced by our collections of word2vec embeddings. For example, in the Gloss2Text setting for NGT→Dutch, there are four experimental settings:

1. *Baseline* (= default OpenNMT encoder/decoder parameters)
2. *Baseline-enc* (= NGT word2vec model encoder, OpenNMT default decoder)
3. *Baseline-dec* (= OpenNMT default encoder, Dutch word2vec model decoder)
4. *Baseline-both* (= NGT word2vec model encoder, Dutch word2vec model decoder)

This repeated for each language, and in the Text2Gloss direction, results in a total of 32 baseline experiments. Each setting is repeated for three runs of 10k epochs, starting at a random seed.

4.2 Pretrain + finetune experiments

Like in Section 4.1, there are also four experimental types: *PT+FT*, *PT+FT-enc*, *PT+FT-dec*, and *PT+FT-both*. However, for these experiments, models are trained on a larger parallel 'silver' dataset which is comprised of utterances in a spoken language alongside *pseudo*-glosses created by rule-based methods of data augmentation (*c.f.* (Moryossef et al., 2021; Chiruzzo et al., 2022; Zhu et al., 2023)).

During the pretraining phase, models are trained for three runs of 10k epochs on the parallel silver data. This phase also follows a warm start strategy by means of joint vocabulary (Nguyen and Chiang, 2017), whereby vocabulary is generated at the start of pretraining containing all tokens from both the *silver* and *gold* datasets. From each run, the best-performing model (BLEU measured for models at every 200 steps, based on the dev set) is chosen. These models are then fine-tuned for a further 5k epochs (three runs each) on the parallel spoken language/SL corpora from the Baseline experiments.

4.3 Datasets

Owing to the way SL datasets are collected, along with their low resource nature, we adopt different strategies for: (a) creating the silver datasets for each language pair, and (b) doing dataset splits in the gold corpora.

This section describes, by language pair, the gold and silver parallel datasets used in translation

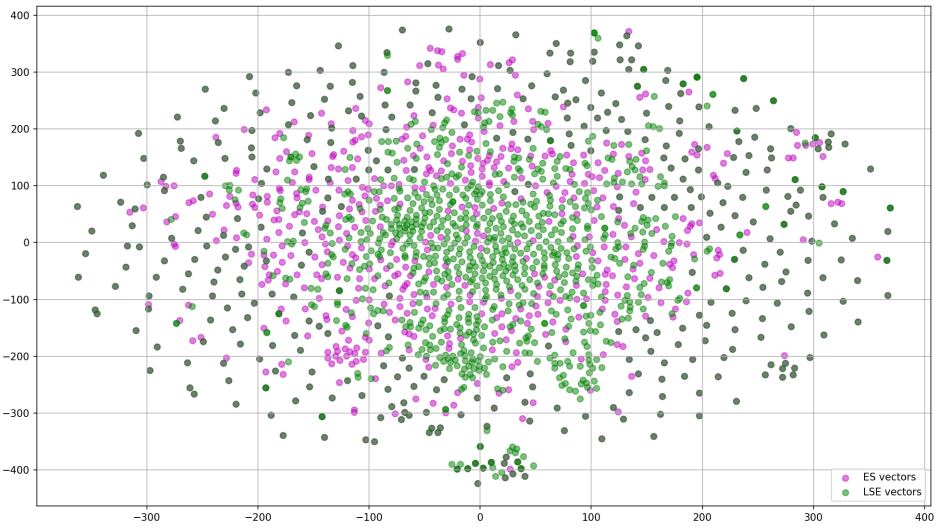

Figure 3: Spanish (purple) and LSE (green) word vectors from the LSE Signbank vocabulary plotted in joint 2d space. Grey points are where the weights are equal for both languages

	ASL/en	**FinSL/fi**	**LSE/es**	**NGT/nl**
Silver	87.7k	24.0k	20.3k	161k
Gold-train	2328	3480	1900	11.9k
Gold-dev	251	449	475	1484
Gold-test	352	534	482	1484
Gold-all	2931	4463	2857	14.8k

Table 2: Number of parallel utterances per language pair divided into dataset splits

	Baseline		**PT+FT**	
	#toks	overlap	#toks	overlap
ASL voc.	2410	75.8%	14.1k	73.7%
FinSL voc.	814	95.2%	2684	66.4%
LSE voc.	1123	60.7%	10.4k	83.5%
NGT voc.	3277	85.7%	25.2k	73.0%
en vocab	2377	95.1%	17.8k	78.6%
fi vocab	4523	24.4%	7450	24.2%
es vocab	2705	65.6%	16.4k	95.7%
nl vocab	11.2k	35.6%	38.0k	49.2%

Table 3: Vocabulary statistics for each language: number of unique tokens, and % overlap of tokens between the word2vec model and vocabulary in the gold (left columns) and silver+gold (right columns) datasets

experiments, dataset splits, and the methods used to generate silver data. Tables 2 and 3 show statistics about these datasets.

ASL/English: The NCSLGR and ASLLRP Corpora (Neidle et al., 2022) are combined as both datasets are relatively small for the present task. This data was accessed through ASLLRP's Data Access Interface. These multimodal datasets contain utterances from twelve unique signers and contain a mixture of storytelling and elicited utterances, similar to the other parallel corpora used in this study. Like in Moryossef et al. (2021), the silver data is the sample set[10] from the ASLG-PC12 dataset - a parallel corpus where the ASL pseudo-glosses are generated with a linguistically-motivated rule-based approach. NCSLGR has been used before on its own in comparable studies (Zhu et al., 2023), but the decision was made to combine the two publicly-available glossed corpora so that as much parallel gold data as possible was available. The gold corpus was split into training-dev-test sets as close to 80%-10%-10% as possible, while also ensuring that the each unique signer only appears in one of these splits.

FinSL/Finnish: Corpus FinSL (Salonen et al., 2020) is used as the gold standard parallel dataset. For the silver data, Moryossef's (Moryossef et al., 2021) language-agnostic rules for synthetic SL gloss generation is performed on 24k monolingual Finnish sentences selected at random (minimum 3 words per original utterances, duplicates removed) from the Tatoeba[11] collection. In addition, all first person pronouns are replaced with the gloss "OS:" (*pointing at self*) and other pronouns with "OS:minä" (*pointing sign*) to mirror the contents of the Corpus FinSL. This dataset was split 78%-10%-12% for train/dev/test.

LSE/Spanish: The iSignos Corpus from CORLSE (Cabeza and García-Miguel, 2019) is used for this language pair. There are 10 unique signers in this corpus, which informed the 64%-17%-19% train/dev/test split which is also used in previous studies (McGill et al., 2023). The silver

[10]https://huggingface.co/datasets/aslg_pc12
[11]https://tatoeba.org/en/

data is also created using the same methodology from these studies, but using Tatoeba monolingual Spanish data to generate pseudo-glosses, and with slight differences in preprocessing decisions as described in Section 4.4.

NGT/Dutch: This language pair uses the largest parallel corpus available in this study, the CorpusNGT (Crasborn and Zwitserlood, 2008). Following SLMT experiments in the SignON project (Saggion et al., 2021), the dataset is split into partitions of 80%-10%-10%. Silver data was taken from a subset of the SONAR dataset for Dutch, and then modified with a rule-based approach (Bram Vanroy, *p.c.*) including gloss reordering[12] orginally devised for Flemish Sign Language (VGT[13]).

4.4 Preprocessing

All four parallel corpora are annotated separately by dominant and non-dominant hand for SL glosses. As ML-based models, including NMT models, typically take linear alphanumeric input - it is necessary to modify the gloss annotations from these datasets. A systematic approach following *e.g.* Östling and colleagues (2017) was taken to linearise and lexicalise glosses:

- If two equal glosses occur simultaneously, only retain one
- If two different glosses simultaneously, place dominant hand gloss before non-dominant hand gloss
- Remove gestures which are not lexical signs
- Remove phonological features, tags indicating fingerspelling/name signs etc. from glosses

However, unlike similar studies which remove most affixes and labels, care was taken to match gloss labels in the parallel utterances to what is present in a given Signbank. In order to do this, glossing conventions and/or style guides such as SLAASh (Hochgesang, 2022) for ASLLRP and RADIS (Pérez et al., 2019) for CORLSE were referred to.

The same approach is taken for silver data generation. For example, pronouns which resemble those in the ambient spoken language, or where the silver dataset has its own gloss conventions, were edited to match what is used in the gold corpus/Signbank. In the synthetic ASL, all adjectives contained the prefix "DESC-". As this does not occur in the gold data, they were removed. All gloss and spoken language text data is tokenised and in lowercase.

4.5 Evaluation

The best models from all runs of each experimental setting are evaluated on the held-out test set in the following way:

BLEU (Papineni et al., 2002) and **CHrF** (Popović, 2015) are the primary means of automatic evaluation in this study, measured using sacreBLEU (Post, 2018). BLEU-4 is calculated with disabled internal tokenisation[14] (Müller et al., 2023). **METEOR** (Banerjee and Lavie, 2005) is also calculated through nltk[15] and reported. As there are three runs per experimental setup, **mean and standard deviation** are reported.

Statistical significance testing is also performed by means of **paired bootstrap resampling** (Koehn, 2004) calculated with Graham Neubig's script[16]. Koehn states that this method of calculating significance at a level of $p < 0.05$ is effective with test sets greater than N=300. In this study, all test sets range between N=352 and N=1484.

Some **qualitative evaluation** is provided in the form of perceptive comments by the authors. Qualitative evaluation is of utmost importance to MT as a field[17], especially low-resource MT where output with reasonable BLEU scores may still be ungrammatical or incomprehensible to the reader. Unfortunately, it was beyond the scope of this study to provide a more formal approach to qualitative assessment such as Direct Assessment (Graham et al., 2017; Zhu et al., 2023).

4.6 Reproducibility

The data, experimental configuration files, preprocessing and data augmentation scripts, scripts to generate embeddings, and model outputs for testing are all openly available[18] for the purposes of transparency and reproducibility.

[12] https://clin2022.uvt.nl/data-augmentation-for-machine-translation-of-sign-language-of-the-netherlands-and-flemish-sign-language/

[13] Vlaamse Gebarental

[14] Signature:

[15] https://www.nltk.org/

[16] https://github.com/neubig/util-scripts/blob/master/paired-bootstrap.py

[17] https://bricksdont.github.io/posts/2020/12/seven-recommendations-for-mt-evaluation/

[18] https://github.com/euan-mcgill/gloss_embeddings

5 Results and analysis

Table 4 summarises the quantitative findings of this study, reporting the best-performing model for each setup. Table 5 in Appendix D reports the best model on average (and standard dev.) across three runs for each setup. For the experimental setup acronyms used in this section and Table 4, refer to their descriptions in Sections 4.1 and 4.2.

For **es→LSE Text2Gloss**, the Baseline models with any kind of embeddings improved over the baseline in CHrF and METEOR, but only *PT+FT-both* performed better on the BLEU metric and this difference was not significant ($p = 0.25$, N=482). All PT+FT conditions had significantly higher BLEU scores than the Baseline. Within the PT+FT experimental setups, all metrics were markedly higher in the embedding setups, and *PT+FT-enc* ($p = 0.03$, N=482) and *PT+FT-both* ($p = 0.03$, N=482) showed a significant improvement. For **LSE→es Gloss2Text**, PT+FT tends to be a better strategy with *PT+FT-enc* being the only setup which performs significantly better than the baseline ($p = 0.03$, N=482), and higher scores in both metrics.

The fact that *PT+FT-both* performs significantly better than *PT+FT* in Text2Gloss, and *PT+FT-enc* than *Baseline* in Gloss2Text, is particularly promising as these conditions include the bootstrapped word embedding models for LSE.

For **nl→NGT Text2Gloss**, using embeddings improves BLEU scores in all setups, but only *PT+FT-dec* (with NGT bootstrapped glosses) in METEOR as well as being the only significant improvement on BLEU ($p < 0.01$, N=1484). The results are the mirror image in PT+FT: All setups except *PT+FT-dec* significantly improve over the baseline, and *PT+FT-enc* with only Dutch word2vec embeddings improves over *PT+FT* ($p = 0.05$, N=1484). However, Table 5 indicates a marked degree of variance compared to other language pairs and setups. This would be interesting to investigate further.

In **NGT→nl Gloss2Text** word2vec embeddings, as well as pretraining and finetuning, seems to damage the performance of this translation direction. Across both of these language pairs, compared to the BLEU scores the METEOR scores are also quite weak (compare LSE and FinSL results). In this language pair in particular, Table 3 shows that there is a large disparity in size between a much larger Dutch vocab than NGT. Moreover, the Dutch word2vec model has a very low token coverage with both the gold and silver+gold vocab used in these experiments (both less than 50%). The consequence of this may be that it is difficult to create links between the lexical items in both languages.

For **en→ASL** Text2Gloss, the use of word2vec embeddings improves performance on most settings on both metrics. In the PT+FT setting, encoder English embeddings and both English and ASL embeddings improve significantly over the baseline (*PT+FT-both*: $p < 0.00$, N=352). For **ASL→en** Gloss2Text, over the baseline, significant improvements are seen when ASL embeddings are used in the encoder to support glosses: *PT+FT-enc* ($p = 0.02$, N=352), and *PT+FT-both*: ($p < 0.01$, N=352). It is therefore reasonable to infer that: (a) richer semantic representations for ASL, and (b) *warm-start* transfer learning on a larger silver dataset and joint vocabulary provides a real boost to translation performance and generalisibilty.

It may be the case that there are no marked improvements within PT+FT as the en/ASL test partition is the smallest between all four language pairs (see Table 2). Also, unusually among parallel corpora, ASL's vocabulary is actually larger than the English one where there is usually a large disparity in the other direction (see Table 3).

In the ASL/en language pair in general, the METEOR scores are much stronger compared to the others in this study. Perhaps a more even total of unique tokens in both ASL and English lexica contributes to this.

As for **fi→FinSL** Text2Gloss and **FinSL→fi** Gloss2Text, PT+FT with silver data provides an improvement in metrics across the board. Significant improvements are only seen in Gloss2Text: FinSL embeddings significantly improve in *Baseline-enc* over the baseline ($p = 0.02$, N=534), and in *PT+FT-enc* over *PT+FT* ($p = 0.04$, N=534), but curiously not over the baseline which has a very low BLEU score < 1.

The results for FinSL/fi may be considered rather unusual on the whole. It is possible that the very low vocabulary size of the FinSL gold utterances (N=814), the disparity[19] between this and the Finnish vocabulary size (N=4523), the replacement of all pronouns with just two lexemes (see

[19]BLEU has a brevity penalty, so the short sentences output for FSL should contribute to low scores

Best models - Text2Gloss	LSE			NGT			ASL			FinnSL		
	BLEU	**CHrF**	**Met.**	**BLEU**	**CHrF**	**Met.**	**BLEU**	**CHrF**	**Met.**	**BLEU**	**CHrF**	**Met.**
Baseline	7.34	0.198	0.089	18.66	0.269	0.115	15.46	0.372	0.286	5.54	0.174	0.109
Baseline+enc	7.32	0.212	0.102	19.31	0.261	0.112	17.56	0.398	0.312	6.41	0.195	0.125
Baseline+dec	7.12	0.202	0.093	19.92*	0.271	0.118	15.98	0.363	0.283	4.88	0.166	0.103
Baseline+both	7.64	0.217	0.106	19.66	0.266	0.114	16.30	0.381	0.303	6.27	0.193	0.128
PT+FT	9.94*	0.240	0.151	22.34*	0.310	0.142	18.46*	0.423	0.344	7.06	0.222	0.150
PT+FT+enc	17.83*†	0.341	0.197	22.67*†	0.306	0.144	20.26*	0.432	0.349	7.09	0.250	0.174
PT+FT+dec	16.48*	0.309	0.184	19.75	0.307	0.139	18.73	0.409	0.331	7.13	0.224	0.150
PT+FT+both	18.15*†	0.347	0.198	21.70*	0.311	0.140	19.67*	0.436	0.354	7.58	0.253	0.177
Best models - Gloss2Text	LSE			NGT			ASL			FinnSL		
	BLEU	**CHrF**	**Met.**	**BLEU**	**CHrF**	**Met.**	**BLEU**	**CHrF**	**Met.**	**BLEU**	**CHrF**	**Met.**
Baseline	7.80	0.193	0.146	4.84	0.220	0.144	14.29	0.352	0.356	0.90	0.116	0.119
Baseline+enc	8.25	0.192	0.159	4.47	0.219	0.144	13.61	0.353	0.357	1.27*	0.128	0.122
Baseline+dec	7.09	0.181	0.147	4.42	0.217	0.143	13.78	0.347	0.345	1.27	0.132	0.116
Baseline+both	8.70	0.197	0.161	4.31	0.214	0.140	13.25	0.347	0.352	1.75	0.151	0.131
PT+FT	8.67	0.201	0.168	3.38	0.200	0.129	16.52*	0.410	0.422	2.30*	0.164	0.145
PT+FT+enc	9.64*	0.211	0.178	3.59	0.208	0.132	16.88*	0.419	0.425	3.14†	0.177	0.155
PT+FT+dec	7.95	0.212	0.165	3.55	0.202	0.128	15.94	0.398	0.412	2.48†	0.164	0.138
PT+FT+both	9.02	0.214	0.179	3.53	0.206	0.131	17.05*	0.421	0.424	3.05	0.177	0.150

Table 4: Results summary for translation experiments in OpenNMT (* = significantly better than Baseline, †= significantly better than PT+FT). For each metric (Met. = METEOR), a higher score implies better performance.

Section 4.3), and the low token coverage of the Finnish word2vec model of Finnish tokens (24% in both the gold and silver+gold datasets) contribute to these results. Besides this, the FinSL dataset contains a few signs corresponding to descriptive markers (Salonen et al., 2019) (*e.g.* "_kvkk" for 'whole object' and "_kvmk" for 'shape and size') that are frequent (around 19% of the total signs in the training set). These signs are not lexical and have no corresponding ambient language lexemes, so an "unknown" random embedding was assigned to them.

These tokens' high frequency may also explain the low performance of the FinSL↔fi experiments. In the future, we want to explore the possibility of creating embeddings for these markers using the average embedding of their corresponding Finnish descriptions.

5.1 Qualitative analysis

After a high-level comparison of model output which uses word vectors from this study, it is possible to observe lexical differences across experimental settings. These include, especially in lower-performing language pairs like es→LSE, the replacement of more similar glosses even when the translation is inaccurate, a more similar distribution of PoS categories compared to the gold translation, and a lower prevalence of garbled output and model hallucination in lower-performing language pairs. Some qualitative examples from experimental settings are shown in Appendix E.

Looking at model output utterances, in tandem with the low BLEU scores, may explain unusual patterns of significance for FinSL/fi experiments. Figure 4 is a BLEU interpretability chart[20] which is useful to refer to when interpreting the quantitative results.

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

Figure 4: Interpretability of BLEU scores

6 Discussion and limitations

So far, this exploratory work shows that using semantic representations tailored to SLs (in this case word2vec embeddings adapted to particular SL settings) is a promising avenue of research. Overall, the results present a positive outlook concerning the effectiveness of including bootstrapped word embedding models in the encoder and/or decoder of OpenNMT for Text2Gloss and Gloss2Text translation. In all *PT+FT-** settings, the use of embeddings improved translation performance in at least one setting. This is also true in the baseline setting, apart from with NGT→nl and

[20] https://cloud.google.com/translate/docs/advanced/automl-evaluate

ASL→en. Many of these improvements were significant, and those where *PT+FT-embedding* significantly improved against *PT+FT* are particularly notable.

However, it is necessary to examine more data augmentation methods, types/sizes of word embedding models, sub-word tokenisation, and techniques to adapt semantic representations for the *extremely* low-resource setting of SL processing. It may also be worthwhile to attempt this approach on low-resource pairs of spoken languages, especially those with little or no written data (Aepli et al., 2023) as anchor word embeddings already exist for spoken languages (Eder et al., 2021). Other practical tasks involving word embedding model support may include the tagging and parsing of SL gloss data (Östling et al., 2017; Yang and Zhang, 2018; García-Miguel and Cabeza, 2020).

Besides the use of OpenNMT for experiments, trying alternative open source translation toolkits such as MarianMT (Junczys-Dowmunt et al., 2018) (such as Perea-Trigo et al. (2024) for LSE).

Pretrained models like mBART (Liu et al., 2020) could also be a fruitful direction of research. Some preliminary experiments following Egea Gómez and colleagues (2022) were also attempted, using a mBART translation approach for LSE↔Spanish. However, some issues were found when applying the present method to mBART: Firstly, the model uses SentencePiece tokenisation, while this study's embeddings are created with simple whitespace tokenisation. Furthermore, the mBART model expects a unified embedding space between source and target languages, which could skew the results for glosses that have the same surface form as ambient language words. It is possible to overcome these limitations, but given time and resource constraints the mBART experiments remain out of the scope of this work, and it is planned to explore them further in the future.

It would also be rewarding to explore other lexical SL resources such as Signpuddle[21] which has been used in work on Text2Notation (Jiang et al., 2023) translation work. In addition, when SignNets (Schuurman et al., 2023) are further developed and contain rich metalinguistic information for many SLs, these will be a crucial resource for further studies in this area.

Some researchers may disagree with the use of glosses as a representation in SL processing altogether, and disprefer splitting SLMT into a pipeline of intermediate tasks instead of treating it as an end-to-end task (Yin and Read, 2020). This is a valid position, and other work involving semantic representations in, for example, Video2Text could be complementary to studies like the present one.

Recent innovations into data-intensive methods such as 0-shot MT and NLP tasks often exclude SLs, because even though messy, unorganised, and seemingly irrelevant text data can be used for tasks in many spoken languages, this is not necessarily the case for the multimodal nature of SLs (Yin et al., 2021; Núñez-Marcos et al., 2023). However, recent research into *true 0-shot* translation; using LLMs to read and interpret reference material about the grammar of a language (Tanzer et al., 2024) - may aid SLMT and SL processing beyond that.

The large amount of experimental settings and limited computing resources available also meant that it was not possible to complete all of the evaluation that was initially planned. For example, from the insights gleaned from NGT→Dutch, it would be interesting to quantitatively investigate the connection between word embedding model's vocab coverage and model performance. Qualitative analysis, though present, was unfortunately minimal and not formal and it would be greatly beneficial to expand it.

7 Concluding remarks

This study cast a wide net in order to devise novel methods to create semantic representations for SL glosses, and test their effectiveness when being used in SLMT. These experiments showed mixed but overall positive results, whereby bootstrapped pre-trained word embeddings from a spoken language can be modified with the present methodology in order to represent the semantic relations between SL glosses. It also provides further evidence that pretraining on silver data is effective across language pairs.

Future work will benefit from further experimentation with the methods undertaken to generate vector representation for signs whether represented by gloss, SL notation system, pose, or video frame. These embedding representations sit at the interface of NLP and computer vision-based approaches to SLMT, and characterise the need to follow both avenues of this field of research in a complimentary manner.

[21] https://www.signbank.org/signpuddle/

Acknowledgements

Amb el suport del Departament de Recerca i Universitats de la Generalitat de Catalunya.

This work is part of Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MICIU/AEI /10.13039/501100011033

This work was supported in part by Oracle Cloud credits and related resources provided by Oracle for Research.

This work is a continuation of research started within the SignON project. SignON is a Horizon 2020 project, funded under the Horizon 2020 program ICT-57-2020 - "An empowering, inclusive, Next Generation Internet" with Grant Agreement number 101017255.

References

Aepli, Noëmi, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A benchmark for evaluating machine translation metrics on dialects without standard orthography.

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

Armengol Estapé, Jordi and Marta Ruiz Costa-Jussà. 2021. Semantic and syntactic information for neural machine translation: Injecting features to the transformer. *Machine Translation*, 35:3:3–17.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Goldstein, Jade, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

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

Cabeza, Carmen and José M. García-Miguel. 2019. iSignos: Interfaz de datos de Lengua de Signos Española (versión 1.0).

Camgöz, Necati Cihan, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR 2020*, pages 10020–10030.

Cassidy, Steve, Onno Crasborn, Henri Nieminen, Wessel Stoop, Micha Hulsbosch, Susan Even, Erwin Komen, and Trevor Johnston. 2018. Signbank: Software to support web based dictionaries of sign language. In Calzolari, Nicoletta, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Chiruzzo, Luis, Euan McGill, Santiago Egea-Gómez, and Horacio Saggion. 2022. Translating Spanish into Spanish Sign Language: Combining rules and data-driven approaches. In Ojha, Atul Kr., Chao-Hong Liu, Ekaterina Vylomova, Jade Abbott, Jonathan Washington, Nathaniel Oco, Tommi A Pirinen, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao, editors, *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 75–83, Gyeongju, Republic of Korea, October. Association for Computational Linguistics.

Cormier, Kearsy, Onno Crasborn, and Richard Bank. 2016. Digging into signs: Emerging annotation standards for sign language corpora. In Efthimiou, Eleni, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch, editors, *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 35–40, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Crasborn, Onno A and IEP Zwitserlood. 2008. The corpus ngt: an online corpus for professionals and laymen.

Crasborn, Onno, Richard Bank, Inge Zwitserlood, Els van der Kooij, Ellen Ormel, Johan Ros, Anique Schüller, Anne de Meijer, Merel van Zuilen, Yassine Ellen Nauta, Frouke van Winsum, and Max Vonk. 2020. NGT dataset in Global Signbank.

De Coster, Mathieu, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1–27.

De Sisto, Mirella, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Confer-*

ence, pages 2478–2487, Marseille, France, June. European Language Resources Association.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Duarte, Amanda, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Duquenne, Paul-Ambroise, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations.

Eder, Tobias, Viktor Hangya, and Alexander Fraser. 2021. Anchor-based bilingual word embeddings for low-resource languages. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 227–232, Online, August. Association for Computational Linguistics.

Egea Gómez, Santiago, Luis Chiruzzo, Euan McGill, and Horacio Saggion. 2022. Linguistically enhanced text to sign gloss machine translation. In *Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings*, pages 172–183.

García-Miguel, José M. and Carmen Cabeza. 2020. Hacia un treebank de dependencias para la lse. *Hesperia: Anuario de Filología Hispánica*, 22:111–143, mar.

Graham, Yvette, Qingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra, and Carolina Scarton. 2017. Improving evaluation of document-level machine translation quality estimation. In Lapata, Mirella, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 356–361, Valencia, Spain, April. Association for Computational Linguistics.

Grushkin, Donald A. 2017. Writing signed languages: What for? what form? *American annals of the deaf*, 161(5):509–527.

Hanke, Thomas. 2004. Hamnosys—representing sign language data in language resources and language processing contexts. In *LREC 2004, WS on RPSLs*, pages 1–6, Paris, France.

Hochgesang, Julie A. 2022. Slaash id glossing principles, asl signbank and annotation conventions, version 3.2.

Jiang, Zifan, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In Vlachos, Andreas and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Kezar, Lee, Jesse Thomason, and Zed Sehyr. 2023. Improving sign recognition with phonology. In Vlachos, Andreas and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Bansal, Mohit and Heng Ji, editors, *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In Lin, Dekang and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

McGill, Euan, Luis Chiruzzo, Santiago Egea Gómez, and Horacio Saggion. 2023. Part-of-speech tagging Spanish Sign Language data and its applications in sign language machine translation. In Ilinykh, Nikolai, Felix Morger, Dana Dannélls, Simon Dobnik, Beáta Megyesi, and Joakim Nivre, editors, *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced*

Languages and Domains (RESOURCEFUL-2023), pages 70–76, Tórshavn, the Faroe Islands, May. Association for Computational Linguistics.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Moryossef, Amit and Yoav Goldberg. 2021. Sign Language Processing. `https://sign-language-processing.github.io/`.

Moryossef, Amit, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In Shterionov, Dimitar, editor, *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.

Müller, Mathias, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada, July. Association for Computational Linguistics.

Neidle, Carol, Augustine Opoku, and Dimitris N. Metaxas. 2022. ASL video corpora & sign bank: Resources available through the american sign language linguistic research project (ASLLRP). *CoRR*, abs/2201.07899.

Nguyen, Toan Q. and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In Kondrak, Greg and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Núñez-Marcos, Adrián, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

Östling, Robert, Carl Börstell, Moa Gärdenfors, and Mats Wirén. 2017. Universal dependencies for swedish sign language. In *Nordic Conference of Computational Linguistics*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Perea-Trigo, Marina, Celia Botella-López, Miguel Ángel Martínez-del Amor, Juan Antonio Álvarez García, Luis Miguel Soria-Morillo, and Juan José Vegas-Olmos. 2024. Synthetic corpus generation for deep learning-based translation of spanish sign language. *Sensors*, 24(5).

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *3rd Conf. on MT*, pages 186–191, Belgium, Brussels. ACL.

Pérez, Ania, José M. García-Miguel, and Carmen Cabeza. 2019. Corpus annotation for studying grammatical expression of events: notes about the design of radis project. *Sensos-e*, 6(1):40–61, Sep.

Saggion, Horacio, Dimitar Shterionov, Gorka Labaka, Tim Van de Cruys, Vincent Vandeghinste, and Josep Blat. 2021. Signon: Bridging the gap between sign and spoken languages. In *Alkorta J, Gonzalez-Dios I, Atutxa A, Gojenola K, Martínez-Cámara E, Rodrigo A, Martínez P, editors. Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021); 2021 Sep 21-24; Málaga, Spain. Aachen: CEUR Workshop Proceedings; 2021. p. 21-5.* CEUR Workshop Proceedings.

Salonen, Juhana, Tuija Wainio, Antti Kronqvist, and Jarkko Keränen. 2019. Suomen viittomakielten korpusprojektin (cfinsl) annotointiohjeet. In *Annotation Convention. Helmikuu: Department of Linguistics and Communication Sciences, Sign Language Center, University of Jyväskylä*, page 40.

Salonen, Juhana, Antti Kronqvist, and Tommi Jantunen. 2020. The corpus of Finnish Sign Language. In Efthimiou, Eleni, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch, editors, *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 197–202, Marseille, France, May. European Language Resources Association (ELRA).

Schuurman, Ineke, Thierry Declerck, Caro Brosens, Margot Janssens, Vincent Vandeghinste, and Bram Vanroy. 2023. Are there just WordNets or also SignNets? In Rigau, German, Francis Bond, and Alexandre Rademaker, editors, *Proceedings of the 12th Global Wordnet Conference*, pages 172–178, University of the Basque Country, Donostia - San Sebastian, Basque Country, January. Global Wordnet Association.

Sennrich, Rico and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *1st Conf. on MT*, pages 83–91, Berlin, Germany. ACL.

Tanzer, Garrett, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book.

van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Walsh, Harry, Ben Saunders, and Richard Bowden. 2022. Changing the representation: Examining language representation for neural sign language production. In Efthimiou, Eleni, Stavroula-Evita Fotinea, Thomas Hanke, John C. McDonald, Dimitar Shterionov, and Rosalee Wolfe, editors, *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 117–124, Marseille, France, June. European Language Resources Association.

Wang, Xinyi, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In Muresan, Smaranda, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland, May. Association for Computational Linguistics.

Wong, R., N. Camgoz, and R. Bowden. 2023. Learnt contrastive concept embeddings for sign recognition. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1937–1946, Los Alamitos, CA, USA, oct. IEEE Computer Society.

Yang, Jie and Yue Zhang. 2018. Ncrf++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Yin, Kayo and Jesse Read. 2020. Better sign language translation with STMC-transformer. In Scott, Donia, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Yin, Kayo, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online, August. Association for Computational Linguistics.

Zhang, Biao, Mathias Müller, and Rico Sennrich. 2023. SLTUNET: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations*.

Zhou, H., W. Zhou, W. Qi, J. Pu, and H. Li. 2021a. Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325, Los Alamitos, CA, USA, jun. IEEE Computer Society.

Zhou, Hao, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021b. Improving sign language translation with monolingual data by sign back-translation.

Zhu, Dele, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural machine translation methods for translating text to sign language glosses. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada, July. Association for Computational Linguistics.

A OpenNMT parameters

(1) To **build** vocabulary:

python build-vocab.py –n_sample 50000

(2a) To **train** translation models (pre-training):

python train.py –feat_merge "concat" –bucket_size 144 –world_size 1 –gpu_ranks [0] –save_checkpoint_steps 200 –train_steps 10000 –valid_steps 200 –log_file "specified.log"

(2b) To **train** translation models (fine-tuning):

python train.py –feat_merge "concat" –bucket_size 144 –world_size 1 –gpu_ranks [0] –save_checkpoint_steps 200 –train_steps +5000 –valid_steps 200 –train_from "specified-pt-model.pt" –reset_optim keep_states –log_file "specified.log"

(3) To **translate** test data for evaluation:

python translate.py –ban_unk_token

B Signbanks and word2vec models used

For **ASL-English**, a combination of the ASL Signbank[22] ASLLRP Sign Bank (Neidle et al., 2022) are used and the GoogleNews word2vec (Skip-gram) model[23].

For **FinSL-Finnish**, it is the Suomen Signbank[24] and the Finnish Text Collection word2vec model[25].

For **LSE-Spanish**, the CORLSE lexicon gathered from the iSignos Corpus' web resource (Cabeza and García-Miguel, 2019) as well as the Spanish Billion Words model[26].

And for **NGT-Dutch**, the Global Signbank (NGT dataset) (Crasborn et al., 2020) and SONAR embeddings (Duquenne et al., 2023).

C Vector similarity plots

Figure 5 shows the ten most similar (cosine similarity) word in the LSE word2vec model for the three glosses based on the lexeme "BLOOD" in LSE mentioned in Section 3.2, represented in 2D vector space.

D Results: Mean and standard deviation

Table 5 shows the best-performing model (number of training epochs shown) on average from three runs in each experimental setup. The *PT+FT* experiments only show one set of experimental runs, as recall that from the pre-training phase, the best-performing epoch from each of the three runs is chosen to fine-tune for another 5000 epochs on *gold* data. Results for FinSL↔fi could not be shown, as only one run per setup was undertaken.

Similar to the findings based on the best model in each setup shown in Table 4, for most language pairs PT+FT performed more strongly than the Baseline. Using features tends to improve translation results on average, but the standard deviation figures show a high degree of variance between settings, particularly when translating from Dutch.

E Qualitative analysis examples

This Appendix shows four utterances from different translation directions and experimental setups which exemplify the use of bootstrapped SL embeddings in the encoder or decoder.

Figure 6 is an example from es→LSE. The original *gold* output sentence from the test set of iSignos is challenging, particularly as it contains a classifier predicate[27] "RECIBIR-MONTÓN". Comparing the *PT+FT-both* hypothesis to *PT+FT*, notice that "child" is rendered more accurately as "HOMBRE PEQUEÑO2" rather than "HOMBRE PERSONA" (a frequent bigram in this corpus). Also, the first person plural pronoun is correctly identified. Whether or not having tailored semantic representations available to the decoder/SL output brings about this improvement is up for debate, but the output is more faithful to the *gold* output nonetheless.

As a counterexample, Figure 7 compares the *gold* output for the given sentence with the *Baseline*, *PT+FT*, and *PT+FT-both* hypotheses. In this case, it appears that *PT+FT* output reflects the semantics of ASL in a better way. The signs "GIVE" and "GIFT" are exemplars of the phenomenon in ASL where signs can be used as nouns, verbs, or adjectives interchangeably, so using either in this instance would be grammatical. As for the model using both word2vec embedding representations, it chooses "GO-OUT" which - while still a verb - would not necessarily be the best choice.

Finally, Figure 8 shows a more challenging example - again from Spanish→LSE where no model can provide a grammatical output. The outputs from *Baseline* and *PT+FT* appear like model hallucinations of frequently-occurring tokens from the training data. The same may be said about the *PT+FT-both* output. However, the connection between "padres" and "PADREˆMADRE" appears to be more robust and appears in its hypothesis. The *PT+FT-both* hypothesis is the only one to include a negative "NO" ("NADA2" appears in the *gold* output) which may imply that using SL-derived embeddings may also be more robust to part-of-speech class.

[22] https://aslsignbank.haskins.yale.edu/
[23] https://code.google.com/archive/p/word2vec/
[24] https://signbank.csc.fi/
[25] http://urn.fi/urn:nbn:fi:lb-2022041405
[26] https://crscardellino.ar/SBWCE/

[27] Signs which are more iconic, which may be unique to a given signer, and do not have a fixed meaning *e.g* in a SL dictionary. These are used to depict or describe actions, entities, and events among other things.

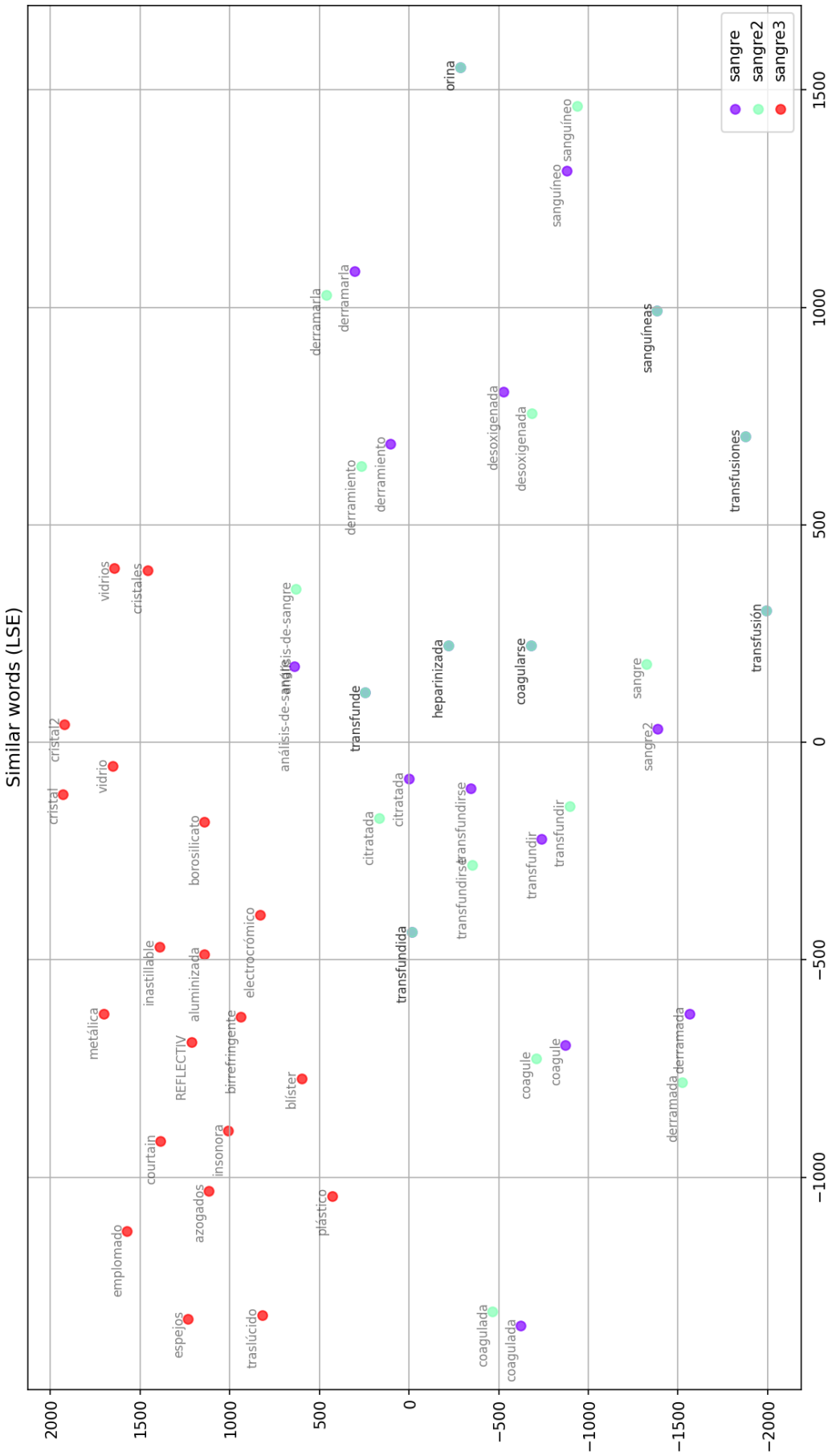

Figure 5: Top ten most similar lexemes to each gloss based on the Spanish word for “blood”

Mean + std. dev Text2Gloss	es→LSE		nl→NGT		en→ASL	
	Epoch	**BLEU**	**Epoch**	**BLEU**	**Epoch**	**BLEU**
Baseline	10000	5.55 ± 1.71	7600	12.63 ± 3.81	9000	14.74 ± 0.63
Baseline+enc	9600	5.47 ± 0.41	9400	11.54 ± 4.17	6400	16.94 ± 0.58
Baseline+dec	9600	4.30 ± 1.27	4600	16.92 ± 2.60	4800	14.30 ± 1.60
Baseline+both	10000	5.07 ± 3.15	8400	13.63 ± 0.90	7800	15.42 ± 0.33
PT+FT	7600+3200	9.12 ± 0.75	9200+1000	17.88 ± 2.34	1200+3200	18.01 ± 0.42
PT+FT+enc	7800+3000	16.50 ± 1.19	9800+3800	18.24 ± 2.11	6200+4400	18.84 ± 1.23
PT+FT+dec	5600+3400	15.40 ± 0.64	6200+1800	16.10 ± 4.64	6400+5000	17.92 ± 1.08
PT+FT+both	3800+4400	16.61 ± 0.47	6400+1600	18.20 ± 1.84	3200+4800	18.87 ± 0.69
Mean + std. dev Gloss2Text	LSE→es		NGT→nl		ASL→en	
	Epoch	**BLEU**	**Epoch**	**BLEU**	**Epoch**	**BLEU**
Baseline	3600	6.96 ± 0.73	7400	4.41 ± 0.43	5400	12.80 ± 1.42
Baseline+enc	3000	7.63 ± 0.29	7400	4.30 ± 0.15	4800	12.89 ± 0.41
Baseline+dec	4400	6.11 ± 0.59	9800	4.12 ± 0.18	5600	12.94 ± 0.75
Baseline+both	4000	7.75 ± 0.88	8400	4.10 ± 0.10	5400	12.68 ± 0.86
PT+FT	8600+2000	7.91 ± 0.69	4600+3200	3.18 ± 0.12	8800+1400	15.57 ± 0.54
PT+FT+enc	9800+3000	9.12 ± 0.53	9200+3200	3.37 ± 0.12	8400+3600	15.86 ± 0.59
PT+FT+dec	7800+1800	7.52 ± 0.29	5200+4000	3.09 ± 0.24	9400+1600	15.01 ± 0.76
PT+FT+both	8200+3400	8.49 ± 0.79	8400+4400	3.25 ± 0.11	9600+3800	16.53 ± 0.42

Table 5: Results summary for translation experiments in OpenNMT - BLEU-4 based mean and standard deviation for three runs in each experimental setup, along with the number of epochs for which the model is chosen. fi→FSL not shown as only underwent one run per setting.

Gold hyp: "HOMBRE PEQUEÑO2 REVISTA RECIBIR-MONTÓN INDX:PRO:1pl"
gloss: man small book receive-stack.$_{CL\text{-}M}$ we
ES: *Los niños nos dan los libros*

PT+FT hyp: "HOMBRE PERSONA REVISTA DAR INDX.PRO:3pl"
gloss: man person book give they

PT+FT-both hyp: "HOMBRE PEQUEÑO2 REVISTA DAR INDX.PRO:1pl"
gloss: man small book give we

Figure 6: Translation output from the Spanish sentence "The children give us the books" into LSE from the original corpus, and two model output hypotheses

Gold hyp: "JOHN NOW i:GIVE:j CHOCOLATE MOTHERwg IX-3p:j"
gloss: John now $_{REF1.}$give$_{.REF2}$ chocolate mother it
EN: *John is right now giving chocolate to mother.*

Baseline hyp: "JOHN NOW FINISH NOW SUE"
Gloss: John now already now Sue

PT+FT hyp: "JOHN NOW FINISH i:GIFT:j MOTHERwg"
gloss: John currently already $_{REF1.}$give$_{.REF2}$ mother

PT+FT-both hyp: "JOHN NOW GO-OUT CHOCOLATE MOTHERwg"
gloss: John now go-out chocolate mother

Figure 7: Translation output from the English sentence "John is right now giving chocolate to mother" into ASL, and three model output hypotheses

Gold hyp: "PADRE^MADRE ENTENDER NADA2"
gloss: parents understand nothing
ES: *A mis padres no los entendía en absoluto*

Baseline hyp: "INDX.PRO:2sg PEQUEÑO2 CÓMO"
gloss: you small how

PT+FT hyp: "PROPIO HOMBRE ESPÍRITU"
gloss: own man mind

PT+FT-both hyp: "INDX.PRO:1SG PADRE^MADRE NO"
gloss: I parents no

Figure 8: Translation model output from the Spanish sentence "As for my parents, I did not understand them at all", and three model hypotheses