Translating Spanish into Spanish Sign Language: Combining Rules and Data-driven Approaches

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

Euan McGill **Santiago Egea-Gómez** **Horacio Saggion**
Universitat Pompeu Fabra, Barcelona, Spain
{euan.mcgill,santiago.egea,
horacio.saggion}@upf.edu

Abstract

This paper presents a series of experiments on translating between spoken Spanish and Spanish Sign Language glosses (LSE), including enriching Neural Machine Translation (NMT) systems with linguistic features, and creating synthetic data to pretrain and later on finetune a neural translation model. We found evidence that pretraining over a large corpus of LSE synthetic data aligned to Spanish sentences could markedly improve the performance of the translation models.

1 Introduction

The widening of access to technology is crucial in today's highly interconnected online world, and it is important that technologies are made available across languages and for people with different needs. The World Federation of the Deaf[1] states that 70 million people communicate in one of the 400 sign languages (SLs) around the world. Jointly with the United Nations, they supported a resolution[2] in order to include the Deaf and Hard-of-Hearing (DHH) community in all matters concerning the provision of technology for them[3], respect the linguistic and cultural identity of signers, and improve access to education and services.

According to the Spanish National Confederation of Deaf People (CNSE)[4], approximately 2.3% of Spain's population experience hearing loss to some degree and a large number of them use *Lengua de Signos Española* (LSE) as their primary means of communication. Also, ethnologue estimates that there are between 45 to 75 thousand LSE signers. LSE was first described in the late 18th Century, while only recently a grammar (Rodríguez González, 2003) has been written to capture the features of the language. In 2011, LSE was recognized as an official language, and there has been a greater focus on providing resources for signers and learners.

As with other SLs, LSE is produced in the visual-spatial modality (Baker, 2015) rather than the oral-auditory modality of spoken languages. Manual and non-manual (facial expression, body position) features including the space around the signer can be articulated simultaneously to produce meaning. Whereas textual forms are well-established in spoken languages, those capturing the spatio-temporal nature of SLs including HamNoSys (Hanke, 2004) are long extant but not widely known or used by signers (Jantunen et al., 2021). The most frequently encountered representation of SLs are glosses – a lexeme-based representation using the ambient spoken language of the region where the SL is native. For example, glosses for LSE are written in Spanish. One criticism of glosses is that a great deal of semantic information is lost (Zhang and Duh, 2021). However, their linearity as text is a beneficial input format for machine learning (ML) models.

Machine translation (MT) has advanced significantly in recent years, specially thanks to the development of methods based on Deep Neural Networks, reaching quality levels comparable to humans (Hassan et al., 2018) for spoken languages. Despite these advances, MT is in its infancy when it comes to translation between spoken and sign languages or between different sign languages. In this paper we address a little researched topic in MT, that of translating between Spanish and LSE using a combination of rule-based and neural approaches. We present experiments on building MT systems between spoken Spanish and LSE using a small parallel corpus of sentences and gloss sequences. We first show a baseline system using

[1] https://wfdeaf.org/our-work/

[2] UN Resolution 72/161: "International Day of Sign Languages"

[3] This resolution emphasises the 'nothing about us without us' method of working with the DHH community.

[4] https://www.cnse.es/inmigracion/index.php?lang=en

Proceedings of the 29th International Conference on Computational Linguistic, pages 75–83
October 12–17, 2022.

only the parallel data, and then present two techniques for improving this baseline: enriching the representation of words and glosses with linguistic information; and using a large corpus of synthetic data for creating a pretrained model, and then finetuning using the original training data. As we will see, this last approach is the one with the most promising results.

The rest of this paper is structured as follows: Section 2 presents related work on LSE and SLs in general; section 3 introduces the dataset we base our research on, the ID/DL corpus; section 4 describes the different experiments we carried out with this dataset; section 5 shows the evaluation of the experiments over the test partition; and finally section 6 presents some conclusions and future work.

2 Related Work

The scarcity of linguistic resources constitutes a major barrier in the adaptation of latest technology to SLs (Yin et al., 2021). In fact, SLs are considered *extremely* low resource languages (Moryossef et al., 2021) for MT models. This section explores computational resources and systems existing for LSE, SL Translation (SLT) and processing.

2.1 LSE technologies and resources

There has been a wide range of work focusing on LSE, including resources such as image and video signbanks and lexica (Cabeza and García-Miguel, 2019; del Carmen Cabeza-Pereiro et al., 2016; Gutierrez-Sigut et al., 2016), language learning resources (Herrero-Blanco, 2009), and corpora containing full utterances for academic purposes (Porta et al., 2014). The largest barrier to create technologies on par with those available to spoken languages, one that is shared with all SLs (Bragg et al., 2019; Holmes et al., 2022), is the size and tendency towards domain-specificity in LSE parallel corpora.

Outside of static reference resources, there also exist rule-based translation systems from Spanish into LSE. Porta and colleagues (Porta et al., 2014) worked with a psycholinguistics-based corpus consisting of one SL interpreter reciting six passages translated from Spanish into LSE in varied domains. There are 229 parallel sentences in total, with 611 unique sign types. The LSE glosses are transcribed to an extent in a convention which incorporates prosodic, morphological and syntactic phenomena. This study leverages knowledge of LSE grammar, a language-agnostic dependency parser, the bilingual corpus, and the DILSE dictionary (Fundación CNSE, 2008) to form the rule-based MT system. The BLEU (Papineni et al., 2002) and Translation Error Rate (TER) (Snover et al., 2006) metrics are commonly used in MT studies. This system reported a reasonable BLEU of 30.0 and TER of 42%, especially coming from a domain-unspecific testbed.

In addition, Vegas-Cañas (Vegas Cañas et al., 2020) outlines their web-based Text2LSE system. This system is also rule-based, and translates between simple Spanish text and LSE text, or LSE videos from the ARASAAC resource[5]. Text2LSE was evaluated on 137 simple utterances, and was shown to be severely limited as 82.5% of output sentences were deemed 'errorful'. The lack of crossover between output glosses and existing signs in an LSE lexicon was the most salient factor. It is therefore important to check whether SLT outputs have a grounding in the real language.

In this work, we focus on the ID/DL corpus created by San-Segundo and colleagues (San-Segundo et al., 2008), based on utterances drawn from Spanish identity card and driving license application data. They also used it to design a statistical rule-based end-to-end (E2E) translation system from speech recognition through translation and outputting to a 3D avatar. They achieved a BLEU score of 49.4 when using them with a phrase-based statistical MT model. Using a rule-based system with 153 linguistically-motivated rules crafter by the authors and tuned specifically to the dataset, they achieved a BLEU score of 57.8. These findings are of importance for the present study, which is comparable as it is trained on the same ID/DL dataset.

2.2 Current methods in Sign Language Translation

SLT is inherently multimodal (Bragg et al., 2019), where it is necessary to incorporate audiovisual processing, speech recognition, and SL generation through technologies such as avatars. E2E systems between text and sign exist (Camgoz et al., 2020), but modular systems with intermediate representations such as Text2Gloss (Yin and Read, 2020) before transforming to a sign appear to currently yield higher accuracy (Zhang and Duh, 2021) in

[5] https://arasaac.org/

translation.

Transformer-based neural machine translation (NMT) (e.g. Klein et al. 2017; Xue et al. 2021) has been instrumental in forming the current state-of-the-art between a wide range of languages, including low-resource spoken languages. Due to the unique multimodal nature of the SLT task, as well as the status of most SLs as *extremely* low-resource languages, further strategies are necessary to perform adequate SLT. One example is data augmentation methods to boost the amount of training data available. These strategies include backtranslation (Zhou et al., 2021), and a rule-based strategies between parallel corpora (Moryossef et al., 2021). Another method is to supplement the encoder of a transformer model with linguistic information (Sennrich and Haddow, 2016). Our previous work on German-DGS[6] using linguistic feature embeddings (Egea Gómez et al., 2021) and transfer learning methods (Egea Gómez et al., 2022) result in an increase in performance of more than 5 BLEU over a baseline not incorporating linguistic information. In the present study, we propose using methods of data augmentation based on the linguistic features of LSE, as well as incorporating part-of-speech and syntactic dependency tags on input data for translation models.

3 Corpus

For our experiments, we use the ID/DL corpus (San-Segundo et al., 2008)[7], made up of 416 parallel Spanish-LSE utterances. Below, we show one example of the parallel text samples composing this corpus:

> **Spanish**: deberá tener preparadas las fotografías y documentos necesarios
> **LSE**: FUTURO TÚ OBLIGATORIO PREPARAR PLURAL FOTOGRAFíA Y PLURAL DOCUMENTO PLURAL NECESARIO

We randomly split the dataset into 266 training utterances, 75 dev utterances and 75 test utterances. Table 1 presents the data composition of the different partitions used in our experiments and the LSE Lexicon (Gutierrez-Sigut et al., 2016) for comparison.

As can be observed, the train partition contains 290 unique glosses, which cover 89.9% of dev glosses and 85.6% of test glosses. Consequently, there are a lot of glosses both in the dev and test sets that the MT models will never see at training time (out-of-vocabulary glosses). The models might overfit train patterns while some of the input sequences may not be properly learnt leading to inaccurate predictions. Also, we notice that the glosses in the LSE Lexicon seem not to be representative enough of the glosses found in the ID/DL corpus, as less than 50% of the glosses found on the train, dev and test sets are in the lexicon, which is in line with the problems mentioned in Section 2.1.

	Train	Dev	Test	Lexicon
Sentences	266	75	75	-
Total words	3153	859	917	-
Unique words	531	312	289	-
Total glosses	2952	803	885	2243
Unique glosses	290	188	181	2243

Glosses coverage between sets in %

	Train	Dev	Test	Lexicon
Train coverage	100	58.3	53.4	37.9
Dev coverage	89.9	100	64.4	35.6
Test coverage	85.6	66.9	100	45.3
Lexicon coverage	4.9	3.0	3.7	100

Table 1: The top part shows the sizes of the training, development and test splits, and the lexicon set. The bottom part shows the coverage of glosses between each pair of sets.

4 Experiments

We have carried out a series of experiments on building MT models between (spoken) Spanish and LSE. Although ID/DL corpus is not a fully comprehensive representation of LSE, it is one of the few available LSE resources with a suitable format to experiment with ML algorithms.

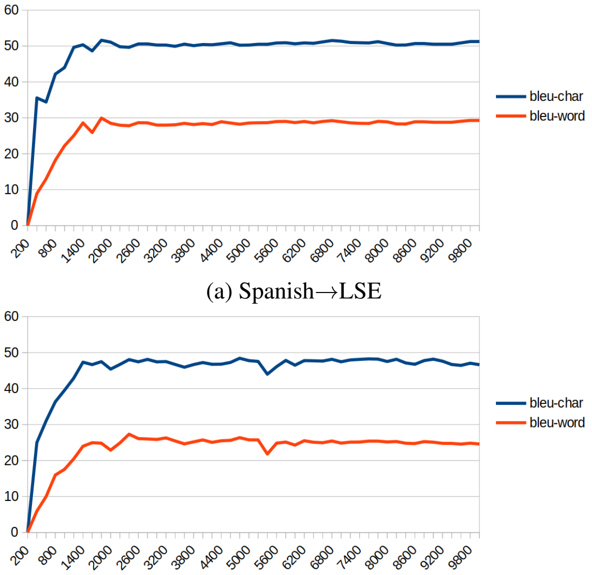
In the present work, we first create a baseline for both translation directions *LSE↔Spanish* (section 4.1); then we incorporate linguistic features to boost our MT model (section 4.2); and finally we pretrain a transformer model on synthetic data generated using data augmentation rules, and fine-tune it with the ID/DL training data (section 4.3). All the results in this section are evaluated against the ID/DL development set, while section 5 shows results over the test set.

4.1 Baseline model

In the preliminary experiment a DL model is trained using only the parallel word and gloss sequences from ID/DL. We used the Open-

[6] Deutsche Gebärdensprache (German Sign Language)

[7] Enquiries about the corpus should be addressed to: `https://www.fundacioncnse.org/`



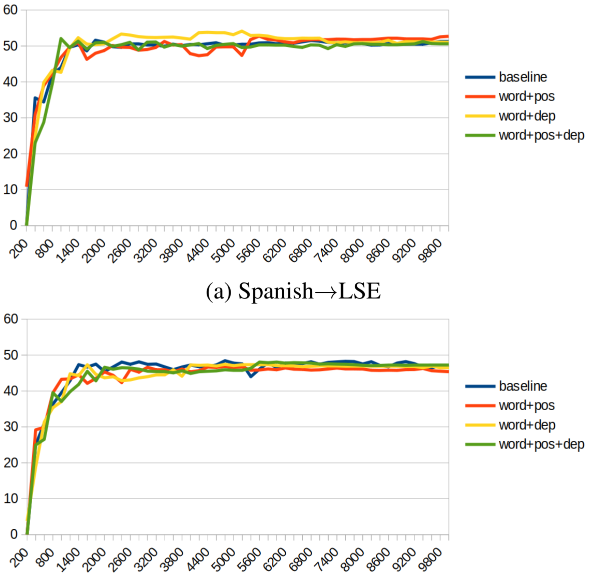
(a) Spanish→LSE

(b) LSE→Spanish

Figure 1: Performance of the baseline experiment during training, calculated over the development set.

NMT (Klein et al., 2017) system in its default configuration, consisting in a stack of Long Short-term Memory (LSTMs) layers with a general attention mechanism. The model was trained for 10,000 steps taking a snapshot every 200 steps to evaluate it against the dev corpus; this training setting is used in all experiments reported here. Fig. 1 shows the performance of this experiment on the dev set for both directions, according to the BLEU (Papineni et al., 2002) metric calculated using SacreBLEU (Post, 2018) both at word and character level, which we refer to as BLEU-word and BLEU-char respectively. The BLEU-char metric is also used in other works related to SLT (Egea Gómez et al., 2021).

Regarding *Spanish→LSE*, convergence is achieved between 2000 and 2500 steps, while for the other direction convergence happens after 2600 but the performance fluctuates more than in the other case. Both metrics (at word and character level) seem to be very correlated, but the best performance is not necessarily achieved at the same time. For example, for the *LSE→Spanish* direction, peak BLEU-char performance is 48.43 at 4800 steps, while peak BLEU-word performance is achieved much earlier, 27.32 at 2400 steps. On the other hand, for the *Spanish→LSE* direction, the best BLEU-char and BLEU-word performances are obtained at 1800 steps, 51.60 and 29.94 respectively. Since both metrics are correlated, and for the sake of better chart visualisation, in the rest of this section we report only the BLEU-char metric for the dev partition, while both metrics will be examined on test data.



(a) Spanish→LSE

(b) LSE→Spanish

Figure 2: BLEU-char performance for the models with linguistic features during training, calculated over the development set. The baseline model (in blue) is also shown for comparison.

4.2 Enriching models with linguistic features

Following (Egea Gómez et al., 2021), linguistic information is incorporated into our model in order to boost translation performances. We used the Spanish spaCy model[8] to analyse the spoken Spanish utterances, obtaining their part-of-speech (POS) and dependency parsing information (DEP). Then we trained three different models where the source text uses these combinations of features: (1) words + POS, (2) words + DEP label, and (3) words + POS + DEP label. The OpenNMT models, in this case, use separate dictionaries for words, POS and DEP features, creating separate embedding models for each of the feature spaces. Then, the embedding vectors are concatenated and fed to the LSTM network.

The experimental setting described so far can only be employed in the *Spanish→LSE* direction; because the Spanish spaCy model manages only Spanish sentences, while sign glosses follow different linguistic rules and the annotation model is not applicable to them. Even the dependency grammars and treebanks for other sign languages are

[8] https://spacy.io/models

still under development or are too small to work with (Östling et al., 2017). Therefore, in order to try the same configuration in the *LSE→Spanish* direction, we transfer POS and DEP features generated for spoken Spanish to glosses using the statistical-based alignment model fast_align (Dyer et al., 2013). We use the following rules to create silver-standard POS and DEP data for glosses:

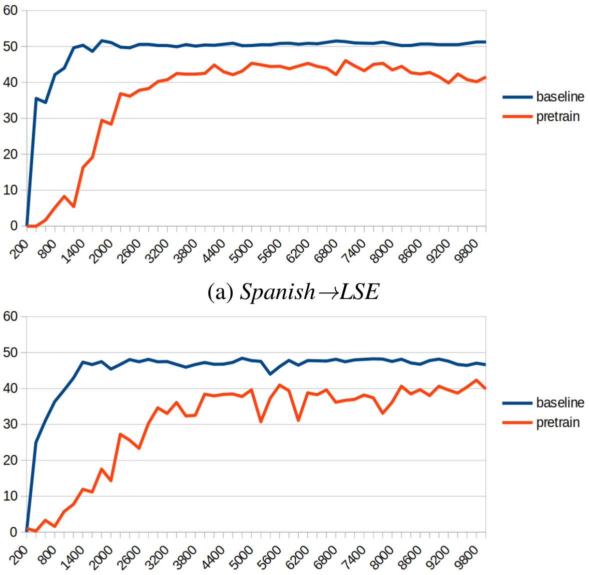
- (1) If gloss `j` is aligned to word `i`, assign the label for `i` to the gloss `j`.
- (2) If gloss `j` is not aligned to any word, assign the most common label for gloss `j` found in the gloss side of the corpus.
- (3) Otherwise, use the label `UNK` for the gloss.

This feature transfer schema is independently applied for each data partition. However, it is important to remark that in a real scenario this process will not be applicable, since DEP and POS features are annotated on gloss utterances based on their corresponding spoken ground truths, which are not available in a real scenario. Consequently, the results on *LSE→Spanish* must be seen as an unrealistic upper bound, and further research is needed to build actual POS and DEP models for LSE.

Fig. 2 shows the evolution of performance over the dev set for these experiments. We can see that all models behave in a similar way, but the word+DEP model overcomes the others in *Spanish→LSE* between steps 2600 and 5600 in up to 3 points, reaching a BLEU-char of 54.1. Conversely, this improvement is not clear for *LSE→Spanish*.

4.3 Augmenting the corpus with synthetic data

Previous work like (Moryossef et al., 2021) have shown that it is possible to use corpus augmentation strategies for improving performance of MT models in sign language scenarios. In our case, we follow a strategy with two steps: first we *pretrain* over a large set of synthetic data, and then we *finetune* using the ID/DL training set. Based on the LSE grammar (Rodríguez González, 2003) and our observations of the training data, we created a rule-based system that tries to mimic the most salient rules for getting the sequence of glosses from the corresponding sequence of words. We first obtain morphosyntactic and dependency information for a spoken sentence using spaCy, and then we use three sets of rules shown in table 2 to create a rough translation. Using these rules already yields somewhat good results on the development set: 63.42 BLEU-char and 29.73 BLEU-word, compared to 51.60 BLEU-char and 29.94 BLEU-word obtained in the baseline MT system described in section 4.1.



(a) *Spanish→LSE*

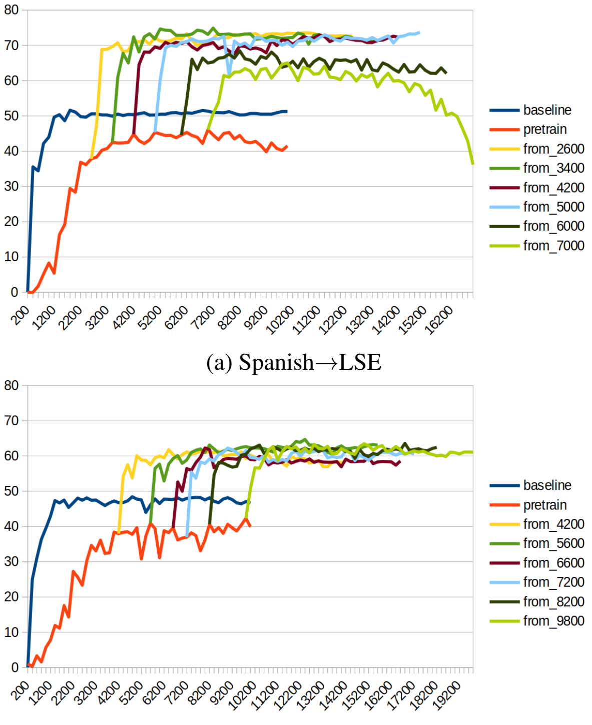
(b) *LSE→Spanish*

Figure 3: BLEU-char performance of the pretrained models (trained over the synthetic corpus), calculated over the development set. The baseline model (in blue) is also shown for comparison.

Using this rule-based system, we translated the whole Spanish set of the Ancora corpus (Taulé et al., 2008). This corpus contains 17k sentences of from newspaper text, around 500k words. After translating all the sentences, the resulting gloss sequences corpus has around 400k glosses. With this, we created a silver-standard synthetic corpus of glosses aligned to their corresponding sentences in spoken Spanish. Then we pretrained neural translation systems with this synthetic corpus for 10,000 steps in both directions. Of course, the results of these pretrained models over the ID/DL development corpus were much lower than for the rest experiments described so far, because even if the synthetic Ancora parallel set is much larger, its sentences are very different from the ones in ID/DL. However, as we will see, we can use this pretrained model as a starting point for finetuning with the ID/DL training data, which achieves much better results. Fig. 3 shows the BLEU-char performance of the pretrained models compared to the baseline model, where we can see that the performance of the pretrained model is always below the baseline model.

Inclusion of explicit morphological markers	Example
1) Add the “PLURAL” token before any plural word.	perros → PLURAL PERRO
2) Add the “FUTURO” token before any verb in future tense.	comerá → FUTURO COMER
3) Add the “TÚ” token before any verb in second person.	vienes → TÚ VENIR
4) Change a possessive determinant to “PROPIO” + the pronoun.	mi madre → PROPIO YO MADRE
Removal of words not used in LSE	**Example**
5) Remove determinants (except the possessive, which are changed by rule 4).	el perro → PERRO
6) Remove prepositions “de” and “en”.	de tarde →TARDE
Particular lexical transformations	**Example**
7) Copula words are changed to the token “SE-LLAMA”.	esto es importante → ESTO SE-LLAMA IMPORTANTE
8) Sequences whose lemmas correspond to the sequence “TENER QUE”, are changed to “NECESITAR”.	tiene que llevar → NECESITAR LLEVAR
9) Instances of “denei” are changed to “DNI”.	llevar denei → LLEVAR DNI
10) All other words are represented as their uppercase lemmas.	perros → PERRO

Table 2: Rules used in the rule-based system for creating synthetic data.



(a) Spanish→LSE

(b) LSE→Spanish

Figure 4: BLEU-char performance of the finetuned models (trained over the training set but starting from different steps of the pretrained model), calculated over the development set. The baseline model (in blue) and the pretrained model (in red) are also shown for comparison.

The pretrained model also seems to converge much more slowly than the baseline, and shows some spikes in performance at some points. We chose some of those points where performance seems to peak (six in each direction) as starting points for finetuning. We then finetuned the model using the original ID/DL training data for 10,000 more steps in each case. Fig. 4 shows the BLEU-char performance of these new models over the development set, the baseline model (blue) and the pretrained model (red) are shown for comparison. Note that there is a considerable leap in performance for all finetuned models, which start from the pretrained line and suddenly jump much higher than the baseline.

In the *LSE→Spanish* direction, the performance of all finetuned models plateau between 55 and 65 BLEU-char.

5 Results and Discussion

We chose the model that yielded the best results according to BLEU-char for each of the described experiments, and we evaluated them over the test set. Table 3 shows the results of this evaluation. The first thing to notice is that all the finetuned models behave much better than the baseline model and the models infused with linguistic features, having as much as 20 more points in BLEU-char or 15 points in BLEU-word in both directions. Besides BLUE-char and BLEU-word, we show other usual MT metrics: Meteor, TER and ROUGE-L. All these metrics also show a similar trend, having substantial improvements when using the finetuned models. Our best result for the finetuned models is a BLEU-word of 58.98, which is higher than any configuration in San Segundo’s work (San-Segundo et al., 2008).

The models that incorporated linguistic features performed similarly for the dev split. However, this performance is not reflected on the test split,

		BLEU	BLEU		ROUGE	
Direction	Experiment	char	word	Meteor	L F1	TER
Spanish→LSE	baseline	49.92	30.87	0.4382	0.4772	0.6785
	word+pos	52.23	31.99	0.4590	0.4914	0.6715
	word+dep	49.80	28.60	0.4296	0.4772	0.6746
	word+pos+dep	43.94	21.46	0.3689	0.4141	0.7387
	pretrain	42.34	9.63	0.2841	0.3748	0.7311
	from 2600	74.16	57.11	0.7139	0.7316	0.3691
	from 3400	70.93	52.12	0.6978	0.7270	**0.3424**
	from 4200	**75.42**	**58.98**	**0.7153**	**0.7351**	0.3438
	from 5000	72.16	53.82	0.6945	0.7250	0.3794
	from 6000	65.78	49.02	0.6360	0.6815	0.4007
	from 7000	67.33	47.20	0.6478	0.6718	0.4425
LSE→Spanish	baseline	46.08	24.97	0.4026	0.4206	0.7387
	word+pos	43.72	22.84	0.3746	0.4037	0.7438
	word+dep	45.03	24.35	0.3834	0.4061	0.7419
	word+pos+dep	45.16	23.66	0.3963	0.4121	0.7359
	pretrain	36.62	4.88	0.2646	0.2974	0.9568
	from 4200	**64.59**	41.02	**0.6037**	0.6047	0.4658
	from 5600	63.23	41.12	0.5946	0.6016	0.4829
	from 6600	62.71	40.15	0.5991	0.6055	**0.4582**
	from 7200	60.29	38.56	0.5773	0.5911	0.4738
	from 8200	61.35	41.46	0.6030	0.6104	0.4612
	from 9800	61.59	**41.63**	0.5940	**0.6108**	0.4700

Table 3: Results for all the experiments over the test set.

where most models achieve a few points less than the baseline. One of the models, though, seems to have some improvement over the baseline: the word+POS model in the *Spanish→LSE*. But the word+DEP model, which was the most promising on dev, did not bring any improvement over test.

In order to understand the big difference in performance achieved by the finetuned models, we measured the vocabulary coverage obtained by the synthetic data corpus created from Ancora. Table 4 shows the main statistics of the Ancora set and the union of Ancora and ID/DL training set, which was the whole set of data used for training.

The dev and test sets coverage obtained using the Ancora and the training split are much higher than using the training split alone. This is because rule 10 in Table 2 is a productive rule that can create any new gloss it needs to accommodate the words seen in the training data. Using this, systems pretrained on the Ancora set will have at least some model for almost all the glosses in the test corpus, which is an advantage over the models that have not seen any of those glosses during training. Note that, as table 1 shows, we had 14.4% out of vocabulary words with the original training corpus, and it dropped to 1.1% with the union of the training and Ancora corpora.

On the other hand, there is no guarantee that the glosses created by rule 10 are indeed valid signs, so this rule is probably fabricating glosses that have no counterpart in LSE. It would be possible to alleviate this problem using some other heuristics. One way of doing this could be obtaining the closest gloss in the embeddings space that is an actual LSE sign, but since the LSE Lexicon coverage is so low, further research is needed to get a larger set of valid glosses and signs that could lead better insights on this process.

	Ancora	Ancora+Train
Lines	17345	17611
Total words	481638	484791
Unique words	39705	39785
Total glosses	402539	405491
Unique glosses	26198	26232

Glosses coverage between sets in %

	Ancora	Ancora+Train
Train coverage	88.3	100
Dev coverage	91.5	99.5
Test coverage	92.3	98.9
Lexicon coverage	62.6	62.7

Table 4: Sizes and coverage statistics for the synthetic data corpus created from Ancora using the rule-based system. We show only the Ancora set, and the union of Ancora and the ID/DL train split.

6 Conclusions and Future Work

We presented experiments to build machine translation models between Spanish and LSE glosses. Our experiments are based on the ID/DL corpus, a small parallel set of Spanish sentences aligned with their corresponding LSE glosses, about the restricted domain of identity card and driving license renovations. Although glosses are not a full representation of all the complexities of a sign language, they are comprehensive enough and suitable for ML purposes.

First we carried out experiments on infusing linguistic features on a neural model for trying to improve its performance. The results of these experiments were mixed: on dev, the use of words combined with dependency labels seemed to improve performance, but on test the best improvements were achieved using a combination of words and POS labels.

Then we took the Spanish Ancora corpus and transformed it using a rule-based system inspired by the LSE grammar to create a synthetic parallel corpus of Spanish aligned with LSE sequences that is considerably larger than the ID/DL corpus. We found that pretraining on this synthetic corpus, and then finetuning with the original ID/DL training corpus achieves a marked performance improvement (around 20 points on BLEU-char and 15 points BLEU-word) over training using only the ID/DL training corpus. This improvement could be explained in part due to the high coverage of glosses achieved by using the synthetic data, but we have to take in consideration that the process could have also created some glosses that may have no real-world counterpart in LSE. We propose some possible improvements on the process, such as using a heuristic to find appropriate sign glosses when a nonexistent gloss is used.

Furthermore, given that the use of linguistic information showed some potential improvements in some scenarios, we would like to try combining both methods by getting linguistic information for the synthetic data as well for pretraining. Also, as the ID/DL corpus we used is rather small, we would like to see to what extent our approach generalises for other LSE corpora that belong to other domains. We also want to try our approaches on other pairs of spoken and sign languages. Finally, as the dataset is rather small, we could try to use simpler a statistical method, such as phrased-based MT, and combine it with the our rule approach to see if there are also improvements in that scenario.

Acknowledgements

This work has been conducted 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

Anne Baker. 2015. Sign languages as natural languages. In Anne Baker, Beppie van den Boegarde, Roland Pfau, and Trude Schermer, editors, *Sign Languages of the World: A Comparative Handbook*, chapter 31, pages 729–770. De Gruyter, Berlin.

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

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

Necati Cihan Camgoz, 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.

María del Carmen Cabeza-Pereiro, José Mª Garcia-Miguel, Carmen García Mateo, and José Luis Alba Castro. 2016. CORILSE: a Spanish Sign Language repository for linguistic analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1402–1407, Portorož, Slovenia. European Language Resources Association (ELRA).

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Santiago Egea Gómez, 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.

Santiago Egea Gómez, Euan McGill, and Horacio Saggion. 2021. Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *14th WS. on BUCC*, pages 18–27, Online.

Fundación CNSE. 2008. Diccionario normativo de la lengua de signos española.

Eva Gutierrez-Sigut, Brendan Costello, Cristina Baus, and Manuel Carreiras. 2016. LSE-Sign: A lexical database for Spanish Sign Language. *Behaviour Research Methods*, 48:123–137.

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

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *ArXiv*, abs/1803.05567.

Ángel Herrero-Blanco. 2009. *Gramática Didáctica de la Lengua de Signos Española*. SM, Madrid.

Ruth Holmes, Ellen Rushe, Frank Fowley, and Anthony Ventresque. 2022. Improving Signer Independent Sign Language Recognition for Low Resource Languages. In *Seventh International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual*, Marseille, France.

Tommi Jantunen, Rebekah Rousi, Päivi Raino, Markku Turunen, Mohammad Valipoor, and Narciso García. 2021. *Is There Any Hope for Developing Automated Translation Technology for Sign Languages?*, pages 61–73.

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

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation.

Robert Östling, Carl Börstell, Moa Gärdenfors, and Mats Wirén. 2017. Universal dependencies for swedish sign language. In *Proceedings of the 21st Nordic conference on computational linguistics*, pages 303–308.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA. ACL.

Jordi Porta, Fernando López-Colino, Javier Tejedor, and José Colás. 2014. A rule-based translation from written Spanish to Spanish Sign Language glosses. *Computer Speech & Language*, 28:788–811.

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

María Ángeles Rodríguez González. 2003. Lenguaje de signos.

Rubén San-Segundo, R Barra, R Córdoba, L Fernando D'Haro, F Fernández, Javier Ferreiros, Juan Manuel Lucas, Javier Macías-Guarasa, Juan Manuel Montero, and José Manuel Pardo. 2008. Speech to sign language translation system for Spanish. *Speech Communication*, 50(11):1009–1020.

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

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. pages 223–231.

Mariona Taulé, M Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Sara Vegas Cañas, Miguel Rodríguez Cuesta, and Alejandro Torralbo Fuentes. 2020. Text2LSE: Traductor Texto a Lengua de Signos Española (LSE).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *NAACL 2021*, pages 483–498, Online. ACL.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing.

Kayo Yin and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *COLING 2020*, pages 5975–5989, Online. ICCL.

Xuan Zhang and Kevin Duh. 2021. Approaching sign language gloss translation as a low-resource machine translation task. In *AT4SSL 2021*, pages 60–70, Online. AMTA.

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