Part-of-Speech tagging Spanish Sign Language data and its applications in Sign Language machine translation

Euan McGill[1] **Luis Chiruzzo**[2] **Santiago Egea Gómez**[1] **Horacio Saggion**[1]
[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]Universidad de la República, Montevideo, Uruguay
euan.mcgill@upf.edu, luischir@fing.edu.uy,
santiago.egea@upf.edu, horacio.saggion@upf.edu

Abstract

This paper examines the use of manually part-of-speech tagged sign language gloss data in the Text2Gloss and Gloss2Text translation tasks, as well as running an LSTM-based sequence labelling model on the same glosses for automatic part-of-speech tagging. We find that a combination of tag-enhanced glosses and pretraining the neural model positively impacts performance in the translation tasks. The results of the tagging task are limited, but provide a methodological framework for further research into tagging sign language gloss data.

1 Introduction

Lengua de Signos Española (LSE)[1] is a Sign Language (SL) of Spain with an estimated 45-75,000 signers (Eberhard et al., 2022) in that country. Creating SL technology is an inherently multimodal task (Bragg et al., 2019) as SLs are produced in the visual-spatial modality (Baker, 2015). It is also a challenging as SLs lack a commonly-used written representation (Jantunen et al., 2021). Instead, researchers rely on non-standardised glosses which may be considered a suboptimal representation of the rich semantics of signs (Núñez-Marcos et al., 2023). However, while there is not enough data available to build large machine learning (ML)-based end-to-end (E2E) models, they are a valuable tool and can be enriched with linguistic information *post-hoc* by researchers (Egea Gómez et al., 2021).

Compared to spoken languages, tools and techniques for natural language processing (NLP) in sign language glosses are less established. This is particularly true with regards to linguistic processing techniques (Yin et al., 2021) such as tagging and parsing. This paper focuses on the part-of-speech (PoS) tagging of LSE, its application in linguistically-informed SL machine translation (SLT), and the next steps in linguistic processing for SLs. It also uses data from the iSignos (Cabeza and García-Miguel, 2019) corpus, and introduces a method to linearise and lexicalise the glosses found there.

In Section 2 we describe previous relevant work on LSE, and information on linguistic tagging for SLs. Next, Section 3 describes the pre-processing of gloss data, before Section 4 proposes a novel approach to incorporate manual PoS tags into neural machine translation models - as well as another experiment training a naïve PoS-tagger on these glosses. This section, and Section 5 provide insight into experimental results, and their implication in the wider field.

2 Previous work on LSE

There exists only a small body of suitable LSE corpora and parallel resources for use in NLP tasks, which means this SL could be considered an *extremely* low resource language (Moryossef et al., 2021). Despite this scarcity, researchers have adopted strategies to mitigate this limitation when building processing and translation tools for LSE.

For example, San-Segundo et al. (2008) explored two approaches for Spanish→LSE translation. Researchers and experts crafted 154 rules to transform Spanish text into LSE glosses, and also built a completely statistical translation model. The linguistically-informed rule-based model outperformed the statistical model on a small, domain-restricted dataset[2]. This corpus is labelled with lexeme-based glosses, which resemble words in the ambient spoken language (Spanish) and are similar to representations found

[1]Spanish Sign Language

[2]< 1k parallel utterances on a national identity card renewal dataset (ID/DL)

Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023), pages 70–76, May 22, 2023

in SL dictionaries and Signbanks - for example DILSE (Fundación CNSE, 2008). In contrast, work by Porta et al. (2014) uses glosses from a storytelling corpus which contain grammatical and phonological information such as handshape and mouthing. Later work (Chiruzzo et al., 2022) also took a hybrid approach for Text2Gloss (T2G) and Gloss2Text (G2T) transformation between Spanish and LSE, using a rule-based synthetic data to pretrain a neural translation model.

2.1 Universal Dependencies and SLs

Treebanks based on the standardised Universal Dependencies (UD)[3] framework are a valuable resource for NLP applications. The first publicly-available UD treebank for a SL was for Swedish Sign Language (Östling et al., 2017), with subsequent work on Italian Sign Language (Caligiore, 2020). Work is ongoing on a treebank for LSE (García-Miguel and Cabeza, 2020) - also based on the iSignos/RADIS dataset. As this work is not yet complete or available, we craft our own PoS labels for this research. In addition, we describe an experiment on running an LSTM-based PoS tagger on our manually tagged data.

3 Dataset and pre-processing

iSignos (Cabeza and García-Miguel, 2019) is a dataset consisting of parallel Spanish and LSE gloss/video data from 12 unique signers in multiple dialogue settings: Spontaneous conversation, storytelling, elicitation and a web drama. The total duration of the SL videos is $02\text{h}45$, consisting of $2.7k$ utterances, $16.5k$ words, $7.5k$ glosses, and $1,356$ unique glosses. Annotation conforms to the RADIS (Pérez et al., 2019) project guidelines.

Each of the 24 iSignos videos are annotated with a gloss channel for each hand, with each utterance being timestamped and having a Spanish and English translation. iSignos glosses contain lexical, semi-lexical, and non-lexical signs.[4]

Linearised, lexical glosses were generated by implementing the following steps:

- Count number of glosses in the left and right-hand channels, and the greater number is assigned as the dominant hand
- Append co-occurring glosses on each hand into one gloss separated by an underscore '_'
- If the glosses are equal, retain only one instance of the gloss name
- If the glosses differ, remove the underscore and retain the dominant-non-dominant gloss ordering as seen in Östling et al. (2017)
- Remove all semi- and non-lexical signs including paralinguistic gestures, buoys, listing buoys, semantic clarifications, unintelligible signs, etc.
- Remove all markers indicating a sign was fingerspelled or a sign name ('DT:, DL:, SN:')
- Replace numbers in brackets with a dashed affix to indicate number inflection (Herrero-Blanco, 2009), e.g. MES(CINCO)→MES-CINCO[5]
- Remove markers for phonological variance and manual quality, e.g. PADRE(2p)→ PADRE, PERDER3→PERDER
- Remove markers indicating classifier predicates, types thereof, and manual quality and then capitalise remaining gloss e.g. cl.m(5d>Q):coger+guardar-fruta→COGER+GUARDAR-FRUTA
- Remove tags indicating an index/pointing sign for pronomials, locatives, and demonstratives and supplete them with an equivalent Spanish-derived gloss - e.g. INDX.PRO(2p):1pl→NOSOTROS, INDX.DEM→ESTE, INDX.LOC→AQUÍ

Then, all lexical glosses were assigned a UD PoS tag by two annotators based on their correspondence with the meaning of their Spanish translation, as well as the PoS tags and dictionary definitions in LSE reference resources (Herrero-Blanco, 2009; Gutierrez-Sigut et al., 2016; Cabeza and García-Miguel, 2019). Inter-annotator agreement was 88.1%, based on a sample of 6.0% of the data. Proper nouns ($N = 165$) were also given a named-entity recognition (NER) tag in the BIES scheme[6] for use in the automatic PoS-tagger experiment. Appendix 1 shows an example of this process start to finish.

More detail, including all necessary scripts and steps to reproduce this experiment are found on Github.[7]

[3] https://universaldependencies.org/

[4] A full discussion of what makes a lexical sign is beyond the scope of this paper, consult Pérez et al. (2019)

[5] English translations in order: "Five months, father, to lose, to pick up and hold fruit, we/us, this/that, here"

[6] Beginning, Inside, End, Singleton

[7] https://github.com/LaSTUS-TALN-UPF/nmt-lse-es

4 Experiments and findings

This section presents our experiments on automatic PoS-tagging for glosses, and the use of these PoS-tags to improve MT systems. For these experiments, we split the corpus 64%-18%-18% for training, development and test.

4.1 Automatic PoS-tagging

On larger datasets, it would be unfeasible to manually PoS tag gloss data. As such, we used NCRF++ (Yang and Zhang, 2018)[8] which is a customisable LSTM-based neural sequence labelling toolkit to generate predicted PoS and NER tags. We experimented with using Spanish word2vec embeddings (Cardellino, 2019) in the sequence labelling process.

The overall accuracy of the NCRF++ PoS tagger on our data is 70.88% with no word embeddings, and 55.67% with Spanish word2vec embeddings. It appears that including embeddings had an adverse effect on tagging.

Table 1 shows accuracy, precision, recall and F1-score statistics for each PoS category in the no embeddings experimental setup. The distribution of tags in the training and dev. sets are shown in Appendix 2. It seems that this naïve model is biased towards the more frequently-occurring grammatical categories (from the manual tags), as NCRF++ only predicts three categories more than once in the iSignos test set data. Nouns and pronouns are markedly overpredicted, with adverbs being underpredicted. This observation is borne out in good recall figures for nouns, verbs and pronouns - but lower precision. As for NER, the test set had a low number of proper nouns of which none were correctly identified by NCRF++. Therefore, it is not possible to do further analysis on this task.

PoS	Man.	NCRF	Acc	Pre	Rec	F1
ADJ	83	0	0.00	N/A	0.00	N/A
ADP	24	0	0.00	N/A	0.00	N/A
ADV	**112**	**1**	**0.00**	**0.00**	**0.00**	**N/A**
AUX	3	0	0.00	N/A	0.00	N/A
CCONJ	1	0	0.00	N/A	0.00	N/A
INTJ	7	0	0.00	N/A	0.00	N/A
NOUN	**573**	**702**	**0.76**	**0.69**	**0.85**	**0.76**
NUM	22	0	0.00	N/A	0.00	N/A
PRON	**96**	**166**	**0.58**	**0.46**	**0.79**	**0.58**
PROPN	6	0	0.00	N/A	0.00	N/A
SCONJ	0	0	N/A	N/A	N/A	N/A
VERB	**467**	**525**	**0.86**	**0.81**	**0.91**	**0.86**
All	1394	1394	0.7088			

Table 1: Comparison of the distribution of PoS tags generated by manual tagging, and the NCRF++ tagger, and performance statistics for each PoS which appears in the corpus

4.2 Text2Gloss and Gloss2Text Translation

We ran a series of experiments to evaluate whether MT systems benefit from injecting PoS tags into neural models. Following (Chiruzzo et al., 2022), we train models using the OpenNMT (Klein et al., 2017) tool, based on an LSTM attention, trying different configurations of data. We carried out experiments in both directions T2G and G2T. As a baseline, we train a LSTM only using word (T2G) or gloss (G2T) tokens as input features; and compare it against models that aggregate our PoS tags. To incorporate the PoS features to the model, OpenNMT generates a new embedding table and combines them and the word embeddings according to three rules: Concatenation, sum or using a multilayer perceptron (MLP). This results in in a total of eight experimental settings.

Finally, for each experiment we also tried a pretraining and fine-tuning with silver standard data approach. This pretraining approach is based on (Chiruzzo et al., 2022), where they describe a way to generate a version of the AnCora (Taulé et al., 2008) corpus in LSE glosses through a simple rule-based system. They pretrain an MT system with this data, and fine-tune it using the ID/DL (San-Segundo et al., 2008) corpus data, obtaining improvements in all metrics compared to the models without pretraining. In our case, we use the same approach for generating an LSE version of AnCora, but we also obtain the original PoS for AnCora words and generate PoS associated to the AnCora LSE glosses. In this way, we can replicate this approach for the eight experimental configurations.

In these experiments, we evaluate the model using BLEU and select the best-performing weights to compute the performances on the test data. The results reported here correspond to the metrics obtained on the test partition assuming the following metrics: BLEU (calculated using SacreBLEU (Post, 2018) with international tokenisation), ChrF (Popović, 2015), METEOR, and ROUGE$_L$-F1.

The results of our experiments are shown in Table 2. First, note that the pretraining and fine-tuning approach improved the scores for all

[8] https://github.com/jiesutd/NCRFpp

Direction	Use of PoS	Finetuning	BLEU	ChrF	METEOR	ROUGE$_L$
T2G	none	no	10.10	0.260	0.140	0.200
	none	yes	11.77	**0.317**	**0.183**	0.269
	concat	no	10.66	0.267	0.144	0.208
	concat	yes	**12.14**	0.282	0.158	0.228
	sum	no	10.19	0.260	0.140	0.198
	sum	yes	11.81	0.315	0.182	0.263
	mlp	no	10.07	0.265	0.139	0.198
	mlp	yes	10.47	0.307	0.179	**0.274**
G2T	none	no	5.88	0.225	0.176	0.211
PoS	none	yes	9.03	**0.266**	0.215	0.244
annotated	concat	no	6.87	0.229	0.176	0.207
manually	concat	yes	**9.91**	**0.266**	**0.217**	**0.247**
	sum	no	2.88	0.163	0.113	0.149
	sum	yes	6.93	0.243	0.193	0.221
	mlp	no	3.56	0.178	0.133	0.170
	mlp	yes	6.89	0.245	0.202	0.233
G2T	concat	no	6.93	0.225	0.171	0.200
PoS	concat	yes	**10.22**	**0.263**	**0.214**	**0.244**
predicted	sum	no	1.87	0.155	0.102	0.137
with	sum	yes	7.39	0.225	0.170	0.192
NCRF++	mlp	no	3.77	0.183	0.140	0.175
	mlp	yes	5.64	0.234	0.185	0.212

Table 2: Results of the MT experiments over the test corpus. The second column indicates is PoS information was used, and how its information was combined. The third column indicates whether the pretraining and fine-tuning approach was used.

metrics in all the variants of the experiments. Although the performance is in general inferior Chiruzzo et al. (2022), probably because the iSignos corpus is larger and less domain-specific than ID/DL, it is interesting to see that pretraining with AnCora silver data still yields marked improvements in this corpus. The best models in both directions, according to BLEU score, are the ones that use PoS information, and combine them through the concatenation method. However, using other combination methods yielded lower performances in the G2T direction. More experiments are needed to understand why this could be the case.

In general, it seems that PoS information could be leveraged by the models in order to make better predictions. These tags possibly also tackle the disparity between the total number of LSE glosses ($7.5k$) and Spanish words ($16.5k$) in the corpus. On top of there being information loss in the gloss representation, such a difference in total tokens is surely a challenge for translation models. When we compare using the manually annotated PoS, with the same experiments with PoS predicted with the NCRF++ method from section 4.1, we can see that the use of NCRF++ predictions in combination with the fine-tuning approach does not hinder the performance, and for some of the methods the results are even slightly better.

5 Future work

This study uncovers some interesting findings, and provides the means of generating a sizeable parallel resource for Spanish-LSE, as well as English for most of the iSignos videos.

By way of future research directions, it would be interesting to test this method on the ID/DL corpus - or non-LSE SL datasets - in order to make these results comparable with San-Segundo et al. (2008) and Chiruzzo et al. (2022). In addition, it would be desirable to run other PoS taggers on this dataset for comparability such as the more recent RoBERTa 0-shot tagger (Bujel et al., 2021) or the HMM-based Apertium PoS tagger (Sánchez-Martínez et al., 2007) which is tailored towards low-resource languages by means of the Baum-Welch algorithm. In order to robustly confirm differences between experimental setups, future studies would benefit from statistical significance testing as advocated in Koehn (2004).

More generally, further work on grammatical parsing in SLs would be beneficial. There is debate as to whether LSE has PoS in the traditional sense (Rodríguez González, 2003), and that these categories are more flexible across SLs. Regardless, further studies are vital to increasing the volume of resources available, and to answer the call for more computational linguistic based resources for these vibrant languages.

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.

This work is partly supported by the Spanish State Research Agency under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M).

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. https://doi.org/10.1145/3308561.3353774 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.

Kamil Bujel, Helen Yannakoudakis, and Marek Rei. 2021. Zero-shot sequence labeling for transformer-based sentence classifiers. In *Workshop on Representation Learning for NLP*.

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

Gaia Caligiore. 2020. https://doi.org/10.13140/RG.2.2.16359.06565 *Universal Dependencies for Italian Sign Language: a treebank from the storytelling domain*. Ph.D. thesis.

Cristian Cardellino. 2019. https://crscardellino.github.io/SBWCE/ Spanish Billion Words Corpus and Embeddings.

Luis Chiruzzo, Euan McGill, Santiago Egea Gómez, and Horacio Saggion. 2022. Translating spanish into spanish sign language: Combining rules and data-driven approaches. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 75–83.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. http://www.ethnologue.com/ *Ethnologue: Languages of the World*, twenty-fifth edition. SIL Int., Dallas, TX, USA.

Santiago Egea Gómez, Euan McGill, and Horacio Saggion. 2021. https://aclanthology.org/2021.bucc-1.4 Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 18–27, Online (Virtual Mode). INCOMA Ltd.

Fundación CNSE. 2008. https://fundacioncnse-dilse.org/ Diccionario normativo de la lengua de signos española.

José M. García-Miguel and Carmen Cabeza. 2020. https://doi.org/10.35869/hafh.v22i0.1657 Hacia un treebank de dependencias para la lse. *Hesperia: Anuario de Filología Hispánica*, 22:111–143.

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

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

Tommi Jantunen, Rebekah Rousi, Päivi Raino, Markku Turunen, Mohammad Valipoor, and Narciso García. 2021. https://doi.org/10.31885/9789515150257.7 *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. https://www.aclweb.org/anthology/P17-4012 OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn. 2004. https://aclanthology.org/W04-3250 Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. https://arxiv.org/abs/2105.07476 Data augmentation for sign language gloss translation.

Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. https://doi.org/https://doi.org/10.1016/j.eswa.2022.118993 A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

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

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Jordi Porta, Fernando López-Colino, Javier Tejedor, and José Colás. 2014. https://doi.org/10.1016/j.csl.2013.10.003 A rule-based translation from written Spanish to Spanish Sign Language glosses. *Computer Speech & Language*, 28:788–811.

Matt Post. 2018. https://doi.org/10.18653/v1/W18-6319 A call for clarity in reporting BLEU scores. In *3rd Conf. on MT*, pages 186–191, Belgium, Brussels. ACL.

Ania Pérez, José M. García-Miguel, and Carmen Cabeza. 2019. https://doi.org/10.34630/sensos-e.v6i1.3527 Corpus annotation for studying grammatical expression of events: notes about the design of radis project. *Sensos-e*, 6(1):40–61.

María Ángeles Rodríguez González. 2003. www.cervantesvirtual.com/nd/ark:/59851/bmc08633 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. https://doi.org/10.1016/j.specom.2008.02.001 Speech to sign language translation system for Spanish. *Speech Communication*, 50(11):1009–1020.

Felipe Sánchez-Martínez, Carme Armentano-Oller, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada. 2007. Training part-of-speech taggers to build machine translation systems for less-resourced language pairs. *Proces. del Leng. Natural*, 39.

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)*.

Jie Yang and Yue Zhang. 2018. http://aclweb.org/anthology/P18-4013 Ncrf++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. https://arxiv.org/abs/2105.05222 Including signed languages in natural language processing.

Appendix 1

(See over)

Appendix 2

Distribution of PoS categories in the training, development, and test sets:

PoS	Train	Dev	Test
ADJ	292	77	83
ADP	18	19	24
ADV	440	158	112
AUX	35	2	3
CCONJ	9	7	1
INTJ	98	30	7
NOUN	1553	391	573
NUM	91	75	22
PRON	465	157	96
PROPN	109	51	6
SCONJ	7	0	0
VERB	1675	353	467
All	4792	1320	1394

Dominant	PERRO	SN:Luis	VER	cl.m(5d)tocar-colmena		
Non-dominant				cl.m(5d)tocar-colmena		
LexGloss	PERRO	LUIS	VER	TOCAR-COLMENA		
EN-gloss	dog	Luis	see	playing_with_beehive.clp-movement		
PoS	NOUN	PROPN	VERB	VERBCL		
Dominant	PADRE^MADRE	G	ENTENDER	NADA2	ABUELO	TAMPOCO(2M)
Non-dominant		G	B:G			TAMPOCO(2M)
LexGloss	PADRE^MADRE		ENTENDER	NADA		TAMPOCO
EN-gloss	parents		understand	nothing	grandpa	neither
PoS	NOUN		VERB	ADVERB	NOUN	ADVERB

Figure 1: Appendix 1 - Example of linearisation and lexical glossing process on two sentences