Machine Translation between Uruguayan Sign Language (LSU) Glosses and Spanish

Traducción Automática entre Lengua de Señas Uruguaya (LSU) y Español

Luis Chiruzzo, Bruno Pérez Rodríguez, Juliana Bianchi-Madera, Martín Dutra, Roberto Aguirre
Universidad de la República
Montevideo, Uruguay
luischir@fing.edu.uy, bperez@fpscico.edu.uy, jbianchi@fpsico.edu.uy,
m.dutra@psico.edu.uy, zimmer20uy@gmail.com

Abstract: This paper presents the construction of a parallel corpus for Uruguayan Sign Language (LSU) represented as glosses and spoken Spanish, and some experiments on machine translation for this language pair. The best MT models in our experiments are obtained by fine-tuning NLLB, and achieved 16.55 BLEU in the LSU to Spanish direction, and 17.90 BLEU in the opposite direction. Our manual analysis of the results shows that the translations are not perfect but are already showing good grasp of the ideas of the text being translated.
Keywords: Sign Language Processing, Uruguayan Sign Language, Machine Translation

Resumen: Este artículo presenta la construcción de un corpus paralelo para Lengua de Señas Uruguaya (LSU) representada como glosas y español oral, junto con experimentos de traducción automática para este par de lenguas. Los mejores modelos de traducción de nuestros experimentos se obtuvieron mediante el ajuste fino del modelo NLLB, llegando a 16.55 de BLEU en la dirección LSU a español, y 17.09 de BLEU en la dirección opuesta. Nuestro análisis manual de los resultados muestra que las traducciones aún no son perfectas, pero se puede ver que el sistema ya captura bien las ideas detrás del texto a traducir.
Palabras clave: Procesamiento de Lenguas de Señas, Lengua de Señas Uruguaya, Traducción Automática

1 Introduction

Sign languages are natural languages that use the visual-spatial modality instead of the oral-auditory modality (Yin et al., 2021; Baker, 2016). They are full natural languages used by millions of people throughout the world, especially relevant (but not only) to the deaf and hard-of-hearing population. Each country has its own sign language (on occasions more than one, like in Spain), and even in countries where the most prevalent oral language is the same (e.g. Spain and Uruguay both use Spanish as main oral language), their corresponding sign languages might be very different, like the Spanish Sign Language (Lengua de Signos Española, LSE) and the Uruguayan Sign Language (Lengua de Señas Uruguaya, LSU).

According to the latest Uruguayan census, there are around 120,000 hard-of-hearing people and more than 30,000 deaf people in the country,[1] and many of them communicate mainly using LSU. It is very important to have tools to process this language that enable signers to be included in the digital world, as there might be many situations in which they cannot access relevant information (e.g. legal documents, medical treatments, etc.), because not all documents are translated to LSU and it is very hard to have LSU interpreters available at all times. It is therefore very important to create new resources for this language, which could collaborate with the social inclusion and narrow the inequality gap for this population.

There is currently an ongoing project for building a relatively big corpus of LSU an-

[1] https://www.gub.uy/ministerio-salud-publica/comunicacion/noticias/personas-sordas-hipoacusia-primera-vez-ministerio-salud-publica

notated with fine-grained linguistic information (Fojo et al., 2024). The aim of this project is to annotate a set of videos that contain audio in spoken Spanish (of the Rioplatense variety, spoken in Uruguay) and LSU signers or interpreters. These videos are being annotated with linguistic information such as sign segmentation, glosses, and other relevant data.

In this work, we present machine translation (MT) experiments between LSU and spoken Spanish using the data we have annotated so far. In order to do this, we built a manually curated LSU - spoken Spanish corpus (see Section 3), and performed several rounds of experiments with different MT technologies and data augmentation techniques to improve performance. Section 4 describes these experiments showing their performance on the development split, while Section 5 discusses the results of the models over the test split.

2 Related Work

Recent advances in NLP such as Large Language Models (LLMs) and their subsequent applications like ChatGPT (Brown et al., 2020) and Gemini (Team et al., 2023) have greatly improved the processing performance for oral languages, but sign languages have not seen a similar level of development (Núñez-Marcos, Perez-de Viñaspre, y Labaka, 2023). For example, high quality MT systems, parsers, or language models, which are taken for granted for many oral languages, simply do not exist for sign languages.

One of the main reasons for the low performance for these languages is the lack of data, especially parallel and annotated corpora. All sign languages are considered low-resourced or extremely low-resourced (Moryossef et al., 2021; Joshi et al., 2020), and more efforts are needed to collect, curate, and build resources for them, and at the same time better methods that could generalize with little data are necessary, for example complementing with data augmentation techniques.

An important issue when working with sign language representations is that there is no universally agreed standard written notation. One technique for writing down these languages is the glossing method. Sign language glosses are an annotation technique that works as a bridge between two languages (in this case one sign language and one spoken language), in which concepts form the spoken language are used to label signs. This technique is far from perfect, as part of the gesture and rhythm information cannot be annotated accurately (Zhang y Duh, 2021), but still contains a lot of rich information that can be used for processing and is widely used for MT and in particular Sign Language Translation (SLT) research (Núñez-Marcos, Perez-de Viñaspre, y Labaka, 2023). However, we must take in consideration that the focus on SLT using glosses should be part of a larger pipeline that includes recognizing signs and labeling glosses (to go from sign language into spoken language), and generating signs from glosses (to go from spoken language into sign language).

In this work, we will focus on the particular stage of the pipeline of translating between gloss segments and spoken sentences, as much of the SLT efforts done in the past. Because of the low-resource nature of sign languages, data augmentation techniques have been tried to enhance the little available data. For example, (Moryossef et al., 2021) uses a rule-based data augmentation to create synthetic data for Americas Sign Language (ASL) and German Sign Language (DGS), translating into English or German respectively. Similary, (Chiruzzo et al., 2022) focuses on LSE, and uses a simple rule-based translation method to build a parallel synthetic corpus of LSE - Spanish by translating the AnCora (Taulé, Martí, y Recasens, 2008) corpus. The MT models were trained was done in two phases: first pretrain with the AnCora synthetic data, then choose some pretraining checkpoints, and finally perform a fine-tuning starting on those checkpoints but with real data instead of synthetic. This two step process yielded interesting improvements, beating the performance of the standard training by several points.

Another interesting technique is described in (McGill, Chiruzzo, y Saggion, 2024), where they focus on using static word embedding models to improve the representations of ASL, LSE, Finnish Sign Language (FinSL) and Dutch Sign Language (NGT). Word embedding techniques are robust techniques to generate static word and token representations with interesting semantic properties, with the most popular ones being word2vec (Mikolov et al., 2013), GloVe (Pennington, Socher, y Manning, 2014), and Fast-

File	Content	Length	Pairs	Glosses	Words
20150101_CERESO_PINOCHO_P1	Literary	4:19	72	245	551
20150101_CERESO_PINOCHO_P2	Literary	5:13	70	288	686
20211004_CERESO_LAZARILLO	Literary	24:56	140	1972	3188
20220513_CERESO_ABSOLUTISMO	Education	14:18	89	1292	1751
20240531_CERESO_EXPRESION_CORPORAL	Education	29:14	187	1718	3628
20220714_CANAL_5_PERIODISTAS	Journalistic	1:36:35	546	6560	8535
20220801_CANAL_5_NOTICIAS_CENTRAL	Journalistic	1:50:00	225	3094	5491
20220508_TV_CIUDAD_SOBRE_CIENCIA_P1	Education	13:58	89	1043	1632
20220508_TV_CIUDAD_SOBRE_CIENCIA_P2	Education	22:53	124	1366	2201
20220508_TV_CIUDAD_SOBRE_CIENCIA_P3	Education	15:20	101	939	1414
20221003_TV_CIUDAD_CIUDAD_VIVA_P1	Journalistic	14:31	119	833	2316
20221003_TV_CIUDAD_CIUDAD_VIVA_P2	Journalistic	20:36	133	1224	2979
Total		6:11:53	1895	20574	34372

Table 1: Composition of the videos used in the construction of the parallel LSU - spoken Spanish corpus.

Text (Bojanowski et al., 2017). But as there is not enough data to build robust embeddings representations for these sign languages, they proposed to leverage existing embeddings for the oral languages to bootstrap embeddings for the sign language, by manipulating the spoken embeddings in several ways to map the signs. Then these representations are used as starting points for training neural MT models, obtaining improvements for some of the languages, considering that whether this method improves the results largely depends on the quality of the embeddings collection.

3 LSU-Spanish Parallel Corpus

The LSU corpus that is under construction (Fojo et al., 2024) contains videos with examples of use of LSU: some of them are TV live streams with simultaneous LSU translation by an interpreter; others are study materials or films created specifically to document LSU. Their characteristics are quite different, as the language produced during simultaneous interpretation and the language produced spontaneously by native signers in a non-controlled situation might work quite differently. However, as they are all valid examples of LSU, we decided to work with them together without distinction.

For this work, we obtained the 12 videos that had been manually annotated with sign segmentation and glosses so far. Table 1 shows a summary of the annotated videos used for the construction of the LSU - spoken Spanish parallel corpus. These are all the videos annotated with LSU glosses that also have a corresponding spoken Spanish audio track or subtitles. When the subtitles were incorporated into the file, we extracted them from the ELAN (Wittenburg et al., 2006) file or manually typed from the video. But in most cases it was necessary to do the extra pre-processing step of transcribing the audio from the file using the Whisper (Radford et al., 2023) tool, in particular the Whisper-X (Bain et al., 2023) version, and then doing a manual inspection to correct errors in the automatic transcription.

In this way, we obtained a set of videos of around 6 hours, where all videos have the following information:

- Signer producing LSU (either native signer or bilingual interpreter).
- Translation to spoken Spanish of the LSU signs.
- Individual annotation of each sign as LSU glosses.

Using this information, we were able to build a parallel LSU - spoken Spanish corpus with sentences in both languages. But there is another problem that needs to be solved before this: timestamps of LSU glosses and spoken Spanish do not always align. This can happen partly because the original spoken subtitle timestamps are not perfect, but the problem gets worse in videos with a LSU interpreter of a live show. In these shows, you first have the auditory signal in spoken Spanish (for example the newscaster narrates a piece of news), and then the interpreter makes a translation to LSU and produces the signs equivalent to that Spanish sentence. This produces a delay that generally goes from some milliseconds to a few seconds between the audio signal and the corresponding

Spanish	LSU
esto revela algunos resultados que encaminan la investigación .	PT:DET VERDAD ALGUNXS RESULTADO INFORMACIÓN CAMINO
así sería más sencillo clasificar , claramente y sin necesidad de ninguna equivocación .	PORQUE FÁCIL DIVIDIR CLARX SEPARAR-PL NO-NECESITAR ERROR NINGUNX
¿ se precisan más recursos económicos o una mejor dirección ?	NECESITAR MÁS COSAS DINERO O MEJOR DIRIGIR GESTO:DUDA
ahora , ¿ cuál es el problema ?	CUÁL PROBLEMA G:DUDA
allí creó la aplicación para que sea funcional a los profesores .	PT:LOC CREAR APLICACIÓN FS:APP QUERER SERVIR PROFESOR

Table 2: Some examples of parallel sentences from the corpus.

LSU signs. This time is also variable, and is made worse by the fact that the interpreter can do other transformations to reorder, summarize, or adapt parts in order to transmit the same information in the available time.

In the past there have been attempts at tackling the alignment problem between sign sequences and oral sequences using automatic methods (Moryossef et al., 2023; Dutra, Aguirre, y Chiruzzo, 2025) but results are still far from perfect. Because of this, we decided to manually align all segments in the corpus, starting from a timestamp based alignment and correcting signs that where attached to the wrong sequence by hand. This was a relatively easy task for some videos, only having to move a few signs that laid on the boundaries between sentences. However in other cases, especially for live TV shows with simultaneous LSU interpretation, the task was more complex and we had to make many decisions about segmenting or regrouping signs based on the lexical units that were used. The final results has 1895 LSU segments aligned with their corresponding spoken Spanish segments (sentences). Note that, as the table shows, the number of spoken Spanish words in a sentence tends to be higher than the number of LSU glosses for the corresponding segment. This is expected, as we have empirically seen that in general the ratio is between 1.5 and 2.5 words per gloss.

Split	Train	Dev	Test	Total
Pairs	1360	291	291	1942
Total Glosses	14347	3032	3265	20644
Total Words	27380	5825	6141	39346
Unique Glosses	2454	1020	1105	2948
Unique Words	4551	1646	1725	5580

Table 3: Statistic of the parallel LSU - spoken Spanish corpus splits.

We decided to split exceptionally long sentences and LSU segments (up to 100 tokens), also based on the content and lexical units being used, so the final corpus contains 1942 parallel sentences. Table 2 shows some examples of sentence pairs extracted from the corpus. As seen in the table, the LSU glosses are mainly Spanish lemmas, losing most of the Spanish morphological information, and also includes some particular linguistic markers (such as determiners "PT:DET" and location adverbs "PT:LOC") and gestures (such as "G:DUDA"). One aspect of morphology that can be seen in the LSU side is that some glosses as marked with a "-PL" suffix, representing the plural version of the sign, and some signs use the suffix "X" to denote the undetermined grammatical gender.

Finally, we split the corpus in training, development and test of around 70 %-15 %-15 %. Table 3 shows the size distribution of each split in the final corpus.

4 Machine Translation Experiments

In this section, we present the MT experiments done using our corpus. For these experiments, we used two neural architectures: First we used the OpenNMT (Klein et al., 2017) library, which provides two neural architectures for MT based on encoder-decoder LSTM (Hochreiter y Schmidhuber, 1997) with attention mechanism or transformer (Vaswani et al., 2017) models, that have to be trained from scratch. Later on we used the NLLB (NLLB Team, 2024) model, which is a pretrained encoder-decoder transformer that has already been trained on a large dataset containing all language pairs for 200 lan-

guages, many of them low-resource, and can be fine-tuned to new language pairs as in our case.

We also tried two data augmentation techniques: Based on (McGill, Chiruzzo, y Saggion, 2024), we created a word embeddings set for LSU and used it to enhance an OpenNMT model. The other technique was inspired in the data augmentation technique described in (Chiruzzo et al., 2022), but using a more modern mechanism of synthetic data generation.

When we discuss the experiments, we write L2S to mean the LSU into spoken Spanish direction, while S2L means the Spanish into LSU direction. In all experiments we use the BLEU (Papineni et al., 2002) and CHRF (Popović, 2015) metrics. These metrics have their drawbacks but are nonetheless widely used in MT research. Both are calculated using n-grams of different orders in the reference and candidates, and both have values between 0 and 100, but the difference is BLEU uses full word n-grams while CHRF uses character n-grams, which makes it a less strict metric and more suitable for morphology rich languages.

4.1 Bootstrapped Embeddings

We followed the method proposed in (McGill, Chiruzzo, y Saggion, 2024) to create LSU embeddings. In our case we used a Glove embeddings collection for Spanish with 300 dimensions trained using the Spanish Billion Word Corpus,[2] and we designed the heuristics to create LSU gloss embeddings based on the Spanish word embeddings as shown below:

- If the gloss corresponds to a Spanish word that has an embedding, use that embedding.
- If the gloss ends in "X", try to find the embedding of the same word ending in "O".
- If the gloss ends in "-PL", try to find the embedding of the same word ending in "S" or "ES", or else try to find the same word without "-PL".
- If the gloss can be split in a set of words separated by "_" (compound notation), use the average of the embeddings that can be found in the Spanish set for those words.
- If the gloss is a number or ordinal, use the embedding of "NÚMERO" in the embeddings set.
- If the gloss starts with a punctuation sign, use a special punctuation embedding.
- Embeddings for personal pronouns "PP:PRO1", "PP:PRO2" y "PP:PRO3" and possessives "PS:POS1", "PS:POS2" y "PS:POS3", are mapped to the embeddings of words "yo", "tú", "él", "mi", "tu" y "su". If they have the plural mark "-PL" or a numeric mark they are mapped to the corresponding plural pronouns in Spanish.
- Glosses "PT:DET" and "PT:LOC" are mapped to the embeddings of determiner "el" and adverb "ahí" respectively.
- We add other particular mappings for interjections that lack fixed notation such as "SHHHH" and "PUUUM" using a representative word that can be found on the embeddings set.
- In other cases, map the gloss to a special unknown word.

With this technique, we mapped most of the glosses to some word embedding from the collection.

We did a first round of experiments to determine a good hyperparameter combination to use with OpenNMT. We used only its encoder-decoder LSTM architecture, as OpenNMT models have to be trained from scratch and the transformers architecture is too data intensive for our scenario. After these first experiments we settled in using 2 LSTM layers, with a hidden size of either 300 or 500, batch size of 256 and bucket size of 2048. We consider these our baseline experiments.

Then we trained the MT models in both directions, either using the initial random initialization of the embedding layers, or swapping the embedding layers of both languages to our bootstrapped embedding sets. The hyperparameter combinations for these experiments are using a hidden size of 300 or 500 units, and with or without our embeddings. All models were trained during 10,000

[2] https://github.com/dccuchile/spanish-word-embeddings

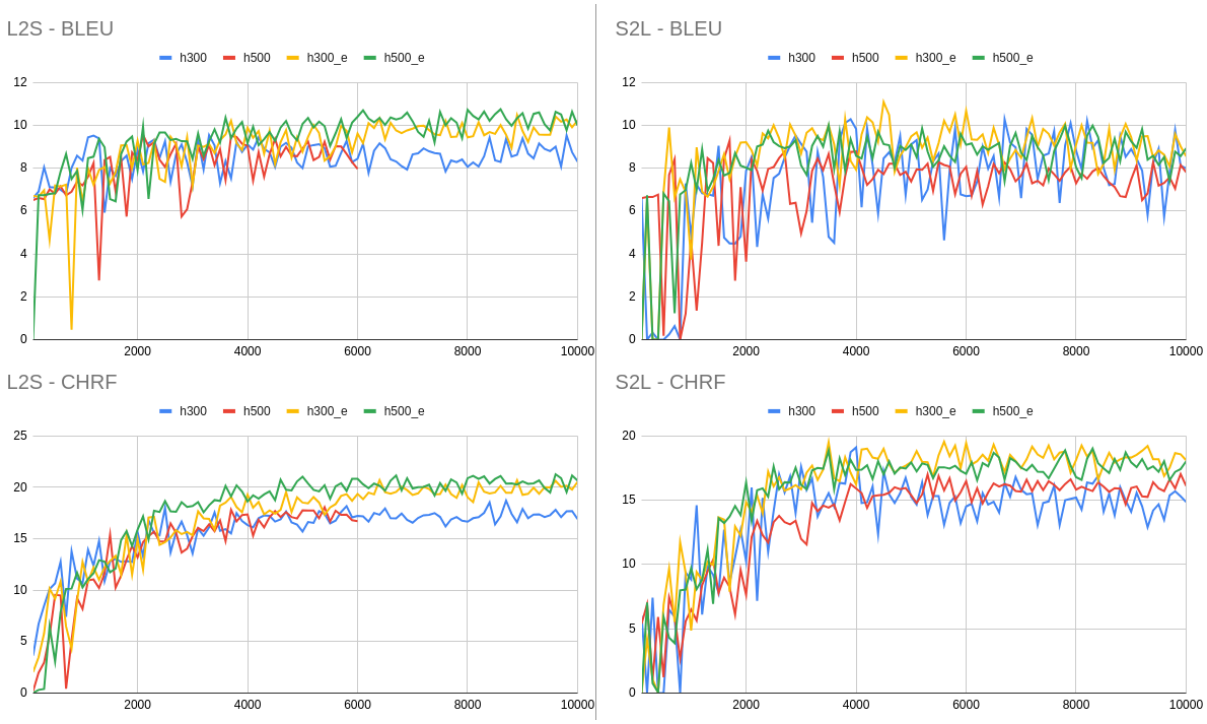

Figure 1: BLEU and CHRF performance of experiments over the dev split with and without bootstrapped embeddings. The "h" numbers represent the LSTM hidden size, the "e" experiments are the ones that use our bootstrapped embeddings.

steps and taking metrics every 100 steps over the dev set.

Figure 1 shows the performance evolution during training. As can be seen in the figure, using our bootstrapped embeddings (yellow and green lines) leads to more stable training and eventually better performance, although there is no clear advantage between using 300 or 500 units for the hidden size. The best model for the L2S direction uses pretrained embeddings and 500 hidden units, with 10.75 BLEU and 21.27 CHRF. In the opposite direction, the best configuration uses pretrained embeddings and 300 LSTM units, obtaining 11.08 BLEU and 19.54 CHRF.

4.2 Synthetic Data

While (Moryossef et al., 2021) and (Chiruzzo et al., 2022) propose using rule-based techniques to generate larger synthetic sets for augmenting the data used for training SLT systems, in this project we propose another way of creating synthetic data. We decided to leverage the linguistic capabilities of LLMs that have already been pretrained with large corpora and several NLP tasks. One thing to note is that these models are in general not very good at handling languages that have not been seen during their training, as is the case for all sign languages, which are very low resource. But considering that the gloss representation of the sign language uses many tokens of the environment language (in this case Spanish), and the models already have good knowledge of Spanish, perhaps they could leverage this knowledge and generalize starting from only a few examples.

We used the GPT model (Brown et al., 2020; Radford et al., 2018) in its versions 5 and 4-o to generate synthetic examples of LSU and their corresponding spoken Spanish, in the following way: We divided the whole training set in batches of 20 samples, and prompted the model to generate 100 new pairs of sentences (LSU and spoken Spanish) for each batch. Once we obtained all the generated examples, we manually cleansed them to remove duplicates and retry the process with random samples until we obtained a total of 10,188 new synthetic samples, totaling around 68,000 words in Spanish and 45,000 LSU glosses.

Using this process we could as well create more synthetic data, but as we keep on asking for more examples, it happens that the new generated examples start to look very similar to the previous ones, and many of them

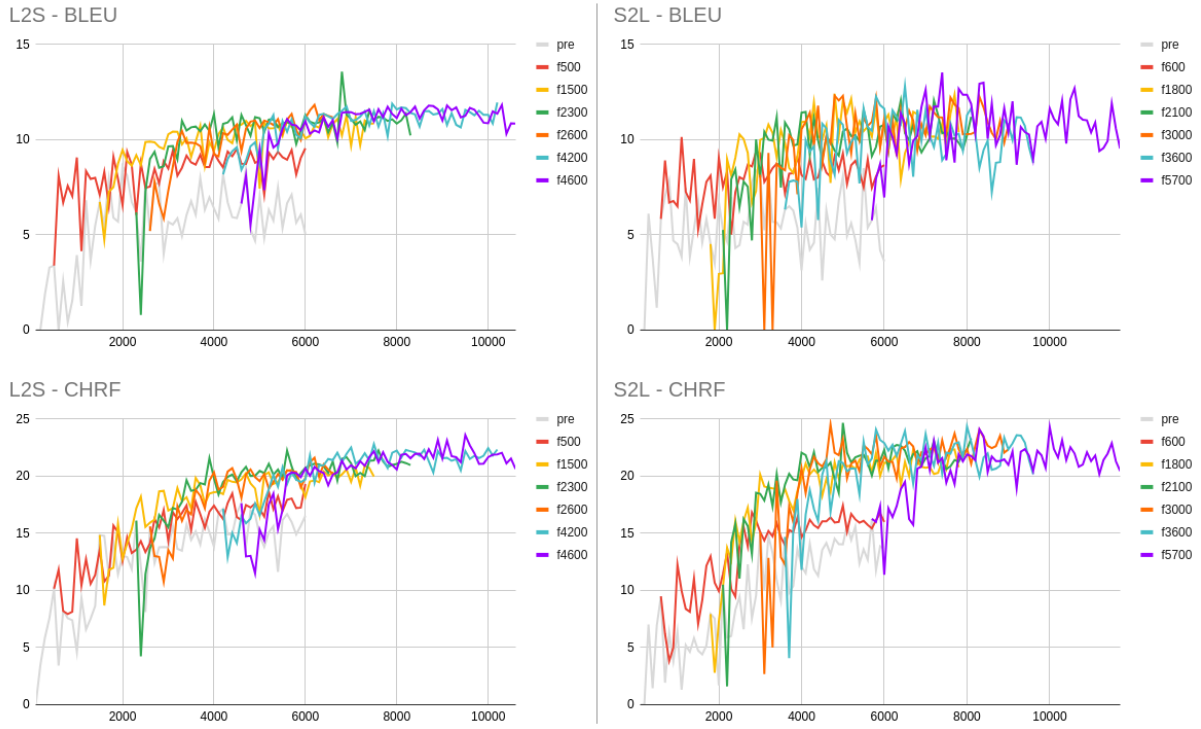

Figure 2: BLEU and CHRF performance of the pretraining+finetuning experiments over the dev corpus. Grey "pre" lines show performance of the pretraining using only synthetic data, colored "f" lines show several models finetuned using the real training corpus starting from different pretraining checkpoints.

are already present in the generated set. This happens because the LLMs do not have real creativity and tend to generate repetitions or variations of the same simple sentences, using the same sets of words and situations. However, we considered the created synthetic set large enough for our experiments, especially compared to our small training set of real data.

4.3 Data Augmentation

We followed the same strategy as (Chiruzzo et al., 2022), first doing a pretraining phase with only synthetic data, and then choosing some checkpoints and continue training (finetune) on the real data. We did this experiment with our two architectures: OpenNMT and NLLB.

For OpenNMT, using the same hyperparameter configuration as before, we did a pretraining phase with synthetic data for 6000 steps. As expected, the performance over the dev split for this model is very low, but we chose six pretraining checkpoints, based on the relative performance at each step. In the L2S direction, the pretraining points are step 500, 1500, 2300, 2600, 4200, and 4600; for the S2L direction they are 600, 1800, 2100, 300, 3600, and 5700. We used the pretraining checkpoints and continued training starting from each of these checkpoints for 6000 more steps, but using our real training data this time instead of the synthetic data.

The performance of these experiments over the dev split can be seen in Fig. 2. In each figure the gray line, in general below the colored lines, represents the performance of the pretrained model, which only sees synthetic LSU - spoken Spanish samples. Then starting from some points of the gray line, the six colored lines represent individual fine-tunings that start from different checkpoints. Some of them are more stable and some have more oscillatory behavior, but finally all stabilize in values higher than the ones achieved for the pretrained model. The best models in the L2S direction are the fine-tuning starting from step 2300, with 13.57 BLEU, and step 4600 with 23.60 CHRF, both values higher than the ones obtained in previous experiments. Something similar happened in the opposite direction, achieving 13.53 BLEU on the fine-tuning of step 5700, and 24.67 CHRF starting from step 2100.

For NLLB, we started first with a baseline experiment using its standard configuration, with the `nllb-200-distilled-600M` model. NLLB has seen many languages during its pretraining, but in our work we are dealing with LSU, and the pretrained model probably has never seen any example of this language. However it has certainly seen a lot of examples of Spanish, and most of the LSU glosses are concepts represented as Spanish words. So it is interesting to see if a model with solid Spanish knowledge can generalize to this other language that shares some tokens but has at the same time many lexical, morphological and structural differences.

We added a new language to the model (LSU) and continued its training focusing only on the LSU-Spanish pair for 1000 steps, taking snapshots every 100 steps. In both directions, the best performance was achieved in the step 900: 15.39 BLEU, 33.23 CHRF for L2S; and 18.16 BLEU, 37.02 CHRF for S2L. Note that these values are already improving on the ones obtained in the OpenNMT experiments, which hints at the robust performance that can be achieved as a result of pretraining on a large dataset with many language pairs.

Then we replicated the pretraining with synthetic data and finetuning strategy with the NLLB model. In this case, the pretraining phase with synthetic data can be seen as a kind of finetuning, a continuation of the original (much more massive) pretraining. We continued the training of NLLB but only with our synthetic data, and we took snapshots at steps 200, 500, 700, and 1000, up from which we did 1000 more steps of finetuning with our gold training data.

In this case, the best models were found finetuning from step 700, at the step 1600 for the L2S direction (16.91 BLEU, 37.15 CHRF), and at the step 1700 for the S2L direction (18.57 BLEU, 39.39 CHRF).

5 Experimental Results

We took the models that obtained the best results for each experiment on the dev split and ran them over the test split. Table 4 shows a summary of these experimental results for the dev and test splits. The first thing to notice is that the values for dev and test are largely similar. For the baseline and embeddings experiments, they are practically the same, for the finetuning experiments test performance is a couple of points below dev, but not by much. This suggests overfit was not a major issue in our experiments, and gives us confidence that these results might be close to what we could find in another settings beyond our test data.

			Dev		Test	
Dir.	Model	Step	BLEU	CHRF	BLEU	CHRF
L2S	h300	3300	9.54	16.10	9.55	16.05
	h300	8700	9.37	18.71	9.26	18.02
	h300_e	8900	10.48	20.68	9.82	19.17
	h300_e	6400	10.32	20.70	10.40	19.96
	pre	2000	8.60	13.05	8.89	13.08
	pre	4600	6.65	17.63	6.74	17.11
	f2300	6800	13.57	21.52	11.56	21.21
	f4200	10200	11.96	22.33	11.40	21.97
	f4600	9500	11.26	23.60	11.15	23.01
	f4600	8900	11.60	23.17	11.35	22.68
	nllb	900	15.39	33.23	13.43	30.73
	nllb_f700	1600	**16.91**	**37.15**	**16.55**	**35.64**
S2L	h300	8200	10.21	16.38	9.68	16.30
	h300	4000	9.83	19.10	9.81	19.19
	h300_e	4500	11.09	19.01	10.49	17.15
	h300_e	5600	10.22	19.54	10.72	18.29
	pre	5500	8.51	13.59	7.87	13.19
	pre	5700	5.77	16.27	5.11	16.05
	f5700	7400	13.53	21.58	12.50	21.23
	f5700	8400	12.99	23.51	11.20	21.54
	f2100	5000	10.42	24.67	9.77	22.42
	f3000	4700	9.80	24.53	9.43	22.88
	nllb	900	18.16	37.02	16.12	35.37
	nllb_f700	1700	**18.57**	**39.39**	**17.90**	**38.91**

Table 4: BLEU and CHRF metrics, over the dev and test splits, for the MT experimetns that had the best performance in each experiment. The best results for both metrics and both architectures are always achieved in fine-tuning scenarios. In general we obtained similar metrics over dev and test, which suggests that overfit was not a major issue.

As for the best performance, we obtained 16.55 of BLEU and 35.64 of CHRF in the L2S direction, and slightly higher values in the S2L direction with 17.90 of BLEU and 38.91 of CHRF. These values are relatively good for being the first experimental values obtained for LSU, and we want to highlight the huge improvements obtained over our baseline experiments (rows h300 in the table).

If we manually analyze the results, we find that the best models are not yet ready to be functional in a real scenario, but they are already showing signs of having learned at least basic features of the language. Table 5 shows some translation examples in the L2S direction, while Table 6 shows examples in the opposite direction. For L2S we can see that, even if translations are not perfect, at least they are on topic. For example the first

Reference	ONMT f2300/6800	NLLB f700/1600
llevaban a que directamente ese poder se le entregara al monarca a los efectos que lo usara todo y plenamente sobre esa misma sociedad .	se viene en primer lugar y la burguesía , no es algo de la monarquía absoluta y de la misma forma que no se puede hacer nada de la tanda .	entonces venimos a designar un que fuera un líder que le pareciera más fácil y que le diera el poder a la sociedad y le dijo que le iba a dar el poder para hacer todo lo que podía .
ahora , ¿ cuál es el problema ?	¿ y eso sí ? hay problemas también .	¿ y cuál es el problema ?
¿ es bueno el sueldo en china ? ¿ es muy desnivelado ?	es el sueldo de allá .	el sueldo en china es muy bueno , es muy desnivelable .
en materia política esto se llama retroversión , retornar el poder a la sociedad , al pueblo . retroversión .	al pueblo , la cantidad de expertos y la cantidad de expertos y la sociedad , no me contenté con el poder que tiene que ver con el poder del poder ,	entonces el mismo tipo de retrocesivo fue dado a las personas con el poder que yo morí , pero no se dio como cambio .
yo sigo todo lo que sea de acá de uruguay .	yo , además de todos los currículum y uruguay .	vamos a perseguir a todos a nuestro lado y a los uruguayos .
un delicado títere	pepe pensó la cuerda de la cuerda de la circulación .	es un hermoso tittere tranquilo .
por ejemplo , el censo dijo que el noventa por ciento de la población está a favor , se dijo que esto estaba manipulado , pero no es así , en verdad las personas apoyan .	por ejemplo , primero veo un inventario , por ejemplo , la negociación por los sectores , por ejemplo , lo que se verá .	por ejemplo en el censo , el 93 % respalda la afirmación de que no se manipulea a las personas , sino que son realmente felices con lo que están de acuerdo .
bueno , es el opuesto a mi vida acá en uruguay	entonces de la misma vida es la vida uruguaya .	es el contrario de la vida en uruguay .

Table 5: Examples of sentences translated with the best OpenNMT and the best NLLB model found, in the L2S direction.

row talks about the monarchy, and the candidates mention monarchy, bourgeoisie, and leaders; the sixth talks about a puppet, and the candidates mention a puppet or strings, keeping in the topic of the tale Pinocchio. The third row talks about salary in China, and the candidates mention salary or China. There is still a lot of room for improvement, but at least the translations start to zone in on the ideas behind the original phrases, and we can clearly see that the NLLB translations are much more fluent and preserve the semantic more faithfully.

One of the annotators of the corpus, expert in LSU, analyzed the results in the S2L direction, highlighting that in terms of fluency and faithfulness, the NLLB model once again obtains better results. But in this case there are other differences that are worth mentioning. As the NLLB model uses a subword tokenizer decoder, it could in principle create new words and glosses that are not present in training data. This is an advantage in L2S, but in S2L it could create glosses (e.g. DELICADXTERE in row 6) that seem to be close in meaning but don't really exist as signs in LSU, or it could use them in invalid contexts like LLAMAR in row 4. On the other hand, the OpenNMT method uses whole word tokens and a separate vocabulary for both languages, which in this case gives it the advantage of only being able to produce glosses that it has seen in the training data. However, it sometimes produces long and more repetitive sequences like in row 7.

6 Conclusion

This paper presents the first attempts at machine translation between LSU and spoken Spanish. In order to do this, we created a parallel corpus of LSU - spoken Spanish, starting with a LSU corpus that is currently under construction. As time goes by this corpus will get bigger, which will enable us to go on improving the quality of our translations. We would also like to try to use our current MT models to feedback and help the annotation process of the corpus.

Reference	ONMT f5700/7400	NLLB f700/1700
ENTONCES VENIR DESIGNAR MONARCA ORGANIZAR LÍDER MEJOR FÁCIL PRO1-DAR-PRO3 PODER SOCIEDAD PRO1-DAR-PRO3 PODER PERSONA MONARCA PT:DET HACER TODX	QUÉ-PASAR PS:PRO1 HACER PODER MONARCA	ENTONCES PT:DET PODER DAR PRO1-DAR-PRO1 SER-QUÉ-PL
CUÁL PROBLEMA G:DUDA	AHORA PT:DET PROBLEMA AHORA	QUÉ PROBLEMA
BUENX SUELDO PT:LOC CHINA CÓMO DESNIVEL-PL	SUELDO PT:DET CHINA	BUENO SUELDO PT:LOC CHINA DESNIVELADO
ENTONCES IGUAL IGUAL FS:RETROVERSIÓN PERSONA-PL DAR PODER MONARCA PS:PRO1 MORIR NO-HABER HIJX DAR PT:DET FS:R INTERCAMBIO ENTONCES	PT:LOC AFUERA TEMA LIBERTAD PODER VERDAD PODER	SOBRE POLÍTICA PT:DET LLAMAR RETROVERSIÓN PODER VOLVER SOCIEDAD PUEBLO
PS:PRO1 SEGUIR PERSEGUIR TODX SÍ PT:LOC URUGUAY TA	PS:PRO1 G:CA TODX FORMA PT:LOC URUGUAY PS:PRO3	PS:PRO1 SEGUIR TODO PT:LOC URUGUAY
PRECIOSO TÍTERE TÍTERE-QUIETO	GRILLO-SALTAR	DELICADXTERE
POR-EJEMPLO VER CENSO NUEVEĈERO POR-CIENTO APOYAR DECIR MANIPULAR NO MANIPULAR PERSONA-PL VERDAD CONTENTX DE-ACUERDO TA	POR-EJEMPLO OBVIO DECIR ANTES VERDAD DECIR PS:PRO1 APROXIMADAMENTE PS:PRO1 APROXIMADAMENTE PS:PRO1 APROXIMADAMENTE PS:PRO1 APROXIMADAMENTE ANTES QUIÉN DECIR PS:PRO1 APROXIMADAMENTE NO AL-REVÉS NO	POR- PERSONA-PL CENSO DECIR UNOÑUEVE POR-CIENTO PUEBLO PRO3-DAR-PS:PRO1 DECIR PT:DET MANIPULAR-PL PERO NO VERDAD PERSONA
BUENO OPUESTO VIDA PT:LOC URUGUAY TA	AHORA VIDA PS:PRO2 PT:LOC URUGUAY	CONTRAPOSICIÓN PS:PRO1 VIDA PT:LOC URUGUAY

Table 6: Examples of sentences translated with the best OpenNMT and the best NLLB model found, in the S2L direction.

The results obtained during these experiments are promising, but there is still a lot of room for improvement. We have identified a few aspects where we could continue the exploration: First of all, it is possible to build more synthetic data by prompting other LLMs and varying the prompts. As well as this, we can try with more neural architectures and more data augmentation techniques. Finally, incorporating the sign morphological information into our dataset could be another way of improving the results. This is currently being explored as part of the annotation process, but also we could explore ways of enriching the data with information from hand and pose keypoints obtained with an automatic method.

As mentioned before, it is important to note that this gloss translation step alone does not constitute the full translation pipeline. To go from LSU to Spanish, we would need a continuous sign language recognition (CSLR) system that outputs glosses. Using our dataset, we are currently making experiments of sign recognition, sign spotting and CSLR (Zuo, Wei, y Mak, 2024), which would complete the translation pipeline. To go in the other direction, we would need sign language generation. One simple method could use a dictionary of pose sequences per gloss, but current methods use gloss-informed diffusion models for obtaining more expressive representations (Tang et al., 2025). In the future we plan to try some of these methods on our data to generate fluid LSU poses and videos.

References

Bain, M., J. Huh, T. Han, y A. Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Baker, A. E. 2016. Sign languages as natural languages. En *The linguistics of sign languages: An introduction*. John Benjamins Publishing Company, páginas 1–24.

Bojanowski, P., E. Grave, A. Joulin, y T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, y others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chiruzzo, L., E. McGill, S. Egea-Gómez, y H. Saggion. 2022. Translating spanish into spanish sign language: Combining rules and data-driven approaches. En *Proceedings of the fifth workshop on technologies for machine translation of low-resource languages (LoResMT 2022)*, páginas 75–83.

Dutra, M., R. Aguirre, y L. Chiruzzo. 2025. Experiments on Automatic Alignment of Spoken Spanish and Uruguayan Sign Language Glosses. En *Proceedings of Jornadas Chilenas de Computación 2025 (JCC 2025)*.

Fojo, A., M. Tancredi, F. Etcheverry, A. Pintos, B. Pérez, J. Bianchi, R. Aguirre, M. Castillo, M. N. Macedo, y L. Chiruzzo. 2024. Corpus de lengua de señas uruguaya y desarrollo de metadatos. En *Congresso Internacional de Pesquisas em Línguística de Línguas de Sinais*.

Hochreiter, S. y J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Joshi, P., S. Santy, A. Budhiraja, K. Bali, y M. Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. En *Proceedings of the 58th annual meeting of the association for computational linguistics*, páginas 6282–6293.

Klein, G., Y. Kim, Y. Deng, J. Senellart, y A. M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

McGill, E., L. Chiruzzo, y H. Saggion. 2024. Bootstrapping pre-trained word embedding models for sign language gloss translation. En *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, páginas 116–132.

Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moryossef, A., Z. Jiang, M. Müller, S. Ebling, y Y. Goldberg. 2023. Linguistically motivated sign language segmentation. *Findings of EMNLP 2023*.

Moryossef, A., K. Yin, G. Neubig, y Y. Goldberg. 2021. Data augmentation for sign language gloss translation. En *Proceedings of the 1st international workshop on automatic translation for signed and spoken languages (AT4SSL)*, páginas 1–11.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Núñez-Marcos, A., O. Perez-de Viñaspre, y G. Labaka. 2023. A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

Papineni, K., S. Roukos, T. Ward, y W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, páginas 311–318.

Pennington, J., R. Socher, y C. D. Manning. 2014. Glove: Global vectors for word representation. En *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, páginas 1532–1543.

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

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, y I. Sutskever. 2023.

Robust speech recognition via large-scale weak supervision. En *International conference on machine learning*, páginas 28492–28518. PMLR.

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever, y others. 2018. Improving language understanding by generative pre-training.

Tang, S., F. Xue, J. Wu, S. Wang, y R. Hong. 2025. Gloss-driven conditional diffusion models for sign language production. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(4), Marzo.

Taulé, M., M. A. Martí, y M. Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. En *Lrec*, volumen 2008, páginas 96–101.

Team, G., R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, y others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wittenburg, P., H. Brugman, A. Russel, A. Klassmann, y H. Sloetjes. 2006. Elan: A professional framework for multimodality research. En *5th international conference on language resources and evaluation (LREC 2006)*, páginas 1556–1559.

Yin, K., A. Moryossef, J. Hochgesang, Y. Goldberg, y M. Alikhani. 2021. Including signed languages in natural language processing. En C. Zong F. Xia W. Li, y R. Navigli, editores, *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)*, páginas 7347–7360, Online, Agosto. Association for Computational Linguistics.

Zhang, X. y K. Duh. 2021. Approaching sign language gloss translation as a low-resource machine translation task. En *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, páginas 60–70.

Zuo, R., F. Wei, y B. Mak. 2024. Towards online continuous sign language recognition and translation. En Y. Al-Onaizan M. Bansal, y Y.-N. Chen, editores, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, páginas 11050–11067, Miami, Florida, USA, Noviembre. Association for Computational Linguistics.