

# Towards Developing Parallel Corpora for Portuguese and Portuguese Sign Language

Ziba Khani<sup>1</sup>[0000–0003–2058–7008], Nuno Escudeiro<sup>2</sup>[0000–0003–3940–3846], and  
Paula Escudeiro<sup>3</sup>[0000–0003–2528–572X]

<sup>1</sup> Department of Computer Science - Sapienza University of Rome, Italy  
`khani.1799920@studenti.uniroma1.it`

<sup>2</sup> Departamento de Engenharia informática - Instituto Superior de Engenharia do  
Porto, Portugal, `nfe@isep.ipp.pt`, `pmo@isep.ipp.pt`

**Abstract.** Low-resource languages, including sign languages, are a challenge for machine translation research. Given the lack of large-scale parallel corpora, researchers must use small parallel corpora for training an automatic translation system. This article aims to address this problem by building artificial parallel corpora for Portuguese sign language in automatic translation systems. In this work, we obtained small parallel corpora of Portuguese text and Portuguese Sign Language gloss from the Metro of Porto. We used these corpora to learn grammar rules in translation between Portuguese text and Portuguese Sign language gloss. Applying obtained rules to our data, we generated artificial parallel corpora for Portuguese and Portuguese sign language gloss.

**Keywords:** Sign Language, Deaf, Natural Language Processing, Parallel Corpora

## 1 Introduction

Sign languages (SL) exist wherever communities of deaf people exist, and they have been developed as a means of communication and form the core of local deaf cultures. SLs use the visual-manual modality to convey meaning. Sign languages are expressed through manual articulations in combination with non-manual elements. SLs are full-fledged natural languages with their grammar and lexicon. SLs are not universal and mutually intelligible, although there are similarities among different sign languages. SL is a language that inherits all characteristics of natural languages and can be analyzed like any other language. Hence, it has the flexibility to create a brand new vocabulary and any new grammatical structure required with a full capacity of abstraction and expression. Each sentence in sign language comprises a series of hand gestures called signs arranged according to a syntax governed by spatial and temporal logic. Five parameters characterize each sign: configuration, orientation, location, movement, and facial expression [16]. They co-occur and allow signs to be distinguished from one another. Generally, glossing is used to represent the signs in the form of text. The difference between “writing in a language” and “glossing of a language” has

to do with the fact that the target language may not have equivalent words to represent the original language.

In this work, we present a new framework to transform a part of Portuguese-speech sentences to Portuguese Sign Language (LGP) gloss. The objective is to build a parallel corpora for automatic translation systems. This paper is organized as follows; First, we discuss the previous related work. Then, we provide details of the methodology proposed to build parallel corpora for sign languages (Section 3). Finally we summarize our work and provide conclusions in Section 4.

## 2 Related Work

Several linguistic resources are collected through national and international projects. They are the data used for automatic processing of sign languages such as machine translation, extraction of knowledge, and each has its own characteristics. Despite initiatives to collect resources, there is still lack of a large corpora that can be used for automatic processing of Portuguese Sign Language (LGP). This limitation arises from the high cost of producing signs by signers [17]. In contrast to LGP, the American Sign Language (ASL) is the richest in terms of data resources, i.e. several research work have been conducted until now, [12, 3, 18] where some focus on unique signs, and others focus on sentences and complete speeches.

For German Sign Language (Deutsche Gebärdensprache, DGS), the first corpus is called “Berlin Corpus,” that contains sign language interviews by 90 signers. This project was the starting point to define the first grammar (phonology, morphology, morphosyntactic, and syntactic) for DGS [10]. British Sign Language Corpus Project created a corpus understandable by automatic translators for the British Sign Language. Video footage were collected by deaf and non-deaf signers across the UK to create this parallel corpora [15]. For the Spanish sign language, there is only one parallel corpora for two specific areas: namely, service for renewing identification documents such as driver’s license service [14]. The “LSCOLIN” project proposes an annotated corpus for French Sign Language [4]. The main objective of this project was to show the iconic structures and their importance in the signing area. The Irish Sign Language corpus “ATIS” contains a collection of videos produced by 40 deaf people [2]. The work described in [1] focus on avatars, and on how to produce avatars signs, based on human signs.

The work done in [7] targets the teaching of LGP; the Virtual Sign Translator [6] contributes with a translator between European Portuguese and LGP, and it was also applied to be used in a game that teaches LGP [5].

Most of the studies for different sign languages emerged in the recent years. Their focus ranges from linguistic and humanistic to automatic translation, and language resources for sign languages. However, their objectives differ from one language to another. For example, we note that video-based corpora are forced to go through an annotation tool to better explain the content and semantics in

certain cases because the images are not satisfactory for recognition and interpretation of the language.

Some resources have been developed to automatically translate a text into sign language using statistical approaches, transfer rules, or based on examples. Other resources have also been published to provide linguistic relativity such as lexical, syntactic structures, morphological analysis, and even semantic analysis for the automatic processing of sign languages. We also note that certain languages had several resources in different formats and they are accessible on the web while others are limited to dictionaries that are not very useful, and there are limitations in translation or transcription tools.

Studying all the works have been done until now, to the best of our knowledge, none of them can be used in our case as we are seeking parallel corpora in text format to be used for training machine translation systems.

### 3 Parallel Corpora Collection

A parallel corpora contains large and structured texts aligned between source and target languages used to do statistical analysis and checking occurrences or validating linguistic rules on a specific domain. The acquisition of a parallel corpora for the use in a statistical analysis typically takes several pre-processing steps. In our case, there is not enough data between Portuguese texts and Portuguese Sign Language.

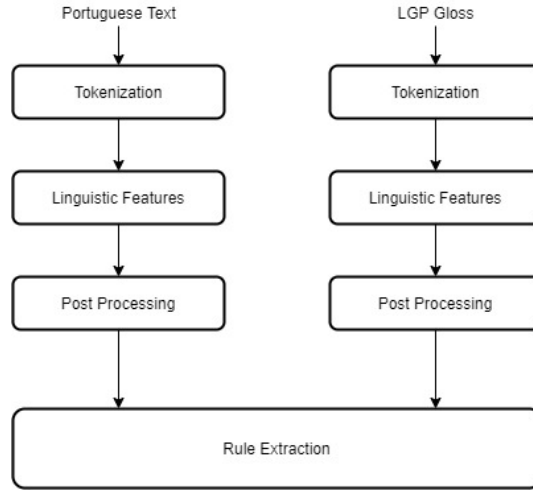
We start collecting only Portuguese data from the Europarl dataset to transform it into LGP gloss. Europarl [11] is one of the most popular multilingual corpora, designed to provide aligned parallel text for Machine Translation (MT) systems. The data has been extracted from the European Parliament’s proceedings, the latest release of the Europarl corpus comprises parallel text for 21 European languages, with more than 743 million tokens overall and over 155 thousand distinct concepts and entities. The data is represented in XML format. Thus we extracted the Portuguese text from that.

On the other hand, we obtained the small parallel corpora from Porto’s metro to extract the rules to transform Portuguese text to LGP gloss. This corpora is translated by experts from the deaf community to assure the correctness of sentences. Using this corpora, we build an immense parallel corpus between these two languages.

Since the translation between two languages concerns the mapping of lexical between a source and destination language, in this work, we revise the translation problem and tailor it for the parallel corpora creation using this small dataset. The idea we used in this work is somewhat inspired by old logical translation systems used before the neural network era.

We define the problem of translation into a few more straightforward problems. The first one is how does the order of words change between a source language and destination language. The order of words in a language usually comes from some grammatical rules. The other sub-problem is that how do words form change. In translation, it is often seen that a pair of words in the

source and the destination languages are very similar, but the semantics are represented in each language context differently. Other sub-problem is defined as how the form of lexical changes when nouns have gender in some languages. These rules are usually raised from grammar rules or the historical background of the language, and usually, linguistics can provide and well define them. We followed the same methodology in this work but automatically obtained such rules from a pair of languages. We obtained two sets of rules, which can be seen as a mapping function, where the input is a representation of language, and the output shows the changes in input. Fig. 1 shows the structure of the proposed system for rule extraction, which is composed of three predominant levels.



**Fig. 1.** Workflow diagram for extracting grammar rules.

LGP gloss uses Portuguese words for each sign or phrase that can be labeled. An example of Portuguese question sentence “Como abrir uma loja no Metro?” would be transcribed into LGP gloss as “Loja Metro abrir como?”.

During the implementation process, lexicons and grammar are compiled before the system uses them. As far as its functioning is concerned, the proposed system takes a text as input, segments it into sentences and words, and generates several analysis levels.

From the previous example, we describe the implementation phase step by step as follows. First, we perform preprocessing and lexical analysis. This phase is important for converting the raw data into a suitable format for automatic processing within our LGP statement generation framework. This begins by segmenting the text into sentences, and performing the same pre-treatment for each sentence. The first preprocessing operation is called “tokenization,” in which the input strings are transformed into “tokens.” This operation applies to the source texts and considers the spaces to separate words, numbers, and punctuation. In

our example, the tokens are:

1. ["Como"; "abrir"; "uma"; "loja"; "no"; "Metro"; "?"]

Then, all characters are converted to lowercase. We obtain the following:

2. ["como"; "abrir"; "uma"; "loja"; "no"; "metro"; "?"]

From the lexical analysis, we proceed to the grammatical analysis of each token. We associate every word to its grammatical category (noun, verb, adjective, adverb, proper noun, etc.). This procedure is called Part-of-Speech Tagging (PoS-Tag). As an example, for each word of our input sentence, we obtain the following:

3. "como" → ADV  
 "abrir" → VERB  
 "uma" → DET  
 "loja" → NOUN  
 "no" → ADP  
 "metro" → PROP  
 "?" → PUNCT

For the labeling of grammatical features, the "SpaCy Part-of-Speech Tagger" tool [9] is used. This tool is a widely used, open-source, and free library in the python programming language. The same process for LGP gloss is applied, for determining the grammatical features of each word in each sentence. Each featurized sentence is passed to post-processing step. This step concerns covering non-essential parts of the sentence. Therefore, extra free spaces, symbols, etc., are discarded.

The output of the two flows are featurized sentences. For translation between Portuguese and LGP, we focus on the first sub-problem, how order of words change, and extract rules that indicates how words order are changed. We name these rules as order-mapping, as it is a map for order of words. In the given example we have the original sentence as:

Words	como	abrir	uma	loja	no	metro	?
Indices	0	1	2	3	4	5	6

Looking at the corresponding LGP of the same sentence, but with the indices from the original language we can have:

Words	loja	metro	abrir	como	?
Indices	3	5	1	0	6

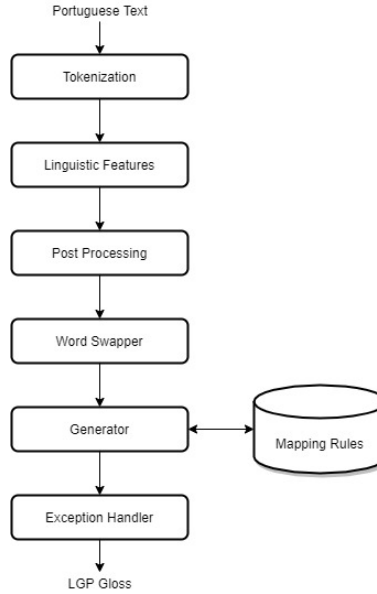
Consider for the given sentence, we have these tables. We can translate the original sentence to LGP by looking at how order of indices are changed. In this

example, the word at index 3 is presented at the index 0 in LGP, index 5 of the original sentence is presented in the index 1 of LGP, index 2 of original sentence does not exist in LGP sentence. The obtained mapping for this pair of sentence is:

Indices of the original sentence:	3	2	-1	0	-1	1	5
-----------------------------------	---	---	----	---	----	---	---

Having this rule for mapping, we can reorder the words in original sentence into LGP sentence. Notice that indices are the indices of the words in original language and values are the indices of the words in the LGP language; value -1 indicate that the particular word is not presented in LGP. This mapping between words indicates how orders are changed, but the problem with this example is that the rules are very specific to the given sentence. Instead of using words for creating this map, we can use features of the sentence that are more general.

In this work, we take the lemma of each word for finding a map, and use PoS-Tags for defining pattern of similar sentences. Using PoS-Tags instead of words does not change the obtained map, but rather for any new sentence with the same pattern (same PoS-Tags), we can use the same map.



**Fig. 2.** Generating LGP gloss from Portuguese text.

In next step, we focus on the second sub-problem, which is transformation of words. The rationale for this step is that not always a pair of words represents the same in both LGP and original language. In another words, a word in original text might be represented with its synonym in LGP. Reminding that for finding

this map we have to find pairs of words in both languages. For finding pairs of similar words we check how similar are the semantic of words, using word embedding.

Word embedding is the process by which words are transformed into vectors of real numbers. After transforming the words, similarity measures is used to find the degree of similarity between words.

One useful metric is cosine similarity, which measures the cosine of the angle between two vectors. It is essential to understand that it measures the orientation and rather than magnitude, meaning that similar vectors will have a similar vector orientation. In more detail, this means that two vectors with the same orientation will have a cosine similarity of 1, two vectors oriented at  $90^\circ$  relative to each other will have a similarity of 0, and two vectors opposed have a similarity of -1. Moreover, this is entirely independent of their magnitude. Cosine Similarity formula is as follows:

$$\cos \theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}}$$

where,

$$\mathbf{a} \cdot \mathbf{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$$

is the dot product of two vectors. We can also focus on using a specific word and use the measure of cosine similarity to understand how this word is used differently across different corpora. For obtaining vector representations for words we used Glove word embedding model [13] for Portuguese language [8] with the size of 300. Once we compute the cosine similarity between the semantic vectors of two words, we consider that they are pairs if their cosine similarity is above 0.9. Also we created a dictionary of word mapping, which indicates which words are translated to which words. Finally, the Europarl dataset [11] is used to build an artificial parallel corpora using the rules extracted in the first steps. The same process illustrated in Fig. 2 is performed for our new dataset, but the input is just Portuguese text for this step. Once the pattern is extracted for each sentence, it is compared with all the patterns extracted from the previous step. If any match is found, then it will be mapped to its counterpart mapping pattern, and the LGP gloss is generated in this way.

## 4 Conclusion

Sign Language Translation is a new research theme because it combines two complex scientific problems: translation and the transcription of sign language. Studies on sign language should include the linguistic, cognitive, and grammar aspects until creating the corpora, automatic translation, and real-time synthesis. Sign Languages are not universal, and in general, the studies are focused on one community of deaf and do not share the same syntactic structures, phonological, lexical, morphological, and semantic aspects. Despite existing tools for

transcription and annotation, each presents drawbacks. However, for the textual annotation in gloss, we proposed an approach that uses the grammatical order of words to generate its counterpart LGP translation. We generated these texts automatically using rule-based approaches using the words' grammatical orders. The accuracy of these texts can be improved with getting access to the LGP dictionary and using linguistic guidance by experts in the future.

## References

1. Bento, J.C.: Avatares em língua gestual portugues (2013)
2. Bungeroth, J., Stein, D., Dreuw, P., Ney, H., Morrissey, S., Way, A., Zijl, L.V.: The atis sign language corpus. In: LREC (2008)
3. Castelli, T.J., Betke, M., Neidle, C.: Facial feature tracking and occlusion recovery in american sign language. In: PRIS (2006)
4. Cuxac, C.: Corpus ls-colin sur plusieurs genres discursifs (josette bouchauveau et henri attia) (2014)
5. Escudeiro, N.: Virtual sign translator in serious games (2014)
6. Escudeiro, P., Escudeiro, N., Reis, R.M., Barbosa, M., Bidarra, J., Baltazar, A.B., Gouveia, B.D.: Virtual sign translator (2013)
7. Gameiro, J., Cardoso, T., Rybarczyk, Y.: Kinect-sign: Teaching sign language to "listeners" through a game. In: eNTERFACE (2013)
8. Hartmann, N., Fonseca, E.R., Shulby, C., Treviso, M.V., Rodrigues, J., Aluísio, S.M.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: STIL (2017)
9. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017)
10. Kaur, R., Kumar, P.: Hamnosys generation system for sign language. 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI) pp. 2727–2734 (2014)
11. Koehn, P.: Europarl: A parallel corpus for statistical machine translation (2005)
12. Neidle, C., Thangali, A., Sclaroff, S.: Challenges in development of the american sign language lexicon video dataset (asllvd) corpus (2012)
13. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
14. San-Segundo-Hernández, R., Pardo, J., Ferreiros, J., Rojo, V.S., Barra-Chicote, R., Lucas, J.M., Sánchez, D., García, A.: Spoken spanish generation from sign language. *Interact. Comput.* **22**, 123–139 (2010)
15. Schembri, A., Fenlon, J., Rentelis, R., Cormier, K.: British sign language corpus project: A corpus of digital video data and annotations of british sign language 2008-2014 (2014)
16. Stokoe, W.: Sign language structure: an outline of the visual communication systems of the american deaf. 1960. *Journal of deaf studies and deaf education* **10** **1**, 3–37 (2005)
17. Su, H.Y., Wu, C.H.: Improving structural statistical machine translation for sign language with small corpus using thematic role templates as translation memory. *IEEE Transactions on Audio, Speech, and Language Processing* **17**, 1305–1315 (2009)
18. Vogler, C., Neidle, C.: A new web interface to facilitate access to corpora: development of the asllrp data access interface (2012)