PE2LGP: translating European Portuguese into Portuguese Sign Language

Matilde Gonçalves, Luísa Coheur and Hugo Nicolau
Instituto Superior Técnico, Universidade de Lisboa
INESC-ID Lisboa
matilde.do.carmo.lages.goncalves@tecnico.ulisboa.pt,nome.apelido@tecnico.ulisboa.pt

Abstract

Like all natural languages, Portuguese Sign Language evolved naturally, acquiring grammatical characteristics different from Portuguese. Therefore, the development of a translator between the two languages consists in more than a mapping of words into signs (which results in a form of signed Portuguese), as it should ensure that the translation it produces satisfy the grammar of Portuguese Sign Language. Previous works use only manual translation rules and are very limited in the amount of grammatical phenomena that they cover, producing signed Portuguese. This thesis presents the first translation system from Portuguese to Portuguese Sign Language based not only on manual rules, but also on translation rules automatically built from grammatical information annotated in a corpus, the reference corpus under development by Universidade Católica Portuguesa. The manual rules deal with grammatical phenomena that the translation rules do not cover, namely morphological phenomena, such as the marking of the female gender and integrate particularities of the language such as facial expressions. It is the first work that deals with grammatical facial expressions that mark interrogative and negative sentences. Given a sentence in Portuguese, the system returns a sequence of glosses with markers that identify facial expressions, spelled words, among others. The thesis reports both a manual and an automatic evaluation. Results show improvements in the quality of the translation compared to the baseline system based on signed Portuguese.

Keywords

European Portuguese, Portuguese Sign Language, automatic translation, annotated corpus, gloss, natural language processing

1 Introduction

Portuguese Sign Language (LGP) is the main form of communication between the Portuguese deaf community. A Portuguese translator for LGP can be used to facilitate communication between deaf and hearing persons, and also for the purposes of LGP learning. However, LGP has several grammatical differences in relation to the Portuguese language. Thus, a translator which avoids producing "signed Portuguese" (translation in which each word in Portuguese is directly transformed into a sign in LGP, without obeying its grammatical rules) must take into account the specifics of LGP. Although there are some linguistic studies about LGP, there is still no official grammar, nor even a consensus on various linguistic features. Perhaps that is why the few computational works related to translation for LGP focus little on the linguistic component, based on small sets of manual rules and excluding facial expressions, resulting in little more than signed Portuguese. In order to fill these gaps and boost the creation of computational resources for the automatic processing of LGP, the project "Corpus & Avatar da Língua Gestual Portuguesa"[1], led by Universidade Católica Portuguesa, is creating the first LGP linguistic reference corpus. In this work, we contributed with a translator for LGP, hereinafter PE2LGP, in which the sentence(s) translated to LGP are represented by sequences of glosses, with markers that identify the facial expressions and fingerspelled words. PE2LGP relies on translation rules and a bilingual dictionary, both created automatically from the aforementioned corpus, as well as manual rules that capture linguistic phenomena related to the morphology of words and facial expressions. Figure 1 illustrates the architecture of PE2LGP. The system starts by extracting information from the corpus and enriching it with linguistic information. Then, the

[1] PTDC/LLT-LIN/29887/2017

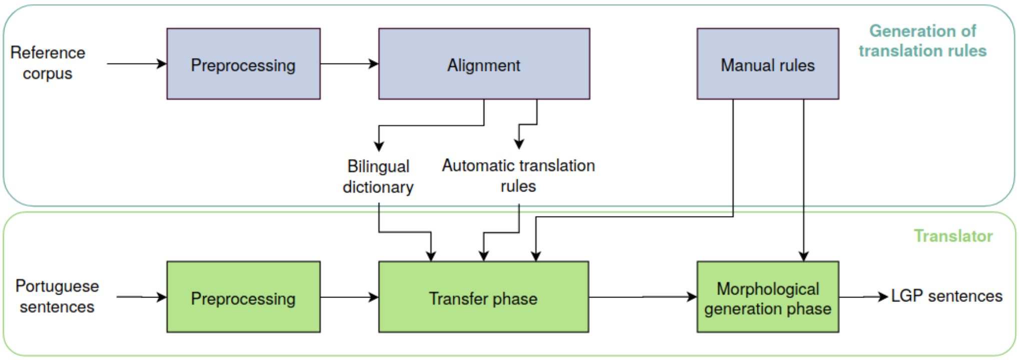

Figure 1: Arquitetura do sistema de tradução PE2LGP

words and signs in the corpus are aligned. From this alignment, the translation rules and a bilingual dictionary of Portuguese and LGP are extracted. When the system is given one (or more) sentence(s) in Portuguese, after linguistic preprocessing, the translation module comes into action, which, based on previously created resources, translates to LGP.

In this paper, we also present two evaluations of PE2LGP, an automatic one, based on a test corpus built by specialists and a manual one, in which speakers of LGP evaluate the quality of the translations. The main contribution of this work is a translator between European Portuguese and LGP, which feeds on a reference corpus to create translation rules and a bilingual dictionary (which can, therefore, grow with the corpus). We also contribute an alignment algorithm based on *string matching* and semantic similarity, a set of manual rules and a module which collects statistical information from the rules extracted from the corpus.

To our knowledge, this is the first translator for LGP with a strong linguistic component and, in particular, which deals with grammatical facial expressions essential to mark interrogative and negative sentences.

This document is organized in five more sections: in the Section 2 some aspects of the grammar of Portuguese Sign Language are described. The literature review can be found in Section 3. In sections 4 and 5 the PE2LGP is described. The evaluation methodology and results are presented in Section 6. Finally, Section 7 summarizes the main conclusions and future work.

2 About Portuguese Sign Language

The first studies about LGP appeared in the 90s, and there is still no official grammar. In this section some grammatical aspects of LGP are described.

2.1 Basic sentence order

There is still no consensus on the basic sentence order of the LGP. Some authors argue that the predominant structure is Subject-Object-Verb (SOV) (Rodrigues, 2018). However, a study carried out by Bettencourt (2015) on the canonical sentences order in LGP concluded that for sentences with non-locative transitive verbs and for declarative sentences, the base sentence order is the same as that of the Portuguese, that is, Subject-Verb-Object (SVO).

2.2 Types of the sentences

The type of sentence, whether it is interrogative or negative, influences the order of its constituents. According to Bettencourt (2015), interrogative sentences are marked by the use of interrogative adverbs and pronouns at the end of a sentence in LGP, accompanied by the interrogative facial expression.

2.3 Feminine gender

The marking of the feminine gender in nouns in LGP is performed by the composition of signs, that is, by the addition of the sign that marks the gender, the sign *MULHER (woman)*, to the base sign. The sign without gender marker is, by default, in the male gender (Bettencourt, 2015). So, the sign *LEÃO (lion)*, because it is a male noun, is represented only by the sign *LEÃO (lion)* while *leoa (lioness)* is composed of *MULHER* + *LEÃO*. However, there are situations in which there is no gender marker in nouns because that specific sign already has a gender associated with it. For instance, the nouns *galo (rooster)* e *galinha (hen)* have their own separate signs (Nascimento & Correia, 2011).

2.4 Diminutive and augmentative

Like the feminine gender, the diminutive and augmentative forms of nouns are made through the composition of signs, specifically with the addition of the signs *PEQUENO* (*small*) or *GRANDE* (*big*), respectively, to the base sign. Thus, the sign for *leoazinha* (*little lioness*) is compound by *MULHER* (*woman*) + *LEÃO* (*lion*) + *PEQUENO* (*small*) (with facial expression).

2.5 Possessive determinants, numerals, adverbs of quantity

Possessive determinants (*my*, *your*, etc.), numerals (*five*, etc.) and adverbs of quantity (for instance, *many*) proceed the noun (Gaspar, 2015; Bettencourt, 2015). For example, *your brother* will originate this sequence of signs: *BROTHER* + *YOUR*.

2.6 Verb tenses

The past and future verb tenses take place in three ways (Nascimento & Correia, 2011):

- by adding facial expressions to the neutral form of the verb (infinitive verb mode).
- by changing the time adverbs (yesterday, tomorrow, etc.) to the beginning of the sentence, if they exist in the sentence.
- or the signs *PASSADO* (*past*) or *FUTURO* (*future*) are added at the beginning of the sentence.

2.7 Negative sentences

According to Carmo et al. (2017) there are two types of negation in LGP, regular and irregular negation. In the first one, the negation is performed by adding grammatical negation markers, such as manual signs like *NÃO* (*no*) or *NADA* (*nothing*) after the verb, or non-manual signs such as *headshake* or facial expressions. In irregular negation, negation is incorporated in the verb, i.e., there are different signs for the negation of a certain verb. For example, the verb *NÃO-QUERER* (*to not want*) is different from the verb *QUERER* (*to want*).

2.8 Articles, copulative verbs and proper nouns

Articles and the verb *to be* are not represented in LGP. Proper nouns are fingerspelled.

2.9 Prepositions

Prepositions are not explicitly represented in LGP (Sousa, 2012), some are incorporated in the sign's movement, for example, using the initial and end positions of the objects (Bettencourt, 2015).

2.10 Conjunctions

According to the preliminary study about interphasic and phrasal connections (Martins & Mata, 2017), the coordinating adversative conjunctions (*but* e *however*) are lexical, which means they are produced manually, whereas the coordenating copulative conjunction *and* is a prosodic connection, produced only with facial expressions. The predominant expression associated with this conjunction is the neutral facial expression.

3 Related work

Sign language translation can be done based on *corpora* and/or manual rules (Chéragui, 2012). If there is a reasonable amount of aligned texts between the source language and the target sign language, computational models can be created based on this data. Examples of these works are the systems of translation for American Sign Language, presented in (Othman & Jemni, 2011) and for German Sign Language, described in (Bungeroth & Ney, 2004).

The first linguistic reference corpus of LGP is under development by Universidade Católica Portuguesa, in which lexical units are transcribed using glosses and grammatical information is annotated. In this work, we take advantage of this corpus to extract a set of translation rules, to which we add a set of manual rules.

Several automatic translation systems for sign language based on manual rules have been proposed in recent years. Here are some examples. The ATLASLang project (Brour & Benabbou, 2019), a hybrid system of Arabic text translation into Arabic Sign Language, based on rules and examples of Arabic sentences (and their translations) defined in a bilingual corpus. If the sentence exists in this corpus, then it is directly translated, otherwise the sentence is processed and manual rules apply. TEAM (Zhao et al., 2000), a prototype of a English text to American Sign Language translation system, translation rules are defined using *tree-adjoining grammars* (Shieber & Schabes, 1990), resolving linguistic differences such as word order in sentences.

The work developed by Su & Wu (2009) stands out. The authors present a statistical translation system of text from Mandarin to Taiwan Sign Language (TSL), which deals with the scarcity of data in a parallel corpus. Grammatical transference is based on grammatical formalism, specifically, on synchronous rules of context-free grammar and on a translation memory that describes the order of the thematic roles between the sentences of both languages. The syntactic structure of the sentences in TSL and the translation memory are extracted from the bilingual corpus through the alignment between the lexicon of the bilingual sentences. Words and signs are aligned using a measure of similarity, rather than probabilistic methods. The strategy implemented for the alignment of words and signs in this work was a source of inspiration for ours, given that the grammar is also extracted from a small corpus.

As for LGP, there are some computational prototypes, recently developed, with different objectives. "Virtual Sign Translator" Escudeiro et al. (2013) (Escudeiro et al., 2015) contributes with a translator between Portuguese and LGP, and is also used in a teaching game for LGP (Escudeiro et al., 2014). In Almeida et al. (2015a,b), (Ferreira, 2016) and (Gaspar, 2015), Portuguese to LGP translation systems are described, with the authors already referring to tools related to natural language processing for the generation of LGP. However, these works are a proof of concept and only cover a small set of phenomena. Thus, we believe that the work proposed here is the first that, with the aim of developing a translator for LGP (and not for signed Portuguese), takes real advantage of an LGP corpus.

4 Construction of translation rules and bilingual dictionary

In the following subsections, the data used in the construction of the translation rules is presented and the main steps that result in the translation grammar used by the translator are described.

4.1 Reference corpus

The corpus under development by Universidade Católica Portuguesa consists of videos of Portuguese deaf people of different age groups and different regions signing LGP formally, informally, spontaneously or according to an established subject. The data annotated in this corpus includes: the translation of the message in the video into Portuguese, the LGP signs (transcribed into gloss) and their respective grammatical classes, the arguments of the sentence (internal and external arguments), and the type of each sentence. The conventions used to annotate the corpus can be found in Table 1. Currently, the data used in the construction of the grammar presented in this work comes from a 5-minute video of an LGP native speaker with informal and spontaneous speech.

Grammatical class	**Convention**
Noun	N
Verb	V
Adjective	ADJ
Adverb	ADV
Syntatic element	**Convention**
Subject	ARG_EXT
Object	ARG_INT
Grammatical Phenomena	**Convention (example)**
Fingerspell	DT(M-A-R-I-A)

Table 1: Conventions for annotating grammatical information.

4.2 Preprocessing

From the corpus, only the grammatical information of sentences in LGP is known, so the Portuguese sentences are analyzed syntactically and morphosyntactically using natural language processing tools. In a preliminary study, the tools that best carried out these tasks were determined. For the first task, it was SpaCy (Honnibal & Montani, 2017); for the second one, FreeLing (Padró & Stanilovsky, 2012). Thus, the grammatical classes and subclasses, as well aspects of inflection and the lemma of the words in Portuguese sentences (and the signs in LGP sentences) are identified through FreeLing. This last step is performed for both words and signs, as it is the basis of the alignment of words and signs described in Section 4.3. In the syntactic analysis, the sentence in Portuguese is divided into its sentence elements (subject, predicate and sentence modifier), based on the dependency relationships identified by SpaCy. At the end of this phase, the tags resulting from the morphosyntactic analysis are converted into the tags of the corpus. For example, FreeLing's tag *NCMS000* refers to a common noun in the singular and male gender, and is converted to *N*, according to the corpus conventions set out in Table 1. In turn, SpaCy and the corpus syntax labels are converted to a simpler notation; for example, subject tags are converted to *S* and those that identify objects are renamed to *O*.

Since LGP does not have articles, these have been removed from the sentence, along with punctuation. Prepositions were also eliminated because they are not explicitly represented in LGP (Sousa, 2012).

4.3 Alignment

In statistical translation systems, the alignment of the lexicon is usually calculated using probabilistic methods (Chiu et al., 2007), however in the case of the Portuguese-LGP language pair, there is not a large enough corpus to train the alignment between words and signs. Thus, we propose a method based on similarity measures (*string matching* and semantic similarity). Words and signs are compared letter-by-letter; if they are equal, they are aligned; otherwise they are compared using *OpenWordNet-PT* and then *word embeddings*. This last step reinforces the semantic alignment, because if some word-sign pairs are not aligned by WordNet, they may be aligned through word embeddings.

OpenWordNet-PT, for being integrated in the NLTK library and for offering several measures of similarity between two concepts, was used to calculate the semantic similarity between a word and a sign. One of the similarity measures is the Wu-Palmer similarity[2]. A word and a sign were considered to be semantically similar if they have a pair of synonyms with a Wu-Palmer similarity value greater than or equal to 0.9. However, this similarity measure is only valid between concepts with the same grammatical class, since there is no common hyperonym between *synsets* from different grammatical classes (Farkiya et al., 2015). Thus, another premise was added: a word and a sign are also semantically similar if they have synonyms with similar radicals, such as the words *art* and *artístico*. Thus, for pairs of synonyms with different grammatical classes and for those with an original similarity value smaller than 0.9, the Jaro-Winkler Distance was calculated[3]. If for a word and a sign there is a pair of synonyms with a value of this measure greater than 0.8, then that word and that sign are aligned. Otherwise, the next step is taken.

In (Hartmann et al., 2017) 31 models of word embeddings were evaluated[4] for European and Brazilian Portuguese. The evaluation revealed that for the semantic analogy and for European Portuguese, the model with the best performance is the one trained with the GloVe algorithm with 600 dimensions. This model converts the word and the sign into vectors, which are then compared using Cosine Similarity. If the word and the sign have a similarity value greater than 0.3, then they are aligned.

In the alignment, we use the lemmas of the signs and words, which allows us to increase the number of exact matches picked up by the first stage.

4.4 Translations rules and bilingual dictionary

The translation rules are divided into two types: those that describe the syntactic structure (henceforth *morphosyntactic rules*) and those that describe the sentence order (*sentence rules*). The first rules are grouped by *sentence element*, that is, separate rules are constructed for sentence modifiers, for the subject and for the predicate. The order of morphosyntactic constituents can be changed according to the type of the sentence. This phenomenon is common in other languages, such as in English, in which the subject appears in interrogative sentences after the auxiliary verb, unlike declarative sentences, in which, normally, the subject appears before verbs. For this reason, the translation rules are also grouped according to the type of sentence (affirmative, negative, interrogative and exclamative) that originated the rule.

The translation rules describe the grammatical transformations necessary for a Portuguese sentence to be converted into an LGP sentence and, therefore, are composed of two "sides", namely the *Portuguese side* and the *LGP side*. The rule examples given hereinafter follow the structure *Portuguese side* → *LGP side*.

The sentence rules were built from the sentence orders of each Portuguese sentence, given by the syntactic analysis, and the sentence order of the respective sentence in LGP was extracted from the corpus. For instance, SVO → SOV, represents a sentence rule constructed from the information of a *interrogative* sentence.

The construction of the morphosyntactic rules is based on the grammatical classes of the elements that make up the word-sign pairs given by the alignment and on the correspondence between the grammatical classes on the Portuguese side and those on the LGP side. This correspondence is marked by a number, called *correspondence number*, which specifies how each element in the sentence should be translated. The rule *V1 N2 ADJ3 N6* → *V1 ADJ3 N6 N2* is an example of morphosyntactic rule of a predicate

[2] Described in www.nltk.org/howto/wordnet.html.

[3] The Python library pyjarowinkler was used for the Jaro-Winkler Distance calculation.

[4] They are available in nilc.icmc.usp.br/embeddings.

which determines that the constituent *N2* should be moved to the end of the sentence. Label *V* represents a verb, *N* is a noun and *ADJ* is an adjective, following the conventions of the corpus in Table 1. Note that without correspondence numbers, it would not be possible to distinguish between the two *N* labels.

In total, 66 morphosyntactic rules were built, 18 of which are related to subjects, 46 to predicates and 2 to phrase modifiers. Additionally, 39 sentences rules were built, 5 associated with interrogative sentences, 3 with negative sentences and 31 with affirmative declarative sentences.

During the construction of the translation rules, the occurrence of each rule was counted, for each type of sentence. These statistics will be used later, in the translation module (Section 5). In addition to their importance in the translator, they present linguistic information which is relevant to the study of some grammatical phenomena of LGP, such as canonical order.

As for the Portuguese and LGP bilingual dictionary, it was built automatically based on the alignment of words with signs in the corpus. This feature assists with the lexical transfer in the translator (Section 5.2), i.e., the mapping between the Portuguese lexicon and the LGP lexicon. In total, 163 word-sign pairs were aligned, most of which correspond to gloss-word pairs (*arte* and *ARTE*), and there are still semantically related pairs, such as *religion* and *IGREJA*. This dictionary was later revised by hand based on the information transcribed from the video, ending up with 102 entries.

4.5 Manual rules

A set of manual rules complements the translation rules previously described. Based on the grammatical characteristics of LGP listed in Section 2, 16 manual rules were built to ensure that the order of constituents with certain subclasses is in accordance with the characteristics of LGP. They also include particularities of the language related to the morphology of words, such as female gender marking and grammatical facial expressions related to negative and interrogative sentences.

Although some grammatical phenomena of LGP are well studied, others are not, such as negation marking. There are several ways to mark negation, which vary in both facial expression and hand sign, depending on the verb. In the research carried out for this article, no studies were found to indicate in which context each of the negation marking options is used. Thus, in this translator, this phenomenon is treated by adding the non-manual marker *headshake* simultaneously with the manual component *NO*, as it is the most frequent manual marker in LGP (Carmo et al., 2017).

To mark facial expressions, a notation was created that identifies the facial expression itself and its duration. The duration is identified by brackets: the open bracket indicates the beginning of the facial expression and the closed bracket, its ending. After this closed bracket, the name of the facial expression appears in parentheses. For instance, the sentence *Amanhã, ela não se vai vestir.* (Tomorrow, she will not get dressed.) would be represented by the translator as *AMANHÃ ELE VESTIR* {*NÃO*}*(headshake)*, the non-manual sign *headshake* is marked by *(headshake)* and the brackets indicate indicate that the non-manual sign is produced simultaneously with the manual sign of negation *NÃO* (no).

5 Translator

The following sections describe the phases of the translation component: first, preprocessing (Section 5.1), followed by the lexical transfer phase (Section 5.2) and the syntatic transfer phase (Section 5.3) and finally the morphological generation phase (Section 5.4). The procedures for each stage will be exemplified by the sentence in 1 and its sentence elements, *subject* in 2 and *predicate* in 3.

(1) A Diana perdeu o seu gatinho ontem. (Diana lost her kitten yesterday.)

(2) Subject: a Diana (Diana)

(3) Predicate: perdeu o seu gatinho ontem. (lost her kitten yesterday.)

5.1 Preprocessing

The Portuguese sentence given to PE2LGP undergoes a preprocessing similar to the one performed in the translation rules construction module (Section 4): the sentence is syntactically and morphosyntactically analyzed, the articles, prepositions and punctuation marks are removed and the labels resulting from previous analyses are converted to the label format used in the corpus, making them consistent with those in the translation rules. Before the punctuation is removed, the type of sentence (affirmative, negative, exclamative or interrogative) is determined and saved, because it will be necessary for syntactic transfer (Section 5.3).

5.2 Lexical transfer

The Portuguese lexicon is mapped to the LGP lexicon based on the bilingual dictionary created in the previous module. If the word is in the dictionary, then it will be replaced by the corresponding sign; otherwise, it will be converted into gloss at the generation stage. Assuming that none of the words in the example sentence in 1 exist in the bilingual dictionary, then the sentence does not change at this stage.

5.3 Syntatic transfer

The conversion of the syntactic structure of the Portuguese sentence to the corresponding syntactic structure in LGP is carried out by applying the translation rules (Section 4.4) and manual rules (Section 4.5). In the case of the first ones, those that best fit the syntactic structure of the Portuguese sentence are applied according to the type of the sentence. For each sentence, the two types of translation rules (morphosyntactic rules and sentence rules) are applied. It is important to clarify that the operations of this phase are not carried out on the sentence but on its sentence elements. Thus, what is received in this phase are the *syntactic structures* of each sentence element, exemplified in 4 for the subject and in 5 for the predicate of the example sentence (the definite articles have been removed in the preprocessing).

(4) Syntatic structure of the subject:
N

(5) Syntatic structure of the predicate:
V DET N ADV

The choice of the best *morphosyntactic rule* is based on the Edit Distance algorithm (Levenshtein, 1966) between the syntactic structure of the Portuguese sentence and the syntactic structure of the Portuguese side of the morphosyntactic rule. Edit Distance is a similarity measure between text strings[5], that lets you know what operations must be done so that the two are the two strings match. The possible operations are insertion, removal and replacement. The costs implemented for these operations are 1, except in the case where the sentence type is replaced, whose cost is 2, since the order of the morphosyntactic constituents can change according to the sentence type.

Before calculating the distances, both the sentence structure and the rules on the Portuguese side are converted to the format *CL1 CL2 CL3 Sentence_type*, in which *CL* are the grammatical classes and *Sentence_type* corresponds to one of the following: exclamatory (EXCL), affirmative declarative (CAN), negative declarative (NEG) and interrogative (INT).

In this way, the subject and predicate structures of the example sentence are converted to:

(6) Subject: N CAN

(7) Predicate: V DET N ADV CAN

The next step is to calculate the Edit Distance between the sentence and the Portuguese side of each of the rules, so the rule with the smallest distance can be chosen. In the event of a tie, the following criteria are followed, in order:

1. The most frequent rule in the corpus is chosen, based on the statistics collected in the previous module;
2. The longest rule is chosen;
3. The rule that comes first in alphabetical order is chosen.

These tiebreaker criteria are arbitrary, but ensure that the choice of rule is consistent.

The translation rules that best fit the syntactic structures of the subject and the predicate of the example are indicated respectively in 8 and 9.

(8) N1 CAN → N1 CAN

(9) V1 N2 ADV3 CAN → V1 ADV3 N2 CAN

The Edit Distance algorithm, also indicates which operations are to be carried out to make the sentence's syntactic structure match the syntactic structure of the Portuguese side of the rule. When this requires inserting an element on the LGP side of the rule, a simple heuristic is followed: the element to be added on the LGP side is inserted after the grammatical class with the correspondence number equal to the correspondence number of the grammatical class before of the inserted value on the Portuguese side. Removal and replacement operations are simpler to perform: the constituent to be removed or replaced on the LGP side of the rule is the one with the same correspondence number as the constituent that was removed/replaced on the Portuguese side. For example, to match the syntactic structures of the predicate in 7 and the rule in 9, it's sufficient to insert a *DET* after *V1* on the Portuguese side of the rule and, following the previous heuristic,

[5] For more details about the algorithm: web.stanford.edu/class/cs124/lec/med.pdf

on the LGP side, a *DET* should be inserted after the element with a correspondence number of 1, which in this case is also the constituent *V1*, with the new constituent being assigned a correspondence number of 4. Thus, the transfer of syntactic structure is determined by the rule V1 DET4 N2 ADV3 → V1 DET4 ADV3 N2, which corresponds to *perdeu seu gatinho ontem* (lost her kitten yesterday). The rule dictates an exchange of the constituent *ADV3* (*ontem* (yesterday)) with *N2* (*gatinho* (kitten)).

This procedure ensures that all input sentences are assigned a morphosyntactic translation rule.

After this, the sentence elements, with a new syntactic structure, are joined to form the LGP sentence. This union is based on the most frequent sentence order in the corpus according to the type of the sentence. For affirmative declarative sentences like the example sentence *A Diana perdeu o seu gatinho ontem. (Diana lost her kitten yesterday.)*, the most frequent sentence order in the corpus is SVO. Thus, the sentence elements are ordered in this way, first the subject (*Diana*), then the verb (*perdeu (lost)*) and at the end the object (*seu gatinho ontem (her kitten yesterday)*).

However, and following the premise of earlier studies, in which it is argued that the most frequent base sentence structure of LGP is SOV, an option was added to the translator, so the user may choose between SVO or SOV word order.

Finally, manual rules are applied, through which the morphosyntactic constituents are reordered according to known grammatical rules of the language. Since, in LGP, temporal adverbs are produced at the beginning of the sentence and possessive determinants come after with the noun, the result of this phase of the sentence in 1 is *Ontem Diana perdeu gatinho seu* (Yesterday Diana losts kitten her).

5.4 Morphological generation

Here, the lexicon is converted into glosses and manual rules related to LGP morphology are applied, such as the marking of diminutive and augmentative degrees in nouns (Section **??**). From this phase comes a sequence of glosses with additional markers that identify facial expressions and spelled words following the reference corpus annotation conventions. So the result of translating the sentence *A Diana perdeu o seu gatinho ontem. (Diana lost her kitten yesterday.)* is *ONTEM DT(D-I-A-N-A) PERDER GATO PEQUENO SEU (YESTERDAY DT(D-I-A-N-A) LOSE SMALL CAT HER)*, where the *DT()* notation indicates that the name Diana is fingerspelled.

6 Evaluation

To assess the quality of the translation of the proposed system, two evaluations were conducted, an automatic one, comparing the translation of the system with a test corpus, and a manual one, based on the opinion of experts.

6.1 Automatic evaluation

The objectives of this evaluation are to ascertain whether the approach explored in this thesis allows the translator to capture linguistic phenomena, producing LGP rather than merely signed Portuguese and to understand the impact of the translation rules on the quality of the translations.

6.1.1 Test corpus

The test corpus was created by a Portuguese interpreter of LGP. It consists of 58 simple sentences in Portuguese (different from those in the reference corpus) and their corresponding translations in LGP. For some Portuguese sentences, more than one possible translation was annotated, but not all possible translations were sought.

6.1.2 Evaluation measures

The 58 Portuguese sentences from the test corpus were translated by the system and the resulting translations were evaluated using the measures *Bilingual Evaluation Understudy* (BLEU) (Papineni et al., 2002) and *Translation Error Rate* (TER) (Snover et al., 2006).

6.1.3 Configurations

The *baseline* system consists of the production of signed Portuguese. The translated sentences follow the grammar of Portuguese and have no facial expressions. For example, the Portuguese translation of the sentence *Quem comeu o bolo?* is *QUEM COMER BOLO*.

In total, 5 experiments were conducted, using different configurations of the translation system. One of the variables in these configurations was whether the system used both the automatic and the manual translation rules, or only the manual rules. Another variable was whether the word order was SVO or SOV. These 5 experiments are shown in Table 2. Configuration I is the baseline system (signed Portuguese), configurations II and III belong to the system based only on manual rules, which form *set 1*, and finally, configurations IV and V belong to the proposed system and form *set 2*.

6.1.4 Results

Table 3 presents the results for the TER and BLEU measures of the configurations of the various systems. The best results were obtained by the translations with the SOV structure translated by the proposed system and by the system based only on manual rules (configurations II and IV).

6.1.5 Discussion

Baseline system vs. other

The results of the developed system surpassed those of the baseline system, reaching 0.29 TER and 0.77 BLEU for the SOV structure. These values show that the application of translation rules and manual rules in grammatical transfer considerably improve the quality of the translations, producing LGP and not only signed Portuguese.

Set 1 vs. Set 2

The results between the configurations belonging to set 1 and those belonging to set 2 are slightly different. The proximity between the values of the two sets is due to the fact that the majority of the morphosyntactic rules applied to the affirmative and negative declarative sentences do not alter the sentence's syntactic structure. However, the application of the translation rules improved the quality of 2 translations, making them equal to the reference.

The comparison of the system's translations with the references allowed us to infer that the errors in the translations are due to: a) flaws in the morphosyntactic analysis, b) limitations in the identification of sentence elements and c) the fact that the morphosyntactic rules, because they only describe the order of the main grammatical classes. This last limitation implies that phenomena related to the order of certain constituents such as adverbs (ADV) are not captured. Consider the following cases: (10) *li muito* (I read a lot) and (11) *li ontem* (I read yesterday). The morphosyntactic rule to be applied will be the same because they have the same syntactic structure (V ADV), assuming they have the same type of sentence. However, the two adverbs are produced in different orders in LGP, *ontem* (yesterday), because it is an temporal adverb, must be produced first, which is not the case for the quantity adverb *muito* (a lot).

6.2 Manual evaluation

The purpose of this evaluation is to know whether the meaning of Portuguese sentence is preserved in the translation, in spite of the differences in grammar and lexicon in relation to the reference. Thus, 11 automatic assessment sentences were chosen, which have significant differences in lexicon and gloss order, which may affect the understanding of the sentence. The assessment was carried out with 4 linguistic experts with knowledge in LGP and Portuguese, who were presented with glosses and asked to translate them into Portuguese and to classify them as to the quality of sentence translation using a *Mean Opinion Score* (MOS) scale, in *poor*, *fair* and *good*. *Poor* when the meaning of the translation is incorrect, *fair* for cases where the meaning of the translation is correct but the grammar fails in some way(s) and *good* when the meaning of the translation and its grammar are correct.

The glosses sequences presented to the participants correspond to translations produced by the PE2LGP system (configuration V).

6.2.1 Results

The quality of the translation of this system for 25% of sentences was *just*, while for the rest (75%) it was classified as *good*.

6.2.2 Discussion

The previous values indicate that the meaning of the sentence was preserved in all translations of the PE2LGP system and 75% of the translations followed the grammar of LGP.

The results of the translation of negative sentences stand out in this evaluation for showing problems in all grammatical aspects (sentence structure, gloss order, facial expressions and lexicon). In all negative sentences, the participants indicated that the verb should be placed before the sign of negation or simultaneously with it, depending on the verb. For instance, the verb *TER* (in English: to have) in the sentence *NAMORADO MEU TER OLHOS VERDES* {*NÃO*}*(headshake)* (in English: *my boyfriend does not have green eyes*) should appear before the sign *NÃO* (in English: no), since

Table 2: Experimental configurations.

Configuration	**Procedure**
Baseline	
I	SVO
Set 1 – only manual rules	
II	SOV structure
III	Structure according to the corpus
Set 2 – manual and translation rules	
IV	SOV structure
V	Structure according to the corpus

Table 3: Results of the 5 experimental configurations.

Configuration	**TER**	**BLEU**	
		1-grama	**2-gramas**
Baseline			
I	0.86	0.5	0.13
Set 1 – only manual rules			
II	0.3	0.75	**0.64**
III	0.4	0.75	0.47
Set 2 – manual and translation rules			
IV	**0.29**	**0.77**	**0.64**
V	0.4	**0.77**	0.49

the negation modifies the verb. For 50% of the participantes, the verb *TER* was considered as a copulative verb, which means that it should be embedded in the object (*OLHOS VERDES*, in English: *green eyes*).

In addition to the order of the constituents, this type of sentence presents errors in manual signs and facial expressions. However, there is no consensus on these two aspects among the participants in any given same sentence. Some said that the manual sign *NÃO* (in English: no) is not correct (it should be the sign *NADA* (in English: nothing)), while others said that the negation is simultaneous to the verb and it is produced only through a facial expression and also that the facial expression *headshake* is not the most suitable for the given context.

Regarding the interrogative sentences, the marking of facial expressions was classified as correct, however, the participants indicated that there are other possibilities that for them are the most correct. These possibilities vary among participants, with no consensus. For example, for the sentence *ESTADO PODER TER ?*, they indicated two following variations of the position of the interrogative facial expression (lifting the chin, tilting the head back and frowning) in the sentence: occuring during the last sign *TER* or from the sign *PODER* gloss until the end of the sentence.

Finally, the observations made during the interview by the participants indicate that the comprehension of the gloss sequences was affected by the lexical ambiguity inherent to the gloss system and by the lack of contextualization of the sentences. For example, most participants interpreted the gloss *SEGURANÇA* (in English: security) in *SEGURANÇA QUERER TAMBÉM RESPEITO* (in English: the security also wants respect) as the feeling of security rather than the profession of a security guard. This is an important aspect to take into account when evaluating gloss sequences.

7 Conclusions and future work

The construction of a European Portuguese to LGP translation system is conditioned by the few computational (and, in the case of LGP, linguistic) resources available for these languages. The main innovation of this translator compared to its predecessors is the exploration of the new corpus under development by Universidade Católica Portuguesa.

The corpus can be used to extract grammatical information about LGP. Thus, in addition to using manual rules, the translation system makes use of this annotated corpus to generate automatic translation rules in order to obtain translations from Portuguese to LGP which reflect the grammar of the language.

The results show that the proposed translation approach is capable of capturing grammatical phenomena and producing sentences in LGP instead of signed Portuguese. The system showed good results in terms of intelligibility, despite the known limitations in the production of negative sentences, identification of sentence elements and in syntactic transfer, caused by the granularity of the morphosyntactic rules. The presented study suggests that this approach should be improved upon and explored in future research work, as it represents a promising strategy in the current context of the resources available for these two languages.

References

Almeida, Inês, Luísa Coheur & Sara Candeias. 2015a. Coupling natural language processing and animation synthesis in portuguese sign language translation. Em *Vision and Lan-*

guage 2015 (VL15), EMNLP 2015 workshop (accepted for publication), Lisbon, Portugal.

Almeida, Inês, Luísa Coheur & Sara Candeias. 2015b. From european portuguese to portuguese sign language. Em *6th Workshop on Speech and Language Processing for Assistive Technologies (accepted for publication – demo paper)*, Dresden, Germany.

Bettencourt, Maria Fernanda. 2015. *A ordem de palavras na língua gestual portuguesa: Breve estudo comparativo com o português e outras línguas gestuais*: Faculdade de Letras da Universidade do Porto. Tese de Mestrado.

Brour, Mourad & Abderrahim Benabbou. 2019. Atlaslang mts 1: Arabic text language into arabic sign language machine translation system. *Procedia computer science* 148. 236–245.

Bungeroth, Jan & Hermann Ney. 2004. Statistical sign language translation. *Workshop on Representation and Processing of Sign Languages, 4th International Conference on Language Resources and Evaluation, LREC 2004* 105–108.

Carmo, Helena, Verónica Milagres da Silva & Elsa Martins. 2017. Os verbos em negação na língua gestual portuguesa. *Cadernos de Saúde* 9. 15–25.

Chéragui, Mohamed Amine. 2012. Theoretical overview of machine translation. Em *ICWIT*, 160–169. Citeseer.

Chiu, Yu-Hsien, Chung-Hsien Wu, Hung-Yu Su & Chih-Jen Cheng. 2007. Joint optimization of word alignment and epenthesis generation for chinese to taiwanese sign synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(1). 28–39.

Escudeiro, Paula, Nuno Escudeiro, Rosa Reis, Maciel Barbosa, José Bidarra, Ana Bela Baltasar, Pedro Rodrigues, Jorge Lopes & Marcelo Norberto. 2014. Virtual sign game learning sign language. Em *Computers and Technology in Modern Education* Proceedings of the 5th International Conference on Education and Educational technologies, Malaysia.

Escudeiro, Paula, Nuno Escudeiro, Rosa Reis, Maciel Barbosa, José Bidarra, Ana Bela Baltazar & Bruno Gouveia. 2013. Virtual sign translator. Em Atlantis Press (ed.), *International Conference on Computer, Networks and Communication Engineering (ICCNCE)*, Chine.

Escudeiro, Paula, Nuno Escudeiro, Rosa Reis, Jorge Lopes, Marcelo Norberto, Ana Bela Baltasar, Maciel Barbosa & José Bidarra. 2015. Virtual sign–a real time bidirectional translator of portuguese sign language. *Procedia Computer Science* 67. 252–262.

Farkiya, Alabhya, Prashant Saini, Shubham Sinha & Sharmishta Desai. 2015. Natural language processing using nltk and wordnet. *International Journal of Computer Science and Information Technologies* 6.

Ferreira, Rui. 2016. *Pe2lgp 3.0: from european portuguese to portuguese sign language*: Instituto Superior Técnico, Universidade de Lisboa. Tese de Mestrado.

Gaspar, Luís. 2015. *IF2LGP-Intérprete automático de fala em língua portuguesa para língua gestual portuguesa*: Instituto politécnico de Leiria, Leiria. Tese de Mestrado.

Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues & Sandra Aluisio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025* .

Honnibal, Matthew & Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear* 7.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Em *Soviet physics doklady*, vol. 10 8, 707–710.

Martins, Mariana & Ana Isabel Mata. 2017. Conexões interfrásicas manuais e não-manuais em lgp: Um estudo preliminar. *Linguística: Revista de Estudos Linguísticos da Universidade do Porto* 11. 119–138.

Nascimento, Sandra & Margarita Correia. 2011. *Um olhar sobre a morfologia dos gestos*, vol. 15. Universidade Católica Editora.

Othman, Achraf & Mohamed Jemni. 2011. Statistical Sign Language Machine Translation: from English written textto American Sign Language Gloss. *International Journal of Computer Science Issues* 8. 65–73.

Padró, Lluís & Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. Em *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey: ELRA.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. Em *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073135. https://www.aclweb.org/anthology/P02-1040.

Rodrigues, Rute Ana Ferreira. 2018. *Compreensão da língua gestual portuguesa em crianças surdas. proposta de um instrumento de avaliação*: Tese de Doutoramento.

Shieber, Stuart M & Yves Schabes. 1990. Synchronous tree-adjoining grammars. Em *Proceedings of the 13th conference on Computational linguistics-Volume 3*, 253–258. Association for Computational Linguistics.

Snover, Matthew, Bonnie J. Dorr, Richard H. Schwartz & Linnea Micciulla. 2006. A study of translation edit rate with targeted human annotation, .

Sousa, Ana Paula de Almeida. 2012. *Interpretação da língua gestual portuguesa*: Tese de Doutoramento.

Su, Hung-Yu & hung-Hsien Wu. 2009. 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(7). 1305–1315. doi:10.1109/TASL.2009.2016234.

Zhao, Liwei, Karin Kipper et al. 2000. A Machine Translation System from English to American Sign Language, 191–193. doi:10.1007/3-540-39965-8_6.