Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2629–2633
Marseille, 11–16 May 2020

Development of a Guarani - Spanish Parallel Corpus

Luis Chiruzzo[1], Pedro Amarilla[2], Adolfo Ríos[2], Gustavo Giménez Lugo[3]
[1]Universidad de la República, Montevideo, Uruguay
[2]Universidad Nacional de Asunción, San Lorenzo, Paraguay
[3]Universidade Tecnológica Federal do Paraná, Curitiba, PR - Brasil
luischir@fing.edu.uy, pj.amarilla@gmail.com, adolforinu@gmail.com, gustavo@dainf.ct.utfpr.edu.br

Abstract

This paper presents the development of a Guarani - Spanish parallel corpus with sentence-level alignment. The Guarani sentences of the corpus use the Jopara Guarani dialect, the dialect of Guarani spoken in Paraguay, which is based on Guarani grammar and may include several Spanish loanwords or neologisms. The corpus has around 14,500 sentence pairs aligned using a semi-automatic process, containing 228,000 Guarani tokens and 336,000 Spanish tokens extracted from web sources.

Keywords: Guarani, Spanish, parallel corpus

1. Introduction

Despite being widely spoken in many regions of South America, with estimates of between 5 and 12 million speakers, Guarani is an under-resourced language in the Natural Language Processing and Computational Linguistics communities. The aim of this work is to present a parallel corpus of Guarani and Spanish with sentence-level alignment. In this work we focus on the Paraguayan dialect of Guarani, called Jopara, which is the widest spoken dialect of Guarani. The corpus contains text extracted from different web sources; we describe the extraction, alignment and evaluation processes.

As stated in the declaration of the 2019 International Year of Indigenous Languages by the United Nations[1], language is a core component of human rights and fundamental freedoms. Focusing on languages that are less resourced and under-represented in the Natural Language Processing community could foster interest in the development of tools for these languages that might help making the lives of their speakers better by narrowing the digital divide. Because of this, we consider generating linguistic tools for less resourced languages such as Guarani is of the utmost importance.

The rest of the document is structured as follows: Section 2 gives a brief introduction of the Guarani language, particularly the Paraguayan dialect; Section 3 discusses some prior linguistic resources and work in the area that exist for Guarani; Section 4 presents the extraction and alignment methods used for building the corpus; Section 5 shows the results of the corpus evaluation; and Section 6 gives some final conclusions and future work.

2. Guarani and Jopara

Guarani is a language of the Tupi-Guarani family spoken by the Guarani people, a group of indigenous people from South America. Unlike other indigenous languages, Guarani did not disappear during the colonization period, and nowadays it is spoken by indigenous and non-indigenous people from several countries in South America. It is an official language in Paraguay and also in some territories of Argentina and Bolivia.

Guarani can be classified as an agglutinative polysynthetic language (Estigarribia and Pinta, 2017; Lustig, 2010), where the words are built by concatenation of affixes, generally around a head (Academia de la Lengua Guaraní (ALG), 2018). A word can act as a noun or a verb depending on the affixes it uses or the context it is in. For example the word *"memby"* could mean the noun "child" or it could be used as the root for the verb "to give birth".

In Paraguay the language has official status since 1992 and is spoken by the majority of the population in the country (Academia de la Lengua Guaraní (ALG), 2018), with only a minority of the population that speaks exclusively Spanish or other languages. However, as a result of the education policies and the demographic movements in the country throughout the years, the language is not spoken homogeneously and different levels of mixture with Spanish can be seen in different parts of the territory or different social strata (Lustig, 2010). The mix between Guarani and Spanish in Paraguay receives the name of Jopara, which is a Guarani word that could be translated as "mixture". The Jopara Guarani can be seen as a form of creole, and the many variants of its use include:

- Pure Guarani sentences:

  "Embohasamína ko marandu umi rehayhuvévape..." / *"Pass this message to the people you love..."*

- Using Guarani grammar incorporating Spanish neologisms adapted to Guarani morphology:

  "Afara orenunsiáta ko'êrõ" / *"Afara will resign tomorrow"*

  In this case *"orenunsiáta"* is based on the Spanish verb *"renunciar"* / *"to resign"*.

- Using Guarani grammar incorporating Spanish loanwords:

  "Ojuhúma 52 allanamiento Argentina gotyo ha 21 detenido, 200.000 munición ha 2.500 fusil ojokóva." /

[1]https://en.iyil2019.org/

"In Argentina there have been about 52 raids, 21 detainees, 200,000 ammunition and 2,500 rifles were collected."

In this example the words *"allanamiento" / "raid"*, *"detenido" / "detainee"*, *"munición" / "ammunition"* and *"fusil" / "rifle"* are written in Spanish.

- Using Spanish sentences incorporating Guarani loanwords:

  "También de ese horno salen humeantes y olorosas sopas paraguayas y chipa guasu." / "Smoky and fragrant Paraguayan soups and chipa guasu also come out of that oven."

  In this case the name *"chipa guasu"*, a type of food that has no translation, is used.

In this work we try to focus on sentences that are based on Guarani grammar, either with or without Spanish loanwords or neologisms.

3. Related Work

Except for a few exceptions, Guarani remains largely unexplored in the Natural Language Processing and Computational Linguistics fields. There is a reference corpus for current Paraguayan Guarani called COREGUAPA (Secretaría de Políticas Lingüísticas del Paraguay, 2019), although this corpus is available for online searches only and it is monolingual, containing no Spanish translations.

For other South American languages there have been attempts to create parallel corpora or treebanks, for example for the Quechua-Spanish pair (Rios et al., 2008). Although there have been no similar attempts for doing this kind of parallel treebank resource in Guarani, there is a small corpus of a Guarani dialect called Mbya Guarani annotated with dependencies (Thomas, 2019; Dooley, 2006) as part of the Universal Dependencies project (Nivre et al., 2016). This corpus contains 98 sentences (around 1,300 words) and has English translations of the sentences. Mbya Guarani is the language spoken by the Mbya Guarani people, and it is significantly different from Paraguayan Guarani.

Other small collections of sentences extracted from Wikipedia and automatically aligned for Guarani-English or Guarani-Spanish also exist. These corpora are very small because the Guarani version of Wikipedia itself is small. There is also an attempt to create a computer aided translation system between Guarani and Spanish (Gasser, 2018), using morpho-syntactic rules for providing translation candidates for the users.

As far as we know, our work presents the first medium sized aligned corpus of current Paraguayan Guarani and Spanish sentences.

4. Development

We collected text from different Paraguayan websites (blogs and news sources) that contained articles in Guarani and Spanish versions. In some cases the Guarani and Spanish text was intertwined in the same article (e.g. at paragraph level) and in other cases there were two different versions of the article accessible via links in the text. Manual inspection was necessary in order to separate the Guarani and the Spanish text in both cases, then a semi-automatic process was used to align the Guarani and Spanish sentences.

4.1. News text

We collected 2,878 news articles written in Guarani between December 2017 and August 2019, out of which 1,793 had a corresponding version in Spanish. However, upon manual inspection it was noticeable that the text in both documents was rather different: Guarani articles were generally much shorter (three or four lines) and represented a translation of only a fraction of the content in the Spanish versions. On the other hand, there were many articles that had Guarani version but no Spanish version.

In total we extracted 312,850 tokens in Guarani and 622,779 in Spanish. Notice the difference between the number of files and the number of tokens: although there are more articles in Guarani, the number of tokens is about half the number for Spanish. This can be explained in part because Guarani sentences tend to have fewer tokens and in part because the Spanish files included more content.

The text in these Guarani news articles uses a colloquial form of Paraguayan Guarani (Jopara) that uses Guarani grammar mixed with several Spanish terms and loanwords.

4.2. Blogs text

We collected 116 blog entries written both in Guarani and Spanish that range from folktales and biographies to descriptions of religious ceremonies. In this case the two versions of the text could be displayed side to side in the article, or intertwined in the same paragraphs. We manually processed these files to extract the sentences in both languages and align them properly.

The blog entries are on average twice as long as the news articles. We extracted 26,461 tokens in Guarani corresponding to 36,149 tokens in Spanish. In this case we are sure that both texts contain approximately the same information, so the difference in token count should be explained because of the differences in the languages.

The text extracted from Guarani blogs uses far less Spanish loanwords, and is more akin to pure Guarani, although in both cases they are forms of Paraguayan Guarani.

4.3. Sentence Alignment

As mentioned before, the blogs text was aligned manually as it was in general less noisy than the news text, but in total it represents about 11% of the corpus. For the latter, we designed a heuristic (similar to (Gale and Church, 1993)) that exploits the property that these texts tend to share many loanwords, and also include many names of people and places that are written similarly in Guarani and Spanish. First of all we used the sentence splitter from the `NLTK` library (Bird et al., 2009) to separate each document in sentences. We then used a dynamic programming algorithm that tries to find the best alignment between pairs of sentences with the following considerations:

- The weight of the alignment between two sentences is based on the Jaccard coefficient of two and three character n-grams.

- There are usually more sentences on the Spanish side, so many of them will have to be discarded: it is more likely to discard a Spanish sentence than a Guarani sentence.
- Assume that no sentences have to be split up or merged on either side.

This process was run over the news articles collected from December 2017 through March 2019 and a manual evaluation was done over the results. During that evaluation we found out that some of the assumptions done during preprocessing and automatic alignment did not hold. In particular:

- The sentence splitter was noisy for both Guarani and Spanish sentences alike. Interestingly, it committed the same kind of mistakes in both cases, for example splitting a sentence when finding not so common abbreviations or titles such as *"Cnel."* (short for *"Coronel"* / *"Colonel"*).
- Sometimes, the sentences in one language were in a different order than in the other language.
- On occasions, the articles collected for both languages did not refer to the same topic at all, probably because of an error in the linking of the articles.

Because of this, we decided to go for a semi-automatic process in a second step, as follows: For all news articles collected from December 2017 through August 2019, we created a pre-calculated version of the alignment matrices that contained information about the number of tokens and Jaccard index for n-grams of different ranges. We used this information to analyze all cases and manually correct errors in sentence splitting and wrong alignments. We also removed content in Guarani files that was written entirely in Spanish, such as verbatim quotes and sentences that had not been translated.

	Docs	Pairs of Sentences	Guarani Tokens	Spanish Tokens
News	1,742	12,077	201,685	300,048
Blogs	116	2,454	26,461	36,149
Total	1,858	14,531	228,146	336,197

Table 1: Composition of the corpus in number of documents, sentences and tokens for each language.

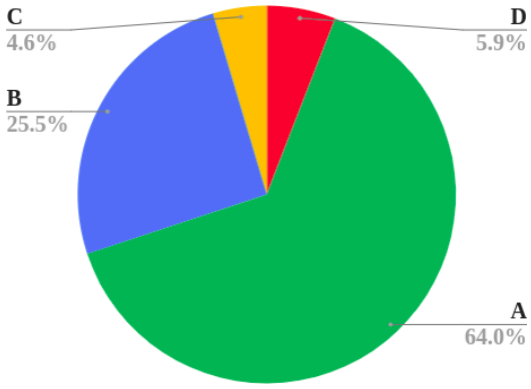
The final corpus consists of a set of text documents containing pairs of sentences. Each sentence in the corpus has a prefix "`gn:`" or "`es:`" depending on its language. Table 1 shows the number of sentences and tokens for each language in the corpus, broken down by source (news or blogs text). Notice that all of the extracted text from blogs is present in the final version of the corpus, but only about two thirds of the news text. The news text that is missing from the final corpus includes articles that only had a Guarani version and also sentences that were pruned because they could not be properly aligned.

5. Evaluation

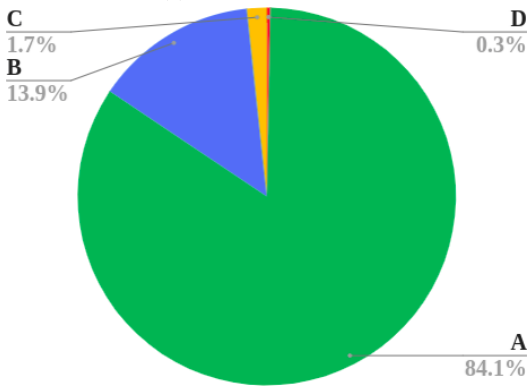
The evaluation of the corpus was done by inspecting a sample of the documents and manually analyzing the quality of the different alignments. In order to make the evaluation process fair, it was done by people that did not participate in the alignment process. The task of the evaluators was to analyze each sentence pair and indicate which of the following classes the pair belongs to:

- **A** - The Guarani sentence contains the same information as the Spanish sentence.
- **B** - The Spanish sentence contains more information than the Guarani sentence.
- **C** - The Guarani sentence contains more information than the Spanish sentence.
- **D** - The sentences do not match at all.

A summary of the results of both methods, the fully automatic one and the manually curated semi-automatic one, is shown in Figure 1.



(a) Automatic method



(b) Semi-automatic method

Figure 1: Estimated proportion of pairs that are correctly aligned (category A), partially correct with more information on the Spanish side (category B), partially correct with more information on the Guarani side (category C) or incorrectly aligned (category D) for (a) the automatic method and (b) the manually curated semi-automatic method.

Category	Original Guarani sentence	English translation for the Guarani sentence	Original Spanish sentence	English translation for the Spanish sentence
A	*Oñemombyta umi obra 17 peve*	*They stop the works until the 17th*	*Paran obras hasta el 17*	*They stop the works until the 17th*
	Oñomongeta oikóvo opavave ndive.	*We are talking with everyone.*	*Se está conversando con todos.*	*We are talking with everyone.*
B	*Opurahéi Patria Querida, péva 100 mitãpyahu poyvi ha pankarta oguerohorývo hikuái okúigui José María Ibáñez (ANR).*	*Signing Patria Querida, about 100 young people with flags and banners celebrated the resignation of José María Ibáñez (ANR).*	*Al son de la canción Patria Querida, poco más de 100 jóvenes con banderas y pancartas celebraron la renuncia de por lo menos uno de los diputados corruptos, en este caso el ladrón confeso José María Ibáñez (ANR).*	*Signing the song Patria Querida, just over 100 young people with flags and banners celebrated the resignation of* ***at least one of the corrupt legislators, in this case the confessed thief*** *José María Ibáñez (ANR).*
	Peteî tenda ojehecháva área de Urgencias.	*A place that caught their attention was the Emergency department.*	*Un sitio que les llamó la atención, a juzgar por sus rostros, fue el área de Urgencias.*	*A place that caught their attention****, judging by their faces,*** *was the Emergency department.*
	Ndorekói proyectos a futuro ni opesa oúvo.; ome'ê aguyke umi familia oipytyvõva ichupe.	*He doesn't* ***have*** *future projects or thinks about coming back; he thanks his family for supporting him.*	*Agradecido, el joven no habla de proyectos a futuro ni de cuándo piensa regresar; solo agradece a su familia por el respaldo incondicional, y a la vida, por las interminables anécdotas y experiencias.*	***Grateful,*** *the young man doesn't* ***speak about*** *future projects or when he plans to come back; he just thanks his family for their unconditional support* ***, and life, for the endless anecdotes and experiences****.*
C	*Mirta Gusinky "ojukaite" plan orekóva hikuái*	*Mirta Gusinky "killed" the plan* ***they had***	*Mirta Gusinky "mató" el plan*	*Mirta Gusinky "killed" the plan*
D	*Oñemyengovia omba'apógui hekopete*	***He was replaced for working correctly***	*Jefe policial fue destituido por atacar a contrabandistas*	***Police chief was dismissed for attacking smugglers***

Table 2: Examples of Guarani-Spanish pairs of sentences for each category, with English translation for each one highlighting the main differences in the meaning of each sentence.

5.1. Automatic method

For the first evaluation, a small sample of 20 documents (around 150 sentences) was used. From this small sample, the evaluation process found that 64.0% of the pairs were a full match, in 25.5% cases there was more information on the Spanish side, in 4.6% cases there was more information on the Guarani side and on 5.9% cases the sentences did not match. This is shown in Figure 1a.

Given this situation, we decided that the automated process was not enough to obtain a high quality alignment, so we used a semi-automatic process to align the rest of the corpus and manually correct the sentence splitting and alignments, using the n-gram overlap as a base to guide the manual analysis.

5.2. Semi-automatic method

After performing the manually curated alignment process for all the corpus, we ran a second round of evaluation with a sample of 38 documents (around 300 sentences). In this case, the evaluation process found that 84.1% of the pairs were a full match, 13.9% contained more information on the Spanish side, 1.7% contained more information on the Guarani side, and the pairs that were a complete mismatch dropped to only 0.3%, as shown in Figure 1b.

Table 2 shows examples for each one of the categories in the final version of the corpus. Sentence pairs in category C are similar to the one shown in the table, where the sentences have approximately the same meaning with some clarifications in Guarani that are not in Spanish but could be derived from context given the rest of the text. However, the most interesting class is category B, where there are more heterogeneous examples: cases with information that could be derived from context, case with differences due to nuances of meaning, but also cases where there is new information that is not present in the Guarani version, like in the first and third examples of category B in the table. Further work is needed to analyze these different classes and find ways to improve the corpus.

6. Conclusions

We presented the development of a parallel corpus of Guarani-Spanish text with sentence-level alignment. The corpus is medium-sized, containing around 228,000 tokens in Guarani along the corresponding 336,000 tokens

in Spanish. The Guarani dialect we use is Jopara, the Paraguayan Guarani dialect, which in many cases includes Spanish loanwords or neologisms. We manually evaluated a small sample of sentence pairs from the corpus and found that 84.1% of the pairs were correctly matched, a further 15.6% were correctly aligned but one of the two sentences might contain more information than the other, while only about 0.3% of the pairs were incorrectly aligned.

As future work, we plan to augment the corpus by extracting text from more sources while using this corpus to guide the alignment process in order to make it less dependent on manual curation. It would also be interesting to design ways to detect and correct examples like the ones in category B that have new information in Spanish that is not present in Guarani, either by pruning this kind of sentences or by editing them.

Furthermore, it would be interesting to use this corpus to develop a machine translation system. Due to the size of the corpus, a statistical or neural machine translation system might not yield optimal results. However, as both Guarani and Spanish are morphologically rich languages, some morpho-syntactic preprocessing might be done in order to extract more information from the text already present in the corpus. Besides this, it would be interesting to develop resources for other pairs such as Guarani-English, which might be easier to do having this initial Guarani-Spanish corpus to begin with and using the resources that have already been created for Spanish.

7. Bibliographical References

Academia de la Lengua Guaraní (ALG). (2018). *Gramática Guaraní*.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Estigarribia, B. and Pinta, J. (2017). *Guarani linguistics in the 21st century*. Brill.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Gasser, M. (2018). Mainumby: un ayudante para la traducción castellano-guaraní. *arXiv preprint arXiv:1810.08603*.

Lustig, W. (2010). Mba'éichapa oiko la guarani? guaraní y jopara en el paraguay. *PAPIA-Revista Brasileira de Estudos do Contato Linguístico*, 4(2):19–43.

8. Language Resource References

Dooley, R. A. (2006). Léxico guarani, dialeto mbyá com informações úteis para o ensino médio, a aprendizagem e a pesquisa lingüística. *Cuiabá, MT: Sociedade Internacional de Lingüística*, 143:206.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Rios, A., Göhring, A., and Volk, M. (2008). A quechua-spanish parallel treebank. *LOT Occasional Series*, 12:53–64.

Secretaría de Políticas Lingüísticas del Paraguay. (2019). Corpus de Referencia del Guaraní Paraguayo Actual – COREGUAPA. `http://www.spl.gov.py`. Accessed: 2019-11-01.

Thomas, G. (2019). Universal dependencies for mbyá guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77.