Experiments on a Guarani Corpus of News and Social Media

Santiago Góngora, Nicolás Giossa, Luis Chiruzzo
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
{sgongora,nicolas.giossa,luischir}@fing.edu.uy

Abstract

While Guarani is widely spoken in South America, obtaining a large amount of Guarani text from the web is hard. We present the building process of a Guarani corpus composed of a parallel Guarani-Spanish set of news articles, and a monolingual set of tweets. We perform some word embeddings experiments aiming at evaluating the quality of the Guarani split of the corpus, finding encouraging results but noticing that more diversity in text domains might be needed for further improvements.

1 Introduction

Guarani is a South American language spoken mainly in Paraguay, but also in some regions of Argentina, Bolivia and Brazil. Despite being an official language of Mercosur and Paraguay, research and resources for Guarani are limited. Our work focuses on the current dialect of Guarani spoken in Paraguay, called Jopara. The Jopara Guarani dialect presents different levels of mixture between Guarani and Spanish, using mainly the Guarani grammar but incorporating many Spanish loanwords (Lustig, 2010).

In this work, we present a Guarani corpus that tries to reflect the nature of this mixed dialect. On the one hand the corpus contains a set of sentences written in a more formal style, composed of news articles, where each sentence also has a Spanish counterpart. On the other hand the corpus has a more informal set of texts extracted from social media; we analyzed them and noticed the different levels of mixture between Guarani and other languages.

This is a work in progress with the aim of creating resources for Guarani that could aid in NLP tasks such as machine translation, so our objective is to make this resource as large as possible but at the same time trying to keep the content quality high. We also show an initial analysis of the corpus based on word embeddings analogies tests and visualization.

2 Related work

Although Guarani remains a little explored language within the NLP community, throughout the years there have been some attempts at creating resources or corpora for this language. COREGUAPA (Secretaría de Políticas Lingüísticas del Paraguay, 2019) is the reference corpus of current Paraguayan Guarani, it can be queried online but it cannot be downloaded in its entirety. Other works have focused on trying to develop machine translation systems or computer aided translation systems for the Guarani-Spanish pair considering the scarcity of NLP resources for the language (Alcaraz and Alcaraz, 2020; Gasser, 2018; Rudnick et al., 2014; Abdelali et al., 2006). Besides the resources focused on the Jopara Guarani dialect, there is a small Universal Dependencies corpus (around a thousand sentences) of the Guarani dialect spoken by the Mbya Guarani people (Thomas, 2019; Dooley, 2006). The work of Chiruzzo et al. (2020) describes the construction of a parallel corpus of Guarani and Spanish sentences built by downloading pages in both languages from web sources and using an automatic process (with manual correction) to align the sentence pairs. We follow a similar approach in the parallel set of our corpus, although we also add a second set of monolingual text extracted from social media. Currently, Guarani is included as one of the target languages in the machine translation Shared Task of the AmericasNLP workshop, which indicates interest in developing resources for this language is on the rise.

3 Construction of the corpus

This section presents the construction of the parallel and the monolingual sets of the corpus.

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 153–158

3.1 Parallel news set

The parallel corpus was built by crawling a set of pages restricted to the Paraguayan top level domain (.py). As a starting point, we took a set of frequent Guarani words from the Chiruzzo et al. (2020) corpus and queried different permutations of this set into a search engine, creating a set of URL seeds. Our crawler started with these seeds and downloaded, processed and cleaned each text, then used the internal links to collect more content. Although Guarani is widely spoken in Paraguay, it is a minority language with respect to the amount of text one can find in the web, where most of the Paraguayan pages are written mainly in Spanish. As noticed in Jauhiainen et al. (2020), it is very difficult to build resources for languages that are under-represented on the web, even if there is a top level domain where it is more likely to find content in that language, as the pages generally point back to content in the majority language rather than the language we are looking for. We manually inspected the early results of this experiment and found that most of the downloaded content was in Spanish. However, we also noticed that there were some Paraguayan websites which regularly publish content in both Guarani and Spanish.

We noticed two main strategies that were used by the websites to present versions of their content both in Guarani and Spanish: links within the pages to the Spanish version, and publishing the page in both languages in a short time frame. The first strategy is easy to deal with: the scraper collects the Guarani versions of the files and extracts the corresponding Spanish version following the link. This link is present in most articles, but not in all of them. If it is not found, the scraper still downloads the Guarani version[1].

For the second strategy, we designed a heuristic process for matching Guarani and Spanish files based on their timestamps. The heuristic clusters the articles by its creation date, pairing up each Guarani article to the Spanish one with the closest creation time in the group. This simple heuristic solved most of the cases, although we found two types of problematic situations:

- On occasions, the number of Guarani articles published on a given date did not match the number of Spanish ones on the same date.
- Some Guarani articles were paired with the same Spanish article due to sharing the same closest article in time in the group.

Since the number of articles affected by these issues were only a small percentage of the total, we used the heuristic for the general case and manually solved these outliers. We evaluated the heuristic results by sampling 100 random pairs and manually inspecting them, resulting in 100% correct pairs.

The parallel set is composed of 2580 news articles published in Paraguayan websites. These articles are aligned at sentence level, following the n-gram overlap heuristic described in Chiruzzo et al. (2020).

It contains a total of 14,792 Guarani-Spanish sentence pairs; including 334,501 Guarani word tokens and 635,226 Spanish word tokens. Table 1 shows a comparison between our parallel set and the one presented in Chiruzzo et al. (2020)

	Chiruzzo et al. (2020)	Parallel set
Documents	1,858	2,580
Sentences	14,531	14,792
Guarani tokens	268,684	334,501
Spanish tokens	380,275	635,226

Table 1: Size comparison between Chiruzzo et al. (2020) and our parallel set.

3.2 Tweets set

We first tried to extract tweets in Guarani using the Twitter API. The first issue was finding which of the texts contained at least some content in Guarani. The API has a language detection option that includes the Guarani language. However, this language detector API is not perfect, as we empirically found that none of the tweets was ever getting the Guarani label, even it they were written entirely in Guarani. We then trained our own language detector with the aim of telling apart between Spanish and Guarani texts, using a Naïve Bayes classifier with 5-gram character features, trained over the Chiruzzo et al. (2020) corpus. The language detector was very good for detecting Guarani in this corpus (99.6% in our test partition), but it proved to be not good enough for the noisy texts found in tweets.

Finally, we decided to use a frequent words based approach. We created two lists of frequent words: a *long list* (314 words) composed of words that appear in the corpus, filtering out dates, numbers, punctuation symbols, words with less than 3

[1]However, since they only represented 2% of the total, these articles were not included in the corpus.

	Chiruzzo et al. (2020)	Reliable text
Total Tokens	268,684	391,102
Unique tokens	31,456	41,813
Exclusive count	18,056	28,413
Overlap	13,400	

Table 2: Size comparison of the monolingual split. *Exclusive count* shows the number of tokens that are not on the other set. *Overlap* is the number of tokens that are on both sets.

characters and words that could be mistaken with other languages in the region such as Spanish and Portuguese; and a more restrictive *short list* (48 words) containing words that appear over 10 times in the corpus. We periodically collected tweets that contained at least some of the words from the short list, which includes the stop-words and many other very frequent Guarani words, both from Paraguay (*local*) and from anywhere in the world (*global*). Then we counted the number of Guarani tokens present in each tweet using the long list, and manually analyzed the extracted sets of tweets based on location and Guarani tokens. During the manual inspection we marked a tweet as a hit if it had at least some Guarani content, and a miss if all the text was in another language and was a false positive. Paraguayan (*local*) tweets with at least two of the frequent words seem to all have reliable Guarani content. However, for the *global* tweets this threshold seems to be at four words, and precision drops to around 85% with fewer words. We defined three categories:

- A (very reliable): *local* tweets with three or more frequent words and *global* tweets with four or more frequent words. (532 tweets; 7,706 tokens)
- B (reliable): *local* tweets with two frequent words. In this case, although usually containing Guarani content, there are also cases of tweets mainly in Spanish with some Guarani expression. (4,199 tweets; 48,895 tokens)
- C (unreliable): *local* tweets with just one frequent word and *global* tweets with three frequent words. This category contains many tweets in Guarani, but other languages may be present as well, such as Portuguese or Filipino. (46,197 tweets; 453,996 tokens)

We define the monolingual split of the corpus as the reliable tweets (categories A and B) plus the Guarani sentences from the parallel set. Table 2 compares the size of our monolingual split with the Guarani data from Chiruzzo et al. (2020).

4 Experiments

We carried on some experiments to try to analyze the quality of the monolingual split of the corpus built so far. We followed the approach described in Etcheverry and Wonsever (2016), where they trained a word embeddings collection for Spanish from Wikipedia text and analyzed its quality based on intrinsic tests and visualization. We trained several variants of 150-dimensional *word2vec* embeddings collections using the Gensim library (Řehůřek and Sojka, 2010). The different variants we trained correspond to using different sets of data. Besides the text collected in this work, in our experiments we also used the Guarani Wikipedia text[2], and the Guarani data from Chiruzzo et al. (2020). All models reported here were trained on some combination of those sets, a summary of the sizes of the sets is shown in table 3.

Corpus	Token count
Wikipedia	582,122
Chiruzzo et al. (2020)	268,684
Parallel news set	334,501
Reliable tweets set	56,601
Unreliable tweets set	453,996
Total (*reliable* tokens)	1,241,908
Total (*all* tokens)	1,695,904

Table 3: Guarani tokens on each set used in the experiments, tokenized using NLTK (Bird et al., 2009).

4.1 Word clustering visualization

Category	Example	Color (legend)
Years	1975	Black (k)
Months	jasyteĩ (*january*)	Black (k)
Days	arakõi (*monday*)	Black (k)
Countries	hyãsia (*France*)	Magenta (m)
Attributes	vai (*bad*)	Red (r)
Colors	hovy (*blue*)	Cyan (c)
Animals	mbarakaja (*cat*)	Green (g)
People	Romina	Yellow (y)

Table 4: Categories for the visualization experiment.

Following Etcheverry and Wonsever (2016), we selected a subset of words and created a visualization by reducing the dimensionality of the vectors. The aim of this visualization is to show that related words tend to cluster together and form regions in the vector space. The set of words contains

[2]Wikipedia dump from February 20, 2021: https://dumps.wikimedia.org/gnwiki/20210220/.

Wiki	Chiruzzo et al. 2020	Parallel News Set	Reliable Tweets	Unreliable Tweets	familiy		ccc	
					Exact	Top 5	Exact	Top 5
X					29.97%	38.89%	4.41%	10.01%
X	X				**41.27%**	**48.41%**	5.27%	11.53%
X	X	X			32.54%	34.92%	**5.53%**	**13.37%**
X	X	X	X		28.57%	36.51%	5.27%	13.04%
X	X	X	X	X	26.98%	35.71%	4.55%	12.25%

Table 5: Results for the analogies tests for the different experiments. For those words that are not in the corpus (and therefore were not trained for the word embeddings) the analogy answer is counted as wrong.

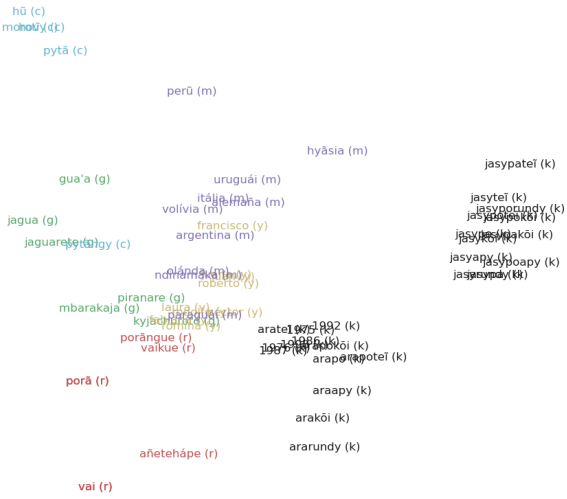

Figure 1: Embeddings visualized over a 2-dimensional space using PCA.

examples from different semantic categories (countries, colors, animals, people names, attributes and dates)[3], and are mostly Guarani translations[4] of the words used in Etcheverry and Wonsever (2016) (see table 4 for details). The embeddings used in this experiment are trained using all the available reliable text (detailed in Table 3). As can be seen in figure 1, words that represent countries (magenta), animals (green) and colors (cyan) form different clusters. Something similar happens with other categories, but notice for example that some proper names show other correlations, such as the name "Francisco" is shown close to "Argentina", probably because it is the name of the Argentinian Pope Francis, a name that appears frequently in the corpus.

4.2 Word analogies task

We did some analogies tests (Mikolov et al., 2013a) based on the vector offset method (Mikolov et al., 2013b). We had to make several simplifications in the analogies test collections due to differences in the language[5], and also because the size of the linguistic resources we use is not enough to cover a great number of the original words used in the tests. However, we were able to translate to Guarani the whole *common capital city* (`ccc`) analogies set from Mikolov et al. (2013a)[6], and we also designed a new `family` set inspired in the original one, but considering the most common family relations in Guarani dictionaries. These two analogies test collections will be made available for future reference and comparison.

Table 5 shows the results of the analogies tests for the five configurations used, corresponding to different combinations of the sets described in table 3. In order to ensure the reliability of the experiments, we ran each configuration three times and averaged the evaluation results. We show exact match and top 5 match for each experiment. First of all notice that including the Guarani part of the parallel corpus described in Chiruzzo et al. (2020) is enough to improve on the results of the Wikipedia embeddings on both categories. Using the parallel set created in this work, we can obtain better results for the *ccc* analogies test, but not for the *family* test (although they are still better for exact match than the vectors using only Wikipedia). One possible reason why the *ccc* tests improve is that this corpus includes more news articles, which frequently speak about political regions and geography, so the semantic generalization in these categories could be improved. However, our new corpus does not include a particular type of text found in Chiruzzo et al. (2020) that is text from blog posts, which includes folktales and biographies that could help the vectors improve their generalization capabilities about family members. On the other hand, includ-

[3]This categories were determined by us based on Etcheverry and Wonsever (2016) before performing the experiment.

[4]The only difference is changing the Spanish word *violeta* (purple) to the Guarani *pytãngy* (pink).

[5]For example, some English pairs do not make sense in Guarani, such as words for some family members, or the ones that change the grammatical number, which is used differently in Guarani.

[6]https://aclweb.org/aclwiki/Google_analogy_test_set_(State_of_the_art)

ing text from the tweets collections (both reliable and unreliable) seems to hinder the performance for the tests (although they still behave better than plain Wikipedia for *ccc* tests). We consider this is because text from social media tends to be much more noisy than news articles. However, it is possible that extracting a larger collection of this type of text could still help the generalization, so more experiments are needed in this regard.

5 Conclusions

We described the construction of a Guarani corpus that contains a parallel news set and a monolingual set of social media texts. We performed word embeddings experiments over different combinations of the data. The visualization experiment showed that the available text is enough to form clusters of words of the same semantic category. The analogies experiments showed that, in some cases, adding our corpus improved the performance, although results for the *family* test might indicate that more diversity of texts is needed, and text from tweets seems to be too noisy for enhancing the embeddings.

As future work, we plan to perform machine translation experiments (in line with the experiments described in Borges et al. (2021)), which might be a better way of validating the dataset. We think it is important to widen the variety of texts in the corpus: currently the crawling process keeps running daily to collect more text, and it could also be used to collect more data from different sources. Now that we have more text available and partially annotated, we can try some statistical approaches such as training a language detector for tweets instead of our keyword list strategy. We think this type of text is relevant, providing a broader and modern usage of Jopara Guarani, which might aid in other NLP tasks such as sentiment analysis.

References

Ahmed Abdelali, James Cowie, Steve Helmreich, Wanying Jin, Maria Pilar Milagros, Bill Ogden, Hamid Mansouri Rad, and Ron Zacharski. 2006. Guarani: a case study in resource development for quick ramp-up mt. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,"Visions for the Future of Machine Translation*, pages 1–9.

NB Alvarenga Alcaraz and PR Alvarenga Alcaraz. 2020. Aplicación web de análisis y traducción automática guaraní–español/español–guaraní. *Revista Científica de la UCSA*, 7(2):41–69.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Yanina Borges, Florencia Mercant, and Luis Chiruzzo. 2021. Using guarani verbal morphology on guarani-spanish machine translation experiments. *Procesamiento del Lenguaje Natural*, 66:89–98.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Robert A Dooley. 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.

Mathias Etcheverry and Dina Wonsever. 2016. Spanish word vectors from Wikipedia. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3681–3685, Portorož, Slovenia. European Language Resources Association (ELRA).

Michael Gasser. 2018. Mainumby: un ayudante para la traducción castellano-guaraní. *arXiv preprint arXiv:1810.08603*.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. 2020. Building web corpora for minority languages. In *Proceedings of the 12th Web as Corpus Workshop*, pages 23–32, Marseille, France. European Language Resources Association.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Alex Rudnick, Taylor Skidmore, Alberto Samaniego, and Michael Gasser. 2014. Guampa: a toolkit for collaborative translation. In *LREC*, pages 1659–1663.

Secretaría de Políticas Lingüísticas del Paraguay. 2019. Corpus de Referencia del Guaraní Paraguayo Actual – COREGUAPA. http://www.spl.gov.py. Accessed: 2021-03-13.

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