Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2098–2107
Marseille, 20-25 June 2022

Jojajovai: A Parallel Guarani-Spanish Corpus for MT Benchmarking

Luis Chiruzzo[1], Santiago Góngora[1], Aldo Alvarez[2], Gustavo Giménez-Lugo[3], Marvin Agüero-Torales[4,5], Yliana Rodríguez[1]
[1]Universidad de la República, Montevideo, Uruguay
[2]Universidad Nacional de Itapúa, Encarnación, Paraguay
[3]Universidade Tecnológica Federal do Paraná, Curitiba, PR - Brasil
[4]Universidad de Granada, Granada, Spain
[5]Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona, Spain
luischir@fing.edu.uy, sgongora@fing.edu.uy, aldo.alvarez@fiuni.edu.py,
gustavogl@utfpr.edu.br, maguero@correo.ugr.es, yrodriguez@fhuce.edu.uy

Abstract

This work presents a parallel corpus of Guarani-Spanish text aligned at sentence level. The corpus contains about 30,000 sentence pairs, and is structured as a collection of subsets from different sources, further split into training, development and test sets. A sample of sentences from the test set was manually annotated by native speakers in order to incorporate meta-linguistic annotations about the Guarani dialects present in the corpus and also the correctness of the alignment and translation. We also present some baseline MT experiments and analyze the results in terms of the subsets. We hope this corpus can be used as a benchmark for testing Guarani-Spanish MT systems, and aim to expand and improve the quality of the corpus in future iterations.

Keywords: Guarani, Jopara, Jehe'a, Machine Translation, language contact, corpus linguistics

1. Introduction

Guarani is one of the few indigenous languages spoken daily by both indigenous and non-indigenous people. Even though it is not a minority language in terms of its speakers, it is under-resourced and under-researched from a computational linguistics perspective. Together with Spanish, Guarani is an official language of Paraguay, making it the only South American nation where an indigenous language has survived as a majority language spoken by the non-indigenous population (Estigarribia, 2015). It is also spoken in Bolivia, Argentina and Brazil, and is one of the official languages of Mercosur[1].

However, as no language remains immutable throughout time, the long history of contact between Guarani and Spanish in South America has resulted in many interesting language varieties, the most salient being Jopara and Jehe'a. Here we will refer to the Jehe'a variant as a mixture between Guarani and Spanish where Spanish loanwords are incorporated with their morphology adapted to Guarani; while we will refer to the Jopara variant as a deeper mixture between Guarani and Spanish, often involving code switching and Spanish loanwords that keep their original morphology.

It is worth mentioning that language contact and its consequences are not exceptional nor new linguistic phenomena. In fact, Guarani-Spanish language mixing can be traced back to colonial times in the Jesuits notes, (e.g. Dobrizhoffer (1783)). Contact between both languages started with colonization during the first initial encounters with Europeans in the 1500s. In spite of this longstanding contact between Guarani and Spanish, there are very scarce bilingual texts to resort to for machine learning purposes.

In this work, we present *Jojajovai* (after the term for "parallel" in Guarani), a parallel corpus aligned at sentence level that can be used for testing machine translation (henceforth MT) systems. It might also be used for training MT systems or using the Guarani section for training monolingual resources, such as language identifiers, word embeddings, etc., and it has the potential of serving as a base for annotating POS, morphology and other tasks. Together with the parallel corpus, we also provide an analysis of the contents of the test set in terms of dialect and correctness, information that can be used to analyze the results of MT systems beyond giving plain scores. Jojajovai is the result of a collaboration between different NLP research groups from different institutions and countries. We tried to be as exhaustive as possible in compiling the available parallel data for modern Paraguayan Guarani varieties. The corpus and its annotations are available on Github[2].

2. Related Work

Interest in development of NLP tools for American indigenous languages, including Guarani, has increased in the last years, but most of these languages are still considered low-resource languages. Although there has been some work for compiling monolingual resources for Guarani (Agüero-Torales et al., 2021; Secretaría de Políticas Lingüísticas del Paraguay, 2019;

[1]Trade bloc and regional organization involving a number of countries in South America.

[2]https://github.com/pln-fing-udelar/jojajovai

Rios et al., 2014), machine translation for the Guarani-Spanish pair, and the development of parallel data to work with this pair, remain largely under-explored topics.

Regarding machine translation, there have been previous efforts for building systems that took into consideration the lack of available data (Alcaraz and Alcaraz, 2020; Gasser, 2018; Rudnick et al., 2014; Abdelali et al., 2006) or used morphology to enhance the translation (Borges et al., 2021). Moreover, other researchers have also worked on the creation of parallel corpora for this language pair. For instance Alvarez (2019b) organized a linguistic hackathon with the aim of translating Spanish sentences into Guarani, and then used these to train MT systems; (Chiruzzo et al., 2020) aimed at creating a larger corpus from web sources, which was continued by Góngora et al. (2021), who explore the difficulties found when scraping Guarani text from the web.

Despite being widely spoken, the presence of Guarani content in the web is scarce, even in Paraguayan websites, where Spanish is the predominant language. This phenomenon has been observed before when trying to build corpora for minority languages in other multilingual contexts, for example, a similar argument is presented by Jauhiainen et al. (2020) when building a web-centered corpus for Uralic minority languages, and it has been documented for the Guarani-Spanish pair as well (Góngora et al., 2021).

The first workshop on NLP for indigenous languages of the Americas (AmericasNLP) took place in June 2021 and included a shared task on machine translation for a handful of American indigenous languages (Mager et al., 2021), translating from Spanish into those languages[3]. The submitted systems were tested using manual translations of a subset of XNLI corpus (Conneau et al., 2018). This XNLI subset is made up of approximately 1,000 parallel sentences for the development set and 1,000 for the test set.

In spite of these valuable contributions, Guarani is still a low-resource language in the NLP community (Joshi et al., 2020), since the amount of resources developed is scarce and has little or no presence in main NLP conferences. Consequently, both monolingual and parallel resources are still far from getting satisfactory results.

3. Guarani Language Features

The language pair of the present paper is representative of, and responds to the linguistic reality of a big part of South America. However, for practical purposes (i.e. availability of resources) we have concentrated on Paraguayan Spanish and Modern Paraguayan Guarani (henceforth Guarani; ISO 639-3 code 'gug'; we also use the language code 'gn' in this document), which is referred to by its speakers by the name *Avañe'ẽ* ("language of men", from *ava* "man", *ñe'ẽ* "language"). Guarani belongs to the Tupi-Guarani family, further classified into the Tupian stock which comprises between 60 and 70 languages. Within the Tupi-Guarani family, Guarani is the one with the most speakers, and among the top three Amerindian languages by number of speakers (Estigarribia, 2015). Morphologically, Guarani can be classified as a language of the Amazonian type, with an agglutinating and incorporating structure (Tovar, 1961). The most outstanding morphosyntactic characteristics of Guarani are detailed in the grammars of Guasch (1983) and in that of Estigarribia (2020).

3.1. Historical Relevance

Guarani has been spoken in the South American territory for thousands of years, coming into contact with Spanish and Portuguese as a consequence of the arrival of the two European Empires. This event marked the beginning of relentless interactions with the several ethnic groups that inhabited the region, including the Guarani (Rodríguez, 2018).

Several religious orders were sent to evangelize the Guarani communities (i.e. Franciscans, Mercedarians, Dominicans and Jesuits), and the preservation of the local language seems to have facilitated this task (Rodríguez, 2019). It was more convenient for the missionaries to study this pre-hispanic language, which was then the continent's lingua franca. Furthermore, the language policy of the colonial indigenous sub-societies favored the preservation of Amerindian languages as central instruments to guarantee their social cohesion. These ecclesiastical and administrative linguistic policies facilitated the permanence of the use of Guarani. However, in the face of events such as the Andean insurrection led by Tupac Amaru, the Spanish power once again insisted on the need of a quick hispanization to better control the indigenous population. Despite this, during the colonial period no real efforts were made in this respect outside the urban centers, and even in populated areas there does not seem to have been any serious attempts to impose the Spanish language, except for the fact that it was always the official language in which bureaucratic activities were carried out and all administrative documents were recorded (Lienhard, 2003). The situation might be different today, as the Paraguayan Languages Law (Ley de Lenguas Nº4251[4]) promotes bilingualism in the spheres of public administration, but in practice this is still not happening.

Paraguay is a bilingual[5] country both in practical and legal terms. Furthermore, Spanish and Guarani are two

[3] Hñähñu, Wixarika, Nahuatl, Guarani, Bribri, Rarámuri, Quechua, Aymara, Shipibo-Konibo and Asháninka.

[4] https://www.bacn.gov.py/leyes-paraguayas/2895/ley-n-4251-de-lenguas

[5] Note that there is no universally agreed upon definition of bilingualism. In this paper we use the term in a broad manner, taking into account proficiency, usage, language history as well as schooling.

of the most used languages in South America: Spanish holds the first place, followed by Portuguese. However, when it comes to native American languages, Guarani and Quechua dispute the first place. Considering all these aspects, working with the Guarani-Spanish pair is geographically, politically and socially pertinent, considering that language technologies should be grounded on culturally relevant needs.

3.2. Varieties and Dialects

As mentioned before, in the last centuries, Guarani has had contact with Spanish and Portuguese, as well as with other indigenous languages (Rodriguez, 2017). The contact of Spanish with the American native languages is 500 years old. In colonial times the indigenous settlers outnumbered the Spaniards. However, the characteristics of the colonization did not always allow situations in which the indigenous languages influenced Spanish (Palacios, 1997). The case of Guarani stands out, since not only did the indigenous language impact Spanish, the contact produced a series of new dialects and mixed languages. Today the Guarani-Spanish contact in Paraguay is a canonical case of language contact in the Americas.

Jopara is the name of the commonly used code that mixes Guarani and Spanish in Paraguay. Its character is still debated amongst language specialists, some consider it a variety of Spanish, some a Guarani dialect and others a mixed language. The metaterm Jopara is derived from Guarani, it is composed by the Guarani morpheme *jo-* (reciprocity) and *-para* (mix), the term encompasses various ways of mixing these two main languages of the country: Spanish and Guarani (Blestel, 2021). In the present paper, Jopara is used to refer to one of the linguistic products or communication systems in which two linguistic codes intervene in some way, i.e. Guarani and Spanish (Penner, 2007).

There is also another product of the contact: Jehe'a, a term referring to the incorporation of Spanish into Guarani, which according to the Ministry of Education is the adaptation of Spanish loanwords into Guarani (Ministerio de Educación y Cultura, 2001).

3.3. A Low-Resource Language

Guarani entails the distinctive feature of being widely spoken by non indigenous people. However, in spite of that, the language is still under-resourced (Krauwer, 2003), since it has only recently stabilized its orthography and has a limited online presence, amongst other characteristics that make it difficult to work with from a computational viewpoint (lack of digital resources for language processing, bilingual electronic dictionaries, transcribed speech data, etc.). It is important to note these low-resource languages (also known as "low-data languages", "low-density languages", "resource-poor languages", and "less-resourced languages"), are not necessarily minority languages, in which a language is spoken by a small population. In fact, some under-resourced languages are actually official languages of their country (such is the case for Guarani) and spoken by a very large population (Besacier et al., 2014).

Finally, there are complex and contextual social, historical, and geographical factors that influence how best to collect a dataset in a manner that is respectful of individuals, hence sometimes dataset creators can benefit from collaborations with experts in other domains (Gebru et al., 2018). For the particular case of NLP, being aware of the linguistic context from which the language samples were extracted becomes mandatory. Because of this, meta-linguistic data has been carefully considered for the present dataset.

4. Composition of the Corpus

The Jojajovai Guarani-Spanish corpus is aligned at sentence level. The corpus is made up of several subsets, trying to include different registers and types of texts (e.g.: news, folktales and articles). There are two main reasons behind this: first of all, we want to make a corpus as diverse as possible, in order to boost MT performance in different types of texts. Secondly, we wanted to give MT developers the possibility to test on the different subsets and know their characteristics, so it is easier to analyze why the performance might work better on some subsets than on others. Our main objective is to create a comprehensive test dataset for modern Guarani that could be used as a benchmark for future reference.

We tried to be exhaustive enough to include every parallel set developed for the Guarani-Spanish pair so far. We aimed at including all the modern parallel content we could find, leaving aside other types of texts that have been historically used for machine translation due to their presence in many languages, such as the Bible or the Book of Mormon, since their register tends to be rather different from modern texts (it should be noted that, although not used for testing, we carried out some experiments using the Bible for training, as seen in section 6). We also did not include monolingual texts, since our aim was to build a purely parallel corpus. As far as we are aware, Jojajovai is the most exhaustive parallel corpus of Guarani and Spanish to date.

In addition, to explore the characteristics of the corpus, we carried out the annotation of a sample of the corpus by native speakers. We will describe this in section 5.

The corpus is structured in the following eight subsets:

abc The `abc` subset is a set of news texts built from the union of the earlier datasets by (Chiruzzo et al., 2020) and Góngora et al. (2021). These resources were extracted from the ABC Color newspaper[6], and were re-aligned with an n-gram overlap based heuristic and then manually corrected for sentence splitting and alignment errors.

anlp The AmericasNLP (`anlp`) subset comprises the development and test sets used for the Americas-

[6] `https://www.abc.com.py/`

	Tokens		Sentence pairs			
Subset	gn	es	Train	Dev	Test	Total
abc	329,476	474,001	11,550	2,470	2,472	16,492
anlp	16,619	24,126	-	996	1,004	2,000
blogs	31,676	41,468	1,712	361	371	2,444
hackaton	2,370	3,607	359	77	77	513
libro_gn	5,388	6,958	992	215	216	1,423
libro_td	3,733	5,525	711	153	152	1,016
seminario	35,624	51,435	1,535	322	322	2,179
spl	113,440	154,692	3,348	720	720	4,788
Total	538,326	761,812	20,207	5,314	5,334	30,855

Table 1: Number of tokens in Guarani and Spanish for each subset, and number of sentence pairs for each subset and split.

NLP workshop shared task (Mager et al., 2021). It is a manual translation of a subset of the well-known XNLI corpus (Conneau et al., 2018). For translating these texts, the human translator tried to keep the interference with Spanish at a minimum. The type of texts present in this subset are dialogues, with many first-person – singular and plural – sentences. We are not including the Guarani training split used in AmericasNLP, as it was a subset of other corpora we already considered here.

blogs This subset consists of blog posts, which include a variety of content like biographies, historical notes, folktales and poems. First introduced in (Chiruzzo et al., 2020), the articles were extracted from different web sources, mainly the *lenguaguarani* blog[7]. The sentences in this set were manually splitted and aligned.

hackathon At the end of 2019, Universidad Nacional de Itapúa organized a marathon (Alvarez, 2019a; Alvarez, 2019b) for translating Spanish text into Guarani, dubbed the Linguistic Hackaton. The original sentences were extracted mainly from Wikipedia and the Tatoeba platform[8]. After 4 hours and many participants, a total amount of 799 sentences were translated from Spanish into Guarani. We removed the duplicate pairs and obtained a total of 513 unique Guarani-Spanish sentence pairs.

libro_gn The `libro_gn` subset comprises the transcription of two books with parallel versions in Guarani and Spanish, edited by *Fundación Yvy Marãe'ỹ* and *Consejo Nacional de Ciencia y Tecnología de Paraguay* (CONACYT). Introduced as corpus in (Alvarez, 2022), these books contain a terminology compilation and translation guidelines for modern terms related to computer science and the internet.

libro_td Also introduced as parallel corpus in (Alvarez, 2022), this subset contains the transcription of an issue of the journal *"Territorio Digital"* where they discussed terms associated with social networks in Guarani and Spanish.

seminario The `seminario` subset consists of the transcriptions of the Guarani translation and terminology workshop *"Tercer Seminario Internacional sobre Traducción, Terminología y Lenguas Minorizadas Jarojera Guarani Ñe'ẽ"* with parallel versions in Guarani and Spanish, also edited by *Fundación Yvy Marãe'ỹ* and CONACYT. Introduced as a parallel corpus in (Alvarez, 2022), it contains the transcription of the workshop, several papers originally written in Spanish and translated into Guarani, or originally written in Guarani and translated into Spanish, and a set of vocabulary used in the workshop.

spl The `spl` subset is a set of news previously obtained by Alvarez (2019b) and Góngora et al. (2021). These news are from the Paraguayan Bureau of Linguistic Policies website[9] (in Spanish *SPL: Secretaría de Políticas Lingüísticas de Paraguay*, in Guarani *PÑS: Paraguái Ñe'ẽnguéra Sãmbyhyha*). Sentence segmentation and alignment were done automatically, with further manual inspection to detect incorrect sentence splits.

For each subset, we split the content in approximately 70%-15%-15% for training, development and test partitions. Table 1 shows the composition of the corpus. In the rest of the paper, especially in section 5, we will focus the analysis on the test partition, but notice that as the partitions were done randomly (and then the samples were taken randomly), what we find in the test set should as well hold for the rest of the contents of the corpus.

5. Analysis of the Test Set

The different subsets that comprise the corpus come from different sources and thus have different characteristics. MT developers might use this corpus as a whole for benchmark comparison, but they also might be interested in testing against each subset obtaining different results. Therefore, it is important to analyze the different subsets to get a feeling of their composition and quality.

[7] `http://lenguaguarani.blogspot.com/`
[8] `https://tatoeba.org/`
[9] `http://www.spl.gov.py/`

In this first study on the matter, three native speakers (also authors of the paper) tagged a set of pairs from the corpus along two dimensions. This follows the approaches previously laid out by (Agüero-Torales et al., 2021) and (Chiruzzo et al., 2020). For each of the sets, we sampled a small subset of pairs. The number varied between 25 and 75 according to the size of the set, so the annotators had to tag 400 pairs in total (about 7.5% of the whole test corpus). For each pair of Guarani-Spanish sentences, the annotators had to answer two questions. The first one had the intent of considering the broad spectrum of the Guarani-Spanish contact; while the other one had the intent of evaluating the alignment quality of the sets, which had been aligned with a variety of manual and semi-automatic methods.
The results of this annotation is also available on the repository, along with the corpus.

5.1. Dialect

The first question the annotators answered concerned the dialect or variety that was used in the Guarani sentence of the pair. This question was inspired by (Agüero-Torales et al., 2021), where they tried to distinguish language mixes as "gn", "gn-es", or "other". In our case, we prompted the annotators with ideas and loose definitions of what they might find (e.g. "pure" or "academic" Guarani, Jehe'a, Jopara) but left them free to indicate any variety or dialect they identified, so as to not limit the many possible answers.
Once the annotation process was ready, we analyzed the answers of all three annotators: Some of them used the expected Guarani, Jehe'a and Jopara tags, while others created new tags, and all of them left interesting remarks on some of the sentences that shed more light on difficult cases. By manually inspecting the tags and examples, we normalized all the annotations to five classes with the following working definitions:

- **Guarani** – All of the words are in "pure" or "standard" Guarani.
- **Jehe'a** – It is mostly Guarani but has some adapted Spanish loanwords.
- **Jopara** – There is a deeper mix between Guarani and Spanish, sometimes not even adapting the Spanish words to Guarani morphology.
- **Academic Guarani** – All of the words are in Guarani but it contains neologisms, newly minted Guarani words that might not sound natural to a native Guarani speaker.
- **Current Guarani** – It has some Spanish loanwords that are currently so widespread they would sound natural to native speakers. See for example Bittar Prieto (2016) and Bittar Prieto (2020).

These last two categories are very interesting, but they were used sparsely in the annotations, so we decided to merge them with the other three more standard classes. There were also a few examples where the sentence contained so many Spanish words that some annotators deemed it as "Spanish", which revealed a quality problem in some sources, which will be addressed in section 5.3.
Finally, we unified the annotations with the following criteria: if two or more annotators agreed on a tag, that is the final tag; otherwise, if the three of them gave different answers, we use the category "other", meaning that there is no consensus. Table 2 shows the unified annotation results for the dialect question.

Subset	Guarani	Jehe'a	Jopara	Other	Total
abc	4	15	55	1	75
anlp	40	7	2	1	50
blogs	50	-	-	-	50
hackaton	24	1	-	-	25
libro_gn	44	3	1	2	50
libro_td	25	-	-	-	25
seminario	45	3	1	1	50
spl	55	14	4	2	75
Total	287	43	63	7	400

Table 2: Number of unified annotations for each dialect category for each subset.

Fig. 1 shows a graphical representation of the dialects found in each subset, and an estimated proportion of the total composition of the corpus as a weighted average considering the relative size of every subset. Notice that the subsets contain mostly Guarani sentences with some examples of Jehe'a and Jopara, except the `abc` subset which is comprised mostly of Jopara examples. The size of this subset pushes the estimated number of Jopara examples in the whole test set, so we estimate the total mix contains about 47.6% Guarani, 35.7% Jopara, 15.1% Jehe'a, and around 1.6% not agreed upon.

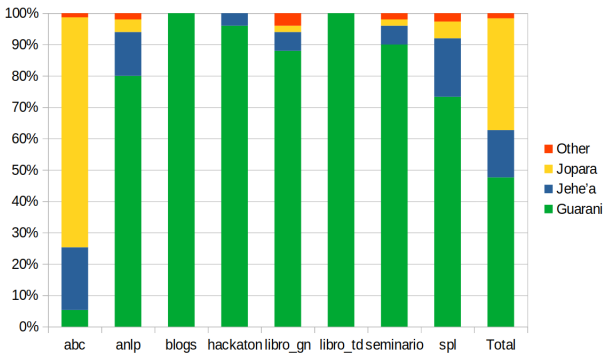


Figure 1: Proportion of dialects present in each subset. "Total" is an estimation of the composition of the whole corpus based on the sizes of each subset.

Here are some examples of each one of the three main categories, where we highlight the spans in Guarani that represent loanwords or borrowed terms from Spanish:

- Guarani

  gn – *Ha'e opyta oiko Augusta-pe upe ñembokapu riréjepe.*

  es – *Él continuó viviendo en Augusta incluso después de los ataques.*

  en – *He continued to live in Augusta even after the attacks.*

- Jehe'a

  gn – *Upe mitâ omanova'ekue* ***onasemavoi*** *hasykatu.*

  es – *El niño que murió había nacido con problemas de salud.*

  en – *The child who died had been born with health problems.*

- Jopara

  gn – *Ha'e oikuaase umi* ***cooperativa de ahorro y crédito*** *rehegua.*

  es – *Ella estaba interesada en las cooperativas de crédito.*

  en – *She was interested in credit unions.*

5.2. Correctness

The second question concerned the quality or correctness of the translation and alignment. Based on (Chiruzzo et al., 2020), we used the following four categories:

- **A** – Both sentences correspond to each other correctly.
- **B** – Sentences match, but the Spanish version has more information than the Guarani side.
- **C** – Sentences match, but the Guarani version has more information than the Spanish side.
- **D** – The sentences do not match at all.

There is a difference between having a pair of sentences that was incorrectly aligned, and having sentences that are correctly aligned but whose contents do not match. Although it would have been interesting to annotate this information as well, it was very difficult for annotators to tell these cases apart because the pairs were sampled without context, so possible misalignments could not be easily detected. Because of this, we decided to keep both cases (misalignment and content mismatch) as category D. In the future it might be interesting to carry further analyses to tell these cases apart.
In this case, we unified the annotations with the following criteria: if two or more annotators agreed on a tag, that is the final tag; otherwise, if the three of them gave different answers, we use the category "D", a sentence mismatch. Table 3 shows the annotation results for the correctness question.
In Fig. 2 the total bar shows a weighted average considering the relative size of every subset, as an estimation of the total composition of the corpus. Notice that most

Subset	A	B	C	D	Total
abc	61	8	-	6	75
anlp	44	2	1	3	50
blogs	45	1	1	3	50
hackaton	23	-	-	2	25
libro_gn	43	2	1	4	50
libro_td	22	1	1	1	25
seminario	43	1	1	5	50
spl	58	10	-	7	75
Total	339	25	5	31	400

Table 3: Number of unified annotations for each correctness category for each subset.

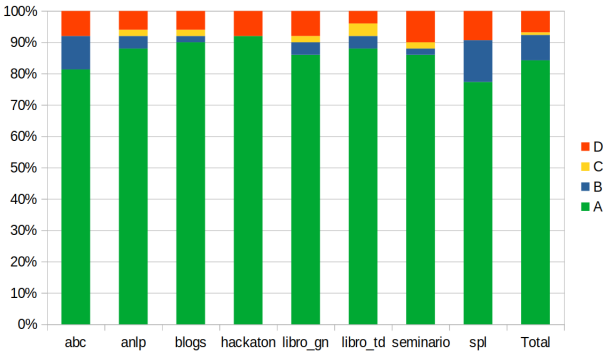


Figure 2: Correctness of the translation pairs in each subset. "Total" is an estimation of the composition of the whole corpus based on the sizes of each subset.

of the pairs are a correct match (83.5%), there are some examples where the Spanish side has more information (8.0%), and very seldom the Guarani side has more information (0.8%). However, there is an important number of examples that do not match (around 7.7%). As mentioned before, these include cases where there is a problem in alignment, and cases in which the sentences are correctly aligned but there is a content mismatch in both sides. This is an important concern that will be addressed in the future, but for now we can state that at least 93.3% of the pairs in the test set seem to be correctly aligned.

5.3. Quality Improvements

After detecting the samples tagged as "Spanish" by some annotators, we noticed there were some cases where the pairs were not translated, and the Guarani version was a verbatim copy of the Spanish version, and other cases in which the Guarani version was almost a verbatim copy of the Spanish version but changing only a few words. This happened particularly in cases of reported speech, e.g. in the following example:

- gn – *"Los consejos de salud indígena, de educación escolar indígena y la formación de los facilitadores judiciales son algunos avances significativos en materia de derechos humanos y atención de los derechos lingüísticos", omombe'u Ladislaa.*

- es – *"Los consejos de salud indígena, de educación escolar indígena y la formación de los facilitadores judiciales son algunos avances significativos en materia de derechos humanos y atención de los derechos lingüísticos", expresó Ladislaa.*
- en – *"The indigenous health councils, indigenous school education and the training of judicial facilitators are some significant advances in human rights and attention to linguistic rights", said Ladislaa.*

In this example, the reported speech is kept exactly as it was in Spanish, and only the reporting verb is translated (*omombe'u* / *expresó* / said or expressed).

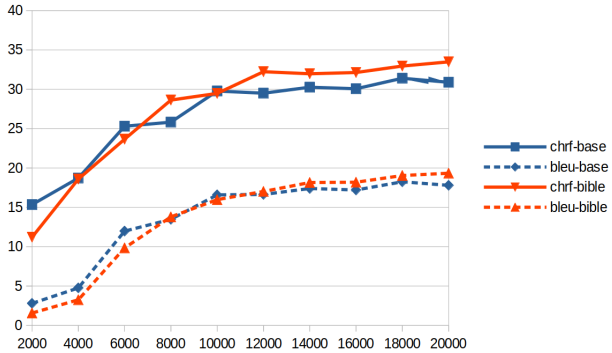
In order to cater for this phenomenon, we did a search across the entire corpus (all partitions) for pairs of sentences where the Jaccard coefficient between the Guarani and Spanish tokens was greater than 0.6. Then we manually analyzed all cases and either removed the pair altogether, or removed the spans that were copied in the cases were the remaining text still made sense. Around 150 pairs were removed or changed in this way, especially from the `abc` and `spl` subsets, which were the ones that had most of these cases.

6. MT Baseline

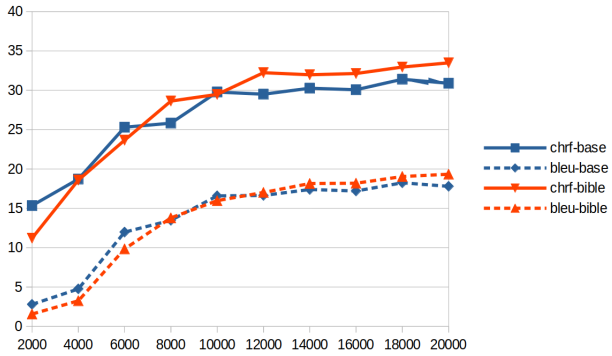
Besides presenting this unified corpus of Guarani-Spanish sentences with the analysis of the test partition, we performed some MT experiments using the training and development partitions in order to check our approach and also establish some baselines for future reference. In this work, we trained baseline models using the OpenNMT[10] tool with its default and most basic configuration, which implements an LSTM approach with an attentional model. All experiments were trained for 20K steps, then evaluated against the development set to find the best performing model. We calculated both the classic BLEU metric for MT (Papineni et al., 2002) and also the ChrF metric (Popović, 2015), which measures similarity at character level, and is considered to be more suitable for languages with rich morphology (Popović, 2017; Mager et al., 2021) such as Guarani.

We tried two different configurations: training using only our training partition (base), and training using our training partition plus the parallel text from the Bible in Guarani and Spanish (bible). The Bible contains 22,818 more sentences, which is more than the rest of our training corpus combined. Consequently, even though the language used is somewhat archaic, it might still bring important performance improvements. Fig. 3 shows the performance over the development set for the directions gn→es and es→gn. In the base case for both directions, the peak seems to be around step 18K. For the bible case, the performance for ChrF seems to still be gaining modest improvements, but the BLEU seems to plateau around 18K as well. Thus, for consistency, we will consider the models for trained 18K steps when comparing against the test corpus.

[10] `https://opennmt.net/`



(a) gn→es



(b) es→gn

Figure 3: Global performance of MT models for 20,000 steps over the development set.

Table 4 shows the performance of the models against the test corpus, broken down by subset, showing both ChrF and BLEU metrics. First of all, notice that the ChrF metric is always higher than the BLEU metric in all cases, we can say that ChrF is more relaxed than BLEU, as it compares characters instead of whole words. Both metrics are not directly comparable, but we can see that they are in general correlated.

As expected, there are huge differences in performance between subsets. This fact could remain hidden if we only took in consideration the global metric that takes the test corpus as a whole. The performance for the subsets `abc` and `spl` is the highest in both directions, which is somewhat surprising given that the main dialect present in `abc` is Jopara, while the main dialect in `spl` is Guarani, but notice that both subsets are the ones that have the highest prevalence of other varieties different than pure Guarani. The reason why the performance for these two subsets is so high might be in part because both make extensive use of shared vocabulary between source and target (for example, using many named entities with little or no changes), and the models might learn to translate those terms very easily.

The worst performance for any of the models is achieved for the `anlp` subset; this is expected as no sentences from `anlp` are used for training, it is only development and test sets. Furthermore, our baseline models do not achieve the performance obtained by the best model for Guarani in the AmericasNLP

Dir	Model	Metric	Global	abc	anlp	blogs	hackaton	libro_gn	libro_td	seminario	spl
gn→es	base	ChrF	31.84	40.25	14.77	24.71	19.35	17.15	24.02	23.15	41.68
		BLEU	19.06	20.84	1.55	11.89	6.45	5.40	10.25	6.37	25.93
	bible	ChrF	33.31	42.03	17.19	25.40	23.58	19.08	26.45	23.05	41.24
		BLEU	19.98	22.14	2.52	12.50	6.48	7.80	8.56	6.80	25.83
es→gn	base	ChrF	29.41	37.44	14.10	21.35	20.02	16.98	24.10	19.83	37.49
		BLEU	16.10	18.24	0.75	7.73	3.09	3.44	5.15	3.02	20.73
	bible	ChrF	35.28	46.14	18.67	25.45	23.39	19.15	28.25	22.32	39.63
		BLEU	20.77	24.48	1.76	11.26	3.06	7.46	3.38	5.15	23.51

Table 4: Performance breakdown of the best models over every subset of the test set.

shared task, which is the Helsinki model (Vázquez et al., 2021), obtaining 33.6 ChrF and 6.13 BLEU for the es→gn direction without using the development data for training. Our baseline results for the `anlp` subset would fit in the middle of the table for that competition.

For the gn→es direction, one important antecedent is shown in Borges et al. (2021), which is evaluated using a test partition from the (Chiruzzo et al., 2020) corpus, roughly a combination of our `abc` and `blogs` subsets. In that paper, the authors trained a system enriched with morphological annotations to improve the performance of the neural model. Although the partition is different and thus not directly comparable, the results achieved in that work were 20.3 BLEU. Our baseline models have at least that performance for the `abc` subset, but it is a few points below for the `blogs` subset.

When comparing the use of the Bible as training data together with our training set, the metrics are almost always better, but this difference is more noticeable in the es→gn direction, and especially for the ChrF metric. This might indicate that the sheer volume of examples present in the Bible, even when the language and style are not modern, helps to generalize better in an agglutinative language such as Guarani.

7. Conclusions

We presented a Guarani-Spanish parallel corpus consisting of 30,855 sentence pairs, split into around 20K pairs for training, 5K for development and 5K for test. The corpus is structured as a collection of subsets coming from different sources and having different characteristics. A sample of pairs from each subset of the test corpus was manually annotated and inspected in terms of the dialect or variety of Guarani used, and in terms of correctness or quality of alignment and translation.

The Guarani-Spanish language pair addressed in this study has been in contact for centuries, generating several contact varieties. In spite of this, bilingual literature is scarce, and distinguishing between the many varieties of Guarani spoken in Paraguay has been a challenging endeavor. We are convinced that the interdisciplinary spirit of this study is also a novelty for the field, as it provides many meta-linguistic details about the corpus. Our ultimate goal is to create a functional corpus that could be used for MT benchmarking purposes, and also to allow for better analysis of results, beyond the simple collection of metrics, taking into account the different varieties present in the corpus.

This is a first iteration of the corpus, and in the future we hope to improve its quality and the depth of analysis. For example, an important question that was left out of the annotation process is the genre of the text. We know some of the subsets contain a mixture between narrative, lyric and article elements, which might have different registers and behave differently in terms of translation. It would be worth the while to categorize these elements to improve the capabilities of the corpus in terms of analysis. We also plan to expand the corpus in the future, incorporating more sources and more dialects, even including other varieties spoken exclusively by native indigenous population, or spoken in other South American countries.

We presented neural translation baselines for our corpus, trained using only the information of the corpus or including the lengthier contents of the Bible. The results of these models were analyzed in terms of the differences between subsets, which would not be possible if we only considered the corpus globally. However, neural models are not the only possible MT models to use, and they are not suitable in all cases (Castilho et al., 2017), especially in low-resource scenarios (Mager et al., 2018). So in the future we would like (and also encourage other MT researchers) to try other approaches, such as a those purely statistical, or the use of rule-based approaches to leverage some information that could be used in statistical models.

8. Bibliographical References

Abdelali, A., Cowie, J., Helmreich, S., Jin, W., Milagros, M. P., Ogden, B., Rad, H. M., and Zacharski, R. (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.

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

Alvarez, A. (2019a). Linguistic hackathon: Accelerating bilingual data generation through collaboration for guarani-spanish language pair. In *Presentation at*

8th Podlasie Conference on Mathematics (8th PCM), Białystok, Poland, December.

Alvarez, A. (2019b). Statistical Machine Translation, an Approach for Guarani, a Low-resource Indigenous Language. Master's thesis, University of Illinois at Chicago, Chicago.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100.

Bittar Prieto, J. (2016). A variationist perspective on Spanish-origin verbs in Paraguayan Guarani. Master's thesis, The University of New Mexico, New Mexico. URL: https://digitalrepository.unm.edu/ling_etds/4.

Bittar Prieto, J. (2020). A Constructionist Approach to Verbal Borrowing: The Case of Paraguayan Guarani. URL: https://www.youtube.com/watch?v=C5XiLqR4onA, 4. The University of New Mexico's Latin American & Iberian Institute 2020 PhD Fellows.

Blestel, É. (2021). Entramados lingüísticos e ideológicos a prueba de las prácticas: español y guaraní en paraguay. In *Prácticas lingüísticas heterogéneas: Nuevas perspectivas para el estudio del español en contacto con lenguas amerindias*, pages 69–86. Language Science Press.

Borges, Y., Mercant, F., and Chiruzzo, L. (2021). Using guarani verbal morphology on guarani-spanish machine translation experiments. *Procesamiento del Lenguaje Natural*, 66:89–98.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120, 05.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. (2018). Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Dobrizhoffer, M. (1783). *Geschichte der Abiponer, einer berittenen und kriegerischen nation in Paraquay*, volume 1. Joseph edler von Kurzbek, Traduccion al español: Dobrizhoffer, Martín (1967-1970), Historia de los Abipones, 3 vols., Resistencia, Universidad del Nordeste., Vienna, Austria.

Estigarribia, B. (2015). Guarani-spanish jopara mixing in a paraguayan novel: Does it reflect a third language, a language variety, or true codeswitching? *Journal of Language Contact*, 8(2):183–222.

Estigarribia, B. (2020). *A Grammar of Paraguayan Guarani*. London, UCL Press.

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

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. (2018). Datasheets for datasets, March.

Góngora, S., Giossa, N., and Chiruzzo, L. (2021). Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online, June. Association for Computational Linguistics.

Guasch, A. (1983). *El idioma guaraní: gramática y antología de prosa y verso*. Loyola, Asunción, Paraguay.

Jauhiainen, H., Jauhiainen, T., and Lindén, K. (2020). Building web corpora for minority languages. In *Proceedings of the 12th Web as Corpus Workshop*, pages 23–32, Marseille, France, May. European Language Resources Association.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world.

Krauwer, S. (2003). The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM*, volume 2003, pages 8–15, Moscow, Russia.

Lienhard, M. (2003). *La voz y su huella*. Ediciones Casa Juan Pablos, Universidad de Ciencias y Artes de Chiapas, Mexico D.F., Mexico.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., and Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online, June. Association for Computational Linguistics.

Ministerio de Educación y Cultura. (2001). *Estudio sobre bilingüismo en el marco de la reforma educativa*. Ministerio de Educación y Cultura de Paraguay, Asuncion, Paraguay.

Palacios, A. (1997). Situaciones de contacto lingüístico en hispanoamérica: español y lenguas amerindias. In *Actas do Simposio Internacional sobre o bilingüismo*, Vigo, Spain.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational*

Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Penner, H. (2007). Se habla. Es guaraní. No es guaraní. Es castellano. No es castellano. Es guaraní y castellano. No es ni guaraní ni castellano.¿ Qué es? *Signos lingüísticos*, 3(05).

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Popović, M. (2017). chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

Rodriguez, Y. (2017). Vestiges of an amerindian-european language contact: Guarani loanwords in uruguayan spanish.

Rodríguez, Y. V. (2018). Language contact and the indigenous languages of uruguay. In *Biculturalism and Spanish in Contact*, pages 217–238. Routledge.

Rodríguez, Y. (2019). Spanish-guarani diglossia in colonial paraguay: A language undertaking. In *The Linguistic Heritage of Colonial Practice*, pages 153–168. De Gruyter.

Rudnick, A., Skidmore, T., Samaniego, A., and Gasser, M. (2014). Guampa: a toolkit for collaborative translation. In *LREC*, pages 1659–1663.

Tovar, A. (1961). *Catálogo de las lenguas de América del Sur*. Sudamericana, Buenos Aires, Argentina.

Vázquez, R., Scherrer, Y., Virpioja, S., and Tiedemann, J. (2021). The helsinki submission to the americasnlp shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264.

9. Language Resource References

Agüero-Torales, Marvin and Vilares, David and López-Herrera, Antonio. (2021). *On the logistical difficulties and findings of Jopara Sentiment Analysis*. Association for Computational Linguistics.

Alvarez, A. (2022). Initial parallel corpus creation and statistical machine translation experiments for spanish-guarani pair of languages. In *(under review) Universidad Nacional de Itapúa, Dirección de Investigación y Ambiente. Facultad de Humanidades, Ciencias Sociales y Cultura Guaraní. Facultad de Ingeniería.*, Encarnación, Paraguay.

Chiruzzo, Luis and Amarilla, Pedro and Ríos, Adolfo and Giménez Lugo, Gustavo. (2020). *Development of a Guarani - Spanish Parallel Corpus*. European Language Resources Association.

Rios, A. A., Amarilla, P. J., and Giménez-Lugo, G. (2014). Sentiment categorization on a creole language with lexicon-based and machine learning techniques. *2014 Brazilian Conference on Intelligent Systems*, pages 37–43.

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