Initial Experiments for Building a Guarani WordNet

Luis Chiruzzo[1] **Marvin M. Agüero-Torales**[2,3] **Aldo Alvarez**[4] **Yliana Rodríguez**[1]
[1]Universidad de la República, Montevideo, Uruguay
[2]University of Granada, Granada, Spain
[3]Global CoE of Data Intelligence, Fujitsu, Madrid, Spain
[4]Universidad Nacional de Itapúa, Encarnación, Paraguay
luischir@fing.edu.uy,maguero@correo.ugr.es,aldo.alvarez@fiuni.edu.py,yrodriguez@fhuce.edu.uy

Abstract

This paper presents a work in progress about creating a Guarani version of the WordNet database. Guarani is an indigenous South American language and is a low-resource language from the NLP perspective. Following the *expand* approach, we aim to find Guarani lemmas that correspond to the concepts defined in WordNet. We do this through three strategies that try to select the correct lemmas from Guarani-Spanish datasets. We ran them through three different bilingual dictionaries and had native speakers assess the results. This procedure found Guarani lemmas for about 6.5 thousand synsets, including 27% of the base WordNet concepts. However, more work on the quality of the selected words will be needed in order to create a final version of the dataset.

1 Introduction

Guarani is an indigenous South American language spoken by around 6.5 million native speakers, mainly in Paraguay and in parts of Bolivia, Argentina and Brazil. Despite being one of the most widely spoken languages in the region, it has received little attention from a computational linguistic perspective. In the latest years, interest in natural language processing (NLP) research for indigenous languages of the Americas has increased, and nowadays, a number of researchers are building tools and resources for many of these languages, such as multilingual corpora (Mager et al., 2021). However, the creation of lexical databases and ontologies, such as WordNets, is only very recently starting to gather attention.

WordNet (Miller, 1995) is a lexical database, originally created for English but later on for many other languages (e.g. Gonzalez-Agirre et al., 2012; Vossen, 1998), that organizes concepts in an ontology of inter-related terms. The basic unit of WordNet is the synset, defined as a set of words that could be used interchangeably, at least in some context, and is similar to the notion of a sense or meaning in a dictionary. Synsets are organized in an ontology with hyponymy as the central relation between concepts, but also including (depending on the POS) other relations such as meronymy, antonymy or implication. The concepts stored in WordNet belong to one of the four lexical categories: nouns, verbs, adjectives and adverbs.

Historically there have been two main approaches to building WordNets (Bosch and Griesel, 2017; Vossen, 1998), which are: manually creating a new set of concepts for each language and establishing links to the original Princeton WordNet (named the *merge* approach), or using the original structure of Princeton WordNet and translating the lemmas corresponding to the different concepts into the target language (named the *expand* approach). In this paper, we present a work in progress for building a WordNet database for the Guarani language using the *expand* approach. We collected different bilingual datasets (i.e. Guarani-Spanish dictionaries), and implemented some heuristics to select the correct Guarani lemmas that correspond to WordNet synsets. Then native speakers annotated a sample of the results obtained by the heuristics in order to assess the quality of the built resource.

2 Related Work

There have been very few attempts at creating WordNets for indigenous American languages. Two of them are about languages spoken mainly in Peru, Shipibo-Konibo (Maguiño-Valencia et al., 2018), and several varieties of Quechua (Melgarejo et al., 2022). Previous attempts for Quechua do not focus on building a WordNet ontology, but include using links to the Spanish WordNet in order to help word sense disambiguation (Rudnick, 2011) or morphological analysis (Gasser, 2010).

Bosch and Griesel (2017) describe an attempt to build WordNets in several indigenous African

languages. These attempts, as well as the ones mentioned above, generally use the *expand* approach to building WordNets, as it is the easiest one to use when at least there are bilingual datasets available.

Our work is, as far as we know, the first attempt to build a WordNet for Guarani. We focus on the modern Paraguayan variety of Guarani. Similarly to Melgarejo et al. (2022), we use the Spanish version of WordNet to support the translation, because there are more Guarani-Spanish bilingual resources available. Guarani is a low-resource language (Joshi et al., 2020) and, like other languages in this category, it lacks large monolingual and parallel corpora to build even some relatively simple NLP applications. There are some small multilingual (Mager et al., 2021) and bilingual (Chiruzzo et al., 2022, 2020) corpora that include Guarani, and even the newest version of the Google Translate tool includes Guarani as one of its options[1], but so far, the size and performance of these resources is not enough to obtain accurate lexical information.

3 Guarani language

Modern Paraguayan Guarani belongs to the Tupi-Guarani family, part of a posited Tupian stock comprising between 60 and 70 different languages. The Tupi-Guarani family is the largest family within the Tupian stock, and within it, Guarani is the language with the most speakers. Tupian languages are spoken in Brazil, Argentina, Bolivia, French Guiana, Paraguay and Peru.

3.1 Historical perspective

Following the arrival of Europeans to South America, Franciscans and Jesuits documented and standardized Guarani (Meliá, 1992). The Jesuits reduced the indigenous language to writing, and cunningly used it as the language of evangelization until they were expelled in 1767 (see Rodríguez, 2019). Guarani was declared a national language by Paraguayan leader Stroessner in 1967 in article 5 of the first chapter of the new constitution. However, it was only in 1992 that it was given co-official status together with Spanish in the Ley de Lenguas (Law of Languages) and bilingual education began to be established in 1994 (see Penner, 2016 for an analysis of the law's practical implications and outreach). Analysis of recent census data confirms previous observations that Guarani-Spanish bilingualism is higher in urban and border areas (e.g. Rubin, 1963; Solé, 1991), while high rates of Guarani monolingualism at home are limited to rural areas (Gynan, 2001).

Five centuries after Guarani was given a written code by means of using the Latin alphabet, Guarani is still not frequently used in writing. When it comes to NLP, the challenge is taken even further, as there are not many digital resources and corpora that could be used for automatic processing. The contact between the two languages and their many varieties, and its repercussions, has been studied by numerous scholars, amongst which Dietrich (2001, 2004), Kallfell (2006), Thun (2006) and Zajícová (2010) stand out. Within the ample scope of the contact scenario and its outcomes, we will constrain to the matter of Jopara, the very commonly used code in Paraguay that resorts both to Guarani and Spanish (for a structural analysis of Jopara see Thun, 2005; Gómez Rendón, 2008 and Kallfell, 2011). Although scholars do not agree on whether Jopara is a variety of Spanish, a variety of Guarani, a new mixed language, the result of code-switching (as Estigarribia, 2015 states) or languages that keep mixing (the latter is argued by Thun, 2005, p. 311), the fact that there is a code in which two languages are being mixed is relevant for the purpose of our work. The features of a mixture of languages (that is what the word *Jopara* actually means in Guarani: mixture) can hinder our work in the manners presented in section 5.3. Interestingly, Guasch (1948) was the first to use the term to refer to the language mixing that had to be avoided (see Blestel, 2021). The sociological and attitudinal significance of such an idea has had repercussions until to date.

3.2 Language features

We now move on to present Guarani's trait characteristics following Estigarribia (2020). The overview is narrowed down to morphology given the aim of this paper. At the level of word formation, most meanings are built into a word as parts of it, as affixes or other particles (i.e. Guarani has an agglutinative morphology). There are remnants of an extensive polysynthetic behaviour, most words are composed of many parts, each with its own meaning to contribute to the whole. As a consequence, what would otherwise be a whole sentence in English is a single word in a polysynthetic language like Guarani. There are two first-person

[1] https://ai.googleblog.com/2022/05/24-new-languages-google-translate.html

plural pronouns, one that includes the addressees (*ñande*) and one that excludes them (*ore*).

Guarani has specific prefixes that simultaneously represent a first-person agent acting on a second-person patient, and there are two kinds of intransitive verbs whose subjects look different (split intransitivity). There is a class of words that take different prefixes when they are in the same phrase with other words and three ways to indicate events where a participant makes another participant do something (i.e. three different morphological causatives). Verbs and other predicates are also negated by a circumfix, that is, a negation that has two parts: a prefix that comes before the verb and a suffix that comes after. For example, the verb "ndaguatái" (I do not walk) can be analyzed as *nd-a-guata-i*: the first person singular affix *a-* and the base verb form *guata* (to walk) surrounded by the negation circumfix *nd-V-i*. We consider that the base form of a verb, without any of the affixes, is the appropriate way to represent Guarani verbal lemmas in WordNet.

When it comes to nouns, they take suffixes that indicate past or future, among other interpretations (nominal temporal-aspectual inflection). Guarani has an extensive system of postpositions that come at the end of a noun phrase to indicate its relation to a predicate. Guarani's lexicon has been influenced by Spanish, however, in Paraguayan Guarani most of the basic lexicon is still of Tupi-Guarani extraction.

Even though we have used traditional grammatical relations to describe Guarani, restraining from using units of observation from other languages as units of analysis could provide a wider hold of how Guarani works, i.e. analyzing Guarani data without relying on antecedently given formal or relational structure (see Otheguy, 2002 for an elaboration on this theoretical matter). For example, there is a class of nouns in Guarani called triform nouns or relational nouns (Estigarribia, 2020; Academia de la Lengua Guaraní (ALG), 2018) which are written with a different prefix depending on their use within a sentence or structure. They take forms prefixed by *t-*, *h-* or *r-* depending on whether they are referring to the generic form of the noun, or if they relate to another participant in the sentence. However, there is a discussion around whether these sets of nouns should be considered as a base form with a set of prefixes, or as sets of three distinct lemmas. As we will see, dictionaries and native speakers tend to consider them as different lemmas, and under this assumption, the three forms would be generally included in the same synsets.

4 Process

This section describes the heuristics we use to select Guarani lemmas for the synsets, and the datasets we obtain the information from.

4.1 Selectors

We follow the selector-based strategy similar to (Pradet et al., 2014; Herrera et al., 2016; Methol et al., 2018). In these works, they define a *selector* as a strategy that takes the set of lemmas in a synset for a source language, and the set of translation candidates for those lemmas in the target language, and chooses which target language lemmas should be assigned to the synset.

The main difference we have is that in those previous works, the source language was always English, which is the best possible scenario as English is the original and most complete language of WordNet. However, there are no bilingual Guarani-English dictionaries available, at least not with a considerable size that could be used for our purposes. Because of this, we resort to the Spanish version of WordNet, which has much fewer lemmas, and Guarani-Spanish dictionaries. The efficacy of the selectors will depend on the quality of the dictionaries, but also on the adequate coverage of the Spanish version of WordNet.

The three selectors we use in this work are the following:

Monosemy Given a lemma sl in the source language that belongs to only one synset s, we consider that the lemma is monosemic. In that situation, assign all the possible translations of sl in the target language $\{tl_1, ..., tl_n\}$, to the synset s. The intuition is that if sl only has one sense, its counterparts tl_i should have the same sense.

Single Translation Consider a lemma sl in the source language that belongs to one or more synsets $\{s_1, ..., s_n\}$, and according to the dictionary, the lemma has only one possible translation tl in the target source. In this case, assign tl to all synsets $\{s_1, ..., s_n\}$. The intuition is that if we had a perfect dictionary with all possible translations and there is only one way to translate sl, that translation should be valid for all senses of sl. Of course, this assumption does not happen in real life, so it will

depend on the quality and coverage of the available dictionaries.

Factorization Given a synset s that has lemmas $\{sl_1, ..., sl_n\}$ in the source language. Each source lemma sl_i has a corresponding set of lemmas in the target language $\{tl_{i,1}, ..., tl_{i,k_i}\}$. This selector takes the intersection of all these sets and assigns all the lemmas in the intersection to s. In this case we also ask that s has at least two lemmas in the source language.

4.2 Dictionaries

As mentioned above, the success of these selectors will be significantly influenced by the quality of the translation resources we can find. Given that Guarani is a low-resource language from the point of view of NLP, and the existing machine translation (MT) systems for this language are still not accurate enough, we relied mainly on bilingual Guarani-Spanish dictionaries. These are the sources we collected:

Avalos The Ñe'ẽryruguasu bilingual dictionary (Ávalos, 2011) contains more than 17,000 entries of Guarani words with Spanish translations and examples in Guarani. It also contains the POS of each Guarani entry, which is very helpful for determining the appropriate synsets. The dictionary was compiled in PDF format, and there were many transcription issues when converting it to plain text format for processing. We used rules to detect full spans that were appropriately transcribed and contained entries with available translations, such as:

"`guarani_lemmas [guarani_pos] spanish_lemmas`"

Not all entries and variants could be converted in this way, but we ended up with a set of 18,698 Guarani-Spanish lemma pairs.

DC Descubrir Corrientes[2] is a web portal that contains an online Guarani-Spanish bilingual dictionary. The entries in this dictionary also indicate the POS (in this case in Spanish) of the words. We processed this dictionary (as in Borges et al., 2021) and compiled a set of 14,164 Guarani-Spanish lemma pairs.

Wiktionary Wiktionary[3] is a project for creating open multilingual dictionaries, part of the Wikimedia foundation. The Guarani language still has very few resources inside the Wiki ecosystem, and Wikipedia and Wiktionary are no exception. In the latest dump of the Guarani Wiktionary (September 1, 2022), there were only 2,499 Guarani-Spanish pairs, 207 Guarani-English pairs, and 113 Guarani-Portuguese pairs. The words in the Guarani Wiktionary also lacked a clear way of determining their POS, so we used the Spanish lemmas lists categorized by POS from the FreeLing project (Padró and Stanilovsky, 2012). We assigned the POS of the Spanish lemma associated with a Guarani word, which is not perfect since a word could have multiple POS but only one of them could be appropriate in the other language, so this is a potential source of noise for these lemmas. After this process, we ended up with 2,276 Guarani-Spanish lemma pairs for this source.

[2]https://descubrircorrientes.com.ar/2012/index.php/diccionario-guarani/

[3]https://www.wiktionary.org/

Category		Unique Pairs	Unique Synsets	Unique Lemmas
POS	Noun	6,618	3,514	2,791
	Verb	3,977	2,110	1,364
	Adjective	1,182	802	391
	Adverb	190	93	146
Rule	Monosemy	3,589	1,716	2,837
	Single Tran.	8,412	5,322	2,182
	Factorization	952	615	592
Source	Avalos	4,082	2,403	1,800
	DC	6,757	4,583	2,678
	Wiktionary	2,088	1,754	653
Base concept	Yes	2,604	1,263	1,550
	No	9,363	5,256	3,837
Overall		11,967	6,519	4,298

Table 1: Number of <synset, lemma> pairs extracted for each POS, by each rule, from each source, and belonging to the base concepts. Notice that the number of pairs for rules and sources do not add up to the overall value because some pairs were found by more than one rule or belonged to more than one source.

5 Results and evaluation

Table 1 shows the number of <synset, lemma> pairs found using the described selectors and dictionaries. We show the number of unique pairs, unique synsets, and unique lemmas. The table also breaks down the information for each POS, each selector rule, and each dictionary source. Note that the selector that yielded the most results was the *Single Translation* selector, while the one with the fewest results is *Factorization*.

The rules also found possible lemmas for 1,263 (around 27%) out of 4,689 synsets considered base concepts of WordNet[4], defined to be high in the

[4]http://globalwordnet.org/resources/gwa-base-concepts/

semantic hierarchy and to have many connections to other concepts.

5.1 Precision of the selectors

In order to evaluate the quality of the lemmas chosen by the selectors, we sampled a set of <synset, lemma> pairs generated by our rules. Two native speakers (authors of this paper) annotated the samples to identify if the selected lemmas were suitable for the corresponding synsets. We then calculated the precision of the selector based on the number of pairs considered correct by the annotators, over the total number of extracted pairs. This can be calculated as an overall measure, but we can also break it down by POS, selector or dictionary source to have a more fine-grained analysis.

The annotators were given the ID of the synset, a Spanish translation of the synset's definition, the known Spanish lemmas, and all the Guarani lemmas found by the rules. They had to indicate, for each lemma, if it was appropriate for that synset, and optionally, they could also indicate other suitable Guarani lemmas and some comments.

For example, one of the synsets to annotate was `play.v.29`, which has the definition "make bets". The Spanish lemmas for this synset are "apostar" (to bet) and "jugar" (to play or to gamble). The rules selected the Guarani lemmas "ha'ã", "ra'ã" and "ñembosarái". In this case, both annotators agreed that "ha'ã", "ra'ã" are appropriate lemmas for `play.v.29`, while "ñembosarái" was not.

Each annotator had to label 106 synsets with approximately 300 lemmas in total, but 40 of these synsets were annotated by both, so we were able to calculate the inter-annotator agreement between them. We calculated the inter-annotator agreement using Cohen's Kappa, which was 0.561 for our sample, which indicates moderate agreement.

In total, they annotated 476 <synset, lemma> pairs, having 172 unique synsets and 412 unique lemmas, approximately 4% of the total number of extracted pairs. We sampled the pairs so that there were at least some samples of each POS, rule and source, and also samples from synsets that belong to the base concepts. We aimed to have at least 60 samples (<synset, lemma> pairs) for each category.

Table 2 shows the number of samples for each category and its precision according to the annotators, calculated as the number of pairs considered correct over the total number of pairs for that category. The overall category considers all sample pairs, which have a precision of 61.3%. From the point of view of rules, the *Factorization* rule seems to work much better than the other heuristics. One possible explanation for this is that it is the most restrictive of the selectors, as we ask that there are at least two Spanish lemmas before doing the factorization process. This means that the selector can only be applied to a reduced number of synsets (see Table 1), but at the same time it helps to achieve more precise results.

Category		Samples	Precision
POS	Noun	171	0.667
	Verb	141	0.638
	Adjective	93	0.484
	Adverb	71	0.606
Rule	Monosemy	213	0.610
	Single Tran.	233	0.579
	Factorization	95	0.758
Source	Avalos	217	0.520
	DC	267	0.708
	Wiktionary	108	0.683
Base concept	Yes	120	0.625
	No	356	0.610
Overall		476	0.613

Table 2: Number of <synset, lemma> sample pairs for each category and their precision based on the annotations. Notice that the number of samples for rules and sources do not add up to 476 because some lemmas were found by more than one rule or belonged to more than one source.

If we take into account the sources, the DC and Wiktionary dictionaries seem to be much more precise than Avalos, even if in the Wiktionary case we did not have the original POS, but we had to assign them automatically from a Spanish dictionary. Additionally, the performance for adjectives is also much lower than for any other POS.

5.2 Coverage of the sources

Given that the annotators were asked to include more Guarani lemmas that they considered suitable for the synsets, we could create a small set of manually curated synsets with lemmas. For each synset, we kept the lemmas selected by at least one annotator as correct, as well as all the lemmas included as extras by them. With this information, we created a collection of 164 synsets with 446 unique lemmas we consider our small *gold standard*. There were only eight synsets for which the annotators

	Noun	Verb	Adj.	Adv.
Avalos	0.494	0.562	0.126	0.155
DC	0.607	0.711	0.116	0.239
Wiktionary	0.274	0.248	0.179	0.141
Union	0.815	0.942	0.305	0.408

Table 3: Coverage of the gold standard created by the annotators in terms of Guarani lemmas for each source. The last line shows the coverage of the union of all the dictionaries.

considered no selected lemma was suitable, and no alternatives were given.

Table 3 shows the coverage of the Guarani lemmas considered in the gold standard for each source. We consider the Guarani lemma as covered if it exists on the source associated to a particular POS, even if it is not associated to a suitable Spanish lemma. So these numbers give us an idea of how good the different dictionaries are at representing the words expected by the annotators, and are consequently an upper bound to the performance we can get when designing selectors that use these dictionaries as sources, as the selectors cannot find lemmas that are not in the sources. When we take the union of dictionaries (last line of the Table 3) the coverage seems very good for nouns and verbs, but it is notably low for adjectives.

5.3 Issues

First of all, as mentioned in section 4.1, the selectors work under some assumptions. The *Single Translation* selector would work best if we had a perfect bilingual dictionary with all possible Guarani-Spanish translations. However, no dictionary is perfect, and this is probably one of the reasons the *Single Translation* selector had poor performance in this experiment.

Furthermore, unlike other works, we use WordNet's Spanish version as starting point instead of the English version. This is not ideal, because the Spanish WordNet has considerably fewer lemmas than the English WordNet. This could have different effects on the different selectors. For example, the *Monosemy* selector relies on finding Spanish lemmas that belong to only one synset, but as the Spanish WordNet is incomplete, it is likely that many possibly polysemous lemmas are erroneously only present on one synset. This hinders the efficacy of the selector.

Finally, the three selectors we chose are very simple, and they only capture certain configurations of synsets and lemmas. We still need to design more and better selectors that could extract more information from the datasets we have, as well as collect more datasets. One way of doing this is using the parallel corpora and MT systems that are being created lately. We could also make use of similarities in some written forms of Spanish loans, similar to the Levenshtein selector described in Pradet et al. (2014), or use gloss information and word vectors as in Maguiño-Valencia et al. (2018).

About the triform nouns mentioned in section 3.2, we noticed the annotators indicated that all forms of a noun should be included as lemmas of a synset. For example, the selectors chose ten possible lemmas for the synset `branch.n.02` with the definition "a division of a stem, or secondary stem arising from the main stem of a plant", and in particular there were two sets of triform nouns selected: {takã, hakã, rakã} and {takãmby, hakãmby, rakãmby}. Both annotators agreed that the first triplet of nouns was appropriate for the synset, but disagreed about the second one. However, it was always the case that the triplets were accepted or rejected together, e.g. {tete, hete, rete} were rejected for the synset `entity.n.01` because they are more suitable to a physical entity or body.

Inconsistencies in orthography are another source of problems for this process. The Wiktionary source was the one with the most problems in this respect. For example, these three words were associated with Spanish "rama" (branch) in Guarani: {taka, hakã, rakä}. This is a triplet of nouns written in three different orthographic conventions for marking a nasal vowel: with no diacritic, with the standard tilde diacritic, and with the diaeresis diacritic, which is not standard.

6 Conclusions

We presented a work in progress on building a version of the WordNet lexical database for Guarani, an indigenous South American language. Our process obtains data from three bilingual Guarani-Spanish dictionaries, and we implemented three simple selectors that decide which Guarani lemmas should be used as the translation of the lemmas present in the Spanish WordNet synsets. The selectors are *Monosemy*, *Single Translation* and *Factorization*.

We extracted lemmas for 6,519 synsets, but the quality of the selected lemmas is highly variable. The *Factorization* method is the one that has the

highest precision according to the human annotators (around 76%), and the sources with the highest precisions are DC (71%) and Wiktionary (68%). However, there is still a lot of room for improvement. As future work, we plan to expand the manual evaluation in order to have a bigger set of curated synsets and lemmas, design new selectors that could extract better information from the sources, and collect or create more datasets, for example, using the existing bilingual corpora or MT systems.

References

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

Celso Ávalos. 2011. *Ñe'ẽryguasu (Gran Diccionario) Guaraní-Español, Español–Guaraní*.

Élodie Blestel. 2021. Entramados lingüísticos e ideológicos a prueba de las prácticas: español y guaraní en paraguay. *Sánchez Moreano, Santiago; Blestel, Élodie (éds). Prácticas lingüísticas heterogéneas: Nuevas perspectivas para el estudio del español en contacto con lenguas amerindias*, pages 69–86.

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.

Sonja E Bosch and Marissa Griesel. 2017. Strategies for building wordnets for under-resourced languages: The case of african languages. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 38(1):1–12.

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. European Language Resources Association.

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A parallel guarani-spanish corpus for mt benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2098–2107.

Wolf Dietrich. 2001. Zum historischen Sprachkontakt in Paraguay: Spanische Einflüsse im Guaraní. *Sprachkontakt und Sprachvergleich. Münster: Nodus*, pages 53–73.

Wolf Dietrich. 2004. La influencia castellana en la sintaxis de la coordinación y subordinación de lenguas tupí-guaraníes. *Paper presented ar the conference Lenguas amerindias en contacto con el castellano: aspectos lingüísticos y sociolingüísticos, Amsterdam, June 24th-25th*.

Bruno Estigarribia. 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.

Bruno Estigarribia. 2020. *A grammar of Paraguayan Guarani*. UCL Press.

Michael Gasser. 2010. Antimorfo 1.1 user's guide.

Jorge Gómez Rendón. 2008. *Typological and social constraints on language contact: Amerindian languages in contact with Spanish*. Netherlands Graduate School of Linguistics.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

Antonio Guasch. 1948. *El idioma guaraní*. Asunción: Imprenta Nacional.

Shaw N Gynan. 2001. Language planning and policy in paraguay. *Current Issues in Language Planning*, 2(1):53–118.

Matías Herrera, Javier González, Luis Chiruzzo, and Dina Wonsever. 2016. Some strategies for the improvement of a spanish wordnet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 115–122.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Guido Kallfell. 2006. Uso de las voces verbales del yopará, en comparación con las del guaraní. *Guaraní y "Mawetí-Tupí-Guaraní": Estudios Históricos y Descriptivos sobre una Familia Lingüística de América del Sur, Wolf Dietrich y Haralambos Symeonidis (eds.)*, pages 333–354.

Guido Kallfell. 2011. *Grammatik des Jopara: Gesprochenes Guaraní und Spanisch in Paraguay*. Frankfurt am Mein: Peter Lang.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez Lugo, Ricardo Ramos, et al. 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.

Diego Maguiño-Valencia, Arturo Oncevay-Marcos, and Marco A. Sobrevilla Cabezudo. 2018. WordNet-shp: Towards the building of a lexical database for a Peruvian minority language. In *Proceedings of the Eleventh International Conference on Language*

Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Nelsi Melgarejo, Rodolfo Zevallos, Héctor Gómez, and John E Ortega. 2022. Wordnet-qu: Development of a lexical database for quechua varieties. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4429–4433.

Bartomeu Meliá. 1992. La lengua guarani del paraguay, historia, sociedad, literatura.

Alfonso Methol, Guillermo López, Juan Álvarez, Luis Chiruzzo, and Dina Wonsever. 2018. Using context to improve the spanish wordnet translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 17–24.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ricardo Otheguy. 2002. Saussurean anti-nomenclaturism in grammatical analysis: A comparative theoretical perspective. In *W. Reid, R. Otheguy and N. Stern (eds.) Signal, Meaning and Message. Perspectives on Sign Based Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *LREC2012*.

Hedy Penner. 2016. La ley de lenguas en el paraguay: ¿un paso decisivo en la oficialización de facto del guaraní? *Signo y seña*, (30):108–136.

Quentin Pradet, Gaël De Chalendar, and Jeanne Baguenier Desormeaux. 2014. Wonef, an improved, expanded and evaluated automatic french translation of wordnet. In *Proceedings of the Seventh Global Wordnet Conference*, pages 32–39.

Yliana Rodríguez. 2019. Spanish-guarani diglossia in colonial paraguay: A language undertaking. *The Linguistic Heritage of Colonial Practice*, 13:153–168.

Joan Rubin. 1963. *National bilingualism in Paraguay*. The Hague: Mouton.

Alex Rudnick. 2011. Towards cross-language word sense disambiguation for quechua. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 133–138.

Yolanda R Solé. 1991. The Guarani-Spanish situation. *Georgetown Journal of Languages and Linguistics*, 2:297–348.

Harald Thun. 2005. 'code switching','code mixing','reproduction traditionnelle' et phénomènes apparentés dans le guarani paraguayen et dans le castillan du paraguay. *Italian Journal of Linguistics, 17-2*, pages 311–346.

Harald Thun. 2006. 'a dos mil la uva, a mil la limón'. historia, función y extensión de los artículos definidos del castellano en el guaraní jesuítico y paraguayo. *Guaraní y "Mawetí-Tupí-Guaraní": Estudios Históricos y Descriptivos sobre una Familia Lingüística de América del Sur, Wolf Dietrich y Haralambos Symeonidis (eds.)*, pages 357–414.

Piek Vossen. 1998. Eurowordnet: A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers*, 10:978–94.

Lenka Zajícová. 2010. Differences in incorporation of spanish elements in guarani texts and guarani elements in spanish texts in paraguayan newspapers. *A new look at language contact in Amerindian Languages*, pages 185–203.