Development of the Parallel Corpus of Mexican Languages (CPLM)

Cynthia Montaño, Gerardo Sierra, Gemma Bel-Enguix
Instituto de Ingeniería
Universidad Nacional Autónoma de México
{cmontanor,gsierram,gbele}@iingen.unam.mx

Abstract

Mexico has a great language diversity. In addition to Spanish, there are 68 language groups and 364 variants (INALI, 2008), divided into 11 families. However, this wealth has been threatened due to discrimination against speakers. Indeed, Spanish has been imposed from the legislative, political and economic point of view, which has interrupted the intergenerational transmission of originary languages and, with it, caused the gradual loss of use spaces and communicative functions. Likewise, few technologies have been developed for these languages, because there are few texts written on the internet. The CPLM is a collaborative parallel corpus that contains texts aligned in Spanish and in six indigenous languages: Mayan, Ch'ol, Mazatec, Mixtec, Otomi and Nahuatl. This article describes the development of the CPLM, as well as the difficulties presented throughout the process.

Keywords: Low-Resources Languages, Parallel Corpus, Indigenous Languages of Mexico

Resumen

México cuenta con una gran diversidad de lenguas, ya que, aparte del español, existen 68 agrupaciones lingüísticas y 364 variantes (INALI, 2008), repartidas en 11 familias. Sin embargo, esta riqueza se ha visto amenazada debido a la discriminación hacia los hablantes. En efecto español se ha impuesto desde el punto de vista legislativo, político y económico, lo que ha interrumpido la transmisión intergeneracional de las lenguas originarias y, con ello, originado la pérdida paulatina de espacios de uso y funciones comunicativas. Así mismo, pocas tecnologías se han desarrollado para estas lenguas, debido a que existen pocos textos escritos en internet. El CPLM es un corpus paralelo colaborativo que presenta textos alineados en español y en seis lenguas indígenas: maya, ch'ol, mazateco, mixteco, otomí y náhuatl. Este artículo describe el desarrollo del CPLM, así como las dificultades presentadas a lo largo del proceso.

Palabras clave: Lenguas de Bajos Recursos, Corpus Paralelo, Lenguas Indígenas de México

1. Introduction

Mexico is one of the most diverse countries linguistically, since it occupies the eighth place worldwide and first in Latin America, followed by Brazil. Despite this, few technological tools have been developed for Mexican languages, which are in danger of extinction, since they have not received the same attention as Spanish, because they have historically been discriminated against. In addition, primary areas for the social welfare of their communities of speakers, such as education and health, have been neglected.

English, French and Spanish, among others, are languages with a large number of speakers, for which numerous linguistic corpus have been built. In contrast, the indigenous languages of Mexico are among the languages of few resources, due to the shortage of written sources to form corpus. To compensate for this, parallel corpus have been constructed in Spanish and in the minority languages of Mexico, since these offer various possibilities that can increase our knowledge about their typological, grammatical and cultural characteristics. In addition, corpora show the differences between genres and their translations.

There are various Natural Language Processing (NLP) tasks that are based on the use of parallel corpora. Some examples are automatic translation, natural language generation, lexical and terminological extraction, morphological segmentation and analysis, part of speech tagging, spelling correction, optical character recognition (OCR), and language identification.

The original languages of Mexico belong to 11 typologically diverse families, each with characteristic features that present particular challenges. Some of the most significant aspects for the treatment of these languages in NLP are the agglutination of morphemes in the Yuto-Nahua family, where Nahuatl is found; the tone in Oto-Mangue languages, which can express both lexical meaning and grammatical function (Suárez, 1973); as well as the ergativity in the Mayan family (Sánchez, 2008). As can be seen, from the perspective of computational linguistics, Mexican languages present a number difficulties.

In general, there is limited production in both digital and printed texts, since in most communities a strong oral tradition is observed, while the written form has not been much encouraged, due to political and social factors that have affected the literacy processes. On the other hand, Mexican languages face a lack of spelling normalization, coupled with great dialect variation, as well as diachronic variation of writing, which represents a challenge in the processing of these texts when you want to work with NLP. According to Mager et al (2018), it is important to point out the challenges of working on the development of linguistic resources and tools for the NLP for the languages of Mexico. Addressing these challenges contributes to creating more computational linguistic models, as well as developing a deeper look at the understanding of human language. Additionally, the creation of language technologies in Mexican languages can have a positive

social impact on language communities, given the scarcity of digital resources in these languages.
The parallel corpus in Mexican languages that we can find online are Axolotl, a parallel Nahuatl-Spanish corpus, which contains documents of classical and modern Nahuatl (Gutiérrez-Vasques, Sierra and Pompa, 2015) and the Tsunkua project, otomí parallel corpus -español, which contains variants from Mezquital and the State of Mexico.
Since these efforts are concentrated in two languages, the UNAM Language Engineering Group proposed to create a parallel corpus that would house several Mexican languages. Thus was born the CPLM.

2. The Parallel Corpus of Mexican Languages (CPLM)

The CPLM is part of an interdisciplinary project whose main objective is to contribute to the development of natural language processing, focused on Mexican languages with limited digital resources -particularly in the task of multilingual lexical extraction- deepening the study of these in terms of models of statistical representation
Among the specific objectives of the project, a methodology for bilingual lexical extraction from parallel corpus of Mexican low-resourced languages is considered. This will allow, for example, to automatically extract bilingual dictionaries and build databases for applications such as machine translation.
Likewise, the project aims to propose one or more types of evaluations that are useful to analyze the effectiveness of the representations and proposed methodology. In addition, we want to explore the development of computational models of various linguistic levels of the treated languages, so that they help in the task of bilingual lexical extraction, mainly morphological segmentation models and syntactic analysis. Finally, it is intended to measure, in quantitative terms, various linguistic phenomena, such as complexity, in order to develop better computational models and contribute from this area to the knowledge and analysis of Mexican languages.

2.1 CPLM Data

The CPLM contains texts in 6 languages belonging to three families: Oto-Manguean, Mayan and Uto-Aztecan. The Oto-Manguean family includes Mixtec, Otomi and Mazatec. Mixtec is spoken in the states of Oaxaca, Guerrero and Puebla and, according to the INALI catalog (2008), presents a total of 81 variants. The Otomi is spoken in the State of Mexico, Hidalgo, Querétaro, Guanajuato, Puebla, Mexico City, Tlaxcala, Veracruz, Michoacán and San Luis Potosí. According to INALI it has 9 variants. Mazatec, spoken in the north of Oaxaca, Puebla and Veracruz, has 16 variants.
Within the Mayan family there are two languages: on the one hand, Ch'ol, which is spoken in the states of Chiapas, Campeche and Tabasco and has two variants: northwest and southeast. On the other hand, the Maya, in the states of Yucatan, Quintana Roo and Campeche. There are some discrepancies regarding the number of Maya variants.
Finally, Nahuatl is the only language of the Yuto-Nahua family present in the corpus. This has 30 variants (INALI, 2008). It spreads through the states of Puebla, Veracruz, San Luis Potosí, Oaxaca, Guerrero, Hidalgo, Colima, Durango, Jalisco, Michoacán, Morelos, Nayarit, Tabasco, Tlaxcala, State of Mexico.
Table 1, shows the languages of the CPLM and the number of variants reported.

Maya	Otomangue	Yuto-nahua
Yucatec Maya (3 variants) Ch'ol (2 variants)	Mazateco (6 variants) Mixteco (30 Variants) Otomí (5 variants)	Nahuatl (5 variants)

Tabla 1: Families, languages and variants

The textual genres that make up the CPLM are: didactic, expository, narrative, poetic, religious, historical and political.
Teaching texts include writing and reading manuals and topics related to language systems. The expository texts include writings of scientific dissemination, for example those dealing with diseases and crops. The stories, traditional fables and of everyday life tales come together in the narrative category. We consider as poetic those texts written in verse. As regards the religious genre, only the Bible is currently available. Historical writings expose the popular history of communities. Finally, the political genre contains articles of the Constitution, as well as explanatory texts on the political-legal field.
Table 2 shows the number of texts for each genre, according to the language.

	Ch'ol	Maya	Mazatec	Mixtec	Nahuatl	Otomí
Didactic	5	5	15	6	5	20
Expositive	7	0	9	12	4	12
Narrative	11	26	28	39	10	66
Poetic	1	5	3	3	11	2
Historic	2	1	1	1	0	1
Politic	2	6	1	5	5	2
Religious	1	1	4	12	10	1

Tabla 2: Genre of the texts

The best represented genre is narrative, since the oral tradition tales are the ones that have been most recorded in the publications of the Summer Linguistic Institute and INALI, the main sources of consultation of the CPLM.
There are three main steps in elaborating this corpus: a) search and compilation of texts, b) digitization and, finally, c) alignment. These steps will be briefly explained in the next section.

3. Elaboration of the corpus

The first step to create the CPLM, was a search of texts published in each of the six languages mentioned above, with their Spanish parallel. Second, the texts were digitized using ABBYY FineReader software, with an OCR that helped prepare the texts for the next stage. Thirdly, the texts

were aligned with their corresponding translation in Spanish.

3.1 Text search and compilation

In this step, texts on the internet and libraries were searched. Although the CPLM intends to be a multilingual parallel corpus, in this first stage only texts in the indigenous languages indicated with their respected translation in Spanish were searched. Most of the texts that make up the CPLM were found in PDF format, however, in some cases, bilingual books were found that contained a significant amount of images, for these they were scanned in high definition to facilitate the use of the OCR. In this first phase a database was also created where information on the textual genre, language, variant, ISO code and community was recorded.

3.2 Digitalization

Once all the texts were in PDF, the files were treated with ABBYY FineReader software. This program is used to more easily recognize the common spellings of indigenous languages, such as superscripts, subscripts and diacritics, thanks to OCR character modeling. This saved a significant amount of time during the review of each text and its correct digital transcription.

3.3 Alignment

An important aspect that allows to exploit the bilingual lexical information contained in a parallel corpus is alignment. Alignment is the process of matching bilingual correspondences at a specific level, for example, at the document level, at the paragraph level or at the sentence level and, finally, at the most granular and difficult level to perform, word level alignment.
In general, in CPLM, the texts are aligned at the sentence level. However, in political and religious texts, the alignment is found, either at the constitutional article level or at the verse level.
Initially, the alignment was attempted automatically with the Gale & Church algorithm (1993), used for other parallel corpus. However, since we worked with languages typologically different from those used by Gale & Church, the algorithm was not totally efficient. For example, paragraphs were deleted in Otomi. For this reason, we decided to manually review each of the alignments.

4. The spellings and their difficulties

As already mentioned, there is no general agreement regarding the orthographic norms of the indigenous languages, since there is still a lot of research on the variants that make up the linguistic groups. On the other hand, the written production collected belongs to different years and authors, so generally, the texts are not orthographically homogeneous and present a large number of spellings. The language that shows more orthographic variants is Otomi, since, apart from being a tonal language, it has a large vocal inventory with 9 oral and 5 nasal vowels. Another example in the spelling change is the use of 'h' for the glottal consonant, but its use has been replaced by the apostrophe (').

In Table 3, we present a compilation of the peculiar spellings found in all the CPLM texts.

Spellings
á, à, ǎ, â, ā, a̱, ą, a̡, a̢, á̱, ą́, ą̀, ą̌, ą̂, ą̄, ä, ä́, æ, ǽ, æ̀, æ̌, æ̂, ǣ, æ̱, æ̨, æ̢, ǽ̱, ǽ̨, æ̨̌, æ̨̂, ǣ̨
é, è, ě, ê, ē, e̱, ę, e̡, e̢, é̱, ę́, ę̀, ę̌, ę̂, ę̄, ë, ë́, ë̀, ë̌, ë̂, ë̱, ɛ, ɛ́, ɛ̀, ɛ̌, ɛ̂, ɛ̄, ɛ̱, ɛ̨, ξ, ɛ̱́, ɛ̨́, ɛ̨̌, ɛ̨̂,
í, ì, ǐ, î, ī, i̱, į, i̡, i̢, í̱, į́, į̀, į̌, į̂, į̄, ï, ḯ, ï̀, ï̱, ɨ, ɨ́, ɨ̀, ɨ̌, ɨ̂, ɨ̄, ɨ̱, ɨ̨, ɨ̢, ɨ̱́
ó, ò, ǒ, ô, ō, o̱, ǫ, o̡, o̢, ó̱, ǫ́, ǫ̀, ǫ̌, ǫ̂, ǭ, ö, ö́, ö̀, ö̱
ú, ù, ǔ, û, ū, u̱, ų, u̡, u̢, ú̱, ų́, ų̀, ų̌, ų̂, ų̄, ü, ǘ, ǜ, ü̱, ʉ, ʉ́, ʉ̀, ʉ̌, ʉ̂, ʉ̄, ʉ̱, ʉ̨, ʉ̢, ʉ̱́
¢

Tabla 3: Indigenous languages spellings

The spellings in Table 3 represent various linguistic characteristics, whether tone, type of oral vowels different from those of Spanish or nasal vowels.

Each of these involved different challenges. First, during the scanning process, the spellings were not recognized with the OCR, so character molding was used, a special tool of the ABBYY FineReader software. With this tool, each of the unconventional letters were recognized and stored in the software for later use. So, every time a new text was digitized, it recognized well-written spellings and accumulated new ones.

For the alignment, we had to be very careful that the spellings were preserved when passing the texts to the .txt files and especially with the UTF-8 encoding. For that we make sure to find the equivalence of the spellings in ASCII code.

The graphical interface is freely available on the website: http://www.corpus.unam.mx/cplm

5. Conclusion

This article describes the creation of the Parallel Corpus of Mexican Languages (CPLM). The different stages of elaboration have been succinctly presented, as well as the most relevant information. The CPLM was created using GECO, a corpus manager that allows the inclusion of several collaborators, since the CPLM intends to invite students or researchers to participate in the feeding of the CPLM with the corpus that belong to them, either with books or elicitations. With the dissemination of this interface in different forums, we intend to give visibility to Mexican languages in the area of the NLP, in addition to promoting the use of the corpus as a tool to create language technologies.
In future work two lines of work are considered. First, we plan to make improvements to the interface, that is, adaptations will be made so that the CPLM can include recordings and increase the number of texts, as well as add more languages along with their variants. Secondly, it our

second goal is to create dictionaries with the vocabulary that many of the texts included in the CPLM contained. Likewise, we will label the texts in Mexican languages in order to perform the search with POS tags.

Regarding the area of the NLP, it is contemplated to work with the analysis and measurement in quantitative terms of the complexity of various linguistic phenomena for each language. The above, in order to understand how to model different types of bilingual relationships depending on the type of languages. Also, another of the future tasks is the creation of bilingual lexical extraction methods based on the distributional vector representations (word embeddings) of word appearance contexts. These models should be able to find word-level correspondences between a pair of languages, based on different statistical approaches of NLP and machine learning techniques. The investigation of these models will be focused on treating typologically distant languages.

9. Acknowledgements

This work is supported by the Mexican Council of Science and Technology (CONACYT) funds A1-S-27780 and FC-2016-01-2225, and PAPIIT IA401219.

6. Bibliographical References

Gale, W. and Church, K. (1993). A program for aligning sentences in bilingual corpora, *Computational Linguistics*, 19:75–102, 01.

Gutiérrez-Vasques, X., Sierra, G., and Hernández, I. (2016). Axolotl: a web accessible parallel corpus for spanish-nahuatl. 05.

Gutiérrez-Vasques, X. (2015). Bilingual lexicon extraction for a distant language pair using a small parallel corpus. pages 154–160, 01.

INALI. (2008*). Catálogo de las Lenguas Indígenas Nacionales: Variantes Lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*

Kann, K., Mager Hois, J. M., Meza-Ruiz, I. V., and Schütze, H. (2018). Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), pages 47–57, New Orleans, Louisiana, June. Association for Computational Linguistics.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza, 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.

Sierra, G., Solórzano Soto, J., and Curiel Díaz, A. (2017). Geco, un gestor de corpus colaborativo basado en web. *Linguamática*, 9(2):57–72.

Suárez, J. A. (1973). On proto-zapotec phonology. *International Journal of American Linguistics*, 39(4):236 249.

Sánchez, M. E. (2008). Ergatividad en la familia lingüística maya. *Memorias del IV Foro Nacional de Estudios en Lenguas*, 19:541–557, 01.