A Mixtec-Spanish Parallel Corpus

Cynthia Montaño **Gerardo Sierra**
Gemma Bel-Enguix **Helena Gómez-Adorno**
Universidad Nacional Autónoma de México
Grupo de Ingeniera Linguística
tzintia.montano@gmail.com, {GSierraM,GBelE}@iingen.unam.mx,
Helena@iimas.com

1 NLP and low-resourced languages

Computational technologies have a key role in Computational Linguistics. Thanks to the capability of compiling and analyzing large collections of texts with computers many resources and applications have been designed that have caused a fast development in Natural Language Processing and Artificial Intelligence. Corpora and parallel corpora are basic instruments for approaching natural language, making it possible the implementation of models for machine translation, automatic summarization, information extraction and other methods for language understanding and analysis.

All these advances in language technologies need large amounts of data. The most spread and best-represented languages in media and internet generate every day Giga Bytes of information that can easily be processed and studied. However, most of the languages in the world are under-represented in social life, the media and, the internet. These are low-resourced languages. An example of this is indigenous languages in Mexico, with a lack of representation in education, government, services, and media.

Regarding linguistic diversity in the American continent, Mexico is placed in the 2nd position, just after Brazil, with 11 linguistic families, 68 linguistic groups and 364 linguistic variants. Building computational resources for these languages is a hard task, provided the scarcity of data. One of the current approaches to tackle this problem is the use of parallel corpora in two languages, using texts that have been translated from one to the other. We are designing parallel corpora including a language that has a large amount of data, Spanish, and a low-resourced language. In a first step of our research, we aim to build parallel corpora of Spanish with most of the called Mexican languages. Among the texts that we have collected for this purpose are legal and biblical data such as Political Constitution of the Mexican United States and the New Testament. This multilingual parallel corpora will be available online for the interested audience. Later, we have the goal of designing methods of lexical extraction and, as a long-term objective, a machine translation system. Regarding Mexican languages and NLP there are a few papers which have studied different aspects of these languages, e.g. we have a summary of the challenges of language technologies for the indigenous language of the Americas (Mager et al., 2018) and a study of the morphological segmentation for polysynthetic languages (Kann et al., 2018).

This paper presents the parallel corpora Spanish-Mixtec.

2 Mixtec

Mixtec is the indigenous language with the largest number of variants in Mexico: the Institute of Indigenous Languages reports 81 different variants (INALI, 2008), while Summer Institute of Linguistics (SIL, 1999) rates about 52 variants. Some researchers (Josserand, 1983) have classified all the Mixtec language diversity in 5 majors dialectal regions. There are around 480,000 speakers of Mixtec (INEGI, 2015) spread across three Mexican States: Oaxaca (around 265,000), Guerrero (around 140,000) and Puebla (around 75,000). Therefore, more than half of the Mixtec speakers are found in the state of Oaxaca. Mixtec belongs to Otomanguean linguistic family which is the most diverse in the Mesoamerican territory.

The prominent phonetic features of the Mixtec language are a) a strong nasal tendency, b) glottalization of vowels, c) presence of tones, which differs from variant to variant in the number and their possible

A.	Mixtec	Cudîi ini lehe ndáhì-si.
	Spanish	*Al gallo le gusta cantar*
	English	The rooster likes to sing
B.	Mixtec	Sa̲ cáhnu va̲ha rí.
	Spanish	*Estaba bastante grandecito y muy bonito*
	English	It was really big and pretty
C.	Mixtec	Tée-de tutū.
	Spanish	*Él escribe*
	English	He writes
D.	Mixtec	Ta catyi ra
	Spanish	*Después siguió diciendo*
	English	Then, he continued saying

Table 1: Different types of ortography and marking tone in Mixtec texts

combinations. Due to this features Mixtec is classified as a highly analytically language. It is commonly claimed that Mixtec distinguishes three different tones: high, middle and low, which express lexical meaning and grammatical function (Macaulay, 1995).

3 Parallel corpus Spanish-Mixtec

3.1 Compilation process

The compilation process of parallel documents Spanish-Mixtec has been challenging. There are not many Spanish-Mixec parallel texts and most of the sources are non-digital books. Due to this, we need to face the errors when digitizing the sources and difficulties in sentence alignment, as well as the fact that does not exist a standard orthography.

For the digitalization process, we used an Optical Character Recognition (OCR) software but it made some mistakes in automatically recognizing Mixtec text. These were mainly associated with the fact that the OCR could not properly identify the Mixtec language and often made fake corrections because it tried to adapt character patterns corresponding to other languages.

An additional problem is the lack of orthographic normalization in Mixtec, a problem that we have to face in every indigenous language in Mexico. Moreover, the orthography of this language considers three types of tones. The high tone for the letter 'a' is written *á*, the middle tone is the neuter one, represented by *a*. The spelling problem is associated with the low tone, that has three different signs: *à*, *ā* and *a̲*.

Spelling variation is another significant difficulty in processing texts. Even though there is an official orthographic norm, most of the texts are not normalized due to diachronic variations, since they were written before the standard norm. The current characters are d, f, g, j, k, l, m, n, ñ, p, r, s, t, v, w, x, y, z, ty, ', a, e, i, i̱, o, u . The most substantial changes that have been introduced are a) the use of k instead of c, qu, b) the use of ty for ch, and finally, c) the use of an apostrophe (') to mark glottalization instead of h. This spelling variation and the different types of marking tone are shown in Table 1.

There are also some difficulties in the sentences alignment process. Since the corpus has different types of text, we have to deal with different levels of alignment from sentences to larger units like paragraphs. Additionally, sentences are not quite exact translations; there can be parts of the Mixtec text that do not appear in the Spanish version.

3.2 Corpus information

Our parallel corpus consists of sixty texts coming from books and digital repositories. These documents belong to different domains: history, traditional stories, didactic material, recipes, ethnographical descriptions of each town and instruction manuals for disease prevention. We have classified this material in five major categories: didactic (6 texts), educative (6 texts), interpretative (7 texts), narrative (39 texts), and poetic (2 texts). The final total of tokens is 49,814 Spanish words and 47,774 Mixtec words.

As we have mentioned before, there is dialectal, diachronic and orthographical variation in the Mixtecan texts. The corpus contains 22 variants of Mixtec: Central Mixtec (9), Mixtec from the Lower Central Coast of Oaxaca (8), Northwest Coast Mixtec (4), Northeast Sierra Sur Mixtec (4), Mixtec of Yosondua (4), Mixtec from the Middle East Part of Guerrero (3), Mixtec of The Frontier Puebla-Oaxaca (3), Mixtec of Xochapa (3), Mixtec of Santa Luca Monteverde (2), Central North High Mixtec (2), Western Mixtec (2), High Western Mixtec (2), Southern Lower Mixtec (2), Mixtec from the High Central Part of Guerrero (1), Mixtec from Northern Part of Guerrero (1), High West Mixtec (1), Central West Coast Mixtec (1), Mixtec of Ñumi (1), Mixtec from the Central Coast of Oaxaca (1), Northwest Mixtec (1), Middle South Mixtec (1), Central Southwest Mixtec (1). Moreover, three of the texts showed features from more than one variant.

The texts belong to the states of Oaxaca (48 texts), Guerrero (9 texts) and Puebla (3 texts).

According to this data, we see that the corpus is unbalanced in what refers to the representation of the different territories. While 55% of speakers are in Oaxaca, 80% of texts come from this region. Guerrero has the 30% of speakers and the 15% of texts and Puebla, with the 15% of the speakers has a representation of the 5% in the corpus.

References

Instituto Nacional de Lenguas Indígenas INALI. 2008. *Catálogo de las Lenguas Indígenas Nacionales: Variantes Lingísticas de México con sus autodenominaciones y referencias geoestadísticas*.

Instituto Nacional de Estadística y Geografía INEGI. 2015. *Encuesta intercensal de población y vivienda. Población de tres aos y ms que habla lengua indígena por sexo y lengua según grupos quinquenales*.

K.J. Josserand. 1983. *Mixtec dialect history*. Ph.D. thesis, Tulane University.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 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.

M. Macaulay. 1995. The phonology of glottalization in mixtec. *International Journal of American Linguistics*, 61:38–61, 01.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Iván Meza. 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.

Summer Institute of Linguistics SIL. 1999. *Elaboración de gramáticas populares de lenguas indígenas: Una breve guía (con referencia especial a las lenguas otomangues)*.