On the development of the Mexican Languages Parallel Corpus

Cynthia Montaño, Gerardo Sierra,
Gemma Bel-Enguix
Grupo de Ingeniería Lingüística
Instituto de Ingeniería
Universidad Nacional Autónoma de México

1. NLP and low-resourced languages

Building computational resources for low-resources languages is a hard task, due to the scarcity of data. One of the current approaches to tackle this problem is the use of parallel corpora in two languages. We also face the lack of orthographic normalization in Mixtec, a problem that we have to face in every indigenous language in Mexico.

Regarding already existing online parallel corpora for Mexican language, it can be mentioned Axolotl with parallel texts in Spanish and Nahuatl, and Tsu̱nkua with parallel texts in Spanish and Otomí.

2. The CPLM

Given the situation of scarcity of resources for Mexican languages, a project for creating different resources for language technologies, e.g. parallel corpora, was carried out by the Linguistic Engineering Group and with the support of the Mexican Council of Science and Technology (CONACYT). The project was called Mexican Languages Parallel Corpus (CPLM for Spanish acronym) and its main goal is to contribute to development of NLP for low-resources Mexican languages.
The CPLM is compound by two modules: the core module and the module conform by a subcorpus of religious and political texts.

.

3. Corpus information

3.1 Core module
The core module of CPLM currently comprises 6 linguistics groups from 3 linguistics families; Mayan: Yucatec Maya and Ch'ol; Otomanguean: Mazatec, Mixtec and Otomí; Uto-Aztec: Nahuatl. Different varieties were considered for each one of these linguistics groups as can be seen in Table 1.

Mayan	Otomanguean	Uto-Aztec
-Yucatec Maya (3 variaties) -Ch'ol (2 variaties)	-Mazatec (6 varieties) -Mixtec (30 varieties) -Otomí (5 varieties)	-Nahuatl (5 varieties)

Table 1. Linguistics families and languages varieties

The core module comprises different kind of texts that were divided in six genres: didactical, expositive, narrative, poetic, historical and dramatic.

For each language, we contabilized different genres of the text and we present the numbers in the Table 2.

	Ch'ol	Maya	Mazateco	Mixteco	Náhuatl	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
Historical	2	1	1	0	0	1
Dramatic	0	0	0	0	1	0

Table 2. Number of text of each genre

As it can been seen in Table 2, there are some differences between the number of texts of each corpus due the size of texts, since some of them had few words. For this reason, we decided to set an average number of words in each corpus. The quantities can been seen in Table 3.

CPLM					
Ch'ol	Maya	Mazateco	Mixteco	Náhuatl	Otomí
Spanish words: 56,722 Cho'l: 67,876	Spanish words: 43,700 Maya: 42,500	Spanish words: 49,700 Mazatec: 48,500	Spanish words: 49,814 Mixtec: 48,375	Spanish words: 213, 133 Nahuatl: 148, 754	Spanish words: 53, 478 Otomi: 56,199

Table 3. Number of words in each corpus

3.2. The Religious and Political Text Subcorpus (STRyP)

The STRyP comprises 34 languages with at least one of is linguistic variants in the religious texts and 62 languages in the political texts.
The STRyP based on 83 translations of the New Testament and 11 translations of three types of text: (1) texts that relate laws and rights, (2) texts explaining the laws and rights and finally, (3) mixed texts, where there are laws and rights with their explanations.

Type of text	Texts	Language	Total of words in Spanish
Religious Texts	83	34 languages	175, 883
Political Texts	11	62 languages	115, 006

Table 4. Information of STRyP

5. Perspectives

- ❖ We aim to build parallel corpora of Spanish with most of the called Mexican languages, thus we expect other people to collaborate.
- ❖ This multilingual parallel corpora will be available online for the interested audience.

References:

-Instituto Nacional de Lenguas Indígenas INALI. 2008. *Catálogo de las Lenguas Indígenas Nacionales: Variantes Lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*.
-Katharina Kann, Jesus Manuel Mager Hois, Iván 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.
-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.

Acknowledgments

This work is supported by the Mexican Council of Science and Technology (CONACYT), funds 370713 and CB-2012/178248 and PAPIIT 401219.