Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2947–2952
Marseille, 11–16 May 2020

CPLM, a Parallel Corpus for Mexican Languages: Development and Interface

Gerardo Sierra, Cynthia Montaño, Gemma Bel-Enguix, Diego Córdova, Margarita Mota
Instituto de Ingenieria
Universidad Nacional Autónoma de México
Ciudad de México
{gsierram,cmontanor,gbele}@iingen.unam.mx
diego.cordova.nieto@gmail.com, mmotam@iingen.unam.mx

Abstract

Mexico is a Spanish speaking country that has a great language diversity, with 68 linguistic groups and 364 varieties. As they face a lack of representation in education, government, public services and media, they present high levels of endangerment. Due to the lack of data available on social media and the internet, few technologies have been developed for these languages. To analyze different linguistic phenomena in the country, the Language Engineering Group developed the Corpus Paralelo de Lenguas Mexicanas (CPLM) [The Mexican Languages Parallel Corpus], a collaborative parallel corpus for the low-resourced languages of Mexico. The CPLM aligns Spanish with six indigenous languages: Maya, Ch'ol, Mazatec, Mixtec, Otomi, and Nahuatl. First, this paper describes the process of building the CPLM: text searching, digitalization and alignment process. Furthermore, we present some difficulties regarding dialectal and orthographic variations. Second, we present the interface and types of searching as well as the use of filters.

Keywords: Parallel corpus, multilingual corpus, low-resourced languages of Mexico

1. Introduction

Collecting corpora for endangered languages is crucial to preserve and expand language diversity, besides the applications for NLP (Amel Fraisse and Fishkin, 2018). Mexico is a linguistically diverse country. There are 11 phyla (e.g. Uto-Aztec) that comprise 68 linguistic groups (e.g. Nahuatl, belonging to the Uto-Aztec) and 364 varieties (Nahuatl from the Huasteca) present all around the country (INALI, 2008). About 6.5% of the population speaks at least one of these varieties, but according to the intercensal survey (INEGI, 2015), about 21% of the total population of Mexico identifies themselves as members of an indigenous group. However, these languages are not represented in education, government, public services and media and, therefore, they show high levels of endangerment. Languages facing this lack of large amount of data are called low-resourced and all linguistic varieties in Mexico are struggling with this situation. One of the current approaches for tackling the scarcity of data is the building of parallel corpora of Spanish, that has a large amount of data, and the low-resourced languages of Mexico.

There are three examples of online parallel corpora regarding Mexican languages. The first one is Axolotl[1] (Gutiérrez-Vasques et al., 2016), a Nahuatl-Spanish parallel corpus also developed by the Language Engineering Group. There is the parallel corpus Wixarika-Spanish[2], that gathers translations of classic Hans Christian Andersen's literature (Mager et al., 2018a). And Tsunkua[3], an Otomi-Spanish parallel corpus developed by Comunidad Elotl.

The importance of the creation of parallel corpora is linked to the analysis of different phenomena that can be found in indigenous languages of Mexico that are not usually common among the traditional ones used in NLP. For example, the Mayan phylum presents ergativity, which is one of the structural distinctive features between them and Spanish (Sánchez, 2008). The contrastive use of the tone in the Otomanguean phylum is one of the most prominent traits and some authors consider it a genetic feature of this macrolinguistic family (Rensch, 1976; Suárez, 1973). There are specific degrees of agglutination regarding the Uto-Aztecan family, e.g. polysynthesis and morphemes that express different functions (Mithun, 2001). All these different characteristics as well as other issues regarding written material constitute a challenge for NLP but also can contribute to the understanding of human language by studying different phonological, morphological and syntax features.

Regarding to Mexican languages and NLP there is some research that has studied different aspects of this subject, e.g. a bilingual lexicon extraction for Nahuatl (Gutiérrez-Vasques, 2015), a development of a web-accessible parallel corpus for Spanish-Nahuatl (Gutiérrez-Vasques et al., 2016), a summary of the challenges of language technologies for the indigenous languages of the Americas (Mager et al., 2018b) and a study of the morphological segmentation for polysynthetic languages (Kann et al., 2018).

The structure of this paper is as follows: Section 2 presents general information about the CPLM and the steps for its creation: compilation of the documents, the digitalization and alignment process, and finally the challenges we found regarding the texts. In Section 3, we describe the interface and the types of searching. Finally, in Section 4, we present the conclusion and future work.

2. The Mexican Languages Parallel Corpus (CPLM)

Given the scarcity of resources for Mexican languages, a project for creating different resources for NLP, e.g. parallel corpora, was carried out by the Language Engineering Group, with the support of the Mexican Council of Science and Technology (CONACYT). The main goal of the

[1] http://www.corpus.unam.mx/axolotl
[2] https://github.com/pywirrarika/wixarikacorpora
[3] https://tsunkua.elotl.mx/

CPLM is to contribute to the development of NLP for low-resourced Mexican languages.

2.1. Corpus information

Nowadays, the CPLM comprises 6 linguistic groups from 3 phylum; Mayan: Yucatec Maya and Ch'ol; Otomanguean: Mazatec, Mixtec and Otomi; Uto-Aztec: Nahuatl. Different varieties were considered for each one of these linguistic groups as can be seen in Table 1.

The relevance in selecting these languages is related to typological and quantitative factors. Nahuatl is the most spoken indigenous language in Mexico with almost 2 million of speakers. The second one is Maya with almost 1 million speakers. The third one is Mixteco with half million speakers (INEGI, 2015). First, we selected the most spoken three languages which belong to different linguistic families, so there are typological differences between them, for example, the presence of tone or a high level of agglutination. We have a wide range of linguistic phenomena to take into account developing NLP. Second, we looked for languages with larger quantities of available written materials like Otomi. Finally, we counted on the help of scholars specialized in Ch'ol and Mazatec.

Seven genres were identified during data gathering: didactic, expositive, narrative, poetic, historical, political and religious. Didactic texts are those regarding handbooks for writing and learning. Expositive ones are those that explain different illnesses or the ways of cultivating some kind of crop. Narrative ones are those that tell a traditional story and present tales of daily life. Poetic texts are the ones that are written in verse. Historical ones present a popular history of the communities and foundations. Political ones are those related to laws and rights, especially the Political Constitution of the United Mexican States. It is important to mention that the translation of political texts is a significant effort done by the federal, regional and local authorities to make it available for indigenous communities. Finally, religious texts are mainly translations of the Bible. We listed the number of text genres found in each language. The totals are presented in Table 2. The genre with the largest amount of texts is the narrative one because most of the texts are tales and stories of an oral tradition of each community.

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

There are three main steps for creating the CPLM: compilation of the texts, digitalization and alignment process. These steps are described in the next section.

2.2. Compilation of the documents

First, we identified the text sources to gather all the available documents in digital and non-digital formats by looking for the material in library catalogs and databases of different institutions. Since Mexican languages face a lack of documentation, it was not easy to obtain the parallel text of them. The main sources for data gathering are the Summer Institute of Linguistics (SIL) and the National Institute of Indigenous Languages (INALI). The SIL is an organization that has worked with the indigenous language of Mexico for a very long time. The materials of SIL are from different dates and comprise different types of texts such as traditional tales, grammar descriptions, teaching materials, etc. The INALI mainly offers texts of oral traditions as riddles, sayings and tales and it also provides translations for the Political Constitution of the United Mexican States, that are available online. Another important source is specialized journals such as *Tlalocan*, a journal for presenting the oral tradition of languages of Mexico. Regarding religious texts, there is a especial website[4] that contains translations of the Bible in languages from all around the world and there is a section especially for Mexico that contains one or more linguistic varieties.

The second step of this phase was the creation of a database for the CPLM. This phase was certainly the most time-consuming due to text selection and metadata registration. During this phase, we registered the usual bibliographic information about texts such as name, author, language and year of publication. Later, the experts in linguistics decided to add more relevant metadata such as the variety of the language, ISO code, country, state, and compiler.

2.3. Digitalization process

Once the text selection is done, the next step is digitalization process from the non-digital sources. For this task, we used ABBYY FineReader, an Optical Character Recognition (OCR) software, however the results were not totally successful, since it made some mistakes in automatically recognizing texts with different characters. These failed results were mainly associated with the fact that the OCR could not properly identify the special written characters of indigenous languages, since the majority of them uses some characters different from those used by dominant languages. Due to this, the OCR often made fake corrections because it tried to adapt character patterns corresponding to other languages.

Once the editable version of the texts was ready, the team verified manually that each character corresponds to the one on the original texts. This phase took a considerable amount of time and we decided to use the character molded available on the ABBYY FineReader software that improved the OCR for the six languages of CPLM, since it recognized automatically the unusual characters. We used Unicode for the diverse characters presented in each language. We double-checked each character in the original text and its Unicode correspondent. This exhaustive task was performed by different people who collaborated in the creation of the corpus.

2.4. Alignment process

As has been mentioned before, the corpus has different types of texts and we had to deal with different levels of alignment: phrases, sentences, paragraphs. The didactic texts offered the smallest level of alignment, since they present short phrases for daily life. Expositive and narrative texts showed length variation: medium, from 2 to 4

[4] https://www.scriptureearth.org

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

Table 1: Linguistics families and language varieties

	Ch'ol	**Maya**	**Mazatec**	**Mixtec**	**Nahuatl**	**Otomi**
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	1	0	1
Political	2	6	1	5	5	2
Religious	1	1	4	12	10	1

Table 2: Number of the texts of each genre

Ch'ol	Maya	Mazatec	Mixtec	Nahuatl	Otomi
Spanish: 56, 722	Spanish: 33, 431	Spanish: 49, 700	Spanish: 49, 814	Spanish : 213, 133	Spanish: 53, 478
Ch'ol: 67, 876	Maya: 36, 495	Mazatec: 48, 500	Mixtec: 48, 375	Nahuatl: 148, 754	Otomi: 56, 199

Table 3: Number of words in each corpus

A.	Mixtec	Cudîi ini lehe ndáhì-si.
	Spanish	*Al gallo le gusta cantar*
	English	The rooster likes to sing
B.	Mixtec	¿Ncha ta-cuu-ni, Pedro?
	Spanish	*¿Cómo está usted, Pedro?*
	English	How are you, Pedro?
C.	Mixtec	Sa̲ cá'nu va̲'a rí.
	Spanish	*Estaba bastante grandecito y muy bonito*
	English	It was really big and pretty
D.	Mixtec	Yee ti ndika yee ti taka ja vixi.
	Spanish	*Come plátano y otras frutas*
	English	He ate banana and other fruits

Table 4: Different types of orthography in Mixtec texts

sentences, and large paragraphs, over 5 sentences. Oral tradition texts are presented in these genres. Political and religious texts showed a special level alignment, since those texts are divided into articles and verses.

Additionally, the sentences are not quite exact translations; since there can be parts of the text that do not appear in the Spanish version. Regarding this, we decided to keep the version as it appears in the original document.

For the alignment process, we explored the use of several algorithms. Finally, we decided to combine the Gale and Church algorithm (Gale and Church, 1993) and a manual check of each line in every text by the coordinators of each corpus. During this process, we found different problems that were solved by consensus. For example, the deletion of orphan sentences.

Once the texts are correctly aligned, we saved the texts in .txt files with UTF-8 codification. For each document, there are two .txt files, one with the Mexican language texts and another one with its respective Spanish translation.

2.5. Challenges of creating the CPLM

We have faced different problems during the process alongside the ones exposed in each one of the steps described before. As it has been mentioned before, digital resources are scarce in indigenous languages of Mexico due to the under-representation of these languages in public life. Furthermore, there is no general agreement regarding a orthographic norms, since there is a lack of research in the language descriptions of different varieties and some speakers do not use the orthographic norm proposed by INALI. Due to this, different kinds of orthography can be found among the communities. Besides, the writing systems can have different orthography depending on the year the texts were written. An example can be seen in Table 4, a sample of parallel text in Mixtec and Spanish. These 4 sentences belong to different years. The A sentence is from a

High tone	Middle tone	Low tone
á	a	à
		ā
		a̲

Table 5: Notations for marking tone in Mixtec

text from 1975. We can see in this sentence the 'h' character, which was used for marking glottal consonants. As well as in example A, the B sentence uses the 'c' character for representing the voiceless velar plosive consonant. This sentence comes from 1968. The sentence presented in C is from 1989 and the character for glottal consonant changed, thus instead of using 'h' they started using an apostrophe (') for that consonant. Nowadays the use of an apostrophe is the most common way of representing glottal consonant in Mixtec language. Finally in sentence C, we can observe the change of 'c' character for 'k' in representing the voiceless velar plosive consonant. This last example is from 1990.
One of the most significant issues about Mixtec orthography is the written representations of tones. As can be seen in examples of Table 4, there is not a homogeneous way of marking tone. The notations for making the three tones found in the texts are presented in Table 5. The most changing feature for marking tone is presented in low tone notation. Table 5 shows examples of these different ways of marking a low tone. Some of them are difficult to recognize by the OCR software.
The orthographic facts presented before are the reason we decided to keep the texts as they were originally written, however, once we decide to conduct NLP experiments the texts need to have a special orthographic treatment.

3. Interface

An information retrieval system is devised to efficiently search for information throughout the corpus. It allows searching words and phrases in one language and retrieves both those parallel sentences that contain the searched word and the translation associated with that sentence in another language in the parallel corpus.
For the text fragments that match the search, additional information of the source will be displayed as well as its dialect variant. This type of web search systems in a parallel corpus are popular tools to assist humans in translation tasks and perform linguistic studies as well as to promote the creation of language technologies.
The interface consists of two parts. The first one is GECO[5], the corpora manager, that provides several user functionalities like registration, projects creation and text uploading (Sierra et al., 2017). Through each project creation, users can select between 18 metadata or choose if the project will be collaborative; that means the user can add more users to the project as long as they're already registered. Otherwise, users are enabled to invite other users via mail on behalf of Language Engineering Group. Once created, projects can be modified and can be public either.
The graphical interface is freely available on the website: http://www.corpus.unam.mx/cplm

[5] http://www.geco.unam.mx/

3.1. Search types

Due to the characters diversity and the type of bilingual or multilingual text (in the case of the new testament and some political texts), it was necessary to create a search engine which could identify these varieties. Therefore, we implemented the search engine in Python with regular expressions.

There are a total of 8 different search types. The default one is a exact form, because the word typed by the user is exactly the word the search engine will look for. Figure 1 shows the exact form example search with the word *mundo* (world). As you can notice, on that example we find the option **Meta** where we can watch all the metadata of the text that contains the lines presented in table. The next column shows Spanish lines that contain the noun *mundo*, on the right column it shows the correspondent lines, on this case from Nahuatl.

The second type of search is the dependence of the character combination language where you can find a word regardless of its spelling correction. It means, for example, if we search the word *avión* (airplane) but we omit accent mark, the search engine would find it anyway.

There are 4 ways of wildcard searches. One is when you want to search for a particular end in a word, for example: ***ito**. The results of this kind of search would show lines of texts where the words could appear:

*perr**ito*** (little dog),
*chiqu**ito*** (very little),
*qu**ito*** (I take away),
*m**ito*** (myth).

In case of the wildcard **ca***, the results could be words like:

***ca**sa* (house),
***ca**rnaval* (carnival),
***ca**sino* (casino),
***ca**cahuate* (peanut).

We can also search complete unknown words with this kind of wild card search: **un * de ***. This type of request could gives us results as:

***un** montón **de** trabajo* (a lot of work),
***un** ave **de** rapiña* (a bird of prey),
***un** hombre **de** negocios* (a business man).

Another type of wildcard request is with a question mark. **P?lo** could bring us results like:

*p**a**lo* (stick),
*p**e**lo* (hair),
*p**o**lo* (pole).

Finally, we can search words in between other words. This can be done using brackets and numbers. **Por {3} de** could

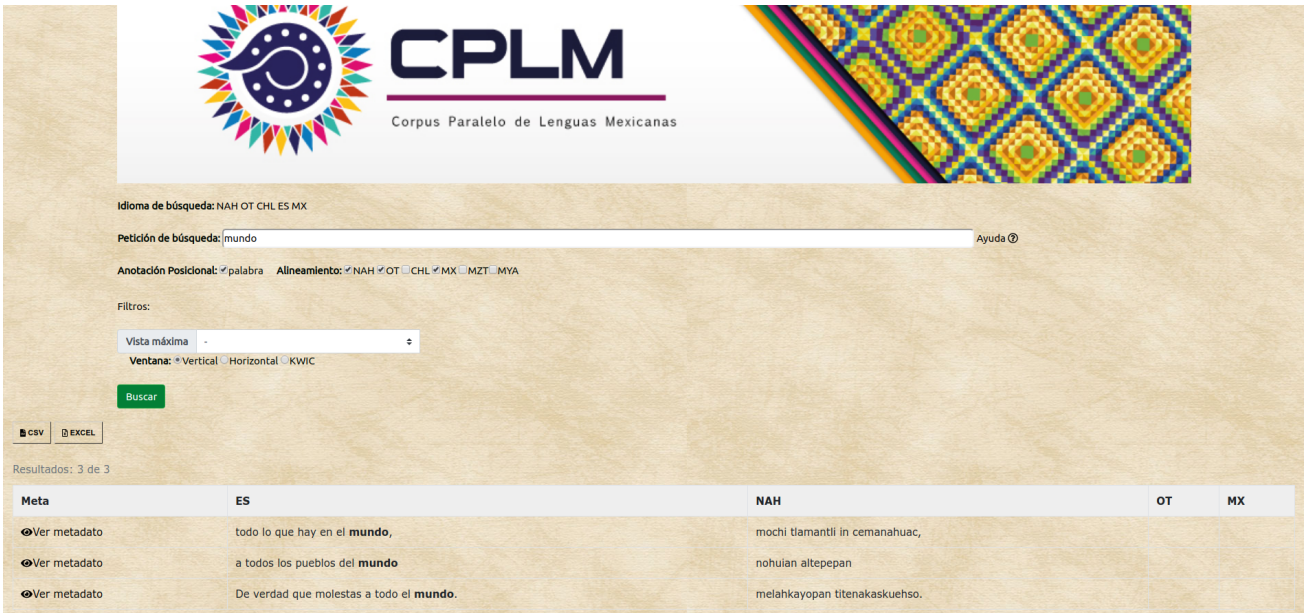

Figure 1: CPLM Interface

show:

por *si nosotros vamos* ***de*** *paseo* (in case we go for a walk),
por *donde la casa* ***de*** *Juan* (where Juan's house),
*¿****por*** *cuánto tiempo has* ***de*** *soportar?* (how long do you have to endure?).

In this example we can notice *por* (for) and de (of) have exactly 3 words of separation. Meanwhile if we write two numbers separated by a middle-dash, this will mean we can expect variation of words.
Example: **un {1-4} en** could bring us:

un *perro* ***en*** *la casa* (a dog in the house),
un *gato gris* ***en*** *la calle* (a gray cat in the street),
un *puente de hierro* ***en*** *la ciudad* (an iron bridge in the city),
un *triste y cansino viejo* ***en*** *agonía* (a sad and tired old man in agony).

Where the separation could be from only one word to four words.

3.2. Filters

In addition to the search types, we can also filter data from the metadata in our project. The captured metadata are recorded as attributes of a label that surrounds the entire document, so it is possible to filter documents based on their values. This effectively allows to create subcorpus at the moment. For this kind of filtering, the interface presents field-value pair selectors to restrict the search domain.
That means we are able to delimiter if we only want results from certain texts with a specific author, main title, secondary title, compiler, translator, editorial, variant in the text, language variants in ISO, year, number, country, state, regional division (municipality), community to which it belongs, classification of textual genres, URLs and notes. Figure 2 shows an example of filters on the CPLM.

There are three types of result views: vertical, horizontal and key-word-in-context (KWIC). These views represent different ways in which we can see results. For example, in a vertical view it is easier to observe results when texts are composed of no more than a few parallel texts. Otherwise, when results are very large, horizontal view can help us not to get lost, since we can see the name of each language for each row. Finally, KWIC is similar to the horizontal one, but the search word appears highlighted in the centre of the screen, with some context to the right and to the left.
It is possible to download all search results in different formats, CSV and Excel, to allow users to store their searches. On the other hand, it is important to make clear that results could appear in two different levels of alignment: sentence level and paragraph level. The former is a level in which each row represents the alignment by sentence, while the later is by paragraph.

4. Conclusions and future work

This paper describes the creation of a parallel corpus in Mexican languages. We expose some of the arising problems in searching, digitalization, and alignment of the texts. We expect to add more languages and increase the number of texts for the six languages. We desire the CPLM to be consulted by different people from language learners, members of different speech communities to researchers and other interested people. We extend the invitation to other researchers to integrate their indigenous language corpora such as text, journals, and elicitation material. We are aware of the difficulties of the alignment. We expect to design a system to aid the researchers to align the texts. The CPLM will contribute providing data for NLP researchings and we hope it can be used for building language technologies (multilingual lexicon extraction, statistical machine translation, typological studies, language complexity, and word embeddings). For future work, we will add audios in the

Modifier	Description	Description	Results
Without modifier	Exact form	avión	"Avión, avión"
Quotation marks	Character combination of language dependence	"avion"	"Avión, avión, aviön, avìòn, avíón"
Asterik	Wildcard	*ito	"chiquito, banquito, chorrito"
Asterik	Wildcard	ca*	"casa, carro, carnaval"
Asterik	Wildcard	de * y *	"de carne y hueso, de oro y plata"
Question mark	Wild card	p?lo	"palo, pelo, polo"
Brackets	Explicit distance of words	un {2} de	"un servicio público de, un amor platónico de"
Brackets	Distance from n to n words	un {1-4} de	"un vaso de, un plato de alimentos de"

Table 6: Search requests

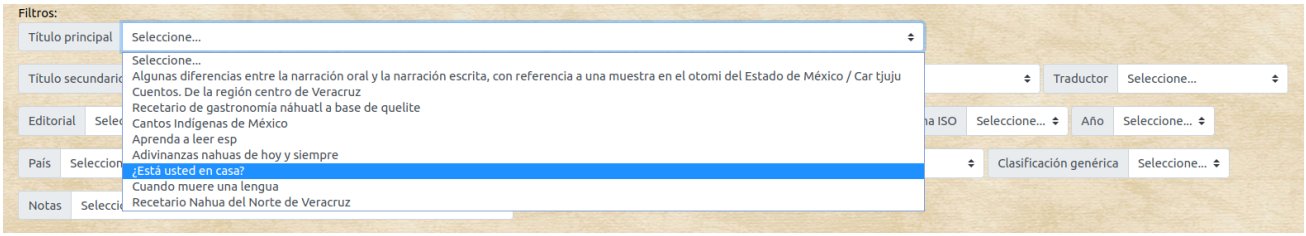

Figure 2: CPLM Filters

web interface and we want to build dictionaries from the vocabulary lists included in the texts of the corpus to enhanced searching.

5. Acknowledgments

This work is supported by the Mexican Council of Science and Technology (CONACYT) funds 370713 and FC-2016-01-2225, and PAPIIT IA401219.

6. Bibliographical References

Amel Fraisse, R. J. and Fishkin, S. F. (2018). Building multilingual parallel corpora for under-resourced languages using translated fictional texts. In Claudia Soria, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

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

INEGI, I. (2015). *Encuesta intercensal de población y vivienda. Población de tres años y más que habla lengua indígena por sexo y lengua según grupos quinquenales*.

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., Carrillo, D., and Meza, I. (2018a). Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Intelligent and Fuzzy Systems applied to Language Knowledge Engineering*, 34(5):3081–3087.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza, I. (2018b). 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.

Mithun, M. (2001). *The languages of native North America*. Cambridge University Press.

Rensch, C. (1976). *Comporative Otomanguean phonology*. Indiana University.

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.