Challenges of language technologies for the indigenous languages of the Americas

Manuel Mager
Instituto de Investigaciones en Matemáticas
Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
mmager@turing.iimas.unam.mx

Ximena Gutierrez-Vasques
GIL IINGEN
Universidad Nacional
Autónoma de México
xim@unam.mx

Gerardo Sierra
GIL IINGEN
Universidad Nacional
Autónoma de México
gsierram@iingen.unam.mx

Ivan Meza
Instituto de Investigaciones en Matemáticas
Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
ivanvladimir@turing.iimas.unam.mx

Abstract

Indigenous languages of the American continent are highly diverse. However, they have received little attention from the technological perspective. In this paper, we review the research, the digital resources and the available NLP systems that focus on these languages. We present the main challenges and research questions that arise when distant languages and low-resource scenarios are faced. We would like to encourage NLP research in linguistically rich and diverse areas like the Americas.

Title and Abstract in Nahuatl

Masehualtlahtoltecnologias ipan Americatlalli

In nepapan Americatlalli imacehualtlahtol, inin tlahtolli ahmo quinpiah miac tlahtoltecnologías ("tecnologías del lenguaje"). Ipan inin amatl, tictemoah nochin macehualtlahtoltin intequiuh, nochin recursos digitales ihuan nochin tlahtoltecnologías in ye mochiuhqueh. Cequintin problemas monextiah ihcuac tlahtolli quinpiah tepitzin recursos kenin amoxtli, niman, ohuic quinchihuaz tecnología ihuan ohuic quinchihuaz macehualtlahtolmatiliztli. Cenca importante in ocachi ticchihuilizqueh tlahtoltecnologías macehualtlahtolli, niman tipalehuilizqueh ahmo mopolozqueh inin tlahtolli.

1 Introduction

The American continent is linguistically diverse, it comprises many indigenous languages that are nowadays spoken from North to South America. There is a wide range of linguistic families and they exhibit linguistic phenomena that are different from the most common languages usually studied in Natural Language Processing (NLP). There are approximately 28 million[1] people who self identify as members of an indigenous group (Wagner, 2016) and they speak around 900[2] native or indigenous languages. This represents an important cultural and linguistic richness. This richness was captured by the following quote from McQuown (1955) "in one small portion of the area, in Mexico just north of the Isthmus of Tehuantepec, one finds a diversity of linguistic type hard to match on an entire continent in the Old World". In spite of this, few language technologies have been developed for these languages, moreover, many of the indigenous languages spoken in the Americas face a risk of language extinction.

The aim of this work is to explore the research in the NLP field for the indigenous languages spoken in the American continent and to encourage research for these languages. We stress the need of developing language resources and NLP tools for these languages and we point out some of the challenges that arise when working on this field. Since indigenous languages are digitally scarce, developing technologies can

[1] Each country has its own methodology and criteria to estimate the amount of speakers. This is the sum of all estimations.
[2] This number varies depending on the classification criteria used on different studies.

have a positive social impact for the communities which depend on these languages. The great diversity of these languages posses interesting scientific challenges, e.g., adapting well established approaches, creation of new methods, collecting new data. Tackling these challenges contributes to building more general computational models of language, and to get a deeper insight into human language understanding. Moreover, many statistical NLP methods seek to achieve language independence, however they often lack of linguistic knowledge or they do not cover the broad diversity of languages (Bender, 2011). In this sense, it is important to acknowledge the characteristics of the indigenous languages of the Americas as a way of complementing the current NLP methods.

Contributions. To sum up, we made the following contributions: (i) we gave a brief introduction into the diverse language space of the indigenous languages of the American continent; (ii) we provide an overview of the existing digital corpora and language sytems that have been developed for some of the languages spoken in the Americas; (iii) and we discusses the advances, used methodologies, challenges and open questions for the most researched NLP tasks in these languages.

In order to maintain the information of this paper updated of the computer readable resources, developed systems and scientific research we made a public available list[3] with the latest advances for the Indigenous Languages of the Americas.

1.1 Languages overview

Linguistic typology is a field that studies the different languages of the world and tries to establish the relationships among them. This is not an easy task since there is still limited knowledge about many languages, specially in highly diverse regions. However, according to several linguistic atlas[4], there are around 140 linguistic families in the world, from these linguistic families almost 40% are native to the Americas. Nowadays, approximately 900 different indigenous languages are spoken of this region, making this continent a linguistically diverse territory.

Americas native languages can exhibit very different linguistic phenomena, these typological features are important to be taken into account when developing language technologies (O'Horan et al., 2016). It would be hard to provide a general description of all the languages spoken in the Americas, however, we would like to highlight some of the linguistic features that usually represent a challenge when doing NLP.

Many languages in North America tend to have a high degree of morphological synthesis, i.e., many morphemes per word (Mithun, 2017). For instance, languages that belong to the Eskimo-Aleut family (native to Canada, Alaska, Greenland and Siberia) are highly polysynthetic suffixing languages. Languages from other linguistic families spoken in North America also show specific degrees of agglutination, polysynthesis, and they have morphemes that express a wide range of functions and nuances of space or direction (Mithun, 2001). Languages with this type of phenomena usually have compact word constructions that are equivalent to whole sentences in other languages like English.

Another linguistic phenomena that is found in many languages spoken in the Americas is the tone, i.e., languages where the pitch is important to distinguish one word from another (the tone can express lexical meaning and grammatical function). Some linguistics families like Oto-manguean (spoken in Mexico) have languages with many types of tones. The orthography of these languages need to mark the wide range of tones, however, many indigenous languages face a lack of orthographic normalization. This can be specially problematic when trying to do NLP and processing text documents.

In general, the native languages of the Americas have a limited digital text production, in some cases they may have a strong oral tradition but not a written one. Due to social and political reasons, alphabetization and education programs are not always available for native speakers. Details about the languages families for which we were able to identify digital and technological resources, are given in Appendix A

[3]Updated list of resources for Indigenous languages of the Americas: `https://github.com/pywirrarika/naki`

[4]Based on Ethnologue (Simons and Fennig, 2017), Glottolog(Nordhoff et al., 2013), WALS (Dryer and Haspelmath, 2013).

2 Corpus and digital resources

Most of the current state of the art methods in NLP are data-driven approaches that require vast amounts of corpora in order to achieve good performance. Widely popular machine learning methods and vector space representations, e.g., neural networks and word embeddings, often rely in big monolingual corpora.

Annotated and unannotated corpora are required for several NLP tasks. For instance, parallel corpora are essential for building statistical machine translation (MT) systems, while morphological annotated data is essential for POS (Part-of-Speech) taggers and morphological analyzers, just to name a few.

In the case of MT, the most common sources for gathering large amounts of parallel data include specific domain texts such as parliamentary proceedings, religious texts, and software manuals that are translated into several languages. Additionally, the World Wide Web represents a good and typical source for finding large-size and balanced parallel and monolingual text (Resnik and Smith, 2003). However, many of the world languages do not have readily available digital corpora. Indigenous languages of the Americas do not have a web presence or text production comparable to richer resourced languages and it is difficult to find websites that offer their content the native languages.

We explored the resources that are digitally available for some of the native languages spoken in the Americas. Regarding parallel corpora, the bible is a common source that contains translations to many of these languages, although it is not always straightforward to extract the content in a digital format. On the other hand, there are some projects that offer parallel content through a web search interface, e.g., Axolotl (Spanish-Nahuatl parallel corpus) that was mainly gathered from non-digital sources (books from several domains), the documents have dialectal, diachronic and orthographic variation (Gutierrez-Vasques et al., 2016). Nahuatl is a Uto-Aztecan language spoken in Mexico (approx. 1.5M speakers) that lack of an orthographic normalization. In fact, this is the case for many languages spoken the Americas: large dialectal variation, and missing standardization.

Also for the same language (Nahuatl), a comprehensive digital dictionary has been collected with information from five previous dictionaries (Thouvenot, 2005), these dictionaries date from 16th to 21st Century (de Olmos et al., 2002; Walters et al., 2002). The query interface of this resource is available online[5].

Another language that belongs to the Uto-Aztecan family is Wixarika or Huichol (approx. 45K speakers). For this language, there is a parallel corpus (Wixarika-Spanish) that gathers translations of classic Hans Christian Andersen's literature (Mager et al., 2018). In this case, the translations belong to one specific dialect (Nayarit). This resource can be fully downloaded from the web [6].

The Shipibo-konibo language (approx. $26,000$ speakers) belongs to the Panoan language family, it is spoken in the Amazon region of Peru and Brazil. Several types of digital resources are available for this language. A parallel corpus between Spanish and this language was constructed using educational and religious documents (Galarreta et al., 2017). Moreover, Shipibo-konibo language has a POS tagged corpus, a set of words and its lemmas and an online-dictionary that has been recently released by Pereira-Noriega et al. (2017).

Also spoken in South America, the Guarani language (approx. 8M speakers) belongs to the Tupi-Guarani family. Abdelali et al. (2006) developed software for collecting Guarani resources from speakers, from this gathering they were able to construct a parallel corpus (Spanish-English-Guarani) and a monolingual corpus. There is also a digital Guarani dictionary (Ramírez and Wolf, 1996) available online[7].

Quechua is one of the most spoken language families on the continent (approx. 9M speakers) but there is scarceness of corpora and language technologies. Espichán-Linares and Oncevay-Marcos (2017) released monolingual corpora in 16 Peruvian languages that belong to different linguistic families (including Quechua).

Regarding speech resources, Guarani has a spoken corpus comprised by $1,000$ phrases from 110 different speakers, it was collected through a web interface (Maldonado et al., 2016), however, this dataset

[5] www.gdn.unam.mx

[6] https://github.com/pywirrarika/wixarikacorpora

[7] https://www.uni-mainz.de/cgi-bin/guarani2/dictionary.pl

has not been publicly released. The Chatino language (approx. 45K speakers) is an Oto-Manguean language spoken in southern Mexico, recently a language documentation and revitalization project has been developed. They use automatic speech recognition and forced alignment tools to time align transcriptions. Parts of this resource are freely available (Cavar et al., 2016).

There are some other types of datasets that are useful for developing language technologies, e.g., morphological annotated data. The CoNLL-SIGMORPHON 2017 Shared Task (Cotterell et al., 2017) released a large morphological database with inflection information of 52 languages, including Haida (7,040 words), Navajo (12,000 words), and Quechua (12,000 words), all of these are indigenous languages spoken in the Americas. In the same way, there is a Oto-manguean inflectional class database which contains over 13,000 verbal entries from twenty Oto-Manguean languages spoken in Mexico, along with information about each verb's inflectional class membership (Palancar and Feist, 2015). For morphological segmentation a data set of four Uto-Aztecan languages[8] were used and released by Kann et al. (2018) (4,468 segmented words from the Mexicanero, Nahutal, Wixarika and Yorem Nokki languages). The UQAILAUT Project contains roots, lexicalized words, infixes, noun and verb endings for the Inuktitut language[9] (Farley, 2012). Plain Cree language, spoken in North America (Algic language family) has also a lexicon databases (16,452 words) collected by Walters et al. (2002) and Wolfart and Pardo (1973).

To improve the data recollection, Dunham et al. (2014) developed tools for annotation of text and audio for Blackfoot, Gitksan, Okanagan, Tlingit, Plains Cree, Coeur d'Alene and Kwak'wala. With these tool they gather 19,187 word forms, 324 texts and 18.8 GB of audio.

A collection of datasets have been developed for the Mapudungun or Mapuche language spoken mainly in Chile (Araucanian language family) (Huenchullan, 2000; Monson et al., 2004). An audio dataset with 170 hours of spoken Mapudungun, that covers three dialects (120 hours of Nguluche, 30 hours of Lafkenche and 20 hours of Pewenche) has been released. This resource contains a word list with the 70,000 most frequent full form words (stem plus inflections) to support a spelling checker for Mapudungun. Also a bilingual Mapudungun-Spanish lexicon was included, containing sample entries 1,600.

Annotated data corresponding to higher linguistic levels is harder to find. For instance, almost no treebanks have been developed for the indigenous languages of the Americas. To our knowledge, the only available dataset is a parallel aligned treebank between Quechua and Spanish (Rios et al., 2008) with 2,000 annotated sentences.

It is important to mention that many of the languages spoken in the Americas have a Wikipedia's set of articles available[10]. This is useful for building monolingual and comparable corpora. Furthermore, Wikipedia can be a helpful resource to predict Part-of-Speech (POS) tags for low resource languages and other tasks (Hoenen, 2016). In any case, one common limitation of the digital resources for these languages is the lack of orthographic standardization and difficulties for processing certain types of characters. Table 1 summarizes the above-mentioned resources.

3 Morphological segmentation and analyses

Morphology has been studied in NLP field focusing mainly on the following tasks: lemmatization, stemming, segmentation, analysis and inflection/reinfection. These tasks serve to other higher level tasks such as machine translation. On this regard, there have been several studies which have been applied to the Americas languages.

In NLP, lemmatization and stemming methods are used to reduce the morphological variation by converting words forms to a standard form, i.e., a lemma or a stem. However, most of these technologies are focused in a reduced set of languages. For languages like English there are plenty of resources, however, this is not the case for all the languages. Specially for languages in the Americas with rich

[8] Available at `http://turing.iimas.unam.mx/wix/mexseg`

[9] `http://www.inuktitutcomputing.ca/DataBase/info.php`

[10] The available languages in wikipedia can be consulted at: `https://es.wikipedia.org/wiki/Portal:Lenguas_indgenas_de_Amrica`. Until the publication of this article there are only entries in: Nahuatl, Navajo, Guarani, Aymara, Klaalisut, Esquimal, Inukitut, Cherokee, and Cree.

Type of resource	Languages	Size	Reference
Parallel	Nahuatl-Spanish	18K sentences	Gutierrez-Vasques et al. (2016)
Parallel	Wikarika-Spanish	8K sentences	Mager et al. (2018)
Parallel	Shipibo konibo - Spanish	11.8K sentences	Galarreta et al. (2017)
Parallel	Spanish-English-Guarani	250K sentences	Abdelali et al. (2006)
Parallel	1259 languages		Mayer and Cysouw (2014)
POS Tagged	Shipibo konibo	217 sentences	Pereira-Noriega et al. (2017)
Lemmatized words	Shipibo konibo	3.5K words	Pereira-Noriega et al. (2017)
Dictionary	Shipibo konibo - Spanish	3.5K words	Pereira-Noriega et al. (2017)
Dictionary	Nahuatl		Palancar and Feist (2015)
Dictionary	Guarani		(Ramírez and Wolf, 1996)
Speech	Guarani	1K phrases	Maldonado et al. (2016)
Speech	Chatino	10 hours with Transcription	Cavar et al. (2016)
Speech	Blackfoot, Nata, Gitksan, Okanagan, Tlingit, Plains Cree, Ktunaxa, Coeur d'Alene, Kwak'wala	19.8 GB	Dunham et al. (2014)
Speech	Mapudungun	170 hours	Huenchullan (2000) and Monson et al. (2004)
Morphological Inflection	Quechua, Navajo, Haida	31K words	Cotterell et al. (2017)
Morphological Inflection	20 Oto-Manguean languages	13K verbs	Palancar and Feist (2015)
Morphological Segmentation	Uto-Aztecan languages (Mexicanero, Nahutal, Wixarika, Yorem Nokki)	4.4K words	Kann et al. (2018)
Morphological segmentation	Inuktitut	2K roots, 1.8K affixes	Farley (2012)
Monolingual	16 Peruvian languages	Unknown	Espichán-Linares and Oncevay-Marcos (2017)
Monolingual	Plain Cree	16K words	Walters et al. (2002) and (Wolfart and Pardo, 1973)
Treebank	Quechua	2K sentences	Rios et al. (2008)

Table 1: Digital available resources of American Indigenous Languages for NLP

morphological phenomena, and not always suffixal, where it is not enough to remove inflectional endings in order to obtain a stem.

Morphological segmentation is the task of splitting a word into the surface forms of its smallest meaning-bearing units, its *morphemes*. On the other hand, Morphological analysis not only focuses in the segmentation of words, but also in assigning tags to each part of the word. There are several approaches to do these tasks, i.e., rule-based, semi-supervised and unsupervised (Goldsmith, 2001; Creutz and Lagus, 2002; Kohonen et al., 2010). Some examples of rule-based methods applied to the Americas languages are the Finite State approaches to model the morphology of a language: plains Cree (Arppe et al., 2017; Harrigan et al., 2017; Wolfart and Pardo, 1973; Snoek et al., 2014), East Cree (Arppe et al., 2017) , for the East Odawa dialect of Ojibwe (Bowers et al., 2017), for Mohawk (Iroquoian language family) (Assini, 2014), for the Bribri (Chibchense language family) (Solórzano, 2017) using the FOMA tool (Hulden, 2009), Quechua (Vilca et al., 2012; Monson et al., 2006), Mapudungun (Monson et al., 2006), and the Argentinian branch of Quechua and Toba (Porta, 2010). More recently a new hybrid ap-

proach of finite-state transducer (FST) with statistical inference is part of the *Basic Language Technology Toolkit for Quechua* (Rios, 2016).

For Uto-Aztecan languages, there exists a computational tool called "*chachalaca*" that performs morphological analysis (Thouvenot, 2011) of Nahuatl. This is a rule-based software focused on Classical Nahuatl, it is able to generate more than one morphological analysis candidate per word. It is based on grammars that describe most of the 16th-century-word constructions. Additionally, Mager et al. (2018) propose a morphological segmentation tool for the Wixarika language, with a supervised approach, using previous given morphological tables and a probabilistic model to infer the inherent morphological rules.

Regarding to unsupervised methods, neural methods have been used to tackle the rich morphology of the languages of the continent. Micher (2017) propose a Segmental Recurrent Neural Network (RNN) for segmenting and tagging Inuktitut. Kann et al. (2018) used a set of extensions to the Encoder-Decoder RNN architecture with Gated Recurrent Units (GRU) for automatically segmenting four Uto-Aztecan languages (Mexicanero, Nahuatl, Wixarika and Yorem Nokki). Semisupervised segmentation approaches like Morfessor have also been successfully applied to Nahuatl (Gutierrez-Vasques, 2017). For the Uto-Aztecan language Tarahumara and the Mayan language Chuj, there are works that try to automatically discover affixes through unsupervised approaches (Medina-Urrea, 2007; Medina-Urrea, 2008; Medina Urrea and García, 2006).

Lately, there has been interest in the reinflection task, i.e., generating an inflected form for a given target tag and lemma. The CoNLL-SIGMORPHON Shared Task (Cotterell et al., 2016; Cotterell et al., 2017) released a dataset for reinflection of 52 languages, including 3 Native American languages. The systems that got the best performance (Kann and Schütze, 2016; Kann and Schütze, 2017; Makarov et al., 2017).

In some cases it is difficult to disambiguate between homonym morphs. To deal with this problem, Rios et al. (2008) used Conditional Random Fields (CRFs) (Lafferty et al., 2001).

Most of the methods that we found that deal with morphology are based on FST. However, the indigenous languages of the continent are far too diverse, it would be expensive to build such analyzers with expert knowledge for each language, besides the fact that the analyzers need to be constantly updated to cope with language change (Creutz and Lagus, 2005).

4 Machine Translation

Machine Translation is a natural task for indigenous languages, since it might provide a communication window with more popular languages. The development of MT systems for indigenous languages have follow the trends in the field, from rule-based, to statistical and neural based approaches.

Rule-Based Machine Translation (RBMT) approaches are sometimes suitable for low resource languages since they do not require aligned parallel corpora. However. In recent years, research on data-driven approaches has increased, with the aim to overcome the scarcity of data using different methods. In any case, translation of low-resource languages represents an interesting and active research problem in the NLP field.

In the case of native American languages, there have been some efforts in building rule-based systems. The Apertium system (Tyers et al., 2009; Forcada et al., 2011) is a big help for this approach, and at least two languages has translation teams working with it. This is the case of Quechua (Eastern Apurimac Quechua and Cusco Quechua)-Spanish (Cavero and Madariaga, 2007) and Spanish-Wayuunaiki (Spoken in Venezuela and Colombia) (Fernández et al., 2013). Other RBMT systems were created for Aymara-Spanish (spoken in Peru) (Coler and Homola, 2014) Wayuunaki-Spanish, Quechua-Spanish and Mapuche-Spanish (Monson et al., 2006). For the latter a web available translator[11](González Hernández, 2016) is available. We found that there are mobile apps for translating indigenous languages, this is the case of Zapotec-Spanish language pair (spoken in Mexico from the Oto-Manguean family)[12].

All of this RBMT systems have a set of shortcomings. The majority is not able to translate complex constructions, specially when the languages are distant from each other, which increases the complexity

[11] `http://142.4.219.173/wt/`

[12] `https://play.google.com/store/apps/details?id=com.SimplesoftMx.Didxazapp&hl=es`

of the machine translation rules. One way to overcome this is to include linguistic information, e.g., morphology, syntax. However, this kind of knowledge or linguistic tools is not always available, especially for low-resource languages. Experiments using the Example Based Machine Translation (EBMT) methodology are not common, we only found the work of Llitjós et al. (2005) and Monson et al. (2006) for the Mapuche-Spanish pair.

Statistical Machine Translation (SMT) Systems are data-driven since they use vast amounts of parallel corpora to model the translations between sentences or subunits. Their performance is highly dependent on the number of training data; they represent a challenge when low resource conditions are faced. In the case of the native languages of the Americas, they tend to be morphologically rich and this must be taken into account to improve the translation and reduce the data sparseness. An example of this is the Wixarika-Spanish SMT system that aligns Wixarika morphemes with Spanish words or tokens (Mager Hois et al., 2016; Mager Hois, 2017)[13]. A similar case can be found for the Nahuatl-Spanish pair. Uto-Aztecan languages can be highly agglutinative with the polysynthetic tendency, Gutierrez-Vasques (2015) extracts bilingual correspondences from a parallel corpus, by aligning the Nahuatl non-grammatical morphs to Spanish words. Another example was collected for the pair Mixteco-Spanish (Santiago, 2017). The same trend can be observed in SMT for Shipibo-konibo (Galarreta et al., 2017).

Regarding commercial systems, Microsoft has targeted some languages spoken in Mexico, Mayan and Otomi (Queretaro variant)[14]. SMT was also applied for the Guarani, it translates to Spanish, but also to English, French, Italian, German and Portuguese[15].

Recently, there has been an increasing interest in Neural Machine Translation (NMT) models, which are also statistical based, but they use neural networks architectures that are feed with very big amounts of parallel corpora. Mager and Meza (2018) showed that in such low-resource scenarios, translating from Mexicanero, Nahuatl, Purepecha, Wixarika and Yorem-Nokki to Spanish, SMT systems achieve better performance than NMT. Even though these architectures are not suitable for low-resource settings, there have been some recent efforts to adapt them. Soriano (2018) experimented with the Mexican Purepecha (an isolated language with about 140 thousand speakers) using the OpenNMT toolkit (Klein et al.,). Tiedemann (2018) took the massive bible corpus (Mayer and Cysouw, 2014) and trained a multilingual NMT model to improve overall translation performance. Experiments included Oto-manguean, Quechua and Mayan families. Moreover, empirical results (Vania and Lopez, 2017) show that problem of data sparsity of rich morphological languages can be handled with subword models: the usage of character level NMT improve performance over token level translation and unsupervised morphological segmentation (Creutz and Lagus, 2002). But their experiment also conclude that a canonical segmentation enhances character level translation.

In order to alleviate the lack of resources automatic data recollection has been proposed, this has been tried for Guarani language (Abdelali et al., 2006). Moreover, it would be very useful to have big repositories of translated texts. One alternative is to create parallel corpora between many languages using manual translations (controlled elicitation) as described for the Mapudungun (Mapuche) language (Probst et al., 2001).

In any case, SMT and NMT systems should be adapted to deal with the scarcity, of the sparseness of word forms and the rich morphology of languages. Although there are works that try to deal with morphology (Virpioja et al., 2007; Popovic and Ney, 2004; Costa-jussà and Fonollosa, 2016; Sennrich et al., 2016; Dalvi et al., 2017), they are rarely applied to Native American languages.

5 Other studies and tools

5.1 Multilinguality and Code-Switching

Most native speakers of indigenous languages are at least bilingual, they have to communicate using the primary or official language of their own country, i.e., Spanish, Portuguese, French or English. Only few

[13] Available at `http://turing.iimas.unam.mx/wix`

[14] `https://www.microsoft.com/en-us/translator/languages.aspx`

[15] `http://www.iguarani.com/?p=traductor`

communities remain completely monolingual in their native language, moreover, modern migrations and the use of social networks contribute to bilingualism and code-switching.

Code-switching occurs when a speaker alternates between two or more languages in a conversation. This adds a challenge when doing NLP for this kind of data. Code-switching is not a new phenomenon, it can be found in historical documents, Garrette and Alpert-Abrams (2016) proposes an unsupervised approach of paired encoding (words and characters) to improve language modeling (Latin, Spanish and Nahuatl) in an Optical character recognition (OCR) task. King and Abney (2013) applies weakly supervised methods for labeling the language of each word in documents that can have many mixed languages. The targeted languages are 30, including Chippewa, Nahuatl, and Ojibwa.

Being able to automatically detect Code-switching could be useful when doing NLP for minority languages, for instance to use the web as a source for a corpus.

From the quantitative linguistics perspective, parallel corpora of an outstanding number of languages have been extracted from the Bible and used to perform typological studies in many languages, included native languages of the Americas. For instance, exploring the tense behavior (Asgari and Schütze, 2017), contrasting the morphological complexity in many languages (Bentz et al., 2016; Kettunen, 2014) just to mention some.

5.2 Language Tools

For some rich resource languages, there are already available NLP tools that deal with several phenomena and linguistics levels of processing. However, for low resource languages, there is still much work to do. In this section we summarize some of the works related with POS Tagging, OCR, Parsing, Spell Checking, Language Identification and other tasks, that we have found for the languages of the Americas.

Speech synthesis and recognition has made some progress for the Raramuri language (Urrea et al., 2009), using a unit selection approach based on function words, suffix sequences and diphones of the language. For speech recognition, Maldonado et al. (2016) applied on Guarani a Hidden Markov Model (HMM) with the CMU Sphinx toolkit (Lamere et al., 2003). Coto-Solano and Solórzano (2016) proposes an automatic aligner of text and voice for the indigenous language Bribri of Costa Rica.

Part-of-Speech (POS) tagging assigns a category from a given symbol set to each token in an input string. It is used as a prepossessing step that serves as input for other tasks or for higher level NLP task. POS Tagging was also incorporated into the Peruvian Shipibo-konibo NLP toolkit (Pereira-Noriega et al., 2017).

Spell checking is not a trivial task for highly agglutinative and polysynthetic languages, that can't rely on a token based evaluation, and need sub-word level models. We found only two tools that handle this issue: Monson et al. (2006) build a dictionary based tool for Quechua and Alva and Oncevay (2017) for Shipibo-Konibo used rule-based analysis and dictionaries.

Another field that is crucial to increase the amount of digitized data for other tasks is **Optic Character Recognizing** (OCR). Maxwell and Bills (2017) studied the challenges regarding the digitalization of Tzeltal-Spanish, Muinane-Spanish, Cubeo-Spanish dictionaries. Garrette and Alpert-Abrams (2016) developed an unsupervised transcription model for dealing with orthographic variation in digitized historical documents, some of them were written in Nahuatl.

In the context of indigenous **Language Identification** (LID), Espichán-Linares and Oncevay-Marcos (2017) studied LID on 17 languages from the Arawak, Aru, Jíbaro, Pano and Quechua linguistic families. The proposed models were flexible enough to handle the lack of orthographic standardization of the language.

Another important NLP task is **parsing**. The research in this area is also weak, however, the Quechua-Spanish Treebank helped to perform some experiments in this topic (Bresnan et al., 2015; Rios, 2016). For other languages parsing experiments were performed on Ayamara (Homola, 2011) using Lexical-Functional Grammar (LFG).

6 Discussion

Study of the languages of the Americas has increased in the recent years, in both the linguistic and the language technology fields. Many factors have contributed to this, such as speakers self-awareness about the importance of their languages and digital inclusion. Figure 1 shows that many of the papers that we reviewed, were published from year 2000 to present day, with a notable increase in the activity in the last five years. The most studied NLP tasks are Machine Translation and Morphology, however, from 2013 upon now, other tasks, e.g., POS-tagging, parsing, speech, spell correction, also received attention.

Despite the fact that we found NLP contributions for around 35 languages, this is still a small number if we take into account the big diversity and number of languages that exist in the continent. Table 2 showed that some linguistic families have concentrated the attention, but even for these languages the developed technology is not enough. We noticed that North American languages are the most studied, despite some of them don't have a big number of speakers compared to other indigenous languages, e.g., Navajo, Haida, Cree, Chippewa, Ojibwa, Blackfoot, Nata, Gitksan, Okanagan, Tlingit, Plains Cree, Ktunaxa, Coeur d'Alene, Kwak'wala, and Inuktitut. Uto-Aztecan language family that includes languages like Raramuri, Nahuatl, Wixarika, Yorem Nokki, Mexicanero (spoken mainly in Mexico) have also received attention from the NLP community. Regarding to South America, the most spoken native languages, Quechua, Mapuche, Guarani and Ayamara, have several resources available. Surprisingly, languages with less speakers such as Shipibo-konibo, Arawak, Aru, Jíbaro, Pano and Wayuunaki have been also studied.

The diversity in the linguistic phenomena of these languages makes developing language technologies a challenging task. In recent years, NLP and Machine Learning fields have paid attention to low resource settings, organizing workshops and special tracks to tackle this issue. Indigenous languages could be greatly benefit from this kind of research in the future. In particular, American languages with rich morphology, e.g., agglutinative and polysyntehthic, seems to benefit from approaches that take into account the morphology and sub-word modeling.

We also noticed that some NLP tasks that are considered almost solved for languages like English, need to be adapted or started from scratch when applied to the languages of the American continent. Moreover, fields like machine translation could enable in the future the creation of multilingual technologies for all of the languages in the world that face a similar situation, this could have a great impact in these communities.

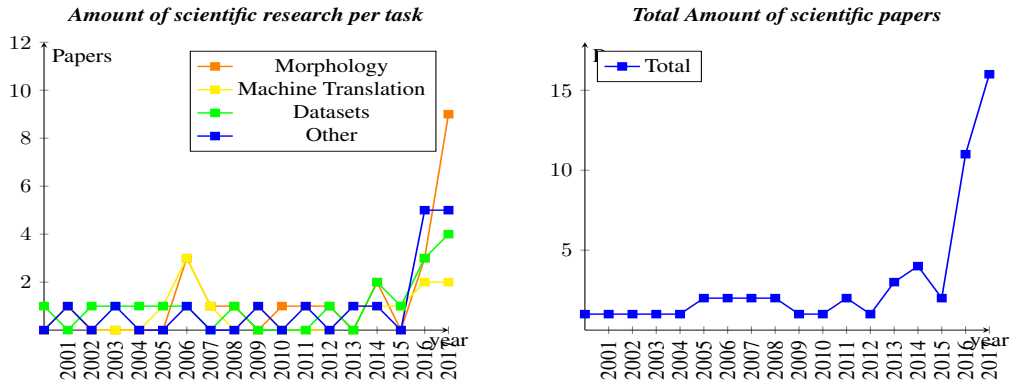

Figure 1: NLP papers and digital resources that contain any indigenous language of the Americas (between 2000 and 2017)

7 Conclusions

In this work, we presented a review of NLP research focused on Indigenous Languages of the Americas. We showed which languages have available digital resources and their related tools. Research has focused in tasks like morphology and machine translation, however, there is still a lot of work to be done since these languages exhibit a wide range of linguistic phenomena while resources are scarce.

Through this work, we discussed some of the challenges that must be taken into account, e.g., small datasets, high dialectal variation, rich morphology, lack of orthographic normalization, scarcity of linguistic preprocessing tools.

NLP research for these languages can broad the understanding of human language structures and help to build more general computational models. Moreover, the development of language technologies can have a positive social impact for the speakers of the indigenous languages, helping to maintain the living cultural heritage that each language represents.

Acknowledgements

This work was supported by the Mexican Council of Science and Technology (CONACYT), fund 2016-01-2225. We will also thank the reviewers for their valuable comments and to Katharina Kann for her comments and support.

References

Ahmed Abdelali, James Cowie, Steve Helmreich, Wanying Jin, Maria Pilar Milagros, Bill Ogden, Hamid Mansouri Rad, and Ron Zacharski. 2006. Guarani: a case study in resource development for quick ramp-up mt. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,"Visions for the Future of Machine Translation*, pages 1–9.

Carlo Alva and Arturo Oncevay. 2017. Spell-checking based on syllabification and character-level graphs for a peruvian agglutinative language. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 109–116.

Antti Arppe, Marie-Odile Junker, and Delasie Torkornoo. 2017. Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. *ComputEL-2*, page 52.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124.

Alicia Alexandra Assini. 2014. *Natural language processing and the Mohawk language: creating a finite state morphological parser of Mohawk formal nouns*. Scholars' Press.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardzic. 2016. A comparison between morphological complexity measures: typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153.

Dustin Bowers, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2017. A morphological parser for Odawa. *ComputEL-2*, page 1.

Joan Bresnan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2015. *Lexical-functional syntax*, volume 16. John Wiley & Sons.

Malgorzata Cavar, Damir Cavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, asr. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Paris, France, may. European Language Resources Association (ELRA).

Indhira Castro Cavero and Jaime Farfán Madariaga. 2007. Traductor morfológico del castellano y quechua. *Revista I+i*, 1(1).

Matt Coler and Petr Homola. 2014. Rule-based machine translation for aymara. *Endangered Languages and New Technologies*, page 67.

Marta R Costa-jussà and José AR Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 357–361.

Rolando Coto-Solano and Sofía Flores Solórzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de costa rica. *Káñina*, 40(4):175–199.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Gėraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, et al. 2017. CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Workshop on Morphological and Phonological Learning*.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and improving morphological learning in the neural machine translation decoder. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 142–151.

Andrés de Olmos, Ascensión H de León-Portilla, and Miguel León Portilla. 2002. *Arte de la lengua mexicana*, volume 9. UNAM.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Joel Dunham, Gina Cook, and Joshua Horner. 2014. Lingsync & the online linguistic database: New models for the collection and management of data for language communities, linguists and language learners. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 24–33.

Alexandra Espichán-Linares and Arturo Oncevay-Marcos. 2017. A low-resourced peruvian language identification model. In *Proceedings of the SIMBig 2017 Track on Applied Natural Language Processing, ANLP 2017*.

Benoit Farley. 2012. The Uqailaut project. *URL http://www. inuktitutcomputing. ca*.

Dayana Iguarán Fernández, Ornela Quintero Gamboa, Jose Molina Atencia, and Oscar Elías Herrera Bedoya. 2013. Design and implementation of an "web api" for the automatic translation colombia's language pairs: Spanish-Wayuunaiki case. In *Communications and Computing (COLCOM), 2013 IEEE Colombian Conference on*, pages 1–9. IEEE.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Ana Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo. In *Proceedings of RANLP*.

Dan Garrette and Hannah Alpert-Abrams. 2016. An unsupervised model of orthographic variation for historical document transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 467–472.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

José González Hernández. 2016. *Herramienta de traduccin automática de wayuunaiki a español. Caso de estudio: sintagmas nominales y verbales simples*. Ph.D. thesis, Universidad de Zulia.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Hernandez Isaac. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *International Conference on Language Resources and Evaluation (LREC)*.

Ximena Gutierrez-Vasques. 2015. Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 154–160.

Ximena Gutierrez-Vasques. 2017. Exploring bilingual lexicon extraction for Spanish-Nahuatl. In *ACL Workshop in Women and Underrepresenting Minorities in Natural Language Processing*.

Atticus G Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4):565–598.

Armin Hoenen. 2016. Wikipedia titles as noun tag predictors. In *International Conference on Language Resources and Evaluation (LREC)*.

Petr Homola. 2011. Parsing a polysynthetic language. In *RANLP*, pages 562–567.

Carolina Huenchullan. 2000. Data collection and language technologies for Mapudungun. In *International Conference on Language Resources and Evaluation (LREC)*.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 29–32. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70.

Katharina Kann and Hinrich Schütze. 2017. The LMU system for the CoNLL-SIGMORPHON 2017 shared task on universal morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 40–48.

Katharina Kann, Manuel Mager, Ivan Meza, and Hinrich Shütze. 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)*, volume 1, pages 47–57.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Ben King and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *ACL Special Interest Group on Computational Morphology and Phonology*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. 2003. The CMU SPHINX-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, volume 1, pages 2–5.

Ariadna Font Llitjós, Roberto Aranovich, and Lori Levin. 2005. Building machine translation systems for indigenous languages. In *Second Conference on the Indigenous Languages of Latin America (CILLA II), Texas, USA*.

Manuel Mager and Ivan Meza. 2018. Hacia la traducción automática de las lenguas indígenas de méxico. In *Proceedings of the 2018 Digital Humanities Conference*. The Association of Digital Humanities Organizations.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for the Wixarika (Huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Jesus Manuel Mager Hois, Carlos Barron Romero, and Ivan Vladimir Meza Ruíz. 2016. Traductor estadístico wixarika - español usando descomposición morfológica. *COMTEL*, (6).

Jesus Manuel Mager Hois. 2017. Traductor híbrido wixárika - español con escasos recursos bilingües. Master's thesis, Universidad Autónoma Metropolitana.

Peter Makarov, Tatiana Ruzsics, and Simon Clematide. 2017. Align and copy: Uzh at sigmorphon 2017 shared task for morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 49–57.

Diego Manuel Maldonado, Rodrigo Villalba Barrientos, and Diego P Pinto-Roa. 2016. Eñe'e: Sistema de reconocimiento automático del habla en guaraní. In *Simposio Argentino de Inteligencia Artificial (ASAI 2016)-JAIIO 45 (Tres de Febrero, 2016)*.

Michael Maxwell and Aric Bills. 2017. Endangered data for endangered languages: Digitizing print dictionaries. *ComputEL-2*, page 85.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. *Oceania*, 135(273):40.

Norman A McQuown. 1955. The indigenous languages of latin america. *American Anthropologist*, 57(3):501–570.

A Medina Urrea and M Alvarado García. 2006. Un experimento de reconocimiento automático de la derivación léxica en el ralámuli. *La lengua y la antropología para un conocimiento global del hombre*.

Alfonso Medina-Urrea. 2007. Affix discovery by means of corpora: Experiments for Spanish, Czech, Ralámuli and Chuj. In *Aspects of Automatic Text Analysis*, pages 277–299. Springer.

Alfonso Medina-Urrea. 2008. Affix discovery based on entropy and economy measurements. *Texas Linguistics Society*, pages 99–112.

Jeffrey C Micher. 2017. Improving coverage of an inuktitut morphological analyzer using a segmental recurrent neural network. *ComputEL-2*, page 101.

Marianne Mithun. 2001. *The languages of native North America*. Cambridge University Press.

Marianne Mithun. 2017. Polysynthesis in north america. In *The Oxford Handbook of Polysynthesis*.

Christian Monson, Lori Levin, Rodolfo M Vega, Ralf D Brown, Ariadna Font-Llitjos, Alon Lavie, Jaime G Carbonell, Eliseo Cañulef, and Rosendo Huisca. 2004. Data collection and analysis of Mapudungun morphology for spelling correction. *Computer Science Department*, page 300.

Christian Monson, Ariadna Font Llitjós, Roberto Aranovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell, and Alon Lavie. 2006. Building NLP systems for two resource-scarce indigenous languages: mapudungun and quechua. *Strategies for developing machine translation for minority languages*, page 15.

Sebastian Nordhoff, Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2013. *Glottolog 2.0*. Max Planck Institute for Evolutionary Anthropology.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. *arXiv preprint arXiv:1610.03349*.

E.L. Palancar and T. Feist. 2015. *Oto-Manguean Inflectional Class Database*. University of Surrey.

José Pereira-Noriega, Rodolfo Mercado-Gonzales, Andrés Melgar, Marco Sobrevilla-Cabezudo, and Arturo Oncevay-Marcos. 2017. Ship-lemmatagger: Building an NLP toolkit for a peruvian native language. In *International Conference on Text, Speech, and Dialogue*, pages 473–481. Springer.

Maja Popovic and Hermann Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *International Conference on Language Resources and Evaluation (LREC)*.

Andrés Osvaldo Porta. 2010. The use of formal language models in the typology of the morphology of amerindian languages. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 109–113. Association for Computational Linguistics.

Katharina Probst, Ralf D Brown, Jaime G Carbonell, Alon Lavie, Lori Levin, and Erik Peterson. 2001. Design and implementation of controlled elicitation for machine translation of low-density languages. *Computer Science Department*.

Gilbert Ramírez and Lustig Wolf. 1996. Interactive guarani dictionary. `https://www.uni-mainz.de/cgi-bin/guarani2/dictionary.pl`. Accessed: 2018-03-16.

Annette Rios, Anne Göhring, and Martin Volk. 2008. A Quechua-Spanish parallel treebank. *LOT Occasional Series*, 12:53–64.

Annette Rios. 2016. A basic language technology toolkit for quechua.

H. Santiago. 2017. Método para la alieación automática de textos entre los idiomas mixteco y español. Master's thesis, Centro Nacional de Investigación y Desarrollo Tecnológico.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Gary F Simons and Charles D Fennig. 2017. Ethnologue: Languages of the world. *SIL, Dallas, Texas*.

Conor Snoek, Dorothy Thunder, Kaidi Loo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42.

Sofía Flores Solórzano. 2017. Desarrollo de un analizador automático de estados finitos para la lengua bribri. `http://morphology.bribri.net/`.

Miguel Soriano. 2018. Traductor Automático español –purépecha mediante OpenNMT. Master's thesis, Universidad de Guadalajara, Mexico.

Marc Thouvenot. 2005. Gran diccionario náhuatl. `http://www.gdn.unam.mx/`.

Marc Thouvenot. 2011. Chachalaca en cen, juntamente. In *Compendio Enciclopedico del Nahuatl, DVD*. INAH.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. *arXiv preprint arXiv:1802.00273*.

Francis M Tyers, Mikel L Forcada, and Gema Ramırez-Sánchez. 2009. The Apertium machine translation platform: Five years on. In *Proc. of the First Intl. Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 3–10.

Alfonso Medina Urrea, José Abel Herrera Camacho, and Maribel Alvarado Garcıa. 2009. Towards the speech synthesis of raramuri: A unit selection approach based on unsupervised extraction of suffix sequences. *Advances in Computational Linguistics*, page 243.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada, July. Association for Computational Linguistics.

Calderon Vilca, Hugo David, Cárdenas Mariñó, Flor Cagniy, and Edwin Fredy Mamani Calderon. 2012. Analizador morfológico de la lengua quechua basado en software libre helsinkifinite-statetransducer (hfst). *COMTEL*.

Sami Virpioja, Jaakko J Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Machine Translation Summit XI*, 2007:491–498.

Claudio Wagner. 2016. Las lenguas indígenas de américa (lenguas amerindias). *Revista Documentos Lingüísticos y Literarios UACh*, (17):30–37.

JCW Walters, M Mirten, P Hernández, E Pérez, and CH Upton. 2002. Diccionario náhuatl de los municipios de mecayapan y tatahuicapan de juárez. *Veracruz. 2ª edición electrónica (cuerpo), Instituto Lingüístico de Verano, AC www. sil. org/mexico/nahuatl/istmo/G020a-DiccNahIst-NAU. htm*.

H Christoph Wolfart and Francis Pardo. 1973. Computer-assisted linguistic analysis. *University of Manitoba Anthropology Papers*, (6).

Appendix A. Language Families

Table 2 summarizes the language families for which we were able to identify some digital and technological resources during this research. We distinguish among only two geographical macroareas: North America (it includes Central America) and South America (Dryer and Haspelmath, 2013).

L. Family	Macroarea	Papers	L. Family	Macroarea	Papers
Uto-Aztecan	North A.	16	Mayan	North A.	4
Oto-Manguean	North A.	3	Arawakan	South A.	3
Haida	North A.	4	Ayamaran	South A.	2
Na-Dene	North A.	5	Aru	South A.	1
Eskimo-Aleut	North A.	2	Jibaro	South A.	1
Algic	North A.	8	Bora–Witóto	South A.	1
Tsimshianic	North A.	1	Tucanoan	South A.	1
Penutian	North A.	1	Araucanian	South A.	7
Salishan	North A.	2	Panoan	South A.	4
Wakashan	North A.	1	Tupian	South A.	5
Iroquoian	North A.	1	Quechuan	South A.	15
Chibchan	North A.	2	Guaicuruan	South A.	1

Table 2: Language families (L. Family) for which some technology was found, and the number of NLP/Computer Linguistic papers referring to each (one paper can reference more than one languages). North A. stands for North America and South A. for South America.