🌽Py-Elotl: A Python NLP package for the languages of Mexico

Ximena Gutierrez-Vasques[1,5] **Robert Pugh**[2,4,5] **Victor Mijangos**[1,5]
Diego Alberto Barriga Martínez[1,5] **Paul Aguilar**[5] **Mikel Segura**[1,5] **Paola Innes**[1,5]
Javier Santillan[5] **Cynthia Montaño**[3,5] **Francis M. Tyers**[2,4,5]
[1]UNAM, México [2]Indiana University, Bloomington [3]University of California, Berkeley
[4]Kaltepetlahtol, A.C. [5]Comunidad Elotl
contacto@elotl.mx

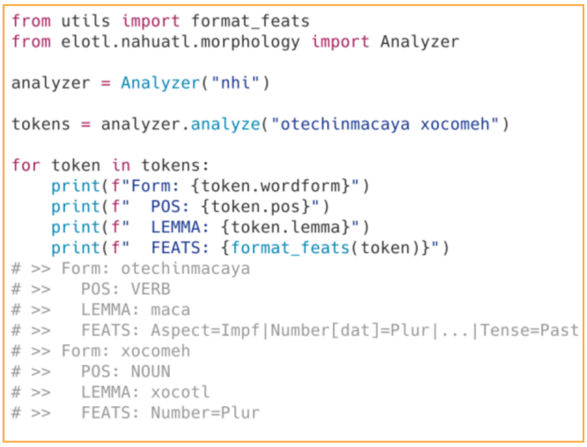
Abstract

This work presents Py-Elotl, a suite of tools and resources in Python for processing text in several indigenous languages spoken in Mexico. These resources include parallel corpora, linguistic taggers/analyzers, and orthographic normalization tools. This work aims to develop essential resources to support language pre-processing and linguistic research, and the future creation of more complete downstream applications that could be useful for the speakers and enhance the visibility of these languages. The current version supports language groups such as Nahuatl, Otomi, Mixtec, and Huave. This project is open-source and freely available for use and collaboration[1].

1 Introduction

Language technologies have become an integral part of daily life for many people around the world. We regularly interact with automatic translators, voice assistants, AI agents, and writing tools, to name a few. These advanced NLP technologies (downstream applications) have only been possible due to the gradual and systematic creation of foundational resources and tools (upstream tasks). This includes the creation of training corpora, linguistic taggers/analyzers, and orthographic normalization tools, among others, all of which play a crucial role in enabling more sophisticated language technology applications.

For many hegemonic languages, these fundamental upstream tasks may appear to have already been solved or are of lesser research interest. As a result, efforts often shift toward advancing more sophisticated technologies, such as large language models (LLMs) capable of generating text, as in commercial assistants like ChatGPT (OpenAI) or Gemini (Google). However, for many other languages, there are still no tools that cover the most basic upstream tasks, so the landscape of language technologies remains uneven (Joshi et al., 2020; Hedderich et al., 2021; Ducel et al., 2022; Blasi et al., 2022).



```python
from utils import format_feats
from elotl.nahuatl.morphology import Analyzer

analyzer = Analyzer("nhi")

tokens = analyzer.analyze("otechinmacaya xocomeh")

for token in tokens:
    print(f"Form: {token.wordform}")
    print(f"  POS: {token.pos}")
    print(f"  LEMMA: {token.lemma}")
    print(f"  FEATS: {format_feats(token)}")
# >> Form: otechinmacaya
# >>   POS: VERB
# >>   LEMMA: maca
# >>   FEATS: Aspect=Impf|Number[dat]=Plur|...|Tense=Past
# >> Form: xocomeh
# >>   POS: NOUN
# >>   LEMMA: xocotl
# >>   FEATS: Number=Plur
```

Figure 1: Example of a morphological analysis (Nahuatl) performed using Py-Elotl.

In order to advance toward a more linguistically-diverse language technology landscape and enable more comprehensive applications for under-resourced languages, it is crucial to start with the fundamental building-blocks, or ”upstream tasks”.

To that end, this work presents a suite of tools and resources for processing text in several indigenous languages spoken in Mexico. Py-Elotl, an open-source Python library, supports several upstream tasks such as parallel corpus loading, orthographic normalization, and morphological analysis (see Figure 1). The name *elotl* comes from the Nahuatl word for “ear of fresh maize”.

This collaborative initiative aims to develop essential resources to support language preprocessing, linguistic research, and the future creation of more downstream applications that could be useful for the speakers and enhance the visibility of these languages.

[1] https://github.com/ElotlMX/py-elotl

Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 38–47
May 4, 2025 ©2025 Association for Computational Linguistics

2 Related work

Although over 7,000 languages are spoken worldwide, most remain largely overlooked in NLP research (Magueresse et al., 2020). The Americas, in particular, are home to immense linguistic diversity, where most of the indigenous languages in the region face varying degrees of endangerment (Moseley and Nicolas, 2010).

In recent years, the NLP community has increasingly focused on the languages of the Americas, promoting specialized forums (Mager et al., 2021b) and shared tasks to advance machine translation, the automatic creation of educational resources, and other applications (Mager et al., 2021a; Chiruzzo et al., 2024). These languages often exhibit high internal diversity and a lack of standardization traditions due to sociopolitical factors, along with other linguistic phenomena that make them particularly challenging to process (Mager et al., 2018).

Previous works have shown that tokenization, data normalization, and cleaning, as well as high-quality corpora, are very important for developing systems for these languages, including machine translation (Vázquez et al., 2021; Attieh et al., 2024). However, it is not that common to find preprocessing tools readily available and easy to use.

Some Python libraries specialize in languages spoken in the Americas. For example, *Chana*[2] is a Natural Language Processing (NLP) toolkit for the Shipibo-Konibo language of Peru, offering tasks such as lemmatization, Named Entity Recognition (NER), and Part-of-Speech (POS) tagging. Another example is *nahuatl-tools*[3], a Python package that supports partial morphological analysis and orthographic normalization for at least one Nahuatl dialectal variant. Furthermore, Apertium (Forcada et al., 2011), an open-source tool for rule-based NLP tasks, provides repositories for several under-resourced languages of the Americas[4], including Guarani, Tzeltal, K'iche', Cusco Quechua, Apurimac Quechua, Nahuatl, Otomi and Huave. The morphological analyzers described in Section 4.3 are also published in Apertium.

Regarding commercial downstream applications, machine translation systems have recently begun supporting some Indigenous languages spoken in the Americas. Google Translate now includes varieties of Zapotec, Nahuatl, Quechua, Guarani, Aymara, Yucatec Maya, Q'eqchi', and Inuktut. Meanwhile, Bing Translator supports translation for a variant of Otomi and Yucatec Maya.

2.1 Digital adoption

The internet, and the digital technologies that underlie it, has become an essential tool for communication, with over 65 % of people in the Americas now using it, and the digital divide between the U.S. and Latin America shrinking rapidly (Martínez-Domínguez and Mora-Rivera, 2020).

While internet adoption has been slow in Mexico, particularly in rural areas, there are indications of a sharp increase in usage. Recent initiatives have pledged to increase the availability of fiber-optic internet access to all municipalities, and cellular service is expanding in rural communities. Given the high concentration of indigenous language speakers in these areas, this growth suggests that a significant and increasing number of indigenous language speakers are gaining access to the internet. These facts highlight the importance of prioritizing language technology research and applications for Mexican indigenous languages.

3 Languages supported by Py-Elotl

Mexico's linguistic landscape. Besides Spanish, the languages spoken in Mexico belong to 11 linguistic families and 68 language groups. All 68 of these groups hold the status of "national languages" alongside Spanish. Despite this linguistic diversity, education and mass media are predominantly in Spanish, which places significant pressure on indigenous languages. All of the indigenous languages of Mexico can be considered at risk of being lost (INALI, 2012b).

Speakers of each language group are immersed in distinct cultural contexts and particularities. However, they share similar conditions that represent a technological challenge for NLP, i.e., significant regional variations at many levels, including a lack of consensus in the orthographic conventions.

In its current version, Py-Elotl provides various functionalities for four language groups: Nahuatl, Otomi, Mixtec, and Huave. The first three are among the most widely spoken in the country; however, many of their varieties face varying degrees of endangerment (INALI, 2012a). Figure 2 illustrates their geographical distribution. Next, we introduce their characteristics and current status.

[2] https://pypi.org/project/chana/

[3] https://pypi.org/project/nahuatl-tools/

[4] https://github.com/apertium/apertium-languages

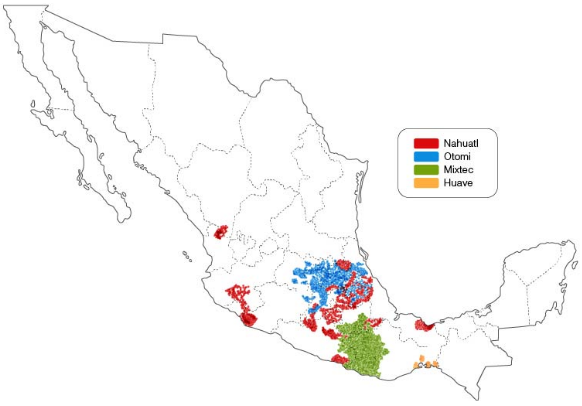


Figure 2: Geographical distribution of Nahuatl, Otomi, Mixtec, and Huave in Mexico.

Nahuatl is a group of languages present in several regions of Mexico (around 1.6 million speakers in total). It belongs to the Nahuan branch of the Uto-Aztecan (or Yuto-Nahua) linguistic family. This language family covers a vast territory; it distributes across the USA, Mexico, and El Salvador. Nahuatl is the Uto-Aztecan language with the most southern geographical distribution, and is spoken in 16 states of Mexico. Some sources recognize 30 dialectal variations (INALI, 2008), while others 28 (Lewis, 2009). Nahuatl has a rich concatenative morphology with polysynthetic and agglutinative tendencies. In particular, verbs can agglutinate many affixes to encode, for example, person and number of subject and objects, tense, aspect, directionality, and reverence.

Mixtec is a group of languages spoken in central and southern Mexico (~500,00 speakers). It belongs to the Mixtecan branch of the Oto-Manguean linguistic family. Mixtec is spoken in three states of Mexico: Oaxaca, Puebla and Guerrero.

This language exhibits the biggest dialectal variation in the country. According to INALI (2008), it has 81 dialectal variants, while Ethnologue (Lewis, 2009) recognizes 52. Due to this, Mixtec is sometimes considered as a 'macro-language'.

One of its main characteristics is the presence of tones. Most varieties distinguish three tones, while some even four (Méndez-Hord, 2017; Mendoza Ruiz, 2016; Palancar, 2016). Its morphology is usually considered isolating/analytic. However, it has the peculiarity that it actually marks many grammatical distinctions, but they are encoded at the suprasegmental level employing the tones.

Otomi is a group of languages spoken in central Mexico (around 300,000 speakers).[5] It belongs to the Oto-Pamean branch of the Oto-Manguean linguistic family (Barrientos López, 2004; Valiñas, 2020). Otomi is spoken in eight states of Mexico, including Guanajuato, Querétaro, Hidalgo, Puebla, Veracruz, Michoacán, Tlaxcala and Estado de México (Lastra, 2001).

INALI (2008) recognizes nine dialectal variants. Ethnologue (Lewis, 2009) recognizes the same number of variants; however, the reported variants are not exactly the same as noted by Valiñas (2020).

Otomi has rich morphophonological phenomena and an elaborated system of inflectional classes (Palancar, 2004). Phonologically, it features a complex vowel system with nine oral vowels and five nasal vowels, as well as a three-tone distinction (low, high, and ascending). Most of the observed orthographic variations occur within the vowel system.

Huave is a language spoken in the coastal region of Oaxaca, near the Isthmus of Tehuantepec. It is a language isolate classified within the Huavean language family and has approximately 37,000 speakers (Valiñas, 2020).

Sources differ on the number of dialectal variations, identifying between two and four distinct varieties. Typologically, Huave exhibits tonal phenomena, although tones are not as productive as in other tonal languages. It is an agglutinative language, where meaning is primarily conveyed through the combination of stems with prefixes and suffixes (Tyers and Castro, 2023).

4 Description of Py-Elotl

In this section we summarize the key components that are currently available in Py-Elotl.

4.1 Corpus loader

The toolkit includes a parallel corpus loader for three of the languages mentioned above. A parallel corpus consists of sentences in a source language paired with their corresponding translations in a target language. This kind of corpus is essential for developing translation technologies and conducting comparative linguistic studies.

In Py-Elotl, the parallel corpora always include Spanish as one of the languages, as it is relatively common to find translations to and from Spanish

[5] http://cuentame.inegi.org.mx/hipertexto/todas_lenguas.htm.

when accessing digital resources for Mexico’s indigenous languages.

This module enables users to load a given parallel corpus directly into a Python data structure, allowing for easy manipulation and analysis of parallel sentences. Additionally, each parallel sentence includes metadata about its source document and the dialectal variety in which it is written.

In all cases, the parallel corpora encompass various dialectal varieties, orthographic conventions, and sources. Below, we describe the characteristics of each corpus.

Spanish-Nahuatl. The data come from the *Axolotl* parallel corpus (Gutierrez-Vasques et al., 2016), one of the largest Spanish-Nahuatl parallel corpora that is also available through a web search interface [6]. It compiles texts from diverse sources, including short stories, history books, and recipe books, among others. These sources cover several dialectal variations, with Classical Nahuatl (nci) being the most common. Additionally, it includes Highland Puebla Nahuatl (azz), Morelos Nahuatl (nhm), Central Nahuatl (nhn), Western Huasteca Nahuatl (nhw), and Eastern Huasteca Nahuatl (nhe). It is important to note that some sources are currently classified as ”unknown” (unk) for various reasons, such as the combination of multiple dialects or difficulties in identification.

Spanish-Otomi. This parallel data comes from the *Tsu̲nkua* corpus[7], which consists primarily of translations from history books, dialogues, grammars, and educational materials. Currently, the corpus includes sources written in three dialectal variations: Hñähñu/Mezquital Otomi (ote), Otomi del Estado de México (ots), and Ixtenco Otomi (otz), with the first being the most prominent.

Spanish-Mixtec The parallel corpus for this language pair was built from educational sources, grammars, and short stories. Although relatively small, it encompasses a wide range of dialectal variations, including Chalcatongo Mixtec (mig), Magdalena Peñasco Mixtec (xtm), Ocotepec Mixtec (mie), Tezoatlán Mixtec (mxb), San Jerónimo Xayacatlan Mixtec (mit), Northern Tlaxiaco Mixtec (xtn). This corpus, named *kolo*, is also available through a web search interface[8].

For a more comprehensive overview of the size and distributions in the parallel corpora, see Table 1 and Figure 4. Additionally, see Figure 3 for an example of using this feature in the Python library.

Corpus	#Parallel sentences	Dialects (ISO-639-3)
Axolotl (Spanish-Nahuatl)	16K	nci, azz, nhm, nhn, nhw
Tsu̲nkua (Spanish-Otomi)	5K	ots, ote, otz
Kolo (Spanish-Mixtec)	2K	mig, xtm, mie, mxb, mit, xtn

Table 1: Parallel corpora currently available in Py-Elotl

```
import elotl.corpus

kolo = elotl.corpus.load("kolo")

for row in kolo:
    print(f"l1={row[0]}")
    print(f"l2={row[1]}")
    print(f"variant={row[2]}")
    print(f"doc={row[3]}")


# >> l1=cuajilote
# >> l2=chite
# >> variant=Mixteco de Magdalena Peñasco (xtm)
# >> doc=Algunos dichos y creencias
#       tradicionales de Magdalena Peñasco
```

Figure 3: Example of the parallel corpus loader in Py-Elotl

4.2 Orthographic normalization

Orthographic normalization is the process of converting written text into a standardized form within a language. While this is not a major issue for languages with well-established writing conventions, it poses a significant challenge for many other languages. Documents written in languages like Nahuatl and Otomi often have multiple orthographic tendencies in use, leading to spelling variation alongside dialectal differences.

Orthographic normalization in Py-Elotl has been implemented explicitly[9] for Nahuatl and Otomí. In both cases, finite-state transducers (FSTs) are used to convert a non-normalized input string, which may or may not conform to a particular orthographic standard, first to a phonemic representation, and subsequently to a user-specified orthographic norm for the language. Therefore, in all cases, this is a two-step process: mapping source text to a phonetic alphabet (IPA) and then generat-

[6] https://axolotl-corpus.mx/

[7] https://tsunkua.elotl.mx/

[8] https://kolo.elotl.mx/

[9] For Huave, there is no explicit orthographic normalization, but the morphological analyzer supports some orthographic flexibility in its input.

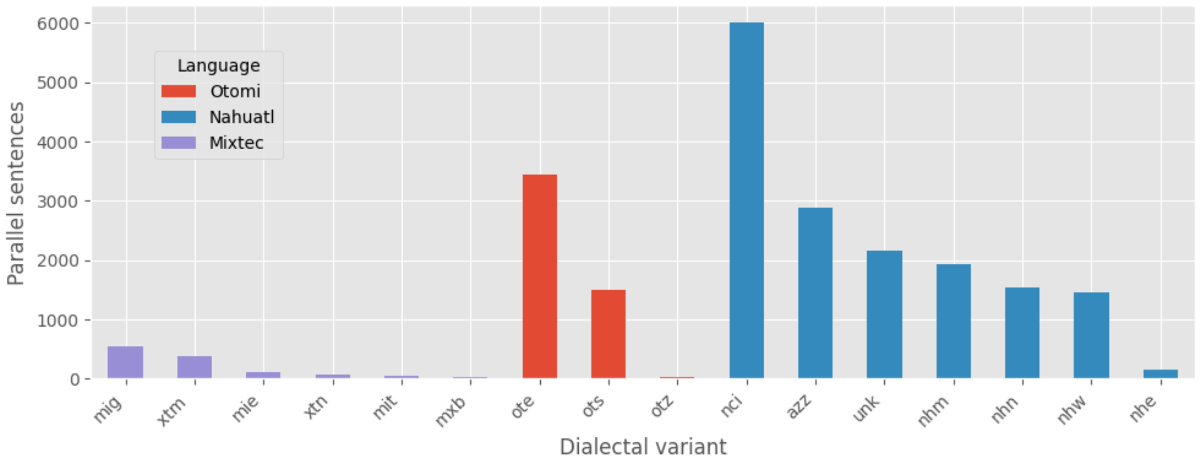


Figure 4: Distribution of dialects for each of the three parallel corpora available in Py-Elotl.

ing the target orthography.

For **Nahuatl**, the input is expected to either follow some combination of the existing writing norms (e.g. k, c or qu for the phoneme /k/), or common patterns observed in Nahuatl writing (e.g. grapheme y for phoneme /i/ word-finally). For the normalized output, four orthographic norms are currently supported, including the orthography often employed by the National Institute of Indigenous Languages (INALI) (INALI, 2018) and the ACK orthography commonly used by academics to write colonial-era Nahuatl (Karttunen, 1992). It is important to note that, given the large amount of linguistic variation within Nahuatl's many variants, a true "phonemic" representation is not possible. Instead, we opt for a generic, approximate phonemic transcription that serves as input for the various output orthographies.

Similarly, in the case of **Otomi**, the system takes input text in a potential source orthographic norm and should be able to convert it to any target orthographic norm. Currently, this module supports four orthographic standards. The transduction rules were informed by a linguist's expertise and existing documentation (Hernández-Green, 2016). Mezquital Otomi (Hñahñu) is the most widely spoken variant and forms the basis of the writing standard proposed by (INALI, 2014). See Figure 5 for an example of using the normalizer in the Python library.

Adding other output orthographies is relatively straightforward, and requires creating an FST that maps the phonemic representation to the norm of interest, and committing the FST in .att format. See Table 3 and Table 2 for a more detailed description of the available norms.

To get a sense of the **performance** of the Nahuatl orthographic normalizer, we used the Universal Dependencies treebank for Western Sierra Puebla Nahuatl, which for each token includes the original orthography and a version written in the INALI norm. We manually converted the normalized forms to the other three output orthographies, deduplicating the original forms[10], and excluding punctuation, Spanish words, and named entities. We then compared the manually-normalized words to the output of the Py-Elotl normalizer given the treebank's original forms. The normalizer correctly normalizes 98% of the 2,142 unique words for all four of the output orthographies.

In the case of Otomi, as a preliminary evaluation, we collected 1,282 word types written in the OTQ and the OTS norm, respectively. Using Py-Elotl, we converted them to the INALI norm: OTS→INALI , OTQ→INALI . We then compared the results to a gold standard, finding that the normalizers correctly processed 81% of word types on average. Performance was affected by code-switching and ambiguity in the dataset, as the current rule set does not yet cover these phenomena.

Given that the presence of named-entities and/or code-switching may mean that certain words should not be normalized or should undergo a different process for normalization. As a first step to support this potential complexity, the orthographic normalizers in Py-Elotl offer the option to provide an exceptions list in the form of a dictionary that maps a

[10] We deduplicate the original forms in order to avoid inflated performance due to the frequent repetition of easy-to-normalize common words such as the determiner/subordinator *in* or the antecessive clitic *o*.

Norm	Description
INALI	Norm used by the National Institute of Indigenous Laguages of Mexico
Ref.	INALI (2018); Flores Nájera (2019),
SEP	Norm used previously by the Secretary of Public Education and for Indigenous Education
Ref.	Various
ACK	Orthography popularized by Nahuatl scholars J. Richard Andrews, Joseph Campbel, and Frances Karttunen
Ref.	Andrews (1975); Karttunen (1992)
ILV	Norm developed by the community of San Miguel Tenango (Western Sierra Puebla Nahuatl) in collaboration of the Summer Institute of Linguistics.
Ref.	Márquez Hernández and Schroeder (2005)
Norm	**Example sentence**
INALI	[...]ihkwak walas mitsitas
SEP	[...]ijkuak ualas mitsitas
ACK	[...]ihcuac hualaz mitzitaz
ILV	[...]ihcuac ualas mitzitas
Phones	[...]iʔk^{w}ak walas mitsitas

Table 2: A description of currently-supported orthographic norms for Nahuatl.

set of words to their preferred normalizations. One possible application of this functionality is to pass a list of common Spanish words so that they maintain the Spanish orthography.

4.3 Finite-state morphological analyzers

The use of finite-state transducers for morphological analysis has a long and rich history in the field of NLP (Kornai, 1996; Beesley and Karttunen, 2003), and is a particularly good option when there is little annotated data with which to train data-driven approaches such as deep neural networks, and can even be useful as a means for generating training data for such approaches (Moeller et al., 2018).

Py-Elotl aggregates free and open-source finite-state transducer morphological analyzers, and currently supports five indigenous Mexican languages: Three variants of Nahuatl (Classical Nahuatl[11], the analyzer for which comes from Tyers et al. (2023) (which in turn leverages the extensive lexicon in Escobar Farfan and Jonathan Irvine Israel (2019)), Highland Puebla Nahuatl (Tyers and Pugh, 2023), and Western Sierra Puebla Nahuatl (Pugh et al., 2021)), San Mateo del Mar Huave (Tyers and Castro, 2023), and Otomí[12].

```
from elotl.otomi.orthography import Normalizer

sentence = "Hindí tsi ra chuni"
# Available norms: ["inali", "ots", "otq", "ref"]
ots_normalizer = Normalizer("ots")
otq_normalizer = Normalizer("otq")

print(f"OTS: {ots_normalizer.normalize(sentence)}")
print(f"OTQ: {otq_normalizer.normalize(sentence)}")

# >> OTS: jindí tsi ra chuni
# >> OTQ: hindí tsi ra txuni
```

Figure 5: Example of orthographic normalization using Py-Elotl. This functionality is currently available for Otomí and Nahuatl. Not featured in the figure is the .to_phones method that return the intermediate, phonemic representation.

Norm	Description
INALI	Norm designed by the National Institute of Indigenous Laguages of Mexico
Ref.	(Inali, 2014)
OTS	Standard used in some texts from variants in the State of Mexico
Ref.	(De la Vega, 2017)
OTQ	Standard proposed mainly for Querétaro variants
Ref.	(Hekking and de Jesús, 1989)
RFE	A phonetic alphabet developed for Spanish. Some Otomi transcriptions follow this standard.
Ref.	(Lastra, 1997)
Norm	**Example sentence**
INALI	[...]bijúgígó escuela pero ndichichithóhó
OTQ	[...]bijúgígó escuela pero nditxitxithóhó
OTS	[...]bikjúgígó escuela pero ndichichitjójó
RFE	[...]bikhúgígó escuela pero ndičičithóhó
Phones	[...]bikhúgígó eskwéla peɾo ndɨt͡ʃit͡ʃithóhó

Table 3: A description of currently-supported orthographic norms for Otomi.

While it is by no means a requirement, currently all of the Py-Elotl morphological analyzers are part of the Apertium project, and are regularly updated to reflect recent changes. The package supports stand-alone FST morphological analyzers as .att files. Since the aggregation of analyzers may result in differing tagsets, we unify tagsets via a rule-based mapping of each analyzer's output to the universal part-of-speech tags and universal morphological features used in the Universal Dependencies project (Nivre et al., 2020).

[11] "Classical Nahuatl" is the name commonly used for the historical literary variety of Nahuatl spoken in central México during the early colonial period.

[12] https://github.com/apertium/apertium-ote

Language group	Parallel Corpus	Orthographic Normalizer	Morphological Analyzer
Nahuatl	🌽	🌽	🌽
Huave	-	-	🌽
Otomí	🌽	🌽	(🌽)
Mixtec	🌽	-	-

Table 4: An overview of the different NLP resources and tools available for each language supported in Py-Elotl. The parentheses around the elote emoji for the Otomí morphological analyzer is used to indicate the "prototype" status of the system, since the coverage and performance of this analyzer has not been published.

5 Free Software

Py-Elotl is freely available as a Python package, allowing users to integrate it into their workflow. Additionally, they can collaborate and contribute through open repositories. As a Free Software tool, it grants users and communities the freedom to run, copy, distribute, study, modify, and improve it. The source code and builds are publicly accessible on GitHub. Table 4 shows an overview of the current functionalities supported.

Releasing source code, models, and data is considered good practice in areas like NLP to ensure reproducibility. Some argue that this is especially important when working with endangered languages due to the ethical implications, i.e., the risk of doing cultural or linguistic appropriation of vulnerable groups (Hämäläinen, 2021; Washington et al., 2021).

Along a similar line, Aguilar Gil (2020) reflects on how the practices of cooperation that the indigenous communities have carried out as means of survival could influence the development of technologies. It should not be a matter of vulnerable groups receiving technology passively but encouraging an intercultural dialogue in how we do technology. She coins the term ”tequiologias” compatible with the free and open-source software philosophy.

6 Conclusions

We introduced a suite of tools and resources focused on facilitating text processing for various under-resourced languages spoken in Mexico. This toolkit integrates: a) three parallel corpora with representation of different dialectal variations within the language groups; b) Orthographic normalization tools where we took on the task of identifying the main orthographic tendencies and wrote FST technology to convert across different standards automatically; c) Morphological analyzers for several dialectal variants that are also available through Apertium.

Currently, we support the following language groups: Nahuatl, Otomi, Mixtec, and Huave. However, adding resources and features for more languages is relatively straightforward. The toolkit is available as a Python package, and the code is openly accessible in public repositories to encourage the development of open and collaborative technologies.

The current scope of Py-Elotl focuses on upstream NLP tasks, including rule-based approaches, as neural and statistical methods are often not entirely applicable. To foster a more linguistically diverse landscape in language technologies and support under-resourced languages, we believe it is essential to first establish strong foundational resources.

Limitations

While we present this Python toolkit as a resource for the languages of Mexico, our coverage is not exhaustive. We currently focus on a few language groups, some with large speaker populations within Mexico. However, many other indigenous languages and dialectal variations remain underrepresented in this release. Expanding coverage to include a broader range of languages and dialects is an important goal for future development, requiring further linguistic collaboration, data collection, and community involvement

The morphological analyzers and orthographic normalization modules used in this work are rule-based, which may limit their flexibility to handle phenomena such as code-switching, ambiguous cases, and non-standard language use, which constitute the linguistic reality many speakers of these languages face.

Finally, the development of technology for underrepresented groups should not only focus on applying the latest NLP techniques but also encouraging diverse groups of work, in a way that the resulting technologies and resources are really aligned with the necessities and context of the speakers.

Acknowledgments

This work was partially funded by the projects PAPIIT TA100924 and TA100725 at UNAM, Mexico, and by the National Science Foundation Proposal #23192462319247, "Collaborative Research: Syntactically-annotated corpora for endangered languages in areal contact." We thank the reviewers for the relevant comments on our work.

References

Yasnaya Elena Aguilar Gil. 2020. A modest proposal to save the world. https://restofworld.org/2020/saving-the-world-through-tequiology/.

J.R. Andrews. 1975. *Introduction to Classical Nahuatl*, 2nd edition. University of Texas Press.

Joseph Attieh, Zachary Hopton, Yves Scherrer, and Tanja Samardžić. 2024. System description of the NordicsAlps submission to the AmericasNLP 2024 machine translation shared task. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 150–158, Mexico City, Mexico. Association for Computational Linguistics.

Guadalupe Barrientos López. 2004. *Otomíes del Estado de México*. Comisión Nacional para el Desarrollo de los Pueblos Indígenas.

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Lázaro Margarita De la Vega. 2017. Aprendiendo otomí (hñähñu). *Ciudad de México, Comisión Nacional para el Desarrollo de los Pueblos Indígenas*.

Fanny Ducel, Karën Fort, Gaël Lejeune, and Yves Lepage. 2022. Do we name the languages we study? the# benderrule in lrec and acl articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 564–573.

Escobar Farfan and Jonathan Irvine Israel. 2019. *Nahuatl contemporary writing : studying convergence in the absence of a written norm*. Ph.D. thesis, University of Sheffield.

Lucero Flores Nájera. 2019. *La gramática de la cláusula simple en el náhuatl de Tlaxcala*. Ph.D. thesis, Centro de Investigactiones y Estudios Superiores en Antropología Social.

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*, 25:127–144.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! *CoRR*, abs/2103.09567.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Ewald Hekking and Severiano Andrés de Jesús. 1989. *Diccionario español-otomí de Santiago Mexquititlán*, volume 22. Universidad Autónoma de Querétaro.

Nestor Hernández-Green. 2016. Misteriosas figurillas de barro de san jerónimo acazulco. *Tlalocan*, 21:19–48.

INALI. 2008. Catálogo de las lenguas indígenas nacionales: Variantes linguísticas de méxico con sus atodenominaciones y referencias geoestadísticas. https://www.inali.gob.mx/clin-inali/.

INALI. 2012a. Catálogo de las lenguas indígenas nacionales en riesgo de desaparición. https://www.cdi.gob.mx/dmdocuments/lenguas_indigenas_nacionales_en_riesgo_de_desaparicion_inali.pdf/.

INALI. 2012b. *México: Lenguas indígenas nacionales en riesgo de desaparición*. Instituto Nacional de Lenguas Indígenas, México.

INALI. 2014. *Njaua Nt't'ot'i ra Hñäñhu Norma de escritura de la lengua Hñähñu (Otomí)*. INALI, SEP.

Inali. 2014. *Njaua nt'ot'i ra hñähñu. Norma de escritura de la lengua hñähñu (otomí) de los estados de Guanajuato, Hidalgo, Estado de México, Puebla, Querétaro, Tlaxcala, Michoacán y Veracruz.* Instituto Nacional de Lenguas Indígenas (inaLi), SEP, Mexico.

INALI. 2018. Breviario: Norma ortográfica del idioma náhuatl, méxico. (conforme al avance preliminar de la norma de escritura de la lengua náhuatl a nivel nacional).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Frances E Karttunen. 1992. *An analytical dictionary of Nahuatl*. University of Oklahoma Press.

András Kornai. 1996. Extended finite state models of language. *Natural Language Engineering*, 2(4):287–290.

Yolanda Lastra. 1997. *El otomí de Ixtenco*. UNAM.

Yolanda Lastra. 2001. *Unidad y Diversidad de la Lengua: Relatos otomíes*. UNAM.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, TX, USA.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 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, USA. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021a. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. 2021b. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics, Online.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Marlen Martínez-Domínguez and Jorge Mora-Rivera. 2020. Internet adoption and usage patterns in rural Mexico. *Technology in society*, 60:101–226.

Esteban I Méndez-Hord. 2017. *Tone in Acatlán Mixtec Nouns*. The University of North Dakota.

Juana Mendoza Ruiz. 2016. Fonología segmental y patrones tonales del tu'un savi de alcozauca de guerrero. *Ciudad de México: Centro de Investigaciones y Estudios Superiores en Antropología*.

Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A neural morphological analyzer for Arapaho verbs learned from a finite state transducer. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 12–20.

Christopher Moseley and Alexander Nicolas, editors. 2010. *Atlas of the World's Languages in Danger*, 3 edition. UNESCO, Paris.

Elizabeth Márquez Hernández and Petra Schroeder. 2005. *Pequeño diccionario ilustrado*, Second edition. Instituto Lingüístico de Verano, A.C., Mexico.

J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.

Enrique L Palancar. 2004. Verbal morphology and prosody in otomi. *International journal of American linguistics*, 70(3):251–278.

Enrique L Palancar. 2016. A typology of tone and inflection: A view from the oto-manguean languages of mexico. In *Tone and Inflection*, pages 109–140. De Gruyter Mouton.

Robert Pugh, Francis Tyers, and Marivel Huerta Mendez. 2021. Towards an open source finite-state morphological analyzer for zacatlán-ahuacatlán-tepetzintla nahuatl. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 80–85.

Francis Tyers and Samuel Herrera Castro. 2023. Towards a finite-state morphological analyser for san mateo huave. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 30–37.

Francis Tyers and Robert Pugh. 2023. A finite-state morphological analyser for highland puebla nahuatl. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 103–108.

Francis Tyers, Robert Pugh, and Valery Berthoud. 2023. Codex to corpus: Exploring annotation and processing for an open and extensible machine-readable edition of the Florentine Codex. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 19–29.

Leopoldo Valiñas. 2020. *Lenguas originarias y pueblos indígenas de México: familias y lenguas aisladas*. Academia Mexicana de la Lengua.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Jonathan Washington, Felipe Lopez, and Brook Lillehaugen. 2021. Towards a morphological transducer and orthography converter for western tlacolula valley zapotec. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 185–193.