A finite-state morphological analyser for Highland Puebla Nahuatl

Francis M. Tyers
Department of Linguistics
Indiana University
Bloomington, IN 47401
ftyers@iu.edu

Robert Pugh
Department of Linguistics
Indiana University
Bloomington, IN 47401
pughrob@iu.edu

Abstract

This paper describes the development of a free/open-source finite-state morphological transducer for Highland Puebla Nahuatl, a Uto-Aztecan language spoken in the state of Puebla in Mexico.[1] The finite-state toolkit used for the work is the Helsinki Finite-State Toolkit (HFST); we use the `lexc` formalism for modelling the morphotactics and `twol` formalism for modelling morphophonological alternations. An evaluation is presented which shows that the transducer has a reasonable coverage—around 90%—on freely-available corpora of the language, and high precision—over 95%—on a manually verified test set.

1 Introduction

This paper describes a new morphological analyser for Highland Puebla Nahuatl, an endangered language spoken in the state of Puebla in Mexico (see Figure 1[2]). The analyser is based on finite-state technology, which means that it can be used for both the analysis and the generation of forms — a finite-state morphological transducer maps between surface forms and lexical forms (lemmas and morphosyntactic tags).

An analyser of this sort has a wide variety of uses, including for automating the process of corpus annotation for linguistic research as well as for creating proofing tools (such as spellcheckers) and for lemmatising for electronic dictionary lookup for language learners — in a language with heavy prefixing and suffixing morphology, determining the stem is not a simple matter.

Our approach is based on the Helsinki Finite-State Toolkit (HFST, Lindén et al. (2011)).

2 Prior art

Finite state transducers (FST) for modeling morphology has a long history within the field of computational linguistics (Kornai, 1996; Beesley and Karttunen, 2003).

Work on morphological analysers for Nahuatl languages includes an effort, inspired by literate programming, to use the code for the transducer as a descriptive grammar of a Nahuatl variety spoken in the state of Guerrero (Maxwell, 2015), and morphological analysers specifically targeting colonial-era Nahuatl, either for the exploration of colonial texts (Thouvenot, 2009), or as a means to evaluate similarity between written Nahuatl varieties (Farfan, 2019). One drawback of these projects is that they are not to our knowledge freely-available or easily-accessible.

Nicolai et al. (2020) describe the development of morphological analysers and generators for more than one thousand languages using the Johns Hopkins University Bible Corpus (McCarthy et al., 2020), including some variants of Nahuatl (however, not Highland Puebla Nahuatl).

Pugh et al. (2021) presents the first open-source morphological analyser for the Western Sierra Puebla Nahuatl variant group. Tona et al. (2023) expand on that system, extending it to support Huasteca Nahuatl. This latter work, however, has not been released.

3 Highland Puebla Nahuatl

Nahuatl (or Nahuat, Nahual) is a polysynthetic, agglutinating Uto-Aztecan language continuum spoken throughout Mexico and Mesoamerica. The Mexican Government's *Instituto Nacional de Lenguas Indígenas* (INALI) recognizes 30 distinct variants (INALI, 2009).

Highland Puebla Nahuatl, (or *Sierra Puebla Nahuatl*, also referred to by INALI as *Náhuatl del noreste central*, `ISO-639-3` *azz*) is a Nahuatl vari-

[1] https://github.com/apertium/apertium-azz

[2] Figure 1 is based on work by users TUBS (https://commons.wikimedia.org/wiki/File:Puebla_in_Mexico_(location_map_scheme).svg) and Battroid (https://commons.wikimedia.org/wiki/File:Mexico_Puebla_Puebla_location_map.svg)

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 103–108
July 14, 2023 ©2023 Association for Computational Linguistics

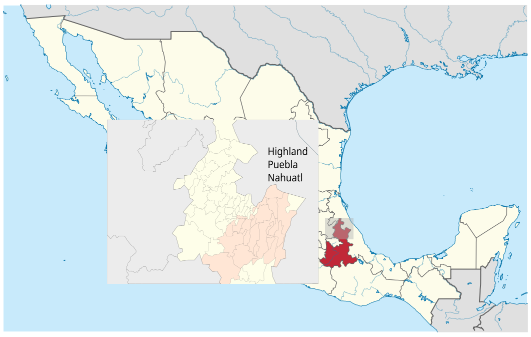

Figure 1: A map highlighting where Highland Puebla Nahuatl (salmon colour) is spoken in Mexico.

ant group spoken in the Northeastern Sierra region of the state of Puebla, Mexico, mainly in the municipalities of Tetela de Ocampo, Zacapoaxtla, and Cuetzalan. According to Ethnologue's 2007 estimate, it is spoken by an estimated 70,000 speakers.

This particular Nahuatl variant has been the subject of a number of descriptive works (Key, 1960; Robinson, 1970; Key and Key, 1953) and dictionaries (Key and Richie de Key, 1953; Cortez Ocotlán, 2017).

4 Data

The source data used to develop the FST comes from three sources: (1) A dataset of transcribed recordings of interviews and conversations, mainly about plants (Amith et al.), (2) a subset of texts in the *azz* variant from the multi-variant parallel corpus Axolotl (Gutierrez-Vasques et al., 2016), and (3) technical publications by the Sociedad Mexicana de Física[3], which consist of translations of various scientific texts. The breakdown of volume for each of these sources is presented in Table 1.

5 Orthography

Writing practices in Nahuatl vary and are characterized by multiple competing views (de la Cruz Cruz, 2014). The most well-known and widely-disseminated orthographic standards for Nahuatl are ACK, a colonial-inspired orthography named after scholars Anderson, Campbell, and Karttunen, who popularized it in their work, the standard from the *Instituto Nacional de Lenguas Indígenas* (INALI) (INALI, 2018), and that used by the Secretaría de Educación Pública (SEP). In practice, Nahuatl writing contains a great deal of orthographic variation, often even within the writing of a single author.

The orthography used for building the analyser follows what was taught in the Nahuatl course for adult learners given in the municipality of Tetela de Ocampo, Puebla in the summer of 2022 (TO). This broadly follows the SEP, but with the addition of the letter *h* which is used before *u* for /w/ after vowels or at the beginning of words. For example SEP *uezueyi*, TO *huehueyi* 'big', SEP *mochiua*, TO *mochihua* "it is made".

We maintain a separate finite-state transducer to account for orthographic and spelling variation. This includes rules for orthographic changes like *ts* (SEP, INALI) → *tz* (ACK) (e.g. *tejuatsin* 'you-HON' → *tehhuatzin*), spelling changes, such as *w$* → *j$* and abbreviations that are found in the transcriptions from the spoken corpora, such as *^t'* → *^tik*.

6 Methodology

In this section, we outline some of the implementation details of the analyzer, including a description of relevant linguistic features.

6.1 Lexicon

The lexicon consists of around 5,000 lexemes which were added in frequency order (calculated using the corpora described in §4) and with reference to the two available dictionaries (Key and Richie de Key, 1953; Cortez Ocotlán, 2017) for part-of-speech classification. The lexicon was created in the `lexc` formalism, which is standard in HFST.

Closed categories (pronouns, conjunctions, etc.) were added manually based on class notes and on existing grammatical descriptions (Key, 1960; Robinson, 1970; Cortez Ocotlán, 2017).

6.2 Tagset

The tagset is based on the tagset of the Apertium project (Forcada et al., 2011), each tag is encased in greater than '<' and less than '>' symbols. The tag names are mnemonic, some of them coming from other analysers in the Apertium project and being based on English, Spanish, or Catalan terms, and some are based on Nahuatl terms. We include a conversion from this Apertium-based tagset to one based on Universal Dependencies (Nivre et al., 2020).

[3] https://site.inali.gob.mx/SMF/Libros2.0/nhtl/index.html

Corpus	Genre	Ortho.	Tokens	Types	Coverage	
					Tokens	*Types*
Puebla-Nahuatl (Amith et al.)	spoken	INALI	353,006	23,174	93.02	44.33
Axolotl (Gutierrez-Vasques et al., 2016)	non-fiction	SEP	18,338	3,492	84.78	48.0
Sociedad Mexicana de Física	non-fiction	SEP	1,649	599	92.05	84.47

Table 1: A breakdown of the three data sources used for developing the analyser, with information about the genre (following Müller-Eberstein et al. (2021), orthography used, data volume, and analyser coverage. Note that the Puebla-Nahuatl dataset's orthography differs slightly from the INALI norms in that it explicitly represents vowel-length with the colon ':'.

Category	Stems	Category	Stems
Verbs	2,937	Other	116
Nouns	1,284	Numerals	42
Adverbs	222	Pronouns	33
Adjectives	202	Conjunctions	27
Proper nouns	160	Determiners	20
Total:			5,043

Table 2: Composition of the stem lexicon in the `lexc` file.

6.3 Morphotactics

The morphotactics of Highland Sierra Nahuatl is very similar to that of other Nahuatl varieties. It is characterised by a concatenative affixing morphology with a large number of inflectional and derivational morphemes. It also features long-distance dependencies between prefixes and suffixes.

6.3.1 Nouns

Nouns inflect for number and possession. They also have very productive derived forms, such as the reverential *-tsin* (1) and less productive derivations, such as *-k(o)* for locative, and can appear as predicates with the addition of subject prefixes. We implement the morphotactics for inflection and for the most frequent subset of the derived forms. Nouns are therefore split into separate continuation classes for their different combinatorial possibilities.

(1) *kikouaj* *in*
ki-koua-j in
O.SG3-buy-S.PL the
tokniuantsitsin
to-kni-uan-tsi~tsin
POSS.PL1-person-PL-PL.HON
"People buy it." (lit. "Our brethren buy it")

In (1), the noun *(i)kni* 'sibling' appears with the first person plural possessive prefix *to-*, the possessed plural marker *-uan*, and the reverential marker *tsi~tsin*, where plurality is further marked with partial reduplication of the *-tsin* morpheme.

Relational nouns: There is also a subcategory of nouns, called "relational nouns," used for expressing spatial and temporal relations, as well as other non-core semantic roles. Unlike common nouns, these nouns have obligatory possession.

(2) *In mochiua* *kuoujtaj,* *in eua*
In mo-chiua kuoujtaj, in eua
O.REFL-make mountains, born
talixko, *amo itech*
tal-ix-ko, amo i-tech
ground-RELN-LOC, NEG POSS.SG3-on
kuapalak.
kuapalak.
tree.trunks
"It grows in the mountains, it comes up from the ground, it doesn't grow in tree trunks."

In (2) we see two methods in which relational nouns can be used. The first is *talixko* where the the relational noun *-ixko* 'in front of / on the surface of' is compounded with the noun *tali* 'ground/earth'. This relational noun itself is composed of *ix* 'face' and *ko* a locative morpheme.

The second method is using a free-standing relational noun with a complement, *itech kuapalak* 'in rotten tree trunks', is composed of a possessive form of the relational noun *-tech* 'on' and the noun compliment *kuapalak* 'tree trunk'.

These relational nouns can also appear separated from their complement, as in (3, where the complement of *iuan* 'with' is *emol* 'beans', but it appears to the right of the verbal complex *se kikua* "it is eaten".

(3) *uan iuan* *se* *kikua* *emol*
uan i-uan se ki-kua emol
and POSS.SG3-with one O.SG3-eat beans
"... and it is eaten with beans"

They can also receive reverential morphology as in one of the typical ways of expressing goodbye, *mohuantsin* ‘with you’ (4).

(4) *mohuantsin*
mo-huan-tsin
POSS.2SG-with-HON
“with you”

Locatives: In addition to compounding with relational nouns there is also a locative derivational suffix *-k(o)* which forms locative nouns from places. For example *ima* ‘her hand’, *imako* ‘in her hands’.[4]

6.3.2 Verbs

Verbs inflect for number and person of subject and object(s), and for tense, aspect and mood. They also can be compounded with auxiliary verbs and can have incorporated adverbial items for both direction of movement and for manner of action. Additionally there is reverential agreement for the second person.

(5) *Xe ma nimitsonchiya huan*
Xe ma ni-mits-on-chiya huan
QST OPT S.SG1-O.SG2-HON-wait and
tisentakuaskej?
ti-sen-ta-kua-s-kej
S.PL1-TOGETHER-O.NN3-eat-FUT-S.PL
“Shall I wait for you and we’ll eat together?”

In (5) we see examples of incorporated adverbials, *ti**sen**takuaskej* “we will eat **together**”, affixal agreement, *ti*-[...]-*kej* for the first person plural subject and *ta-* for the indefinite object and the future tense suffix *-s*. The verb *nimitsonchiya* has the *on-* prefix, indicating reverentiality towards the addressee.

(6) *se mokouilia komo se*
se mo-kou-ilia komo se
one O.REF-buy-APP if one
kikuasneki.
ki-kua-s-neki
O.SG3-eat-FUT-want
“One goes and buys it if one wants to eat it.”

```
^Ixua/<s_sg3>ixua<v><iv><pres>$
^uan/huan<cnjcoo>$
^moskaltia/<s_sg3>moskaltia<v><iv><pres>$
^,/,<cm>$
^ijuak/ijhuak<cnjsub>$
^motamiti/<s_sg3>motami<v><iv><and>$
^peua/<s_sg3>pehua<v><iv><pres>$
^xochiyoua/<s_sg3>xochiyohua<v><iv><pres>$
^./.<sent>$
```

Figure 2: Example output of the analyser for the sentence *Ixua uan moskaltia, ijuak motamiti peua xochiyoua* “It sprouts, grows and later starts to flower”.

(7) *se kiualkui*
se ki-ual-kui
one O.3SG-VEN-bring
“It is brought.” (lit. One brings it (here))

6.4 Morphophonology

Phonological processes are implemented via `twol` rules. There are relatively few of these, and they include degemination (/kk/ →[k]) and nasal assimilation (/n/ →[m] // m).

7 Results

To evaluate the analyzer, we calculate the naïve coverage for both tokens and types. The naïve coverage is reported for each data source in Table 1. Naïve coverage is the percentage of surface forms in a given corpus that receive at least one morphological analysis. Forms counted by this measure may have other analyses which are not delivered by the transducer.

7.1 Evaluation

Since we don’t have a large, annotated dataset for evaluation, we performed a manual inspection of two random samples of data to get a sense of the the system’s precision and to understand the reasons for any missed words.

First, we sampled 100 random analyses from the corpora and identified any mistakes. The precision on this sample was 95%. Next, in order to find out where the most work remains to be done with respect to coverage, we randomly sampled 100 types that are currently not recognised by the system. These words were categorised by part of speech, and in addition we marked each with one or more of the following seven error categories: (1) missing morphotactics, (2) missing orthographic normalisation, (3) missing compound word, (4) reduplication, (5) loan word / code-switching, (6) tokenisation error, and (7) missing lexicon entry.

[4] Although the name is the same, these locatives are unlike those found in other languages as inflection because: (1) not every word can take a locative suffix, (2) they are not selected for by argument structure, (3) the resulting meaning can be idiosyncratic. For this reason we categorise them as derivation as opposed to inflection.

Over half of all unknown words were verb forms. Of these, five were caused by missing orthographic normalisation rules, for example *t'titipitstoti* is an abbreviated form of *tiktitipitstoti* 'you will be blowing the fire', and 10 were due to missing stems in the lexicon.

Around ten percent of the sampled unknown words were caused by errors in tokenisation. The speech corpus contains false starts, for example *amo nike..., amo nikmati* "I don't kn..., I don't know", and these do not currently receive any analysis.

8 Concluding remarks

We have described a robust finite-state morphological analyser for Highland Puebla Nahuatl. This work contributes to the recent increased focus in language technologies for Nahuatl, and may play an important role in supporting further Nahuatl language technology in the future.

In future work we would like to expand the lexicon to include more stems, to increase the coverage of all of the corpora, and to obtain new corpora for testing. We intend to include support for compounding and incorporation and for weighting the transducer. We already have 10,000 tokens manually disambiguated and will use these to weight more probable analyses.

Acknowledgements

We would like to thank Patricia Aguilar Romero, don Pedro Rivera, and Mitsuya Sasaki for their help with the work described in this manuscript. In addition we would like to thank the anonymous reviewers for their helpful comments.

References

Jonathan D. Amith, Amelia Dominguez Alcántara, Hermelindo Salazar Osollo, Ceferino Salgado Castañeda, and Eleuterio Gorostiza Salazar. Audio corpus of Sierra Nororiental and Sierra Norte de Puebla Nahuat(l) with accompanying time-code transcriptions in ELAN.

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

Pedro Cortez Ocotlán. 2017. *Diccionario Nahuat–Español de la Sierra Nororiental del Estado de Puebla*. Tetsijtsilin, Tzinacapan, Cuetzalan.

Victoriano de la Cruz Cruz. 2014. La escritura náhuatl y los procesos de su revitalización. *Contribution in New World Archaeology*, 7:187–197.

J.I.E. Farfan. 2019. *Nahuatl Contemporary Writing: Studying Convergence in the Absence of a Written Norm*. University of Sheffield.

Mikel L. Forcada, María 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.

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.

INALI. 2009. *Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*. Instituto Nacional de Lenguas Indígenas, México, D.F.

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).

Harold Key. 1960. Stem construction and affixation of Sierra Nahuat verbs. *International Journal of American Linguistics*, 28(2):130–145.

Harold Key and Mary Richie de Key. 1953. *Vocabulario Mejicano de la Sierra de Zacapoaxtla, Puebla*. Instituto Lingüístico de Verano, México, D.F.

Mary Key and Harold Key. 1953. The phonemes of sierra nahuat. *International Journal of American Linguistics*, 19(1):53–56.

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

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi Pirinen, and Miikka Silfverberg. 2011. HFST—framework for compiling and applying morphologies. *Communications in Computer and Information Science*, 100:67–85.

Michael Maxwell. 2015. Grammar debugging. In *Systems and Frameworks for Computational Morphology: Fourth International Workshop, SFCM 2015, Stuttgart, Germany, September 17-18, 2015. Proceedings 4*, pages 166–183. Springer.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu, and David Yarowsky. 2020. Fine-grained morphosyntactic analysis and generation tools for more than one thousand languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3963–3972, Marseille, France. European Language Resources Association.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Samppo Pyysalo, Sebastian Schuster, Francis Tyers, and Dan 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.

Robert Pugh, Marivel Huerta Mendez, and Francis M. Tyers. 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*.

Dow F. Robinson. 1970. *Aztec studies 2: Sierra Nahuat word structure*. Summer Institute of Linguistics.

Marc Thouvenot. 2009. *CEN juntamente : compendio enciclopédico del Náhuatl*. Instituto Nacional de Antropología e Historia, México, D.F.

Ana Tona, Guillaume Thomas, and Ewan Dunbar. 2023. A morphological analyzer for Huasteca Nahuatl. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 112–116.