Towards an Open Source Finite-State Morphological Analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl

Robert Pugh
Independent
robertpugh408@gmail.com

Francis Tyers
Indiana University
ftyers@iu.edu

Marivel Huerta Mendez
Independent
marivelhm1@gmail.com

Abstract

In this paper, we describe an in-progress, free and open-source Finite-State Transducer morphological analyzer for an understudied Nahuatl variant[1]. We discuss our general approach, some of the technical implementation details, the challenges that accompany building such a system for a low-resource language variant, the current status and performance of the system, and directions for future work.

1 Introduction

Mexico is home to more than 68 indigenous languages, virtually all of which are considered endangered. Nahuatl, which served as the lingua franca of Mesoamerica in the centuries prior to the Spanish invasion and the primary indigenous language used in Mexico for long afterward, is today the most widely spoken indigenous language of the region at 1.5 million speakers. Despite this, the significant reduction of intergenerational transfer as a result of socioeconomic pressures and marginalization of indigenous communities has resulted in a precarious outlook for the future (Olko and Sullivan, 2014).

Computational tools for endangered languages can play a useful role in maintenance and revitalization efforts (Reyhner, 1999; Ben Slimane, 2008). Nahuatl, like many other languages of the Americas, however, has received relatively little attention in the fields of Computational Linguistics and Natural Language Processing (Mager et al., 2018).

Automatic morphological analysis is fundamental to many language technology applications. A few example applications of a robust morphological analyzer for polysynthetic languages include intelligent dictionaries, spelling correction, and rule-based machine translation.

1.1 Related Work

The Finite-State Transducer (FST) is a powerful formalism for modeling natural language morphology (Beesley and Karttunen, 2003), and has been applied to the computational processing of countless languages. It is particularly attractive in the low-resource case since it requires substantially less data than data-driven statistical approaches. Thouvenot (2009) released a morphological analyzer for Classical Nahuatl (CN), the urban variant of Nahuatl attested from the time of Spanish invasion through the 19th Century, called *Chachalaca*. This analyzer has some limitations, however, since it can only handle single-word inputs and only runs on Microsoft Windows. More recently, Farfan (2019) designed an FST analyzer based on Classical Nahuatl grammar in order to analyze the similarities and differences of writings in contemporary Nahuatl varieties. Maxwell and Amith (2005) described a project for leveraging Finite-State Transducers to document Nahuatl grammar. Neither of these last two works have been made publicly available.

Our motivation stems from a personal connection of the first and third authors with a ZATN-speaking community. The third author is a native ZATN-speaker.

2 Zacatlán-Ahuacatlán-Tepetzintla Nahuatl

Nahuatl 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).

Zacatlán-Ahuacatlán-Tepetzintla Nahuatl (ZATN), (also referred to by INALI as *Náhuatl de la Sierra Oeste*) is a Nahuatl variant spoken in the Northwestern Sierra region of the state of Puebla, Mexico, mainly in the municipalities of Zacatlán, Ahuacatlán, and Tepetzintla. As of 2007, ZATN

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

had an estimated 17,100 speakers.

While literature about CN abounds (Carochi, 2001; Andrews, 1975; Launey and Mackay, 2011), and there has been substantial documentary and descriptive work on some contemporary Nahuatl variants (Langacker, 1977, 1979; Hill et al., 1999), ZATN has until recently received relatively little focus (Sasaki, 2015). Petra Schroeder released an unpublished partial grammar and some descriptive work of the ZATN variant spoken in San Miguel Tenango, Zacatlán (Schroeder and Tuggy, 2010; Schroeder, 2014, 2015). Mitsuya Sasaki published a sketch of the ZATN variant spoken in Ixquihuacan, Ahuacatlán (Sasaki, 2014), as well as dialectological overview of the Northern Sierra region (Sasaki, 2015).

Most ZATN speakers today also speak Spanish. Economic pressures, migration, and educational language policy have left many Nahuatl variants, ZATN included, facing pending endangerment (Olko and Sullivan, 2015).

2.1 Morphology

With the exception of a closed class of Particle words, Nahuatl is generally considered to have two open word classes, nouns[2] and verbs, each containing one or more stems and numerous inflectional and/or derivational affixes. Both nouns and verbs take subject prefixes, and can behave as predicates (1 and 2).

(1)

ni-	*quiza*	*-h*
s2PL	go.out	PL

'You (pl) go out.'

(2)

ni-	*cihua*	*-me*
s2PL	woman	PL

'You (pl) are women.'

Some scholars of CN (Andrews, 1975; Launey, 1994, 2004) note strong parallels between the two word classes, and suggest that both are clauses, either principal or subordinated as arguments (Andrews calls these morphologically-complex predicates *wordal clauses.* Launey coined the term *omnipredicativity* to describe this characteristic of Nahuatl). Alternatively, Sasaki (2012) suggests that the omnipredicative interpretation may not be necessary. The extent to which ZATN shares omnipredicative features with CN is unclear, and such an investigation is beyond the scope of this paper. Throughout the paper we refer to nouns and verbs as nominal clauses and verbal clauses, respectively.

Nominal stems may take subject and possessive prefixes, and absolutive, diminutive, and possessed or unpossessed number suffixes. Since they take a subject prefix (and the 3rd-person singular subject prefix is null), nominal clauses can often be understood as predicates, "he/she/it is N". Noun stems can be compounded with other noun or adjective stems to form new nominal clauses (4).

(3)

no-	*-telpoca-*	*-cone-*	*-uh*
POSS.1SG	boy	baby	SG

'(He is) My baby boy.'

(4)

itzcuin-	*-tomahuac*
dog	fat

'(It is a) Fat dog.'

Verbal stems, like nominal stems, take a subject prefix. Direct and indirect definite object, indefinite object, reflexivity, directionality, and the honorific register are all expressed via prefixation. Tense and aspect are expressed via suffixes, though the past tense also has an associated required prefix /o-/. Verbal stems can undergo a number of derivational morphological processes. The causative (5) and applicative suffixes change the valency of the underlying stem, and verb stems can be nominalized via multiple morphological processes, such as the addition of an absolutive suffix (6). Verbal stems can also be combined to form new verbs (7). Noun incorporation is also possible (8).

(5)

ni-	*mitz-*	*tzahtzi*	*-tia*
s1SG	o2SG	yell	CAUS

'I cause you to yell.'

(6)

mayan	*-tl*
be.hungry	ABS

'Hunger.'

(7)

y-	*o-*	*ni-*	*mayan*	*-calac*
PERF	PST	s1SG	be.hungry	enter

'I've begun feeling hungry'. (lit. 'Hunger has entered me'.)

(8)

ni-	*xoco-*	*cua*	*-s*
s1SG	apple	eat	FUT

'I will eat apples'.

2.2 Morphophonology

There are several phonological processes that accompany the concatenation of affixes with the stem.

[2] We include adjectives in the noun class, as they behave syntactically and morphologically like nouns. For an investigation of the status of adjectives in a contemporary Nahuatl variant, see Pharao Hansen (2011)

```
"/i/-elision"
i:0 <=> o >:> _ [[Cns Cns] - [t z | t l | c h]
                |t l Cns | t z Cns | c h Cns] ;

!@ n o >:> i t z i n
!@ n o >:> i:0 t z c u i n
```

```
"Simple honorific vowel-harmony"
i:o <=> [ t | x ] _ >:> {q}:c {u}:0 {i}:0
        >:> o n >:> ;

!@ t i:o >:> {q}:c {u}:0 {i}:0 >:> o n >:> i t a
!@ x i:o >:> {q}:c {u}:0 {i}:0 >:> o n >:> o n i
```

Figure 1: Two-level implementations of the morphophonological processes described in section 2.2. The rule on the left elides /i/ in subject prefixes when the adjacent morpheme starts with /o/ followed by two consonants. It accounts for two-character consonants *tz* /ts/, *tl* /tɬ/, and *ch* /tʃ/. The rule on the right defines the simplest version of regressive vowel-harmony (only one syllable to the left) triggered by the honorific prefix /on-/. The sequence *qui* /ki/ corresponds to the 3rd-person singular object prefix, and realizes as either *c* or *qui*, depending on phonological context.

Below, we provide two example morphophonological processes that should be accounted for in a ZATN morphological analyzer:

1. Elision of stem-initial /i/ when preceded by /o/ and a morpheme boundary, and followed by two consonants (/no-itzcuin-tli/ →[no.tzcuin]).
2. Regressive vowel harmony triggered by the honorific prefix /on-/[3][4] (/ti-c-on-niqui/ →[to.c.on.niqui]).

2.3 Distinguishing Characteristics of ZATN

All of the linguistic features of ZATN described above are also true of Nahuatl in general. Below, we briefly mention some of the features that distinguish ZATN from other variants.

Raising *e* →*i* Many of the vowels which are realized in most Nahuatl variants as short /e/ are virtually all raised to /i/ in ZATN (*nicnequi* vs. *nicniqui* 'I want it', *tetl* vs. *titl* 'stone'). For a more thorough treatment of the phonological isoglosses relevant to ZATN, see Sasaki (2015).

Noun endings In some ZATN varieties, many multisyllabic noun stems that in other variants take the absolutive *-li* lose this ending (*comalli* 'griddle' vs. *comal*), and those that in other variants take the absolutive *-tli* shorten to *-tl* (*tipoztli* 'metal' vs. *tipoztl*).

Durative *-to* ZATN is one of a number of Nahuatl variants with the durative suffix *-to*. It can be followed by aspect suffixes (9).

(9) *o- ni- nihnin -to -ya*
PST S1PL walk DUR IMP
'I was walking'.

Subject Prefix *ce-* ZATN exhibits an impersonal subject prefix, *ce-* (10).

(10) *ce- ya -z*
IMPRS go FUT
'We will go'.

(11) *ti- ya -z -que*
S1PL go FUT PL
'We will go'.

In some areas, such as Ixquihuacan (Sasaki, 2018) and Omitlán (personal communication), the impersonal subject prefix is in free variation with the 1st-person plural prefix *ti-* (9).

3 Data

Although the FST approach to morphological modeling requires less data than statistical approaches, written works and lexical resources are critical both for sourcing stems and evaluating completeness.

With respect to published texts in ZATN, there are a few. A translation of the New Testament, *In Yancuic Tlahtolsintilil* (Wycliffe Bible Translators, 2012) is freely available for distribution but not licensed for use in linguistic work. The Summer Institute of Linguistics collaborated with authors from the Tenango region to produce 17 short written stories and an illustrated dictionary containing approximately 250 entries. None of these resources are straightforwardly machine-readable, and they require explicit consent to use as part of a corpus.

In addition to obtaining permissions to use the SIL texts (as of writing we have obtained such permissions for only one text), we see the development of the ZATN analyzer as an opportunity to collaborate with the language community to develop new written material. Our collaboration has resulted in two additional documents which are transcribed and edited from oral histories of a ZATN speaker from Omitlán, Tepetzintla. We hope to expand this corpus by encouraging the production of more original texts and working with other speakers from diverse locations within the ZATN-speaking region.

[3] In some Nahuatl variants, /on-/ is the andative directional prefix. In ZATN as well as many other varieties, it has evolved to become a reverential marker.

[4] This vowel harmony generally extends only to the prefix immediately to the left of /on-/: ti-c-on-tlali →[to-c-on-tlali]. In some variants of ZATN it can extend to every /i/-containing prefix left of /on-/: /ni-mitz-on-pahpacho/ →[no-motz-on-pahpacho] (personal correspondence).

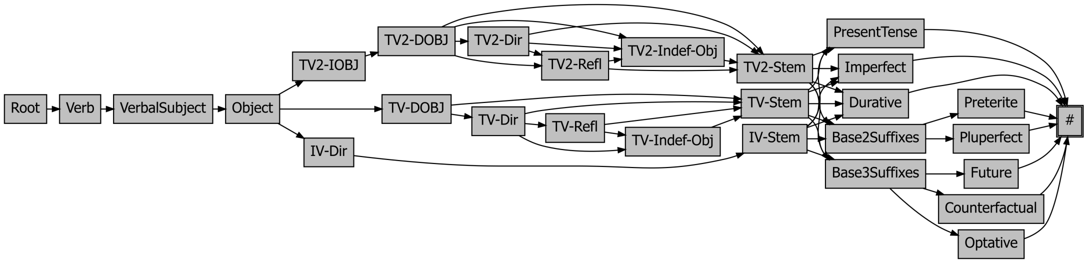

Figure 2: A graph of the current state of the system's verbal morphotactics. Each node represents a continuation lexicon. In the names of the continuation lexicons, `IV`, `TV`, and `TV2` indicate Intransitive, Transitive, and Ditransitive, respectively. As an example, the word *ninechinmacazque*, "You(pl) give them(pl) to me" is analyzed with the following path: VerbalSubject("ni-") → VerbalObject → TV2-IOBJ("nech-") → TV2-DOBJ("in-") → TV2-Stem → Base3Suffixes("maca") → Future("-zque") → #.

4 Approach

Our approach combines the usage of reference grammars for Classical Nahuatl and the Tenango Grammar, the available written works in ZATN, and collaboration with native speakers for word form examples and new corpus development.

We implement our analyzer with HFST (Lindén et al., 2011). A `.lexc` file defines the stems, morphemes, and the transitions between them. We account for morphophonological processes via two-level rules (Koskenniemi, 1983). Figure 1 displays two such rules, corresponding to the morphophonological processes described in Section 2.2. To handle long-distance morphological dependencies (e.g. the number suffix of a nominal clause depends on the presence/absence of a possessed prefix), we use *flag diacritics*, which allow storage and access of values across different states of the FST.

Stems are collected from available ZATN corpora and dictionaries and word lists from other variants in combination with ZATN native-speaker input, which helps ensure proper conversion from the other variants. In the future, with enough collected data, we hope to automate this process of lexical conversion from well-studied variants to ZATN.

4.1 The Nominal Clause

We group noun stems by class according to their absolutive endings: *-tl* (e.g. *cihuatl* 'woman'), *-tli* (e.g. *tochtli* 'rabbit'), *-li* (e.g. *calli* 'house'), *-in* (e.g. *ticpin* 'flea'), and a class for nouns that do not take an absolutive suffix (e.g. *tlahuical* 'husband'). These stems are preceded by subject, honorific, and possessive prefixes, and followed by number and diminutive suffixes. Adjectives are listed as a separate continuation lexicon, take the same prefixes as the noun stems, and can follow noun stems to form compound nouns. Since the 3rd-person subject prefix is null, nominal clauses that do not have a subject prefix are given two analyses: as a plain noun, and as a 3rd-person nominal predicate.

4.2 The Verbal Clause

We separate verbal stems according to their valence (intransitive, transitive, and ditransitive) to ensure the correct number of object prefixes for a given stem. Central to the handling of the verbal clause is the generation of stem alternations corresponding to phonological changes of the stem when combined with affixes. Currently, we generate 5 distinct stem alternations from a canonical form, accounting for the present, preterite, and future tenses, imperfect, durative, and pluperfect aspects, and counterfactual and optative moods. More stem alternations will be added to accommodate causative and applicative affixes. The alternations are generated programmatically with a Python script that implements a series of rules based on a description of CN verb bases in Launey and Mackay (2011), and modified with information from Schroeder's Tenango grammar (Schroeder, 2014) and native speaker input. Figure 2 illustrates the paths through all of the continuation lexicons used for verbal clauses.

4.3 A Note About Orthography

The usage of the Latin alphabet to write Nahuatl has been commonplace since the 16th Century. Classical Nahuatl Orthography used primarily Spanish orthographic rules, variably with diacritics to represent the *saltillo* /ʔ/ glottal consonant and long vowels. The ACK orthography, named after the three scholars—Andrews, Campbell, and Karttunen—considered fundamental in its development, is similar to Classical Nahuatl orthography, except that the *saltillo* is consistently written with 'h', and vowel-length contrasts are omitted. We use ACK orthography for all of the examples in this paper. Mexico's *Secretaría de Educación Pública* and the Summer Institute of Linguistics have encour-

Title	POS	N Tokens	N Types	Token Cov.	Type Cov.
Amo Niniqui Niyaz Escuela	All	738	285	0.83	0.75
	Nominal Clauses	90	56	0.90	0.89
	Verbal Clauses	151	105	0.79	0.73
Nochcahuan	All	296	168	0.72	0.66
	Nominal Clauses	47	27	0.85	0.89
	Verbal Clauses	64	54	0.70	0.67
Tiquitini	All	362	177	0.57	0.38
	Nominal Clauses	48	27	0.67	0.59
	Verbal Clauses	82	71	0.30	0.30

Table 1: Word-level coverage of the analyzer on three ZATN texts. In addition to total coverage, we list the coverage of the two main open word classes, nominal and verbal clauses (see Section 2.1 for details). All stems from the texts were added to the analyzer, and as such these numbers exclusively reflect the state of the system's morphotactics (as opposed to the coverage of the lexicon).

aged the adoption of the *Sistema Práctico* (SP) orthography, purporting greater faithfulness between graphemes and Nahuatl phonemes. There is ongoing debate about orthography and progress in language maintenance and revitalization efforts (de la Cruz Cruz, 2014).

Our initial implementation uses the ACK orthography, but we plan to support all orthographies, and have implemented a FST to convert between ACK and SP. Straightforward orthographic conversion is complicated by the prevalence of Spanish loan words as well as code-switching.

5 Evaluation

In this section, we report on the coverage of our analyzer on our existing corpus[5]. The corpus consists of three short texts: two original texts created in collaboration with a ZATN-speaker from Omitlán, entitled *Amo Niniqui Niyaz Escuela* "I don't want to go to school" and *Nochcahuan* "My sheep", and one text published by an author from San Miguel Tenango, *Tiquitini* "Hard-worker", in collaboration with the Summer Institute of Linguistics. The token- and type-coverage numbers are presented in Table 1.

Coverage was calculated after adding all stems from the texts to the analyzer. The reason for doing this was that we are interested in evaluating the morphotactics of the system. There is no existing published lexicon for ZATN, so testing the lexicon size at this point would not have been informative.

The drop in type coverage on *Tiquitini* reflects the variation between subvariants of ZATN from different areas. For example, the San Miguel variety of ZATN is unique in its use of metathetic subject prefixes (for example, the word realized as *ni#huili* in most ZATN varieties is realized as *in#huili* in San Miguel) (Schroeder and Tuggy, 2010). Until now the analyzer has been developed in collaboration with a single speaker from Omitlán, and thus could be overfit to that specific subvariant.

Our error analysis also reveals a lack of support for directional verbal suffixes (for example, *ni#qu#ita#qui* → "I come and see it.") and extended vowel-harmony (see footnote 4 for a description). We also do not yet support locatives like *i#tich* 'on/in' which may undergo contraction, e.g. *ich*.

6 Future Work and Direction

The current status of the analyzer handles many, though not all, of the fundamental inflectional morphological processes of ZATN. A primary focus of our future work is the collection of corpora and subsequent enhancement of our lexicon. We plan to obtain the rights to and incorporate as many of the available ZATN texts as possible into our development corpus and continue collaboration with ZATN-speakers to produce more original texts that capture the linguistic diversity of ZATN varieties.

On the technical side we will include supporting directional suffixes, and common derivational processes such as the applicative and causative suffixes as well as deverbalization. Noun incorporation is infrequent in our data, and as such is yet to be supported. This will be an important area of future improvement as the corpus expands.

[5] Our corpus is available in the project repo: https://github.com/apertium/apertium-nhi/tree/master/corpus

References

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

Beesley, K. R. and Karttunen, L. (2003). Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

Ben Slimane, M. (2008). *Appropriating new technology for minority language revitalization: The Welsh case*. PhD thesis.

Carochi, H. (2001). *Grammar of the Mexican Language with an explanation of its adverbs (1645)*, volume 89 of *UCLA Latin American Studies*. Stanford University Press, Stanford. Arte de la lengua Mexicana 1645.

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

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

Hill, J. H., Hill, K. C., Farfán, J., and Cruz, G. L. (1999). *Hablando mexicano : la dinámica de una lengua sincrética en el centro de México*.

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.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Number 11 in Publications. University of Helsinki. Department of General Linguistics, Finland.

Langacker, R. W. (1977). *Studies in Uto-Aztecan grammar*, volume 1 of *An overview of Uto-Aztecan grammar*. Summer Institute of Linguistics and the University of Texas at Arlington, Dallas.

Langacker, R. W. (1979). *Modern Aztec grammatical sketches: Studies in Uto-Aztecan grammar 2*. Summer Institute of Linguistics Publications in Linguistics, 56(2).

Launey, M. (1994). *Une grammaire omniprédicative*. CNRS-Editions.

Launey, M. (2004). The features of omnipredicativity in Classical Nahuatl. *STUF - Language Typology and Universals*, 57(1):49–69.

Launey, M. and Mackay, C. (2011). *An Introduction to Classical Nahuatl*. Cambridge University Press.

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T., and Silfverberg, M. (2011). HFST—framework for compiling and applying morphologies. *Communications in Computer and Information Science*, 100:67–85.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (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.

Maxwell, M. and Amith, J. D. (2005). Language documentation: The Nahuatl grammar. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 474–485, Berlin, Heidelberg. Springer Berlin Heidelberg.

Olko, J. and Sullivan, J. (2014). Toward a comprehensive model for Nahuatl language research and revitalization. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, volume 40.

Olko, J. and Sullivan, J. (2015). Empire, colony, and globalization. a brief history of the Nahuatl language. *Colloquia Humanistica*, pages 181–216.

Pharao Hansen, M. (2011). Adjectives in Hueyapan Nahuatl: Do they exist? and if they do what kind of adjectives are they?

Reyhner, J. (1999). Some basics of indigenous language revitalization. http://jan.ucc.nau.edu/~jar/RIL_Contents.html.

Sasaki, M. (2012). R-marking: Referential person affixes in Classical Nahuatl nouns. Master's thesis, University of Tokyo.

Sasaki, M. (2014). A dialectological sketch of Ixquihuacan Nahuatl. 東京大学言語学論集, 35(TULIP):139–170.

Sasaki, M. (2015). A view from the Sierra : the Highland Puebla area in Nahua dialectology. 東京大学言語学論集, 36(TULIP):153–165.

Sasaki, M. (2018). Cityless air makes free: Characteristics of free variation in modern nahuatl. [Draft].

Schroeder, P. (2014). *Gramática del Náhuatl de San Miguel Tenango, Zacatlán, Puebla*. Summer Institute of Linguistics. [Draft publication].

Schroeder, P. (2015). *Phonology of Nahuatl de San Miguel Tenango, Zacatlán, Puebla*. Summer Institute of Linguistics. [Draft publication].

Schroeder, P. and Tuggy, D. H. (2010). The consonantal prefixes of San Miguel Tenango Nahuatl, Zacatlán. *Etnografía del estado de Puebla, zona norte*, pages 112–117.

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

Wycliffe Bible Translators (2012). *In Yancuic Tlahtolsintilil*. Wycliffe Bible Translators.