A morphological analyzer for Huasteca Nahuatl

Ana Tona and **Guillaume Thomas** and **Ewan Dunbar**
University of Toronto

Abstract

We present a morphological analyzer of Huastecan Nahuatl, which builds upon an existing analyzer of the Zacatlan-Ahuacatlán-Tepetzintla variety of Nahuatl (ZAT). We extend the range of morphological phenomena covered by the ZAT analyzer, and we adapt it to the Huastecan variety of Nahuatl, which broadens its potential userbase. The new analyzer has high coverage on a corpus of Huasteca Nahuatl texts (HN) and has improved coverage on a ZAT corpus.

1 Introduction

Nahuatl has been well documented by religious and academic institutions since colonial times, resulting in numerous descriptive and pedagogical grammars, dictionaries, published texts of various genres, and even digital resources. However, the overwhelming majority of these resources are for Classical Nahuatl, a prestige variety that is no longer spoken today. These resources are at best limited in their utility—and more often simply misleading—for native speakers, learners, and teachers, and for linguists interested in modern Nahuatl. While the situation is changing, the resources that exist for modern Nahuatl remain limited.

Modern Nahuatl consists of four dialectal areas: Western periphery, Huasteca, Center, and Eastern periphery. This paper presents an open-source morphological analyzer for Huasteca Nahuatl, one of the three modern varieties with a large number of speakers. Huasteca Nahuatl is spoken as a first language by over one million people in Mexico, and is often studied as a second language. The analyzer builds upon Pugh et al.'s (2021) analyzer of Zacatlan-Ahuacatlán-Tepetzintla (ZAT) Nahuatl, and accepts both the Huasteca and the ZAT varieties, while providing broader coverage of morphological phenomena than the previous analyzer.

2 Finite-state morphological analyzers

Digital resources facilitate access to language, creating opportunities for learners and providing rich supporting materials for teachers (Galla, 2021). Morphological analyzers play a key role in learning resources for morphologically-rich languages, as well as aiding in the construction of annotated textual corpora. Morphological analyzers built using finite-state transducers (FSTs) have the advantage of not requiring any training corpora, which, unlike Classical Nahuatl, are virtually non-existent for modern Nahuatl varieties. Although modern Nahuatl has some presence on the internet, the total amount of data available is not enough to use a data-driven approach. Our analyzer was developed with the HFST toolkit (Lindén et al., 2009). Morphotactic rules were described using a form of right-recursive grammar (Karttunen, 1993), while phonological and orthographical alternation rules were described using a declarative system of rule constraints (Karttunen et al., 1987). These descriptions are compiled out to FSTs by `lexc` (Lexicon Compiler) and `twolc` (Two-Level Compiler) respectively, and the resulting FSTs are composed and inverted to serve as an analyzer. Conversely, a morphological analysis can be given as input to generate the corresponding word form. An example is shown in 1.

(1) Input and parsed output:

a. `nicalchihqui`

b. `ni<Suj_sg1>calli<n><incrp> chihua<tv>qui<pret>`

3 Previous works

Several morphological analyzers exist for Classical Nahuatl, though not all are publicly available. These include *Chachalaca*[1] (Thouvenot,

[1] http://www.sup-infor.com/program/Chachalaca/Chachalaca-txt.htm

2011) and that of Martínez-Gil (1981). Relatedly, Escobar Farfán (2019) applies a morphological analyzer for Classical Nahuatl to modern Nahuatl varieties in order to assess overlap between modern varieties, and between modern varieties and Classical Nahuatl. For modern Nahuatl varieties, only the FST-based analyzer of Pugh et al. (2021) exists, which is open source and freely available. This analyzer is limited in three ways. First, it covers only Zacatlán-Ahuacatlán-Tepetzintla (ZAT) Nahuatl, a variety spoken in Western Puebla. Second, a number of common morphological phenomena are not handled, including causative and applicative suffixes and noun incorporation (see sections 4 and 5 below for more details). Finally, it accepts only the academic Andrews-Campbell-Karttunen orthography (ACK) orthography used for Classical Nahuatl, rather than that of the Secretaría de Educación Pública (SP) typically used in Mexico for modern varieties.

4 Features of Huasteca Nahuatl

General grammatical features are shared among Nahuatl varieties, but there are some phonological, morphosyntactic and lexical differences. Differences among varieties have not been thoroughly and systematically studied. The main resources consulted for the Huastecan variety were the grammatical descriptions of Hasler (2011) and Beller and Beller (1977), as well as Cruz (2018)'s dictionary and class notes from a Nahuatl course (Santos, 2019). Here we give a short summary of Huasteca Nahuatl morphology.

Nahuatl has a rich morphological system. As an omnipredicative language (Launey, 2004), Nahuatl verbs, nouns and adjectives can take subject prefixes to function as predicates, but other morphology is not shared among the three categories.

Non-possessed nouns obligatorily take a suffix called the *absolutive* which varies as a function of the noun class (e.g., *-li, -tli, -tl*).[2] Possessed nouns may take a *possessive* prefix indicating the possessor, which is in some cases obligatory (e.g., body parts). It is possible to form new nouns via compounding and derivational affixation (e.g., honorific, diminutive).

(2) *tepoz-toto-tl*
metal-bird-ABS
'airplane (lit. metal bird)'

(3) *pil-cone-tzi*
DIM-child-DIM
'baby (lit. little child)'

Verbs agree in number with the subject. Plural is marked as a suffix, while the singular is unmarked. Other verbal morphology includes subject (person) and object prefixes (person, plus undefined human and non-human object markers *te-* and *tla-*), a reflexive marker *mo-*, negative prefixes, and derivational prefixes. In non-Huastecan varieties (including ZATN), there is also a past-tense prefix (*o-*). Suffixes include tense and aspect markers. These are suffixed to the root or after causative/applicative suffixes, if they are present:

(4) *ø-cochi-qui*
3SG.SUBJ-sleep-PST
'(he/she) slept'

(5) *ø-tequiti-c*
3SG.SUBJ-work-PST
'(he/she) worked'

Tense and aspect markers are absent when there are directional affixes (which comprise aspect in their meaning) present. Unlike in other varieties, in Huastecan, directionals are suffixes. Deverbal nouns do not have tense and aspect markers, but do take object prefixes.

(6) *tla-mach-ti-quetl*
1OBJ.UNDEF.NH-learn-CAUS-NMLZR
'teacher'

Transitive and intransitive verbs may get (additional) objects by adding causative or applicative suffixes. The applicative marker does not always trigger a valency change, and may be used to focalize the (recipient) object (Peregrina et al., 2017).

Verbs may incorporate nouns and adverbs (Launey, 1999).

(7) *sesem-itta*
each-see
'watch closely'

It is also possible to incorporate other verbs, in which case the incorporating verb functions as an auxiliary.

(8) *ø-cochi-z-nequi*
3SUBJ-sleep-FUT-want
'he/she wants to sleep.'

Infrequently, verbs can be derived by partial reduplication of a verb root, or by adding causative morphology to a nominal root.

[2] Despite the name—an Uto-Aztecanist term—Nahuatl is not an ergative language.

(9) *ϕ-po-poca*
3SUBJ-RED-smoke
'to be smoking or giving off steam'

(10) *ni-h-xochi-tia*
1SG.SUBJ-3SG.OBJ-flower-CAUS
'I (put) flowers on it.'

Huastecan Nahuatl deletes final /n/ on nouns with absolutive *-in* and diminutive *-tzin* and deletes /j/ between vowels (e.g., *chiya* > *chia* 'wait'). The Huastecan variety does not show the extensive /i/-deletion nor the raising of /e/ to /i/ found in other varieties, including ZAT Nahuatl (Pugh et al., 2021).

Tense and aspect markers vary in form across varieties (notably for Huasteca, imperfect: *-yaya*, plural future: *-zeh*), as do pronominal forms (Huasteca, 1sg: *na*, 2sg: *ta*, 3sg: *ya*, 1pl: *tohuantih*, 2pl: *inmohuantih*, 3pl: *inihuantih*).

5 Implementation

Taking the system of Pugh et al. (2021) as a starting point, the new analyzer adds an analysis of reflexive/impersonal, causative, and applicative affixes—all of which typically change argument structure—nominalization, verb incorporation, noun incorporation, directional suffixes, noun-noun compounding, and the diminutive prefix. Because these features are common to many varieties, including ZAT, we chose to build an analyzer that would accept both Huasteca and ZAT forms. As with this earlier analyzer, ours accepts the ACK orthography.

As noted in section 2, the analyzer is implemented using HFST (Lindén et al., 2009) and consists of a morphotactic description in the `lexc` formalism and a morphophonological description in the `twolc` formalism. In `lexc`, morphotactics is described using a set of lexicons. Lexicons contain entries, each of which consists of a form and a continuation class. Figure 1 shows the lexical continuation classes for verbs and the graph that defines verb morphotactics. (In the interest of space, we do not detail the way the analyzer handles nouns here.) In order to handle valency-changing suffixes, the system weakens the traditional separation between inherently intransitive, transitive, and ditransitive verbs. There are no "intransitive" classes in which verbs are restricted to appearing without object prefixes. Rather a distinction is made between ditransitive (prefixed with **2-**) and transitive verbs, neither of which strictly enforces the presence of an object prefix. Similarly, ditransitive verbs can appear with only indirect object morphology, allowing more flexibility for valency changes (namely, reduction).

To handle noun and verb incorporation, additional lexicons were created for verbs that can incorporate other verbs, and for nouns as they appear in noun-incorporation (parallel to the main noun lexicon). The new lexicon of incorporating verbs also has access to the main verb list for incorporation.

Finally, two classes, **TAM-1** and **TAM-2,** allow us to distinguish two groups of inflectional suffixes marking tense, aspect, and mood, which require different verbal stems.

Differences between the Huastecan and the ZAT varieties are primarily in the realization of individual morphemes. Thus, to handle both varieties, we extended the lexicon of Pugh et al. (2021) to include Huastecan forms. The past tense prefix *o-* is not used in the Huastecan variety, and so, in our analyzer, this prefix is optional. ZAT also has several phonological rules which do not exist in the Huastecan variety, which we also made optional.[3]

6 Evaluation and performance

To evaluate coverage, a small corpus of short stories from the Huasteca region was collected and passed to the analyzer (285 tokens, 142 types). We calculate *token coverage* as the percentage of all tokens successfully parsed, and *type coverage* as the percentage of all types successfully parsed. As a baseline, we compare the system of Pugh et al. (2021). As many of the forms in our Huastecan corpus differ from their ZAT Nahuatl equivalents, we also assess performance of both systems on the ZAT corpus for which Pugh et al. (2021) report the best performance (*Omitlan*: 296 tokens, 168 types). In order to ensure the evaluation is of the analyzer, rather than of the lexicon, we add all morphemes in both corpora to the lexicon for the proposed analyzer. As the lexicon of Pugh et al. (2021) contains all of the types in the ZAT corpus, but not those of the new Huastecan corpus, we evaluate that system only on ZATN. Results are presented in Table 1.

As can be seen from Table 1, the proposed analyzer has high coverage on both Huastecan and ZAT Nahuatl. Comparison against the analyzer of

[3]The morphological tags used were also revised from the ZATN analyzer with Nahuatl- and Spanish-speaking users in mind and for consistency with modern grammatical descriptions of Nahuatl.

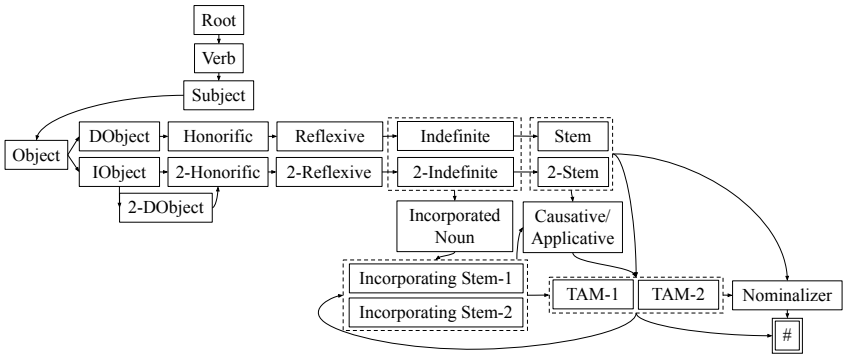

Figure 1: Morphotactics for verbs in terms of lexicon (LEXC) continuation classes. Arrows to (or from) the outside of dotted boxes indicate links to (or from) both of the classes inside the box.

	Huastecan		ZAT	
	Token	Type	Token	Type
Proposed	95.1%	90.8%	**92.6%**	**87.5%**
Pugh et al.	—		88.2%	80.9%

Table 1: Token- and type-level coverage of the proposed analyzer and that of Pugh et al. (2021) on Huastecan and Zacatlan-Ahuacatlán-Tepetzintla Nahuatl corpora.

Pugh et al. (2021) also demonstrates in improvement on ZAT, presumably due to the addition of novel morphological phenomena.

7 Limitations

This analyzer for Huasteca (and ZAT) Nahuatl is currently the broadest-coverage morphological analyzer for any modern variety of Nahuatl. Several phenomena are missing but are straightforward to add: verbal derivational prefixes such as *ix-* (e.g., *ix-ketsa* 'to build (lit. to make stand)') and the negative verbal prefix *ax-*, locative and other derivational suffixes attaching to nouns, and adverbial incorporation. Currently, the analyzer also fails to restrict plural markers to the appropriate noun/verb classes.

Other phenomena may require more reorganization, such as N+Adj/Adv compounds (*a-huelic* 'tasty water', *a-huehca-pan* 'sea bottom, abyss'), Adj/Adv+N compounds (*hueyi-cali* 'skyscraper (lit. big house)', *momoztla-moxtli* 'newspaper (lit. every-day book)'), denominal verbs (*ni-h-xochitia* 'I (put) flowers on it') and reduplication (*poctli* 'smoke' > *po-poca* 'smoking or giving off steam'). Since reduplication has limited productivity, with some instances being fossilized, reduplicated verbs may well be added into the lexicon.

The approach taken to incorporation introduces a cycle in the graph to avoid duplicating parts of the lexicon. As only one root can ever be incorporated, the system will accept unattested forms. The same is true for noun-noun compounds. Similarly, the approach taken to having a two-variety analyzer—by simply rendering variety-specific features optional—allows the analyzer to overgenerate, parsing forms that combine features in ways that would be ungrammatical in either variety. This approach was taken largely to be able to compare the coverage on variety-non-specific phenomena to the previously existing ZAT analyzer. Nevertheless, we believe that an FST-based morphological analyzer for Nahuatl should be multi-varietal, as many linguistic features are often subject to inter- and intra-speaker variation in a single text, and different varieties have overlapping features, as they form part of a continuum. Additionally, learners may not always know which variety they are dealing with and having a system that handles several varieties is helpful in such cases. Still, the problem of overgeneration brings up the question of whether having a multi-varietal system is the right approach. With this in mind, we believe that the current system would be best used for analysis rather than generation. A more robust system would be able to give information about which forms are typical of different varieties, which can be done using flags.

Three factors limit the usefulness of the current system for broad use. First, the current lexicon is limited to words needed to cover the corpora above. Second, it only accepts the ACK orthography, which is not widely used among speakers. Finally, it lacks an accessible and user-friendly interface that would allow the community to take advantage of it. We plan to make the system available via a web interface.

8 Summary of contributions

We have presented an open-source morphological analyzer for Huasteca Nahuatl, and which also accepts Zacatlan-Ahuacatlán-Tepetzintla (ZAT) forms. It has the broadest coverage of any existing modern Nahuatl analyzer, thus contributing to a small but growing body of resources for modern Nahuatl varieties.

References

Richard Beller and Patricia Beller. 1977. *Huasteca Nahuatl*. Modern Aztec Grammatical Sketches. Summer Institute of Linguistics, Dallas.

Marcos Hilario Cruz. 2018. *Diccionario bilingüe náhuatl-español. En defensa del náhuatl en el México moderno*.

Jonathan Escobar Farfán. 2019. *Publications Manual*. Ph.D. thesis, University of Sheffield.

Candace Galla. 2021. *Practical Educational Tools and Applications Used Within Native Communities*. Ph.D. thesis, University of Arizona.

Andres Hasler. 2011. *Gramática moderna del nahua de la Huasteca*.

Lauri Karttunen. 1993. Finite-state lexicon compiler. Technical Report ISTL-NLTT-1993-04-02, Xerox Palo Alto Research Center, Palo Alto, CA.

Lauri Karttunen, Kimmo Koskenniemi, and Ronald Kaplan. 1987. A compiler for two-level phonological rules. *Tools for morphological analysis*.

Michel Launey. 1999. Compound nouns vs. incorporation in classical nahuatl. *STUF - Language Typology and Universals*, 52(3-4):347–364.

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

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. HFST tools for morphology–an efficient open-source package for construction of morphological analyzers. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 28–47. Springer.

Martínez-Gil. 1981. Computer systems for analysis of nahuatl. *Research in Computing Science*, (47):11–16.

Manuel Peregrina, Albert Alvarez, and Zarina Estrada-Fernández. 2017. *Transitivity and valency-changing operations in Huasteca Nahuatl: Theoretical and typological perspectives*, pages 81–106.

Robert Pugh, Francis Tyers, and Maribel 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: Vol. 1*, pages 80–85.

Valeriano Santos. 2019. Nahuatl course notes. Taught at Centro Cultural de Tijuana, Mexico.

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