Developing finite-state language technology for Maya

Robert Pugh[†] and **Quetzil Castañeda**[‡◇] and **Francis Tyers**[†]
pughrob@iu.edu, quetzil@osea-cite.org, ftyers@iu.edu
[†]Department of Linguistics, Indiana University
[‡]Center for Latin-American and Caribbean Studies
[◇]Open School of Ethnography and Anthropology (OSEA-CITE)

Abstract

We describe a suite of finite-state language technologies for Maya, a Mayan language spoken in Mexico. At the core is a computational model of Maya morphology and phonology using a finite-state transducer.[1] This model results in a morphological analyzer and a morphologically-informed spell-checker. All of these technologies are designed for use as both a pedagogical reading/writing aid for L2 learners and as a general language processing tool capable of supporting much of the natural variation in written Maya. We discuss the relevant features of Maya morphosyntax and orthography, and then outline the implementation details of the analyzer. To conclude, we present a longer-term vision for these tools and their use by both native speakers and learners.

1 Introduction

Maya[2] is a member of the Yucatecan branch of the Mayan language family (Figure 2[3]). It is the second most widely-spoken indigenous language of Mexico, with around 800,000 speakers primarily in the states of Yucatan, Quintana Roo, and Campeche in southern Mexico (Collin, 2010) (See Figure 1[4]), including a substantial speaker population in California (Mattiace and de Mola, 2015) and a modest population in Belize.

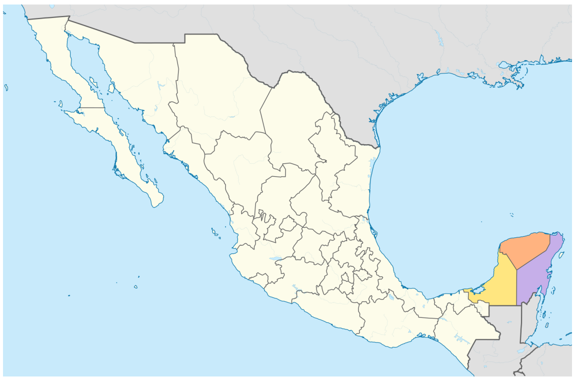


Figure 1: A map highlighting the three Mexican states where Maya is spoken: Yucatan (Orange), Quintana Roo (Purple), and Campeche (Yellow).

Text-based language technologies, ubiquitous for a small number of "mainstream", mostly colonial languages such as English or Spanish, facilitate human-computer interaction and to a large extent computer-mediated communication, and can aid in language learning (Shadiev and Yang, 2020). Furthermore, language technology for endangered languages can play a useful role in language maintenance and revitalization efforts (Reyhner, 1999; Ben Slimane, 2008; Zhang et al., 2022). Unfortunately, there is a paucity of such technology for most of the world's languages, leaving speakers and language learners without potentially valuable resources. Consequently, monolingual speakers face additional barriers to entry in the digital domain, and speakers who are bilingual in a dominant, colonial language for which such technology exists will be more likely to use that language online and on digital devices, further contributing to language shift.

This paper outlines the design and implementation of a finite-state morphological analyzer for Maya. Developed in concert with Maya language educators, the analyzer is intended for use as a writing tool for authors, educators, and students

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

[2] We follow the recommendation of the Open School or Ethnography and Anthropology and the Community Institute of Transcultural Change (see §1.1) with respect to terminology, using the term "Maya," the autonym of the Maya-speaking people, when referring to the language or cultural/ethnic group, instead of "Yucatec Maya," commonly used by linguists, or "Mayan", which should be reserved for referring to the language family or proto-Mayan (Castañeda and Dzidz Yam, 2014).

[3] Figure 2 was created by user Madman2001 (https://commons.wikimedia.org/wiki/File:Mayan_Language_Tree.svg)

[4] Figure 1 is based on work by user Kmusser (https://commons.wikimedia.org/wiki/File:Mexico_States_blank_map.svg)

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 30–39
July 14, 2023 ©2023 Association for Computational Linguistics

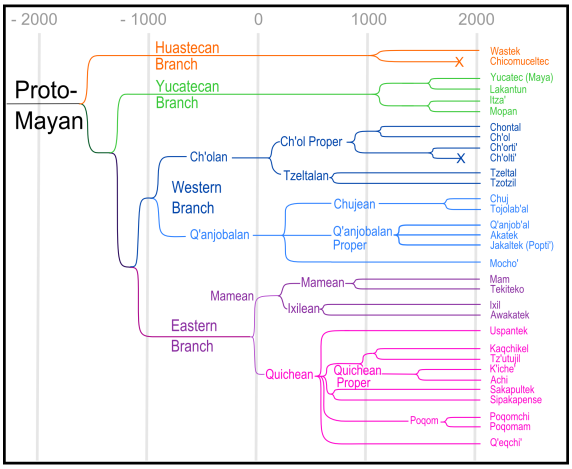


Figure 2: The Mayan language family. Maya ("Yucatec Maya", the focus of this paper) is located on the green "Yucatecan Branch".

(to ensure consistent written resources via a spell-checker), and as a reading-aid that can provide students with lexical information (e.g. the root and/or grammatical features) about an unknown word in a text. We focus primarily on the grammar of Maya and the implementation of the analyzer, and present a prototype of a working spell-checker.

1.1 Motivation and OSEA-CITE

The motivation for the present work stems from a collaboration with the Open School of Ethnography and Anthropology and the Community Institute of Transcultural Change (OSEA-CITE, henceforth OSEA), a Pisté-based organization whose stated focus is "language revitalization, sustainability, cultural ownership, heritage rights, community health and well-being, the innovation of tradition, and the interconnections between local, national, and transnational communities and social forces." While designed with Maya speakers, learners, linguists, and language activists in mind, the technologies described below are particularly informed by and aligned with OSEA pedagogical materials (Castañeda, 2014) for use in the classroom as reading and writing tools for both learners and educators in OSEA programs.

2 Related work

The use of finite-state transducers (FSTs) for modeling human language has a long tradition spanning multiple decades (Kornai, 1996) and proving effective in areas such as morphological analysis (Beesley and Karttunen, 2003), spell-checking (Pirinen et al., 2014), among others. It is particularly attractive in the low-resource case since it requires significantly less data than popular statistical approaches. Furthermore, finite-state systems can also be leveraged in order to generate data to train better statistical machine-learning models (Moeller et al., 2018).

The application of finite-state language technology to indigenous languages of Mesoamerica also has some precedent, with morphological analyzers developed for Nahuatl (Maxwell and Amith, 2005; Pugh and Tyers, 2021; Tona et al., 2023), Zapotec (Washington et al., 2021), Huave (Tyers and Castro, 2023), and K'iche' (Richardson and Tyers, 2021). Nicolai et al. (2020) present the large-scale development of morphological analyzers and generators for over one thousand languages using the Johns Hopkins University Bible Corpus (McCarthy et al., 2020), including some Mayan languages.

Kuhn and Mateo-Toledo (2004) is perhaps one of the earliest published works focused on the development and application of language technology to assist in documenting a Mayan language, Q'anjob'al (spoken in Guatemala), training a maximum-entropy part-of-speech tagger. Palmer (2009) and Palmer et al. (2010) also apply techniques from machine learning and computational linguistics to the documentation of a Mayan language (Uspanteko, also spoken in Guatemala). More recently, Tyers and Henderson (2021) and Tyers and Howell (2021) developed an annotated linguistic corpus of K'iche' and explored approaches to automated tagging and parsing. Maya is also included as one of six Mexican languages aligned with Spanish in the Parallel Corpus for Mexican Languages (Sierra Martínez et al., 2020).

There has also been interest and some work leveraging computational technology to annotate and analyze Classic Maya heiroglyphic writing (Prager et al., 2018; Vertan and Prager, 2022).

Particularly relevant to motivation and aims of the present project, Gasser (2011), outlines useful applications of computational morphological analyzers for learners of morphologically-rich indigenous languages of the Americas.

3 Orthography

The Latin alphabet has been used to write Yucatec Maya since the 16$^{\text{th}}$ century, but the first organized efforts to standardize the orthography took place in the mid-20$^{\text{th}}$ century (Brody, 2004). The colonial-era writing practices are described thoroughly in Shigeto (2011), and a variant of this orthographic approach is also used in Bolles and Bolles (2001). Many linguistic resources for Maya also use an orthography inspired by the Americanist Phonetic Alphabet (Bricker et al., 1998; Blair and Vermont-Salas, 1965), (e.g. using ʔ for the glottal stop). Today, the commonly (though by no means unanimously) adopted "contemporary orthography" is laid out in the publication *Normas de Escritura Para la Lengua Maya* (SEP & INALI, 2014).

In the classroom, OSEA teaches a writing system similar to the contemporary one, with a few pedagogically-motivated changes, like the explicit marking of low tone on long low vowels. Additionally, there are some differences related to the spelling of specific words. In order to offer students a consistent source for spelling questions (primarily with respect to vowel quantity and tone), OSEA uses Bricker et al. (1998) as an authoritative reference. This is not to say that alternative spellings are incorrect from OSEA's perspective, but rather that it is valuable for students to have a thorough and consistent guide to reference when making spelling decisions[5].

Since the project presented here is intended to be used by students and teachers in the OSEA Maya language program, we follow these orthographic norms while still supporting both the colonial and contemporary orthographies. Details about this are provided in section 6.5.

4 A brief overview of Maya morphosyntax

An important linguistic property of Maya worth mentioning at the outset is that it does not have tenses, per se. Instead, it inflects verbs for aspect to reflect whether a given action has been completed, or how long ago it began (Bricker et al., 1998). Details about this system are explored in greater depth in section 4.2.

Maya is a split-ergative language, i.e. it follows ergative-absolutive alignment in all but the imperfective aspect, where it follows nominative-accusative alignment.

As will become obvious in the discussion below, Maya has a complex derivational system. Most word classes can be derived from other word classes, and the transitivity and voice of a verb is derived morphologically as well.

4.1 Pronouns

Maya has three sets of pronouns: one set (the "independent pronouns") is syntactically independent of verbs while two, called "Dependent Pronouns" are affixes or clitics on the verb.

Independent pronouns, as the name suggests, are independent words (e.g. not affixes or clitics). They may be used to emphasize (Example 1) or topicalize (Example 2) a verbal argument, or after prepositions to express indirect objects.

(1) *k'abéet a bin-e'ex te'ex*
OBLIG S2 go-S2PL PRON2PL
'You all (emph.) must go'.

(2) *te'ex-e k'abéet a bin-e'ex*
PRON2PL-TOP OBLIG S2 go-S2PL
'As for you all, you must go'.

Set A pronouns (*a* in examples 1 and 2) which come before the verb, typically written separated from the verb, and are sometimes written as merged or contracted with a preceding aspectual auxiliary. With respect to case, Set A pronouns correspond to the A argument (as defined in Dixon and Dixon (1994)) except when in the imperfective, in which case they are the subject of both transitive and intransitive verbs, except in copular clauses where a Set B pronoun is used to mark the subject. Set A pronouns are also the possessive pronouns.

Set B pronouns are suffixes used to express the S and O arguments of the verb, i.e. the subject of an intransitive verb and the object of a transitive verb, except in the imperfective. They are also used as the subject in copular clauses.

4.2 Verbs

Verbs are by far the most morphologically complex words in Maya. The specific components of the "verb compound" depend on the verb's transitivity and the aspectual class of the conjugation. The aspectual auxiliaries and Set A pronouns are often written as separate orthographic words from the verb itself.

In the imperfective, verbs typically must be preceded by an aspectual auxiliary followed by a Set A pronoun. For example, *k* (habitual), *táan* (progressive aspect), *laili'...e'* ("still doing X"), etc. Note that some of these auxiliaries, such as *laili'* above,

[5] It should be noted that our implementation is also flexible and can be easily-updated to be applied to other writing conventions and pedagogical environments.

Orthography	Notes	Example text
Colonial-style	*cħ*, *pp*, *dz* for /tʃ'/, /p'/, and /ts'/, no tone or length marking on vowels	Le cħocħlin, tumen chen kay cu betice ti le yax kino, ma tu bin u caxte u yoch.
Contemporary (INALI)	*j* for /h/, marks long high and rearticulated vowels	Le ch'och'lin, tumen chen k'aay ku beetike' ti' le yáax k'iino', ma' tu bin u kaxtej u yo'och.
Modified contemporary (OSEA)	Similar to Contemporary. Marks long high, long low, and rearticulated vowels, *h* for /h/	Le ch'och'lin, tumèen chen k'àay ku bèetike' ti' le yáax k'ìino', ma' tu bin u kaxteh u yo'och.

Table 1: An example of three different orthographic styles in written Maya. The original text is from Bolles and Bolles (2001) and is written in a style inspired by colonial-era orthography, which we refer to here as "Colonial-style." Note the differences in character choice (e.g. *j* vs. *h*), as well as minor spelling differences like *tumen* vs. *tumèen* (the latter's vowel quantity and tone coming from a particular reference dictionary). The descriptions of the orthographies are by no means exhaustive, as a complete breakdown of the similarities and differences of each is beyond the scope of this paper.

Person/Num.	Set A	Set B	Indep.
1Sg	*in*	*-en*	*tèen*
2Sg	*a*	*-ech*	*tèech*
3Sg	*u*	Ø	*leti'*
1Pl	*k*	*-o'on*	*to'on*
2Pl	*a...-e'ex*	*-e'ex*	*te'ex*
3Pl	*u...-o'ob*	*-o'ob*	*leti'o'ob*

Table 2: A table of the three sets of Maya pronouns: dependent pronouns (Set A, Set B) and independent pronouns. Note that the second- and third-person plural Set A pronouns consist of both a prefix and a corresponding suffix

Root	Deriv.	Asp. status	SetB	SetA Pl.
hóok	*-s*	*-ah*	*-en*	*-e'ex*
go.out	CAUS	PERF	O.SG1	S.PL2

Table 3: A simplified template of the verbal compound in Maya, with each slot's corresponding value for the word *hóoksahene'ex* "You (pl) took me out." Not all of the possible verbal morphemes are represented in this table. To the left of the verb, the verbal compound can also include a negation marker, an aspectual auxiliary, and/or a Set A pronoun. These are omitted from the template above since they are typically written as separate orthographic words, and thus are treated as such in our analyzer. On the right side, there can also be a "terminal enclitic" (Bricker et al., 1998) corresponding to a previous part of the phrase, such as a negation or locative particle (*-i*).

have a corresponding terminal enclitic that is attached to the end of the verb (Example 3). The aspectual auxiliaries often combine with the adjacent Set A pronoun to form a contraction, e.g. *táan+in* →*tin*.

(3) *laili' u xòok-o'ob-e'*
still s3 study.APS-3PL-CONT.
'They (pl.) are still studying'.

There are three important features of verbs that determine how they are inflected: transitivity, the derivational processes undergone to achieve that transitivity (e.g. is the verb a transitive root or an intransitive/nominal/adjectival root that has become transitive via derivation), and voice (Maya has four distinct voice categories: active, passive, antipassive, and middle).

Intransitive verb stems often take one of a set of aspectual "status" suffixes[6] depending on the aspect and/or mood: *-Vl* suffix in the imperfective, where *V* matches the vowel in the root, a null suffix in the perfective, *-a'an* in the present perfect, and *-Vk* in the subjunctive.

Transitive verb stems in the active voice take aspectual status suffixes *-ik*, *-ah*, and *-mah* in the imperfective, perfective, and present perfect aspects, respectively. In the subjunctive mood, no suffix is added, unless the verb is phrase final, in which case it takes *-eh*.

The majority of root transitive verbs follow a CVC phonological template, which changes systematically to produce changes in voice: CVVC for

[6] Bohnemeyer (1998), Brody (2004), and others have referred to these suffixes as "status suffixes", and they go by various other names in the literature. In the OSEA-CITE pedagogical literature, these suffixes are referred to as "primary suffixes". We use the term "status" in this paper for the sake of consistency with previous linguistic work.

antipassive, CV'VC for passive, and CVVC for the middle voice. The status suffixes for these verbs are listed in Table 5. Transitive verbs can become reflexive with the addition of a suffix of the formula 'Set A + bah' (Example 4).

(4) *táan in wil-ik-**in-bah***
PROG S1SG.A see-STATUS-S1SG.A-REFL
'I am seeing myself.'

Intransitive roots can be transitivized with either the *-t* suffix or the causative *-s* suffix. They typically use the same status suffixes as transitive roots.

A third class of verbs with a distinct morphological pattern is that of Positional verbs. These verbs take status suffixes *-tal*, *-lah*, *-la'an*, and *-lak* in the imperfective, perfective, present perfect, and subjunctive, respectively.

Note that the discussion here is limited only to regular intransitive roots, regular transitive roots, and positionals. There are other verb root classes that follow slightly different inflectional patterns, but a complete description of them is beyond the scope of this paper.

4.3 Nouns and adjectives

Nouns and adjectives have notably less morphologically-complex than Maya verbs. They inflect for number, with the suffixes *-o'ob* and *-tak* (the latter for expressing a plurality of types vs. simply plural in number). Both Nouns and Adjectives can also behave as intransitive predicates, taking a Set B pronoun as the subject (Example 5. Commonly, Nouns that are core arguments of the verb can be topicalized by placing them at the front of the sentence with the topic suffix *-e*. Deixis can also be expressed using nominal morphology. Gender, while not a required feature of Nouns, can be indicated with the prefixes *x-* and *h-* (*x-* is also used as an instrumental nominalizer on verbs). Verbs can be derived from either nouns or adjectives using *-tal / -chahal* for intransitives (e.g. *ma'alob* "good" →*ma'alobtal* "to improve") and *-kuns / -kins* for transitives (e.g. *wíinik* "man" →*wíinikkunsik* "make someone into a man/human").

(5) *kòolnáal-o'on*
farmer-S1PL
'We are farmers'.

4.4 Phrase-level morphology

There are a number of cases of words in Maya which require a corresponding terminal suffix at some point later in the phrase. These include the negation marker *ma'a*, which typically requires that the end of the negated word or phrase have a *-i* suffix, certain aspectual auxiliaries like *laili'* which has a corresponding *-e* at the end of the verb phrase, and numerous other cases. Deictic suffixes *-a* "this", *-o* "that", and *-e* "this right here" also correspond to a prenominal article *le* (See Example 6).

Title	Sentences	Tokens
Simple Sentences	103	553
Tsikbalo'ob	200	1,099
Xkùuruch	85	710
Mam Ku'ukeba	41	376
Hun túul xnùuk òoch	11	129
Ch'och'lin yéetel síinik	11	166
Total	451	3,033

Table 4: A breakdown of the different works that make up the corpus.

(6) *ti' **le** yáax k'iin**o'***
ADP ART first day-DEM3
'At the beginning of that day'.

5 Data

For development, we use a small corpus consisting primarily of pedagogical texts used in the classroom by OSEA. They include lists of sentences and a number of *tsikbalo'ob* (dialogues). We also include four short stories from Bolles and Bolles (2001), for which we changed the orthography to reflect the writing norms of OSEA-CITE (with permission from the author). Sentence and token counts are listed in table 4.

6 Implementation

The morphological analyzer is developed within the Apertium project (Forcada et al., 2011; Khanna et al., 2021), and is made up of three principle components: a model of Maya morphotactics, a model of phonological processes, and an analysis disambiguation step. A sample of the type of analysis that is produced can be seen in Table 6.

One major advantage of using the Apertium platform is that a single morphological model can trivially be extended to additional applications, such as spell-checking and machine translation. Here, we describe the development of the morphological analyzer, and briefly discuss a spell-checking application prototype.

Aspect/Mood	Trans.	Intr.	Positional	Aps	Derived Trans Pss
Imperfective	*-ik*	*-Vl/Ø*	*-t-al*	Ø	*-a’al*
Perfective	*-ah*	Ø	*-l-ah*	*-nah*	*-a’ab*
Present perfect	*-m-ah*	*-a’an*	*-l-aha’an*	*-naha’an*	*-a’an*
Subjunctive	*-Ø/-eh*	*-Vk*	*-l-ak*	*-nak*	*-a’ak*

Table 5: Some of the common aspectual “status suffixes” (“primary suffixes”) for different types of Maya verbs. Trans. and Intr. refer to transitive and intransitive root verbs, “Aps” = Antipassive, and “Derived Trans Pss” refers to intransitive roots that are transitivized and then passivized (e.g. *hóoken* “I went out” →*a hóoksahen* “You took me out” →*hóoksa’aben* “I was taken out.”)

Word	**Analysis**
Ma’	"ma'" neg
ta	"t" aux pfv
	"a" s_sg2 pron
kaxtah	"kax" v tv pfv o_sg3
ba’al	"ba'al" n sg
hanteh	"han" v tv subj o_sg3

Table 6: An example of the output from our analyzer for an example sentence from the story “Ch’och’lin yéetel síinik”: *Ma’ ta kaxtah ba’al hanteh?* “You didn’t find something to eat?” The tagset corresponds with common abbreviations used in Apertium: neg=Negation, aux=Auxiliary, pfv=Perfective s_sg2=Second-person singular subject, pron=Pronoun, v=Verb, tv=Transitive, o_sg3=3rd-person singular object, n=Noun, sg=Singular, subj=Subjunctive.

6.1 Morphotactics

Morphotactics are defined using `lexc`. For verbs, we separate intransitive roots, transitive roots, and positionals. We encode lexical information about the root, e.g. whether an intransitive root takes the *-Vl* ending in the imperfective, in the lexicon entry. When a word undergoes derivation, we maintain the original lemma. For example, the CVC transitive root *xok* has in its lexicon entry the two additional voice derivations:

```
! Study, read
xok<v><tv>:xok TransActive;
xok<v><iv><aps>:xòok TransAps;
xok<v><iv><pss>:xo'ok TransPss;
xok<v><iv><mv >:xóok TransMed;
```

Each continuation lexicon reflects the specific set of status suffixes for the given root, aspect, and mood.

The lexicon entries for intransitive verbs also include lexical information, e.g. whether a given verb’s transitive derivation takes the transitivizer -*t*, the causative -*s*, or nothing.

Noun stems are optionally preceded by the gender/agentive prefixes *h-* or *x-*, and are followed by either the nominal inflections (e.g. diminutive, plural, possessive suffixes) or by denominalizing verbal morphology (e.g. the *-tal* / *-chahal* status suffixes).

Since the aforementioned terminal clitics can be appended to most words, each word optionally ends with them.

6.2 Phonology

Phonological processes are modeled with `twol` rules (Karttunen et al., 1987). As an example, take vowel harmony, a common process in Maya. In cases where a morpheme’s vowel harmonizes with that of the previous morpheme (e.g. the *-Vl* suffix for many intransitive roots), we represent these vowels as archiphonemes, and define the harmony process in `twol` as follows:

```
"Vowel harmony"
V:Vx <=> Vx [Cns | >:0 | ']+ >:0 _
; where Vx in UnaccVow ;
```

This component is also where we handle common contractions. For example, the intransitive verb *tàal* “come”, when transitivized with the causative *-s*, usually drops the last consonant in the root (*tàal-s-ik* →*tàasik*). There are a number of verbs for which this is the case, irrespective of which transitivizer they take. For these verbs, we represent the root with an archiphoneme (e.g. `{l}` as the last consonant of the root, which is surfaced as either ‘l’ or ‘Ø’).

6.3 Analysis disambiguation

Given the complexity of Maya morphology, our model of morphotactics often produces a number of potential analyses for the same form. As a simple example, take the second-person Set A pronoun *a*. This is used for both singular and plural subjects/possessors, and the number of the subject is determined by the presence or absence of the second-person plural suffix on the adjacent verb/noun. Similarly, the phrasal terminal suffixes *-i* and *-e* on a verb could signify negation, agreement with one of a subset of aspectual auxiliaries, or a locative analysis.

We use Constraint Grammar (Karlsson et al., 2011) to disambiguate analyses using the analyses and lemmas of words in the surrounding context. For example, to disambiguate the *a* Set A pronoun, we use the following rules:

```
REMOVE PRO + 2Sg IF (1 VN + SPl2);
REMOVE PRO + 2Pl IF (1 VN - SPl2);
```

Any time the Set A pronoun *a* is seen, it will include both plural and singular analyses. The first rule above removes the singular analysis if the following (right-adjacent) word is a verb with a second-person plural subject analysis. The second rule removes the plural analysis if the right-adjacent word is a verb without a second-person plural analysis.

The example above is one of a large number of Constraint Grammar rules needed to effectively narrow-down the morphological analyses using the surrounding words as context.

6.4 Spell checking

While the ability to automatically provide a morphological analysis is both interesting and valuable in itself, our system, thanks to the infrastructure set up by the Apertium project, is also easily extensible to a number of other applications. Here, we briefly discuss how we integrated the morphological analyzer to make a spell-checker and spelling-corrector for a word processor.

The use of finite-state models for efficient spell-checking of morphologically-rich languages has a long history (Beesley and Karttunen, 2003; Pirinen et al., 2014). As a prototype spell-checker and corrector, we use an FST which transduces incorrectly-spelled words within a fixed edit-distance to the words in our model. This FST can then be integrated with a spelling and grammar extension developed by the Voikko[7] project to be used with LibreOffice Writer[8], a free and open source, multi-platform word processor that is part of the LibreOffice suite of software[9]. Figure 3 shows a screenshot of the spell-checker in action. Its current status is a working prototype, but we plan to improve it by adding common misspellings to the model and weighting it using proofread written text.

6.5 Supporting variation in written Maya: normative and descriptive models

An important intended feature of our model is the ability to simultaneously support a normative model for pedagogical purposes, and a descriptive model for other natural language processing tasks. Specifically, the spell-checker, insofar as it is used by a language teacher to write pedagogical material or to encourage uniformity in writing practices among students, should adhere to the principles taught and followed by the educators. The morphological analyzer on the other hand, which can be used to help understand, analyze, or segment a Maya text from a number of potential sources/authors, should be flexible to common written variation in the language.

The Apertium platform allows for precisely this flexibility via "Direction" flags in our morphotactics file, and a `spellrelax` file. The "Direction" flags are simply commented annotations on a specific line in the `lexc` file that specify which direction that line should be included in at compile time. As an example, take the case of the nominal classifier. It is commonplace to see the number, such as *hun* "one", and the following nominal classifier, e.g. *p'éel* for inanimate nouns, written as a single orthographic word (in this case with nasal place assimilation): *hump'éel*. The OSEA program teaches its students to write these as two separate words: *hun p'éel*. Thus, we would like for our spell-checker to identify hump'éel as "incorrectly" spelled, while still recognizing this form in the analyzer so as to cover common variation in contemporary Maya writing. We can achieve this by including the annotation `Dir/LR` on the entry for this variant. This is a very minor example, but is one of many, and is illustrative of the type of flexibility we want to maintain in our system.

The `spellrelax` file allows for orthographic variation in the input of the morphological analyzer, and the ability to map it to the canonical written forms used in our lexicon. We use this file to support the large amount of orthographic variation that is characteristic of Maya writing.

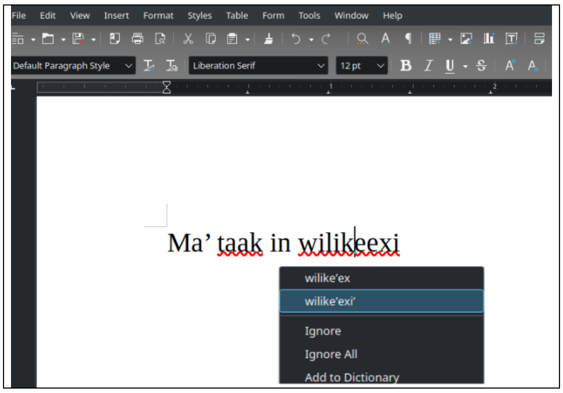
[7] https://voikko.puimula.org/

[8] https://www.libreoffice.org/discover/writer/

[9] https://www.libreoffice.org/

The following three lines illustrate how we handle (1) the common use of [j] where the OSEA orthography uses [h], (2) the omission of tone marking on long low vowels also characteristic of the contemporary INALI orthography but dispreferred for pedagogical purposes by OSEA, and (3) the use of [dz] for [ts'] in texts using the colonial style:

```
[ h (->) j ] .o. ! j for h
[àa (->) aa] .o. ! opt low mark
[ts' (->) dz] ! colonial ts'
```



Ma' taak in wilikeexi

Figure 3: Screenshot of spell-checking for Maya based on the analyzer discussed in this paper. The spellchecker correctly identifies the misspelling of "tàak" (which in our normative spelling requires marking of the long low vowel) and "wilike'exi'" in the incorrectly-spelled sentence "ma' taak in wilikeexi" "I don't want to see you (pl)." Note that the form *wilike'exi'* is not explicitly listed in a spelling lexicon. Instead, the analyzer contains the root verb *il*, and the morphological model enables the suggestion of the correct inflected form *w-il-ik-e'ex-i'*.

7 Coverage

On our modest-sized corpus, the morphological analyzer's coverage is about 96% on tokens and 85% on types. Of the forms currently not covered by the analyzer, many are interjections that may be author-specific (e.g. "kikiriki", the sound of a rooster crowing), and foreign loans (e.g. "cinco", "greedy"). Currently, all of the missed words by our analyzer are hapax legomena.

	N	Coverage (%)
Tokens	3,033	96
Types	734	85

Table 7: Current coverage of our analyzer on the corpus. All of the words not yet covered have a frequency of one.

8 Concluding remarks and future work

We have described in detail a finite-state morphological analyzer for Maya, and demonstrated its utility outside of merely performing morphological analysis by using the model to build a spell-checker.

For the near future, our first priority is growing the corpus. We are in the process of normalizing the orthography for a number of additional texts which we will then add and use to update the analyzer lexicon. Outside of simply improving the vocabulary and coverage of the analyzer, we plan to explore the numerous ways this tool can be of use to students by incorporating it into a browser-extension that aids the user's understanding of Maya texts read in the browser.

We also hope to improve the spell-checker by adding a better-informed error model that takes into consideration common spelling mistakes. Adding support for the spell-checker in other popular word processors is a longer-term goal, as this would greatly improve accessibility of the tool for teachers and students.

9 Acknowledgements

We are grateful to David Bolles for his permission to use, modify, and release his texts. We also extend a sincere thank you to Meesum Alam, Laura Merino Hernández, Héctor Figueroa, Matthew Fort, and Alex O'neil for their careful reading and thoughtful feedback of drafts of the present paper, and to the anonymous reviewers for their helpful comments.

References

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

Mourad Ben Slimane. 2008. *Appropriating new technology for minority language revitalization: The Welsh case*. Ph.D. thesis, Freie Universität Berlin.

Robert Wallace Blair and Refugio Vermont-Salas. 1965. *Spoken Yucatec Maya*. University of Chicago Library, Chicago.

Jürgen Bohnemeyer. 1998. *Time relations in discourse. Evidence from a comparative approach to Yukatek Maya*. Ph.D. thesis, Tilburg University.

David Bolles and Alejandra Bolles. 2001. *A grammar of the Yucatecan Mayan language*. Labyrinthos.

Victoria Reifler Bricker, Eleuterio Po ot Yah, and Ofelia Dzul de Po ot. 1998. *A dictionary of the Maya language: As spoken in Hocabá, Yucatán*. University of Utah Press.

Michal Brody. 2004. *The fixed word, the moving tongue: Variation in written Yucatec Maya and the meandering evolution toward unified norms*. The University of Texas at Austin.

Quetzil E Castañeda. 2014. Ko'ox Tsíikbal Màayat'àan: Beginning and Intermediate Level Maya. Volume One (Introduction Essential Grammar), Volume Two (Beginning Maya Workbook), Volume Three (Intermediate Maya Workbook).

Quetzil E Castañeda and Edber Dzidz Yam. 2014. Ko'ox Kanik Màaya: Grammar and Workbook for First Year Maya.

Richard Oliver Collin. 2010. Ethnologue. *Ethnopolitics*, 9(3-4):425–432.

Robert MW Dixon and Robert MW Dixon. 1994. *Ergativity*. Cambridge University Press.

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.

Michael Gasser. 2011. Computational morphology and the teaching of indigenous languages. In *Indigenous Languages of Latin America Actas del Primer Simposio sobre Enseñanza de Lenguas Indígenas de América Latina*, page 52.

Fred Karlsson, Atro Voutilainen, Juha Heikkilae, and Arto Anttila. 2011. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

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

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hèctor Alòs i Font. 2021. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.

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

Jonas Kuhn and B'alam Mateo-Toledo. 2004. Applying computational linguistic techniques in a documentary project for Q'anjob'al (Mayan, Guatemala). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Shannan Mattiace and Patricia Fortuny Loret de Mola. 2015. Yucatec maya organizations in san francisco, california: Ethnic identity formation across migrant generations. *Latin American Research Review*, 50:201 – 215.

Mike Maxwell and Jonathan D Amith. 2005. Language Documentation: The Nahuatl Grammar. In *Computational Linguistics and Intelligent Text Processing: 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005. Proceedings 6*, pages 474–485. 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.

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.

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.

Alexis Palmer, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation: A case study in creating interlinear texts for uspanteko. *Linguistic Issues in Language Technology*, 3.

Alexis Mary Palmer. 2009. *Semi-automated annotation and active learning for language documentation*. Ph.D. thesis, University of Texas, Austin.

Tommi Pirinen et al. 2014. *Weighted Finite-State Methods for Spell-Checking and Correction*. Ph.D. thesis, Helsingin yliopisto.

Christian Prager, Nikolai Grube, Maximilian Brodhun, Katja Diederichs, Franziska Diehr, Sven Gronemeyer, and Elisabeth Wagner. 2018. 5 the Digital Exploration of Maya Hieroglyphic Writing and Language. *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*, pages 65–83.

Robert Pugh and Francis 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 Volume 1 (Papers)*, pages 80–85.

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

Ivy Richardson and Francis M Tyers. 2021. A morphological analyser for K'iche'. *Procesamiento del Lenguaje Natural*, 66:99–109.

SEP & INALI. 2014. *U nu'ukbesajil u ts'íibta'al maayat'aan (Normas de escritura para la lengua maya)*.

Rustam Shadiev and Mengke Yang. 2020. Review of studies on technology-enhanced language learning and teaching. *Sustainability*, 12(2):524.

Yoshida Shigeto. 2011. *Guía gramatical del maya yucateco para los hispanohablantes*. Ministerio del Educación, ciencia y Cultura del gobierno japonés, Japón.

Gerardo Sierra Martínez, Cynthia Montaño, Gemma Bel-Enguix, Diego Córdova, and Margarita Mota Montoya. 2020. CPLM, a parallel corpus for Mexican languages: Development and interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2947–2952, Marseille, France. European Language Resources Association.

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.

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 Henderson. 2021. A corpus of k'iche'annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20.

Francis Tyers and Nick Howell. 2021. A survey of part-of-speech tagging approaches applied to k'iche'. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 44–52.

Cristina Vertan and Christian Prager. 2022. From inscription to semi-automatic annotation of maya hieroglyphic texts. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 114–118, Marseille, France. European Language Resources Association.

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.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. *arXiv preprint arXiv:2204.11909*.