Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 99-109 recibido 2-12-2020 revisado 22-01-2021 aceptado 31-01-2021

A morphological analyser for K'iche'

Un analizador morfológico para el idioma k'iche'

Ivy Richardson,[1] **Francis M. Tyers,**[12]
[1] Department of Linguistics, Indiana University, Bloomington, IN
[2] National Research University Higher School of Economics, Moscow
{ivrichar, ftyers}@iu.edu

Abstract: This paper describes the development of a free/open-source computational morphological description for K'iche', a Mayan language spoken in Guatemala. The language is of the agglutinative morphological type, with both prefixing and suffixing morphology. Both the nominal and verbal morphology are moderately complex. K'iche' is under-resourced and this is the first publication describing a computational tool for the language, and one of the first publications describing a computational tool for any language of the Mayan group. We use the Helsinki Finite-State Toolkit (HFST) for implementing the finite-state transducer. An automatic evaluation of the coverage of our implementation shows that the coverage is adequate, between 86% and 96% on range of freely available corpora. A manual evaluation gives a recall of over 90% over a hand-annotated test set. Both the analyser and the hand-annotated test set are available under a free/open-source licence.
Keywords: k'iche', morphological analysis, finite-state.

Resumen: Este artículo describe el desarrollo de un modelo computacional de la morfología quiché. La lengua quiché es una lengua maya que se habla en Guatemala. Es un idioma del tipo aglutinante con morfología de prefijos y sufijos. Tanto la morfología verbal como la morfología nominal son complejos a un nivel moderado. El quiché es una lengua de pocos recursos computacionales y esta publicación es la primera que describe una herramienta computacional para el idioma, y alguna de las primeras para cualquier lengua del grupo maya. La herramienta está desarrollada con HFST, una caja instrumentos para implantar transductores de estados finitos. Una evaluación indica que la cobertura de vocabulario está adecuada, entre 86% y 96% calculado sobre diversos corpus libres. Una evaluación manual indica una sensibilidad por 90% sobre un conjunto de pruebas anotadas a mano. Tanto el analizador como el conjunto de pruebas están disponibles bajo una licencia de software libre.
Palabras clave: quiché, análisis morfológico, transductores de estados finitos.

1 Introduction

3 This paper describes a morphological analyser for K'iche', a Mayan language spoken in southwestern Guatemala. Though K'iche' is the most widely spoken indigenous language in Guatemala, with nearly one million native speakers as of 2002 (INE, 2018) it is still categorised as a threatened language by the UNESCO *Atlas of the World's Languages in Danger* (Moseley, 2010).

K'iche' is a language with highly inflectional and derivational verbal morphology, which makes morphological analysers vital for both further computational research and the creation of tools such as spell-checkers for K'iche' speakers. Morphological analysers can both generate and analyse words based on a set of morphological and morphographemic rules and a list of lexemes for a language. Our morphological analyser is based on finite-state technology, which is able to map between surface forms, e.g. *nutinamit* 'my town' and lexical forms, e.g. `<px3sg>tinamit<n>` 'SG3-town-N'.

For the creation of the morphological analyser, we chose to base our analyser on the Helsinki Finite-State Toolkit (Lindén et al., 2011) due to its support for weighted finite-state transducers and the `twol` formalism (Koskenniemi, 1983). We took a freely available K'iche'–English dictionary, converted it into a machine readable format, and then converted the words into HFST-compatible lexemes. We then input morphophonemic rules from existing K'iche' grammars and

ISSN 1989-7553. DOI 10.26342/2021-66-8

teaching resources.

Most resources covering K'iche' are meant to be used as pedagogical tools rather than as sources for linguistic research, so they frequently did not cover in-depth the morphological and phonological rules of the language. In creating the analyser, we found that many aspects of the verbal morphology were not addressed in sufficient detail. In order to fill this gap, we present a diagramme of K'iche' verbal morphotactics.

The remainder of the paper is laid out as follows: section 2 describes the grammar of K'iche', section 3 describes prior computational work on K'iche' and other Mayan languages, section 4 describes the methodology of completing the analyser, section 5 provides an evaluation of the analyser, and looks qualitatively at the remaining issues. Sections 6 and 7 describe some future directions and offer some concluding remarks.

2 K'iche'

K'iche' is a language within the Quichean-Mamean branch of the Mayan language family. As of the 2018 Guatemalan census, it is documented to have over 1.5 million native speakers, however the number is likely higher now and does not account for speakers in the diaspora. There are roughly 23 dialects of K'iche' spoken throughout southwestern Guatemala (cf. Figure 1). Our work is based primarily on the Christenson dictionary (Christenson, 2006), which is based on the West dialect spoken in Totonicapan and Momostenango, and the Ixcoy grammar (Ixchajchal Batz, Cumez, and López Ixcoy, 1996), which is based on the Central dialect spoken in Santa Cruz del Quiche.

K'iche', like other Mayan languages, follows ergative-absolutive alignment. The subject and object of a given sentence are marked within the verb using what are called 'set A' markers, for the ergative, and 'set B' markers, for the absolutive. Set A markers indicate the subject in transitive verbs, as well as possessors for nouns. Set B markers indicate the subject for intransitive verbs and the object for transitive verbs. In addition, both set A and set B markers have null morpheme when referring to a formal second-person, and the verbal form is followed by a formality marker (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). Table 1 gives the forms of the two sets of markers.

Verbs in K'iche' are inflected for aspect, subject, object, and voice. Verbal inflection consists of both prefixing and suffixing, although most inflectional verbal morphemes are prefixes. Finite verb forms may also contain infixes for incorporated movement. These morphemes indicate the direction of an action, for example towards or away from the speaker.

K'iche' follows a Verb-Object-Subject word order (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). Nouns are not inflected for case, so K'iche' relies on a fixed word order to indicate noun function. Instead of a copula that inflects, K'iche' places a set B marker and a certain particle in place of a verb. K'iche' has a complex set of voices, including a passive, an instrumental, and various forms of antipassive. These voices are depicted through a complex system of verbal inflection (cf. Figure 2).

Most nouns that refer to human beings (as well as some nouns that refer to animals) inflect for plurality, while all inanimate nouns do not. All nominals inflect for possession (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). K'iche' contains a class of nouns called relational nouns, which are used to introduce purpose clauses, show causation, and form the comparatives of adjectives, among other functions (Can Pixabaj, 2017). Relational nouns either carry Set A markers, which index the complement, or prepositions, or both.

2.1 Orthography

There is a recognised standard orthography for K'iche',[1] developed by the *Academia de las Lenguas Mayas de Guatemala* (AMLG), and most texts we have used are written in this orthography. However, the precise orthographical form, or 'spelling' of individual word forms can still vary greatly among resources and dialects, particularly with regards to vowels and the glottal stop. Some dialects distinguish between tense and lax vowels, with tense vowels lacking diacritics and lax vowels being marked with diaeresis, i.e. tense *a* vs. lax *ä*. The distinction is represented in the ALMG orthography, however many K'iche' speaking communities lack the distinction. Most of the resources we used were not written in dialects that distinguish

[1]The standard is defined in the *Acuerdo Gubernativo Número 1046-87* of the 23rd November 1987.

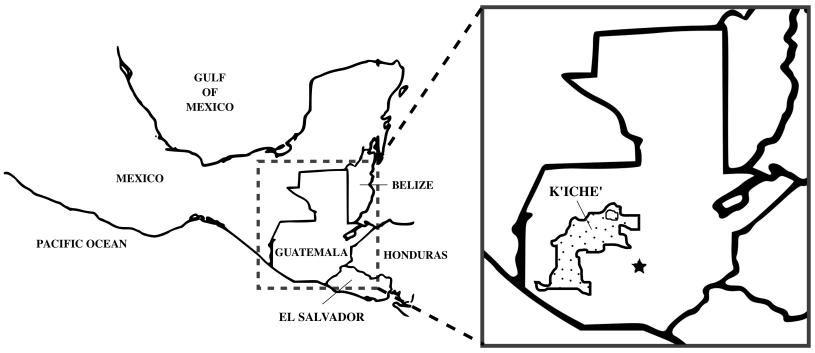

Figure 1: Dotted area represents approximate extent of the K'iche' speaking area in Guatemala.

	Singular			**Plural**		
	1	2	3	1	2	3
Set A	*nu-*, *inw-*	*a-*, *aw-*	*u-*, *r-*	*qa-*, *q-*	*i-*, *iw-*	*ki-*, *k-*
Set B	*in-*	*at-*	*ø-*	*oj-*	*ix-*	*e-*

Table 1: The Set A (Ergative) and Set B (Absolutive) person and number agreement markers for K'iche'. Set A markers are used on nouns to indicate possession and on verbs to indicate a transitive subject, and Set B markers are used on nouns for predication and on verbs for transitive object or intransitive subject. The third person singular Set B marker is null. The Set A markers have phonological variants before consonants (on the left) and vowels (on the right).

between tense and lax vowels, however the dictionary by Christenson (2006) does. We retain the diaereses where they are available in the original resource, but in order to deal with the variation, we implemented a spell relaxer module along with the analyser to accept input both with and without diaereses. The spell relaxer is implemented as a set of finite-state optional replacement rules. These are composed with the surface side of the transducer to produce the final analyser.

Another difference is between short and long vowels. Some works, e.g. Can Pixabaj (2017), indicate this distinction orthographically, by writing short vowels a single time and long vowels twice, like 'i' or 'ii' for /i/ and /iː/ respectively, but we found that most works do not make an orthographic distinction regarding vowel length.

The character for glottal stop /ʔ/ and for the ejective series of consonants is widely written using a number of punctuation symbols, i.e. ' U+0027 *Apostrophe*, ' U-2019 *Right Single Quotation Mark*, ´ U+00B4 *Acute Accent* and ‘ U+2018 *Left Single Quotation Mark*. We standardise on using the Unicode symbol ʼ U+02BC *Modifier Letter Apostrophe* and using the same spell relaxer module to accept input using any of the symbols.

3 Prior work

There is very little prior computational linguistic or natural-language processing work on K'iche', or any of the Mayan languages. Here we describe some of the research we have found. For K'iche', a limited analyser of verbs, containing 408 verb stems, was implemented as part of the Morfo project[2] (Gasser, 2009; Gasser, 2011), but was unpublished and has been unmaintained for over ten years (Gasser, p.c.). Kuhn and Mateo-Toledo (2004) describe some preliminary work on developing natural language processing tools for help in language documentation of Q'anjob'al, another Mayan language of Guatemala. They describe creating a basic finite-state grammar for the language, and train a maximum-entropy part-of-speech tagger on 4,100 tokens of manually annotated data. Their tagger gets an accuracy of 80% when doing 10-fold cross validation. Furthermore they describe some initial experiments in creating a language-model-based spellchecker. A prototype machine translation system from Spanish to Tseltal, a Mayan language of Chiapas in Mexico is

[2] https://github.com/hltdi/morfo

described in Morales Mancilla et al. (2011). The system takes a pipeline-based approach first analysing Spanish text lexically, looking up the Tseltal translations in a bilingual dictionary and then using a context-free grammar to generate Tseltal from Spanish.

4 Methodology

4.1 Lexicon

The lexicon was constructed both semi-automatically, using the dictionary by Chrinstenson (2006) and manually, based on a frequency list. When adding words manually we referred to two other dictionaries, the *Diccionario K'iche'–Español* (Conferencia Episcopal de Guatemala, 2011) and *K'iche' Choltzij* (Academia de Lenguas Mayas de Guatemala, 2004). It contains around 6,000 entries (see Table 2) categorised by part of speech and morphological paradigm.

4.2 Morphotactics

4.2.1 Nominals

Nominals in K'iche' may inflect for possession. To indicate possession, a Set A marker is added as a prefix to the possessed noun e.g. *nutinamit* 'my town' from *tinamit* 'town'. Some nouns also contain differences between their possessed and non-possessed forms. For example, the unpossessed word *kik'* 'blood' gets a suffix when possessed, for example *nukik'el* 'my blood' (Romero et al., 2018).

Relational nouns have functions similar to prepositions and some pronouns (e.g. object and direct object) in Spanish (Romero et al., 2018), but take possessive Set A markers just like nouns. Relational nouns must be possessed. Relational nouns can be combined with prepositions to form adpositional phrases. Phrases with the preposition *chi* 'to' and the relational noun *-ech*, which can have multiple definitions, are contracted to form a single word.

4.2.2 Verbs

As was previously mentioned, K'iche' verbs display a complex morphology. They inflect for the person and number of the subject (and object, in the case of transitive verbs), tense, and aspect. Additionally, they may contain infixes for incorporated movement.

There are three basic types of verb stems in K'iche': intransitive, root transitive, and derived transitive. In addition, there are morphemes of movement which can act either as intransitive verbs or infixes for incorporated movement alongside a verb stem. There are also positional stems, which function similarly to verbal stems.

Conjugated intransitive verbs contain a tense/aspect/mood prefix (hereon referred to as a TAM marker) (Can Pixabaj, 2017), a Set B marker for the subject, the intransitive stem, and a status suffix, depending on the verb's location within the phrase. Occasionally, a finite intransitive form will contain a movement morpheme between the Set B marker and the stem. Intransitive verbs cannot have passive/antipassive forms.

Root transitive verbs follow a consonant-vowel-consonant phonological structure e.g. *b'an* 'to do' (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). In their most basic active conjugated forms, they contain a TAM marker, a Set B marker for the object, a Set A marker for the subject, the verb stem, and an optional phrase-final suffix. Like intransitive verbs, transitive verbs can contain a movement morpheme, although this goes between the Set A and Set B markers when both are present.

The other type of transitive verbs, derived transitive verbs, function like root transitive verbs in many cases. However, derived transitive verbs are typically longer that root transitive verbs and have infinitive forms that end in *-j* e.g *ch'ab'ej* 'to talk to'. The verb stem can be derived from the infinitive form by removing the *-j*. The verb stem for *ch'ab'ej* would be *ch'ab'e*. Basic active forms of derived transitive verbs are similar to root transitive active forms, but they add a *-j* after the stem and do not take phrase-final suffixes. Like root transitive verbs, derived transitives can contain a movement morpheme between the Set A and Set B markers (Ixchajchal Batz, Cumez, and López Ixcoy, 1996).

The valency of transitive verbs in K'iche' can be reduced in the case of passive and antipassive forms. In these cases, the subject and object respectively may still be expressed with the use of an adpositional phrase.

As was previously mentioned, verb forms may contain incorporated movement (cf. Figure 2). In the imperative mood, the TAM marker differs between verb forms with incorporated movement and verb forms without incorporated movement.

Other voices in K'iche', such as passive,

Word class	Subclasses	Entries	Word class	Subclasses	Entries
Nouns	14	2,642	Numerals	2	42
Verbs	3	1,777	Conjunctions	2	21
Adjectives	6	701	Pronouns	3	19
Proper nouns	5	365	Directionals	–	14
Prepositions	–	102	Determiners	–	4
Adverbs	2	155	Other	–	142

Table 2: The lexicon split by word class, there are a total of 5,984 lexical entries in the file, including contractions.

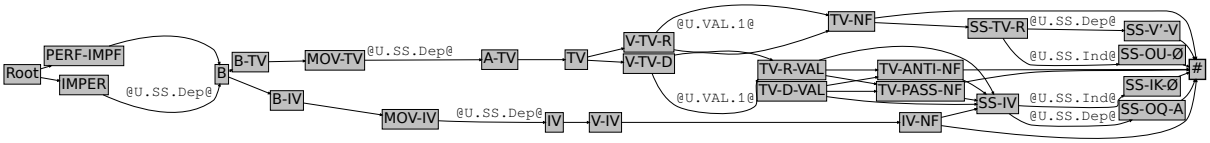

Figure 2: A graph of continuation classes modelling K'iche' verbal inflection. The node labels are the names of the continuation lexica, for example `PERF-IMPF` for the first level aspectual prefixes, *k-* and *x-*, `IMPER` for the imperative prefixes, *ch-*, and *j-*, and `TV` for transitive verb stems. The arc labels are flag diacritics which control non-adjacent morphotactic constraints. For example, in the `IMPER` and `PERF-IMPF` lexica a status suffix variable is set, either dependent, `Dep` or independent, `Ind` and this variable is used to choose the correct status suffix at status suffix lexica word finally. The graph has been lightly simplified for presentation reasons.

antipassive, and imperative, are indicated by differing TAM markers. The passive voice is used to express situations where the agent of an action is unknown/irrelevant (Can Pixabaj, 2017). In passive voice, the subject is omitted and the object is expressed with a Set B marker. The antipassive voice is used to express situations where the recipient of an action is unknown/irrelevant. Similarly to the passive voice, when forming an antipassive verb form, the object is omitted and the subject is expressed with a Set B marker (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). Both antipassive and passive verb forms can contain inherent movement morphemes, but the movement always refers to the semantic agent (Romero et al., 2018).

Participles can be formed from any verb type, although the derivation process is slightly different for each verb type. For intransitive verbs, the participle is derived by adding *(i)naq* to the end of the stem. For both types of transitive verbs, the participle is formed by adding either *um* or *om* to the end of the stem, depending on the root vowel (Romero et al., 2018).

4.2.3 Other categories

There are two prepositions in K'iche', *chi* (approximately 'to') and *pa* (approximately 'in'). These are often contracted with relational nouns to form complex adpositions. For example, *chirij nutinamit* 'about my town' (lit. *chi-* 'to', *-rij* 'its-back', *nutinamit* 'my-town'). K'iche' adverbs do not inflect. Adverbs can be used to introduce purpose, temporal, reason, manner, and conditional clauses (Ixchajchal Batz, Cumez, and López Ixcoy, 1996).

4.3 Morphophonology

We used morphographemic rules in the *twol* formalism to model the phonological alternations. This formalism was first proposed by Koskenniemi (1983) and consists of finite-state constraints over possible input–output string pairs. These constraints are applied in parallel via the composition operator and the output of each of the constraints is intersected. There were a total of 10 rules, covering the preconsonantal and prevocalic forms of the agreement markers, and vowel harmony processes in suffixes. Figure 3 gives two rule examples.

4.4 Example output

Figure 4 gives the output of the system for a sentence from the Universal Declaration of Human Rights. The sentence has been analysed and tokenised by the analyser. Each token starts with a circumflex `^`. This is followed by the surface form and a forward slash

```
"Vowel harmony in root transitive status suffix"
%{U%}:Vx <=> Vy [ ? - Vow ]* %>: _ ;
      where Vx in ( u o )
            Vy in ( u o )
      matched ;
```

```
"Third-person singular possessive alternation"
%{r%}:r <=> _ %>: Vow ;
```

Figure 3: Two phonological constraints: The first deals with the vowel harmony in the status suffix applying to root transitive verbs. The second deals with the form of the third person possessive, which is *u-* preceding consonants and *r-* preceding vowels. Archiphonemes are encoded with `{…}` and `%>` is the symbol for morpheme boundary.

```
^Maj/maj<adv>$
^jun/jun/<det>/jun<adj>/jun<n>/jun<num>$
^winaq/winaq<n>$
^ya'tal/ya'tal<adj>$
^ta/taj<neg>$
^chech/chi<pr>+<px3sg>ech<n><rel>$
^xaq/xaq<adj>/xaq<adv>/xaq<n>/xäq<n>$
^k'ate'/k'ate'<pr>$
^kachapatajik/<impf><s_sg3>chap<v><tv><pass><stat>+ik<mark>$
^,/,<cm>$
^xuq/xuq<adv>$
^kokisax/<impf><s_sg3>okisaj<v><tv><pass>$
^pa/pa<pr>$
^che'/che'<adj>/che'<n>$
^we/we¹<cnjsub>/we²<cnjsub>/we<det>$
^maj/maj<adv>$
^umak/<px3sg>mak<n>$
^ub'anom/<s_sg3>b'an<v><tv><pp>$
^;/;<sent>$
^xuqe/xuqe<adv>$
^kelesaxik/<impf><s_sg3>elesaj<v><tv><pass>+ik<mark>/<s_pl3>elesaj<v><tv><pass><inf>$
^,/,<cm>$
^xuq/xuq<adv>$
^koqatax/<impf><s_sg3>oqataj<v><tv><pass>$
^b'i/b'i<adv><dir>/b'i<n>$
^chupam/chi<pr>+<px3sg>pam<n><rel>/chup<v><iv><inf>$
^pa/pa<pr>$
^ri/ri<det>/ri<cnjsub>$
^utinamit/<px3sg>tinamit<n>$
^./.<sent>$
```

Figure 4: Example output from the analyser for Article 9 of the Universal Declaration of Human Rights, *Maj jun winaq ya'tal ta chech xaq k'ate' kachapatajik, xuq kokisax pa che' we maj umak ub'anom; xuqe kelesaxik, xuq koqatax b'i chupam pa ri utinamit.* 'No one shall be subjected to arbitrary arrest, detention or exile'.

`/`. This is then followed by sequence of analyses delimited by forward slashes. The token ends with the dollar sign `$`. The analysis is composed of a lemma and a sequence of morphological tags which are surrounded by `<` and `>` symbols. The tagset used is based on that of the Apertium project (Forcada et al., 2011). A single token may be split using the `+` symbol, as in the case of contractions, e.g. *chech* is split into *chi* 'to' and a form of the relational noun *-ech* 'to, for'. Tokens are delimited with `^` and `$`, tags are encapsulated by `<` and `>` and contractions are split using the `+` symbol. The tags used are given in Table 3

5 Evaluation

We have evaluated the analyser in three ways. First we calculate the *naïve* coverage over a range of corpora to determine how many tokens receive at least one analysis.[3] Then we manually annotate a subset of 100 tokens and

[3]We consider this naïve as a token is counted as *covered* if it receives a single analysis, however it may not receive all potential analyses and the analysis it receives may not be correct.

Tag	Description	Tag	Description
<adj>	Adjective	<num>	Numeral
<adv>	Adverb	<pass>	Passive
<cm>	Comma	<pass><stat>	Stative passive
<cnjsub>	Subordinating conjunction	<pp>	Perfect participle
<det>	Determiner	<pr>	Preposition
<dir>	Directional	<px3sg>	Possession, 3rd pers. sing.
<impf>	Imperfective	<rel>	Relational
<inf>	Infinitive	<sent>	Sentence marker
<iv>	Intransitive	<s_pl3>	Subject agreement, 3rd pers. plur.
<mark>	Marker	<s_sg3>	Subject agreement, 3rd pers. sing.
<n>	Noun	<tv>	Transitive
<neg>	Negative	<v>	Verb

Table 3: The list of tags used in the analysis in Figure 4 with their descriptions. This is a subset of the full tagset.

calculate the precision and recall. Finally we analyse a randomly selected sample of tokens which do not receive any analysis and categorise the errors.

5.1 Corpora

The analyser was developed principally using the K'iche' translation of the New Testament, *Ru Loq' Pixab' Ri Dios* (Wycliffe Bible Translators, 2011). This was chosen as it was both the largest single text and also fairly orthographically and dialectally consistent. For this reason coverage of the Bible is likely to be better than texts found 'in the wild'. To account for this we also calculated coverage over two texts which we did not use in developing the analyser. The first was the K'iche' translation of the Law on Access to Public Information of Guatemala, *Q'atojtzij rajilib'al 57-2008* (Gobierno de Guatemala, 2008) and the second was a collection of traditional stories, *Tzijob'elil K'aslemal* (Tol Ciprián et al., 2016).

5.2 Naïve coverage

Our first method of evaluation was to calculate the naïve coverage and mean ambiguity on freely available corpora. Naïve coverage refers to the percentage of surface forms in a given corpora that receive at least one morphological analysis. Note that forms counted by this measure may have other analyses which are not delivered by the transducer. The mean ambiguity measure was calculated as the average number of analyses returned per token in the corpus. The results can be found in Table 4.

5.3 Precision and recall

In order to test the precision and recall of the analyser we used a test corpus created from sentences from the *Chqeta'maj le qach'ab'al K'iche'!* course (Romero et al., 2018). We first copied all the example sentences and analysed them using our transducer. We then went through and added missing analyses and removed invalid analyses according to the translations and glosses. This gave us a disambiguated corpus of 337 sentences where each of the 2,021 tokens was assigned the appropriate analysis in context.

To calculate precision and recall, for each of the tokens in the corpus we collected the valid analyses and made a gold standard where each token was associated with a list of valid analyses.

We define true positives, tp, as those analyses which were in both the transducer's output and in the gold standard list of analyses. We define false positives, fp, as those analyses that were in the transducer output but not in the gold standard list of analyses. And we define false negatives, fn as those analyses which were in the gold standard list, but not in the transducer output. Tokens which received no analyses were counted as false positives. We defined precision, P (1), recall, R (2) and F_1-score (3).

$$P = \frac{\text{tp}}{(\text{tp} + \text{fp})} \tag{1}$$

$$R = \frac{\text{tp}}{(\text{tp} + \text{fn})} \tag{2}$$

Corpus	Genre	Tokens	Coverage	Average ambiguity
Ru Loq' Pixab' Ri Dios	Religion	206,827	95.49	1.55
Q'atojtzij rajilib'al 57-2008	Legal	18,853	90.69	1.89
Tzijob'elil K'aslemal	Folklore	5,477	86.89	1.49

Table 4: The naïve coverage of the analyser over a range of texts and text types.

	Precision	Recall	F_1-score
Tokens	76.53	98.22	86.03
Types	67.98	93.17	78.61

Table 5: Precision, recall and F_1-score for the test set. The metrics are substantially higher for tokens as more frequent tokens appear more frequently in the evaluation corpus and exhibit more of the valid analyses.

$$F_1 = 2\frac{PR}{P+R} \tag{3}$$

Intuitively, precision is the likelihood of an analysis presented by the transducer being an analysis found in the gold standard, while recall is the likelihood of an analysis found in the gold standard being in the transducer. Table 5 presents the results of the evaluation.

Note that this method is only an approximation of the precision and recall of the analyser as the corpus may not contain all valid analyses for a given token. For example, the corpus has several mentions of the word *juyub'* as a noun 'mountain', but the lexicon also contains an entry as an adjective meaning 'steep'. Thus completeness in the lexicon will be penalised by the precision metric.

A more thorough evaluation would be to ask a native speaker to supply all and only the valid analyses for a given token with the aid of a concordance for each token.

We inspected the list of false negatives and found that there were some errors which were repeated. For example, in the gold standard the form of the first person singular set B pronoun, *-in-* was assimilated with a following nasal to *-im-* as in the verb form *kimb'e* 'I go', instead of the form *kinb'e*. This assimilation was not found in any of the other corpora we used and accounted for over a quarter of all false negatives. Another phonological difference between our gold standard and the other corpora we used was the form of the antipassive *-Ow-* after the verb *-to'-* 'help'. In the test corpus the suffix vowel was deleted leaving only the *-w-* of the suffix. There were also a number of idiosyncratic forms of the verb *-aj-* 'want', and some forms of other verbs which did not follow the regular patterns. A more thorough study of phonological variation would allow us to resolve these errors.

5.4 Unanalysed forms

In addition to the previous methods, we have also done an evaluation of forms (types) that do not receive any analysis, sorting them into five categories: missing stem, morphotactic error, morphophonological error, orthographic variation and tokenisation error. These forms were selected pseudo-randomly[4] from a concatenation of all of the evaluation corpora.

As can be seen from Table 6, missing stems make up the bulk of the errors. The coverage of the available corpora is impressive, given the small size of the lexicon, but there is a lot of lexicographic work to continue with. The largest number of missing stems was found in the categories of verbs and nouns.

In terms of morphotactic error, we count incorrect paradigm assignment, missing or incorrectly formed morphemes, mistakes in the way the continuation lexica are linked together, or mistakes in use of flag diacritics. For example, the wordform *ech'oko'ib'* 'cripples' was not analysed because the word *ch'oko'* 'cripple' was not assigned to the paradigm of words that have a plural form in *-ib'*.

We consider as orthographic variation any word that is equivalent to one already in our lexicon but with a different orthographic form. This does not imply any judgement as to normativity of the form, and items counted in this category could be anything from dialectal variation to typographical errors. For example, the word *rajilib'al* 'A3SG-date' appears in our lexicon as the entry, [**ajilabal** *(n)* date (calendar); number] — with the

[4] Using the GNU *sort* utility.

Error category	Frequency	Percentage (%)
Missing stem	65	61.9
- Verb	34	32.3
- Noun	18	17.1
- Proper noun	7	6.6
- Adjective	4	3.8
- Other	2	1.9
Orthographic variation	24	22.8
Morphotactic error	10	9.5
Morphophonological error	3	2.8
Tokenisation error	3	2.8
Total:	105	100

Table 6: Proportion of errors by category. Note that although there were only 100 words selected, the number of errors adds up to more than 100 as some words evinced more than one kind of error. For example if a word was both written in a way not found in our lexicon and in generation of the form from our lexicon there was a phonological error, we counted it in both categories.

vowel *a* in place of *i* before the *-b'al* (instrumental, locative) derivational suffix. In another instance we found the wordform *jastasq*, where our lexicon contains [**jastaq** *(n)* things; goods].

6 Future work

There are a number of avenues for future work. First of all we intend to fix all of the errors that we found during the evaluation. Secondly, the lexicon certainly needs to be expanded, both in terms of lexemes and dialect coverage, and improved for consistency in labelling forms as to the dialect they pertain to. We have included lexical information from a number of different dialects and although some of the entries are marked, it was not possible to mark all of them.

There are certain lacunae in terms of non-finite verb forms, it is not clear to us how the various infinitives should be categorised.

Given the fairly high ambiguity exhibited we would like to work on disambiguation for K'iche'. We have started work on manually disambiguating texts and would like to use the analyser as groundwork for bootstrapping a treebank for K'iche' under the Universal Dependencies project (Nivre et al., 2020).

The analyser can also be used to generate training data for machine learning applications, morphological analysis using data generated from finite-state machines to train neural networks has already been used in e.g. (Silfverberg and Tyers, 2019).

In terms of applications, we forsee that this work could be used in developing spellchecking and predictive text software that supports K'iche' as well as providing the basis of further language technology.

7 Concluding remarks

We have presented the first morphological analyser for K'iche', a Mayan language principally spoken in Guatemala. The analyser comprises a finite-state transducer based on the Helsinki Finite-State Tools. It covers a reasonably high percentage — 90–96% of tokens in running text over a number of freely available corpora of K'iche'. The analyser is available as free/open-source software under the GNU General Public Licence.[5]

Acknowledgements

We would like to express our thanks to Allen Christenson for the use of his K'iche'–English lexicon as the initial lexical base of the analyser. We would also like to thank Robert Henderson and Pedro Mateo Pedro for fruitful discussions about the analyser, and the anonymous reviewers for their helpful suggestions. This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

[5] https://github.com/apertium/apertium-quc/

References

Academia de Lenguas Mayas de Guatemala. 2004. *K'iche' Choltzij*. ALMG. 2nd Edition.

Can Pixabaj, T. A. 2017. K'iche'. In J. Aissen, N. C. England, and R. Zavala Maldonado, editors, *The Mayan Languages*. Routledge, Oxford.

Christenson, A. 2006. K'iche'–English dictionary. `http://www.famsi.org/mayawriting/dictionary/christenson/index.html`.

Conferencia Episcopal de Guatemala. 2011. *Diccionario K'iche'–Español*.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Gasser, M. 2009. Semitic morphological analysis and generation using finite state transducers with feature structures. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 309–317, Athens, Greece, March. Association for Computational Linguistics.

Gasser, M. 2011. Computational morphology and the teaching of indigenous languages. In S. Coronel-Molina and J. McDowell, editors, *Proceedings of the First Symposium on Teaching Indigenous Languages of Latin America*, pages 52–63, Indiana University, Bloomington.

Gobierno de Guatemala. 2008. Q'atojtzij rajilib'al 57-2008: Q'atb'altzij re ukujik che uya'ik ub'ixik uwach tinamit. *[Decreto Número 57-2008: Ley de acceso a la información pública]*.

INE. 2018. XII Censo Nacional de Población y VII de Vivienda. `http://redatam.censopoblacion.gt/bingtm/RpWebEngine.exe/Portal?BASE=CPVGT2018`.

Ixchajchal Batz, E. A., L. M. Cumez, and C. D. López Ixcoy. 1996. *Gramática del idioma k'iche'*. Proyecto Lingüístico Francisco Marroquín, Guatemala.

Koskenniemi, K. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, Helsingin yliopisto.

Kuhn, J. and B. Mateo-Toledo. 2004. Applying computational linguistic techniques in a documentary project for Q'anjob'al (Mayan, Guatemala). In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisboa, Portugal.

Lindén, K., E. Axelson, S. Hardwick, T. Pirinen, and M. Silfverberg. 2011. HFST—framework for compiling and applying morphologies. *Communications in Computer and Information Science*, 100:67–85, 08.

Morales Mancilla, J. A., H. Guerra Crespo, G. B. Nango Solís, I. Valles López, and A. G. Cossio Martínez. 2011. Traductor del lenguaje español a la lengua tseltal. *Revista Tecnología Digital*, 1(1):27–39.

Moseley, C. 2010. Atlas of the world's languages in danger. `http://www.unesco.org/culture/en/endangeredlanguages/atlas`.

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

Romero, S., I. Carvajal, M. Sattler, J. M. Tahay Tzaj, C. Blyth, S. Sweeney, P. Kyle, N. Steinfeld Childre, D. G. Tambriz, L. E. Tambriz, M. Tahay, L. Tahay, G. Tahay, J. Tahay, S. Can, E. I. Xum, E. Guarchaj, S. M. G. Can, C. M. T. Cotiy, T. Can, T. Kingsley, C. Hayes, C. J. Walker, M. A. I. Sohom, J. Sandler, S. G. Ixmatá, M. P. Tahay, and S. Smythe Kung. 2018. Chqeta'maj le qach'ab'al K'iche'! `https://tzij.coerll.utexas.edu/`.

Silfverberg, M. and F. M. Tyers. 2019. Data-driven morphological analysis of nominal morphology for Uralic languages. In *Proceedings of the 5th International Workshop on Computational Linguistics of Uralic Languages*, pages 1–15.

Tol Ciprián, C., D. D. Oxlaj Tistoj, E. Velásquez Vicente, H. Calel Vicente, J. G. Calva Loarca, J. A. Vásquez Ajpop, J. J. Menchú Ordóñez, L. M. Calderón, M. Hernández Pocol, M. M. Batz Socop, O. O. Baten Sarat, R. L. Puac Yac, R. Gómez Pérez, S. C. Mejía Paxtor, S. Gómez Par, T. Castro Gutiérrez, and V. M. Alvarez Poncio. 2016. *Tzijob'elil K'aslemal*. USAID.

Wycliffe Bible Translators. 2011. *Ru Loq' Pixab' Ri Dios*. Wycliffe Bible Translators. `https://ebible.org/Scriptures/details.php?id=qucNNT`.