A corpus of K'iche' annotated for morphosyntactic structure

Francis M. Tyers
Department of Linguistics
Indiana University
ftyers@iu.edu

Robert Henderson
Department of Linguistics
University of Arizona
rhenderson@arizona.edu

Abstract

This article describes a collection of sentences in K'iche' annotated for morphology and syntax. K'iche' is a language in the Mayan language family, spoken in Guatemala. The annotation is done according to the guidelines of the Universal Dependencies project. The corpus consists of a total of 1,433 sentences containing approximately 10,000 tokens and is released under a free/open-source licence. We present a comparison of parsing systems for K'iche' using this corpus and describe how it can be used for mining linguistic examples.

1 Introduction

For some time, one of the fundamental resources for language technology has been a part-of-speech tagged (or morphologically annotated and disambiguated) corpus. Creating these resources has traditionally been a lengthy process, from defining an annotation scheme to collecting texts, training annotators and performing the annotation. Recently however advances in annotation schemes and end-to-end linguistic processing pipelines mean that the development of a single resource, a treebank can enable a whole pipeline of language analysis tools from tokenisation to dependency parsing from a single resource.

In this paper we describe the annotation of such a corpus for K'iche', a Mayan language of Guatemala and outline how the corpus can be used to train systems for linguistic annotation.

The remainder of the paper is laid out as follows: Section 2 gives a brief grammatical overview of K'iche'; Section 3 gives an overview of related work on K'iche' syntax; Section 4 describes the corpus and preprocessing steps; Section 5 describes the annotation process; Section 6 describes a range of syntactic constructions in K'iche' and how they were annotated. We evaluate parsing performance using the corpus in Section 7 and show how models trained on the corpus can be used in finding linguistic examples. Finally, we describe some future work (Section 8) and present some concluding remarks (Section 9).

2 K'iche'

K'iche' (ISO-639-3: `quc`, also *K'ichee'*, previously *Quiché*) 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 variants of K'iche' spoken throughout southwestern Guatemala.

K'iche' is a language with ergative-absolutive alignment, basic verb-initial order of constituents, and prefixes for agreement. The language is both prefixing (for inflection) and suffixing (for derivation and some inflection). Neither subject nor object need be overtly expressed when recoverable from context.

An important part of the K'iche' grammatical system are the sets of agreement markers. These are traditionally split into set A and set B. Set A, or the ergative (ERG) markers, are used on nouns to cross-reference, that is, agree with, their possessors and on verbs to indicate a transitive subject. Set B markers, or the absolutive (ABS) markers, are used to cross-reference the transitive object or intransitive subject. Table 1 shows the markers.

K'iche' verbal morpho-syntax, like other Mayan languages, is organised around transitivity. Root verbs, i.e., verbs of the form CVC, and their derived non-CVC counterparts are classified as either transitive or intransitive, and this classification has implications for the kinds of morphology the verb can take. It controls the distribution of Set A and Set B morphology that we have already seen, but it also constrains what kinds of nominalisations a verb stem allows (Can Pixabaj, 2009), a well as which 'Status Suffixes' a verb stem takes (see section 6.9

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 10–20

Person	Set A (ERG) _ C	Set A (ERG) _ V	Set B (ABS) _
SG1	*nu-*	*inw-*	*in-*
SG2	*a-*	*aw-*	*at-*
SG3	*u-*	*r-*	*ø-*
PL1	*qa-*	*q-*	*oj-*
PL2	*i-*	*iw-*	*ix-*
PL3	*ki-*	*k-*	*e-*

Table 1: The Set A and Set B 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, C and vowels, V. There are also formal forms which appear as a combination of one of the prefixes with a following particle, *lal* or *alaj*. The Set B first person plural morph may also be *uj-*.

for more discussion of this unique aspect of Mayan morphology).

While the basic word of K'iche' is VOS, all possible word orders are attested, conditioned by discourse factors, the most important of which are topic and focus. Focus involves marking the focused expression with a focus particle, and then preposing it to a position before the verb. Topicalisation involves morphologically unmarked preposing of the topicalised expression before the verb. If a clause contains both topicalised and focused expressions, the topic comes before the focus.

3 Related work

Broadly, this work is a corpus of K'iche' sentences, morphosyntactically analysed and annotated in a way to support downstream natural language processing tasks like machine translation, relation extraction, etc. While there are annotated corpora of K'iche', like the K'iche' segment of the *Oxlajuuj Keej Maya' Ajtz'iib'* Mayan Languages Collection (Oxlajuuj Keej Maya' Ajtz'iib', 2021) of Telma Can Pixabaj's 2018 annotated collection of ceremonial discourse in K'iche', these are not in easily parsable formats that can be fed directly into existing NLP pipelines. The nearest analogs to the work presented here are Sachse's 2016 XML standard for morphological annotations of Mayan languages, including K'iche', and Palmer's 2010 IGT-XML corpus of the related language Uspanteko. While parseable, and annotated with grammatical information like part-of-speech, these are not treebanks like the present work. In fact, ours is the first treebank of any Mayan language.

4 Corpus

The corpus is composed of sentences from a range of text types. Around two thirds are example sentences either from a published dictionary (Medrano Rojas, 2004) or from linguistic research (Can Pixabaj, 2015; Henderson, 2012). To this we added some language learning materials (Romero et al., 2018), and religious, medical and legal texts (Wycliffe Bible Translators, 2011; Wikimedia Incubator, 2017; Méndez López, 2020; Gobierno de Guatemala, 2009). The remainder was from a collection of folk tales (Ministerio de Educación, 2016a,b). The majority of the texts came with a translation either in Spanish or in English. Some texts, such as the linguistic examples additionally came with interlinear glosses. For the texts that did not have translations, we performed a rough-and-ready glossing into Spanish with the aid of a prototype machine translation system.[1]

The texts were chosen for their availability and for the range of linguistic phenomena they exhibited, as one of the aims of the work was to create annotation guidelines that can be used in further annotation and adapted to other Mayan languages, this was an important consideration.

4.1 Preprocessing

The texts were preprocessed using a freely-available finite-state morphological analyser (Richardson and Tyers, 2021). The morphological analyser returned, for each token the set of possible morphological analyses, including multiple output tokens in the case of contractions. These analyses were then disambiguated by hand, and missing analyses added.

This disambiguated output was then converted to the ten-column CoNLL-U format.[2] Morphological tags were converted to Feature=Value pairs by using a deterministic maximum-set-overlap matching algorithm.

5 Annotation process

The annotation guidelines are based on Universal Dependencies (Nivre et al., 2020), an international

[1] apertium-quc-spa: https://github.com/apertium/apertium-quc-spa

[2] https://universaldependencies.org/format.html

Source	Description	Sentences	Words	Avg. length
Medrano Rojas (2004)	Dictionary examples	657	4081	6.21
Romero et al. (2018)	Language learning material	301	1838	6.11
Can Pixabaj (2015)	Linguistic examples	268	1612	6.01
Ministerio de Educación (2016a,b)	Folk tales	104	1525	14.66
Henderson (2012)	Linguistic examples	57	286	5.02
Wycliffe Bible Translators (2011)	Religious scripture	16	211	13.19
Wikimedia Incubator (2017)	Encyclopaedic text	12	213	17.75
Gobierno de Guatemala (2009)	Legal text	7	113	16.14
Méndez López (2020)	Medical guidance	6	87	14.50
Total:		1433	10002	6.97

Table 2: Composition of the corpus. It is notable, but unsurprising that the example sentences and learning materials are around three-times shorter than the other texts.

collaborative project to make cross-linguistically consistent treebanks available for a wide variety of languages. At time of writing, data for over 111 languages is available through the project in a standardised format and with a standardised annotation scheme.

We chose the UD scheme for the annotation as it provides pre-defined recommendations on which to base annotation guidelines. This reduces the amount of time needed to develop annotation guidelines for a given language, as where the existing universal guidelines are adequate, they can be imported wholesale into the language-specific guidelines.

The treebank was annotated by the first author and difficult cases were determined by discussion between the first author and the second author.

6 Constructions

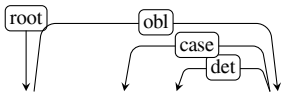
In the following subsections we describe some particular features of K'iche' that are interesting or novel with respect to the Universal Dependencies annotation scheme, and our approach to annotating them. Inline examples are given on three lines, with the original text, a segmentation showing the inflectional morphs, and an approximate translation in English. Glosses are provided when necessary for explaining some particular feature or construction.[3] Where contractions are split, the split is indicated with a hyphen on the both sides of the split, so for example *ch-* followed by *-we* should be read *chwe*.

[3]The following is a list of glossing tags: Question particle QST, Passive PASS, Perfective PERF – also called completive, Imperfective IMPF – also called non-completive, Negative NEG, Classifier CLF, Relative REL, Relational noun RELN, Active ACT, Antipassive AP, Status suffix SS, Directional DIR.

The focus is primarily on the relation between syntactic words, so for example constructions such as the morphological expression and annotation of agreement, tense-aspect-mood prefixes, incorporated movement, and possessive prefixes are not outlined here. It suffices to say that these are encoded with Feature=Value pairs.

6.1 Relational nouns

K'iche' has two prepositions with locative meaning *chi* 'in' and *pa* 'in, at, on, to, towards, from'. Following the guidelines these are attached using the `case` relation to their complement, as in (1).



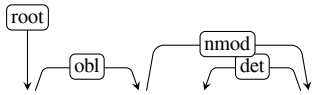
	Kinch'aw	*pa*	*le*	*ch'aweb'al.*
(1)	K-in-ch'aw	pa	le	ch'aweb'al.
	I speak	on	the	telephone.

All other adpositional phrases are made using either relational nouns or combinations of relational nouns with these two prepositions.[4] For readers familiar with Indo-European languages, these relational nouns are similar in function to nouns of the

[4]The fact that we can have relational nouns co-occurring with prepositions — cf. (4) overleaf — is a strong argument that they should not be treated as sharing the category preposition. Instead, bona fide prepositions take nouns as complements, including this special subclass of relational nouns which must bear agreement. Another argument for keeping prepositions and relational nouns separate concerns their behaviour under questioning. Relational nouns can undergo pied-piping with inversion—i.e., the question *ruk' jachin* 'with whom' can also be *jachin ruk'* lit. *whom with*. This inversion is impossible with simple prepositions, which is unexpected if they were structurally equivalent. We direct the reader to Svenonius (2006) for a crosslinguistic survey of preposition-like expressions that are not, in fact, prepositions.

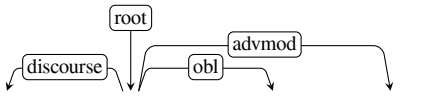
type *front*, *top*, *side* in English or *frente* 'front', *cima* 'top', *lado* 'side' in Spanish (e.g. *al **lado** de la casa* 'at the **side** of the house'). However, they are more extensive, used for encoding relations that in Indo-European languages are encoded with prepositions, such as *with*, *by*, *of*, etc. or even determiners or pronouns, e.g. *-onojel* 'all'.

Relational nouns agree with their complements using possessive markers (set B affixes) and may have an complement or not. For example, in (2) the relational noun *-uk'* 'with' is used with a complement *le nunan* 'my mother'.



	Kinch'aw	*ruk'*	*le*	*nunan.*
(2)	K-in-ch'aw	r-uk'	le	nu-nan.
	I speak	with	the	my mother.

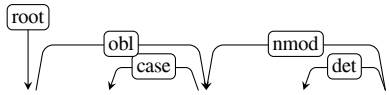
In (3) the same relational noun *-uk'* 'with' is used without a nominal complement.



	¿La	*katpe*	*quk'*	*chwe'q?*
(3)	La	k-at-pe	q-uk'	chwe'q
	QST	will you come	with us	tomorrow

To maintain language-internal consistency these are annotated with the relational noun as the head of the construction, attached to predicates with the `obl` oblique relation and to nominals with the `nmod` relation.

It is worth noting that relational nouns can also be used in conjunction with the true prepositions, as in for example (4).

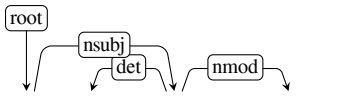


	…	*kyajon*	*chi*	*kech*	*ri*	*ak'alab'.*
(4)	…	k-yajon	chi	k-ech	ri	ak'al-ab'.
	…	tells off	PR	B3PL-RELN	the	children.

In this sentence, *[Ri ajtij,] kyajon chi kech ri ak'alab'.* "[The teacher,] tells off the children." (4), the relational noun *-ech* is introduced by the true preposition *chi*.

6.2 Nominal possession

In terms of nominal possession, K'iche' is a head marking language. The schema for possession is a noun with a possessive prefix followed by the possessor, POS-N_1 N_2 = N_2 of N_1. For example, *utzij ri ajq'ij* "the daykeeper's word" (lit. "his word the daykeeper".

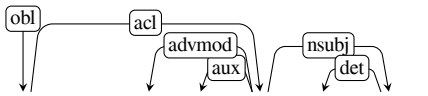


	K'ax	*ri*	*ub'aqil*	*nuq'ab'.*
(5)	K'ax	ri	u-b'aqil	nu-q'ab'.
	Bad	the	its bone	my arm.

Possession can also be expressed on multiple nouns in series, as in the sentence *K'ax ri ub'aqil nuq'ab'.* "The bones of my arms hurt" (5).

6.3 Relative clauses

Following Can Pixabaj (2021), relative clauses in K'iche' are post-nominal and come in two broad types, headed (6) and headless (7). For the headed example we can examine the sentence *[Osea pa taq wa' le] komunidades jawi e k'ow le winaq* "[That is to say that in these] communities where these people are in..." (Can Pixabaj, 2021, ex. 31).

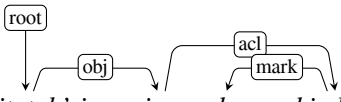


	…	*komunidades*	*jawi*	*e*	*k'ow*	*le*	*winaq*
(6)	…	komunidad-es	jawi	e	k'o-w	le	winaq

In headless relatives, the head becomes the relative itself and the verb is attached to it as an adnominal clause, as in the sentence *Kojtzalijoq jawi ri xojkanaj wi kan [junab'iir].* "Let's go back where we stayed [last year]." (Can Pixabaj, 2021, ex. 39)

	Kojtzalijoq	*jawi*	*ri*	*xojkanaj*	…
(7)	K-oj-tzalij-oq	jawi	ri	x-oj-kanaj	…
	Let's return	where	that	we stayed	…

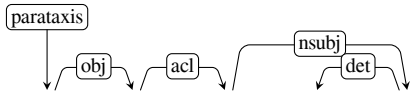
Relative clauses embedded under a head nominal, like (6), can be further split into those that contain an interrogative relative pronoun and those that contain a determiner acting as a subordinating conjunction. The reason for treating the latter as a subordinating conjunction and not a relative pronoun, pointed out by Bridges Velleman (2014), is that the two can co-occur, as in (8).



	Chitatab'ej	*jas*	*le*	*kimb'ij.*
(8)	Ch-ø-i-tatab'ej	jas	le	k-ø-im-b'ij.
	Listen	what	that	I say.

In (8), the relative clause *jas le kimb'ij*, lit. "what that I'm saying" is introduced by the interrogative relative pronoun *jas* which is given the relation of `object`. It is then followed by a relative clause

complementiser we give the `mark` relation. The predicate in the relative clause is then attached to the nominal it modifies with the relation `acl`, adnominal clausal modifier.

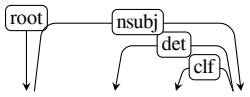


	…	*xuto*	*ri*	*xub'ij*	*ri*	*rati't*
(9)	…	x-ø-u-to	ri	x-ø-u-b'ij	ri	r-ati't
	…	listen	the	say it	the	her mum

In addition to headed and headless relatives, Can Pixabaj (2021) also discusses so-called light-headed relatives. In these, the noun head is usually modified by relative not expressed, leaving only a determiner. As shown in (9), in this case we promote the determiner as head of the construction, and treat the light-headed relative as adnominal clause modification (namely `acl`).

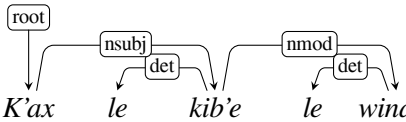
6.4 Non-verbal predicates

In non-verbal predication, for example with nouns or adjectives, the predicate is the root, and the subject, as the example *B'ixonel ri a Lu'* "Lu' is a singer" (10) and *K'ax le kib'e ri winaq.* "The road of the people is difficult" (11).



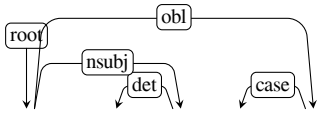
	B'ixonel	*ri*	*a*	*Lu'.*
(10)	B'ixonel	ri	a	Lu'
	Singer	the	CLF	Lu'

Note that there are three definite determiners in K'iche', *ri*, *le* and *we*. They are distinguished by degree of definiteness and familiarity and proximity/visibility to the speaker (Can Pixabaj, 2015).



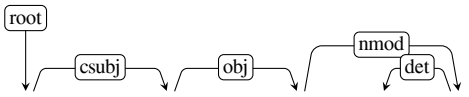
	K'ax	*le*	*kib'e*	*le*	*winaq.*
(11)	K'ax	le	ki-b'e	ri	winaq.
	Difficult	the	their road	the	people.

For existential sentences in the affirmative and in the negative, two non-inflecting words are used *k'o* in the case of existence and *maj* in the case of non-existence. In these constructions, the non-inflecting word is the head and the thing existing is the subject, as in *K'o jun tz'i' pa b'e.* "There is a dog in the street." (12)



	K'o	*jun*	*tz'i'*	*pa*	*b'e.*
(12)	K'o	jun	tz'i'	pa	b'e.
	There is	a	dog	in	street.

Another set of non-verbal predicates involve forms such as *rajawaxik* 'necessary', *k'ax* 'difficult' with verbal subjects. These are analysed as nominals (nouns or adjectives), and the complement is an embedded clausal subject.



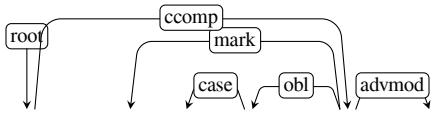
	Rajawaxik	*kqakoj*	*utzij*	*ri*	*Ajq'ij.*
(13)	Rajawaxik	k-ø-qa-koj	u-tzij	ri	Ajq'ij.
	Necessary	we listen	his word	the	Ajq'ij.

In this example *Rajawaxik kqakoj utzij ri Ajq'ij.* "We need to listen to the Ajq'ij."[5] (13) we see a non-verbal predict with a single argument which is itself a predicate.

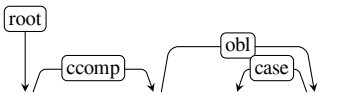
6.5 Complement clauses

Our analysis of complement clauses is based on research done by Can Pixabaj (2015), whose thesis gives a thorough treatment of the topic. This section is based on Chapter 3 of (Can Pixabaj, 2015, p.85). In K'iche', complements can be split into three subcategories: finite with complementiser, finite without complementiser and non-finite.

In UD, the distinction in complements is between those with obligatory control, `xcomp` and those without control, `ccomp`. Each of the three types defined in K'iche' may have control or not. In (14) the subordinate clause is introduced by a subordinator, while in (15) there is no subordinator.

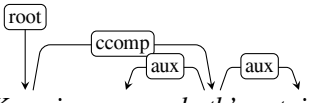


	Weta'm	*chi*	*p-*	*-ulew*	*xwar*	*wi.*
(14)	Ø-w-eta'm	chi	pa	ulew	x-war	wi
	I know	that	on	floor	he slept	



	Kawaj	*kimb'e*	*pa*	*tinamit.*
(15)	Ka-ø-w-aj	k-im-b'e	pa	tinamit
	I want	I go	to	village

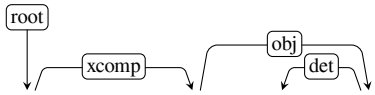
Although in (15) the subjects happen to agree, the fact that this is not a control construction can be seen in (16) where the subordinate clause has a subject not controlled by the matrix clause.



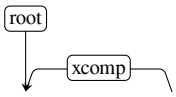
	Kawaj	*na*	*katb'e*	*taj*	.
(16)	Ka-ø-w-aj	na	k-at-b'e	taj	
	I want	NEG	you go	NEG	

[5] *Ajq'ij*, sometimes translated as 'daykeeper', a Maya spiritual guide or shaman-priest.

In (17) and (18) we see examples of obligatory control.



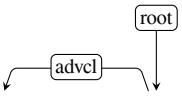
Xenutaqch'ij *kekil* *le* *ak'alab'.*
(17) X-e-nu-taqch'i-j k-e-k-il le ak'al-ab'
I forced them take care the children



Xuchap *nukunaxik.*
(18) X-u-chap nu-kuna-x-ik
She began to cure me

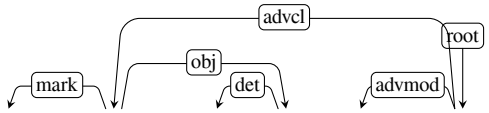
6.6 Adverbial clauses

There are a number of types of adverbial clauses in K'iche', including those introduced using word order, by a subordinator (e.g. *we* 'if' or *are taq* 'when'), and using a relational noun (e.g. *-umal* 'because', *-ech* 'in order to').



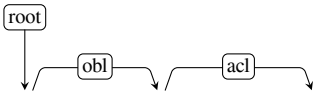
Kinb'inik *xin'ek.*
(19) K-in-b'in-ik x-in-'e-k.
I was walking I left.

In (19) a manner clause *k-in-b'in-ik* 'IMPF-B3S-walk-SS' precedes its main clause. This ordering is mandatory for manner clauses as is the lack of subordinator.



We *keqb'an* *ri* *q'uch* *utz* *kujelik.*
(20) We k-e-q-b'an ri q'uch utz k-uj-el-ik.
If we practice the *q'uch* well we come.

Other kinds of adverbial clauses may precede or follow the main clause. In *We keqb'an ri q'uch utz kujelik.* 'If we practice q'uch[6] it will be good for us.' (20) the conditional clause introduced by the subordinator *we* 'if' appears before the main clause.



Xinkosik *rumal* *xinchakunik.*
(21) X-in-kos-ik r-umal x-in-chakun-ik.
I am tired because I worked.

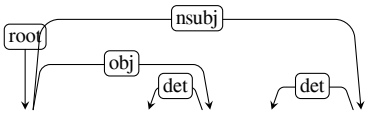
Adverbial clauses can also be introduced by relational nouns, as in (21) where the relational noun *-umal* 'by' has the function of `obl` standing in for a manner oblique and the clause is dependent on it as a adnominal clause.

[6] *Q'uch*, mutual aid, or a group of persons who agree to help each other at certain times

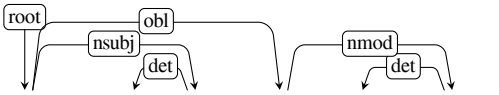
6.7 Valency changing

Transitive verbs in K'iche' are subject to two main valency changing operations, the passive and the antipassive. These are morphological processes which involve suffixation. For the passive, either the final vowel is lengthened, or the suffix *-x* is added. For the antipassive the suffixed morpheme is *-Vn* or *-n*.

In the passive, the subject is omitted and the object promoted to subject position. This can be seen in the comparison between the sentence *Xkikunaj le ali ri ixoqib'.* "The women cured the girl." (22) where the verb *x-ø-ki-kuna-j* 'PERF-B3S-A3P-cure-ACT' has agreement for both subject and object and the sentence *Xkunax le ali kumal ri ixoqib'.* "The girl was cured by the women." (23) where the verb *x-ø-kuna-x* 'PERF-B3S-cure-PASS' agrees only for the subject (previously object) and the subject is demoted to oblique using the relational noun *-umal* 'by'.

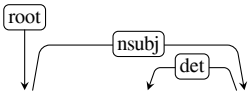


Xkikunaj *le* *ali* *ri* *ixoqib'.*
(22) X-ø-ki-kuna-j le ali ri ixoq-ib'.
Cured the girl the women.

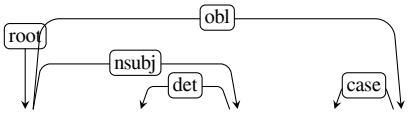


Xkunax *le* *ali* *kumal* *ri* *ixoqib'.*
(23) X-ø-kuna-x le ali k-umal ri ixoq-ib'.
Was cured the girl by the women.

In the antipassive, the subject is retained, but encoded with the absolutive, and the object is demoted to oblique status using the preposition *chi* 'to' and the relational noun *-e(ch)*.



Kinuloq'oj *le* *nutat.*
(24) K-in-u-loq'o-j le nu-tat.
He loves me the my father



Kaloq'on *le* *nutat* *ch-* *-we.*
(25) Ka-ø-loq'o-n le nu-tat chi we.
He loves the my father to me

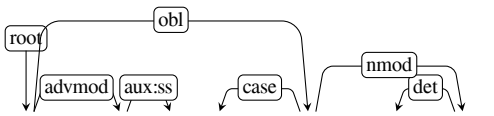
Compare the transitive sentence *Kinuloq'oj le nutat.* 'My father loves me.' (24) where the verb *k-in-u-loq'o-j* 'IMPF-B1S-A3S-love-ACT' has agreement for both subject and object with the antipassive version in (25) which exhibits agreement only for the subject, *ka-ø-loq'o-n* 'IMPF-B3S-love-AP'.

6.8 Directionals

In Mayan languages there is a category of words called directionals, which are grammaticalised forms of intransitive verbs of motion (Can Pixabaj, 2017). Some examples are *b'i(k)* < *-b'e* 'go', *qaj(oj)* < *-qaj* 'go down', and *kan(oq)* < *-kan* 'stay'. The part in parentheses after the directional is the status suffix (see §6.9). They usually follow verbs and other predicates to express movement, deictic or aspectual information and are related to the incorporated movement prefixes *e'-* < *b'e* 'go' and *ul-* < *ul* 'arrive'. Despite being derived from verbs, these are not full predicates, being either modifiers or co-predicates. We analyse them as adverbial modifiers and provide a feature `AdvType=Dir` for linguists interested in querying the corpus for this phenomenon.

6.9 Status suffixes

Status suffixes are a particular feature of the Mayan languages. These are suffixes that appear on verbs (and directionals which historically come from verbs). The particular status suffix a verb bears is conditioned by an amalgamation of morphosyntactic facts about the clause, including the transitivity of the verb, whether the verb is a root verb (i.e., CVC form) or has undergone derivation, the tense-aspect-mood of the clause, and whether the clause is an independent or dependent clause. In K'iche' there are four status suffixes, *-ik*, *-oq*, *-u* ∼ *-o* (with vowel harmony) and *-u'* ∼ *-a'* ∼ *-o'* (with vowel harmony).[7]



	Kattzaq	*b'i-*	*-k*	*chi*	*upam*	*ri*	*jul.*
(26)	K-at-tzaq	b'i	ik	chi	u-pam	ri	jul.
	Fall	DIR	ss	in	its inside	the	hole.

In this example, the directional, itself derived from a verb, bears the status suffix *-ik*, which indicates that the verb is intransitive and non-dependent. One might wonder why *tzaq* 'fall', the main verb does not bear its own status suffix. This is because, in K'iche', these suffixes only appear at the edges of certain prosodic phrases (Henderson, 2012). These is no such phrase break between the verb and directional, and so only the latter bears the status suffix.

[7]Some linguists, e.g., Kaufman (1986) also treat the suffix verbs bear in the perfect as a status suffix. We do not do so here, instead treating these suffixes as deriving stative predicates.

We have chosen to link status suffixes to their verbs with a flavour of the `aux` relation. The reason is that status suffixes are function words accompanying the verb that express aspect and mood information like verbal auxiliaries do in more familiar languages. For instance, swapping the *-ik* and *-oq* status suffixes on an intransitive verb (in certain aspects) is enough to change the interpretation from conditional mood to imperative mood.

7 Experiments

Here we present two experiments using the corpus. The first is an evaluation of three different parsing pipelines and the second is an experiment in using automatic parsing for mining linguistic examples.

7.1 Automatic parsing

In order to test the usage of the corpus for automatic parsing, performed three experiments using three off-the-shelf natural-language processing pipelines: UDPipe 1.2 (Straka et al., 2016), UDPipe 2.0 (Straka, 2018) and UDify (Kondratyuk and Straka, 2019). Version 1.2 (Straka et al., 2016) of UDPipe is a pipeline-based model where tokenisation is performed by a BiLSTM, morphological analysis and part-of-speech tagging are performed using an averaged perceptron model and dependency parsing uses a transition-based non-projective parser, where transitions are predicted by a neural network. Version 2.0 (Straka, 2018) is a complete rewrite of the UDPipe parser. It implements a joint model for part-of-speech tagging, morphological analysis, lemmatisation and parsing. The parsing model is graph-based using the Chu-Liu/Edmonds algorithm for decoding. Finally, UDify (Kondratyuk and Straka, 2019) is a multilingual model that supports parsing 75 languages. This is also a joint model, with a shared BERT representation for all 75 languages. The pre-trained model can be fine-tuned on language data from a new language, and we provide the results for fine-tuning on K'iche'. All parsers were trained with default hyperparameters.

As there was not enough data to maintain a held out test set of sufficient size, we performed ten-fold cross validation. Table 3 presents the results of the comparison. The evaluation was carried out using the official evaluation script from the 2017 *CoNLL Shared Task* (Zeman et al., 2017).

As can be seen from the results in Table 3, UDPipe 2.0 performs significantly better than UDPipe 1.2 and UDify for all of the tasks. This comes at a

	Straka et al. (2016)	Straka (2018)	Kondratyuk and Straka (2019)
Training time	20:22 ± 00:32	636:19 ± 28:56	618:27 ± 18:49
Model size	2.3M	64M	760M
Tokens	**99.8** ± 0.3	—	—
Words	**97.6** ± 0.4	—	—
Lemmas	88.3 ± 1.1	**94.9** ± 0.5	88.3 ± 0.9
UPOS	91.4 ± 1.4	**96.5** ± 0.7	94.2 ± 1.1
Features	92.0 ± 1.2	**96.6** ± 0.8	93.5 ± 0.7
UAS	82.8 ± 1.9	**91.1** ± 2.0	85.2 ± 2.8
LAS	76.7 ± 2.5	**86.5** ± 2.4	78.9 ± 2.5

Table 3: Results on tasks from tokenisation to dependency parsing. Standard deviation is obtained by running ten-fold cross validation. The columns are F_1 score: **Tokens** tokenisation; **Words** splitting syntactic words (e.g. contractions); **Lemmas** lemmatisation; **UPOS** universal part-of-speech tags; **Feats** morphological features; **UAS** unlabelled attachment score (dependency heads); **LAS** labelled attachment score (dependency heads and relations). Model size is in megabytes, training time is in mm:ss, as run on a consumer-grade laptop.

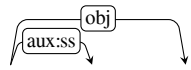
substantial increase in model size and training time compared to UDPipe 1.0, but results in a model that is still tractable on a consumer-grade laptop.

7.2 Linguistic example mining

Using corpora of under-resourced languages to test predictions pertinent to linguistic theory is often difficult. The reason is that the predictions are usually highly structurally dependent, making it hard, or even impossible, to search for relevant examples via string matching. We show the utility of the present treebank through a case study probing the distribution of phrase-final status suffixes (see section 6.9). Henderson (2012) proposes that the status suffixes that only appear phrase-finally are sensitive to intonational phrase boundaries, which roughly map onto clause boundaries. The generalisation is that a phrase final status suffix should only appear if the verb / directional bearing it is (i) utterance final, (ii) directly before an embedded clause, (iii) directly before a functional head that itself embeds a clause. Notice that to find counterexamples to this generalisation, one must search for sentences that do not satisfy a structural description—e.g., give me sentences containing a status suffix that is not directly followed by an embedded clause. This is impossible to do without a treebank. It is not even possible to do via string matching over a corpus with grammatical annotations like part-of-speech tags.

We used the corpus to test the generalization in Henderson (2012) against a larger set of K'iche' texts. In order to produce a larger corpus of examples, we took all of the texts we had available from the sources mentioned in Section 4 and to that added the *Crúbadán* corpus of K'iche' (Scannell, 2007) and processed them with the UDPipe 2.0 model described in the previous section.

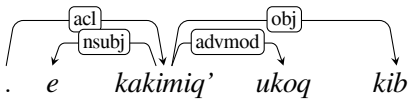
We used the *Grew* (Guillaume, 2019) corpus query language to extract all sentences where a verb had both a dependent that was an auxiliary with the relation of `aux:ss` and a noun with the relation `obj`. The query can be seen schematically in (27).



(27) VERB AUX NOUN

This lead to a total of 16,196 sentences containing 352,509 tokens. Note that the annotation for these sentences was not hand annotated, but simply the output of the data-driven parser. Although the output contained errors, the number of false positives due to errors in the parse tree was unexpectedly low.

The result is that we discovered a series of examples with structures that have not yet been considered in the literature on status suffixes, including direct counterexamples to Henderson (2012). For instance, we see in the following example a directional bearing the phrase-final dependent status suffix *-oq*. Yet, the directional is not at clause boundary or before a functional head that embeds a clause. Instead, it occurs before a reflexive pronoun, which in K'iche' is a relational noun construction.

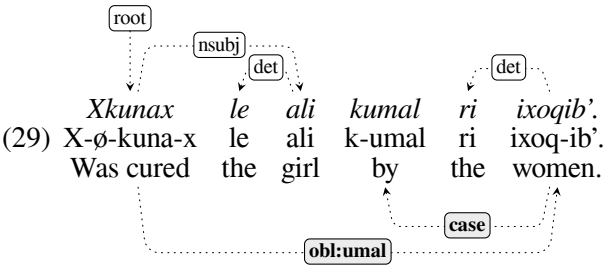


…	*e*	*kakimiq'*	*ukoq*	*kib'*.
(28) …	e	ka-ø-ki-miq'	uk-**oq**	k-ib'.
…	B3PL	they warm	DIR-SS	themselves.

An example like *Kekanaj kan kuk' chila' [e kakimiq' ukoq kib']*. "They remained over there with those [that were warming themselves]." (28) is intriguing because while a counterexample, there are plausible stories one could tell. For instance, these reflexives are prosodic clitics. Perhaps the requirement that the status suffix be phrase final ignores expressions that are prosodically deficient because they do not count as independent phonological words. While arguing for this account would take more work, the fact that we have very quickly found a theoretically interesting counterexample to a prominent generalisation in literature shows the utility of the treebank for example mining.

8 Future work

We would like to investigate the use of *enhanced dependencies*[8] to provide a more semantics-oriented encoding of relational nouns. For example if we take example (23), we could envisage an enhanced `obl` link from the verb *Xkunax* 'was cured' to the semantic head of the agent phrase *ixoqib'* 'the women' (29) where we indicate the differences with respect to the basic tree in boldface. This would fall under *Case information* in the enhanced schema and would be an additional layer on top of the basic syntax. The process could be partially automated using the *Grew* tool.



(29)

We also intend to expand the treebank and apply the lessons learnt and annotation solutions to other Mayan languages, this is a large group and we would like to start with languages related to K'iche' such as Uspanteko and Kaqchikel.

9 Concluding remarks

We have presented the first syntactically annotated corpus of sentences in K'iche'. Both the corpus and the documentation of the annotation scheme are freely available[9] through the Universal Dependencies project.[10] It is our hope that the work we describe here will facilitate the annotation of, and promote language technology for other Mayan languages.

[8] https://universaldependencies.org/u/overview/enhanced-syntax.html

[9] https://github.com/UniversalDependencies/UD_Kiche-IU

Acknowledgements

We would like to thank Telma Can Pixabaj for the permission to use the examples from her PhD thesis. The Academia de las Lenguas Mayas de Guatemala kindly allowed us to use the example sentences from their K'iche'–Spanish dictionary. Dan Zeman provided helpful feedback on specific constructions at an early stage of the treebank, and Gus Hahn-Powell gave feedback on the pre-review manuscript. We would additionally like to thank the two anonymous reviewers for their kind comments and helpful suggestions. Any errors remain our own.

References

Leah Bridges Velleman. 2014. *Focus and movement in a variety of K'ichee'*. Ph.D. thesis, University of Texas at Austin.

Telma Angelina Can Pixabaj. 2009. Verbal nouns in K'iche'. Master's thesis, University of Texas at Austin.

Telma Angelina Can Pixabaj. 2015. *Complement and purpose clauses in K'iche'*. Ph.D. thesis, University of Texas at Austin.

Telma Angelina Can Pixabaj. 2017. K'iche'. In Judith Aissen, Nora C. England, and Roberto Zavala Maldonado, editors, *The Mayan Languages*. Routledge, Oxford.

Telma Angelina Can Pixabaj. 2018. Documentation of formal and ceremonial discourses in K'ichee'. London: SOAS University of London, Endangered Languages Archive. Handle: http://hdl.handle.net/2196/00-0000-0000-000F-B63F-4. Accessed on Feb 3, 2021.

Telma Angelina Can Pixabaj. 2021. Headless relative clauses in K'iche'. In Ivano Caponigro, Harold Torrence, and Roberto Zavala Maldonado, editors, *Headless Relative Clauses in Mesoamerican Languages*. Oxford University Press, Oxford.

Gobierno de Guatemala. 2009. Taqanik b'elejeb' junab' joq'o' (9-2009): Taqanik rech uq'atuxik ri eqelenik ruk' chuq'ab'il, ch'akow pwaq xuquje' ub'anik k'ax chi kech ri winaq. *[Decreto Número 9-2009: Ley contra la violencia sexual, explotación y trata de personas]*.

Bruno Guillaume. 2019. Graph matching for corpora exploration. In *JLC 2019 - 10èmes Journées Internationales de la Linguistique de corpus*, Grenoble, France.

[10] https://universaldependencies.org

Robert Henderson. 2012. Morphological alternations at the intonational phrase edge. *Natural Language & Linguistic Theory*, 30(3):741–789. Doi:10.1007/s11049-012-9170-8.

Terrence Kaufman. 1986. Some structural characteristics of the Mayan languages, with special reference to Quiché. *Unpublished ms., University of Pittsburgh*.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

José Miguel Medrano Rojas. 2004. *K'iche' choltzij: K'iche'-kaxl'an tzij, kaxl'an tzij-k'iche*. Academia de Lenguas Mayas de Guatemala, Guatemala. [Vocabulario k'iche'].

Ministerio de Educación. 2016a. *Tzijob'elil K'aslemal*, volume I. USAID Leer y Aprender. [Antología de cuentos: I].

Ministerio de Educación. 2016b. *Tzijob'elil K'aslemal*, volume II. USAID Leer y Aprender. [Antología de cuentos: II].

Tomás Alberto Méndez López. 2020. Tajin kraqpux le xk'ulmatajem pa uwi' le yab'il covid-19 (esam pa le oms). https://covid-no-mb.org/?page_id=2009.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis M. Tyers, and Dan Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.

Oxlajuuj Keej Maya' Ajtz'iib'. 2021. Oxlajuuj Keej Maya' Ajtz'iib' Mayan languages collection. The Archive of the Indigenous Languages of Latin America, ailla.utexas.org. Access: public. PID ailla:124456. Accessed Feb 3, 2021.

Alexis Palmer, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology*, 3(4):1–42.

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

Sergio Romero, Ignacio Carvajal, Mareike Sattler, Juan Manuel Tahay Tzaj, Carl Blyth, Sarah Sweeney, Pat Kyle, Nathalie Steinfeld Childre, Diego Guarchaj Tambriz, Lorenzo Ernesto Tambriz, Maura Tahay, Lupita Tahay, Gaby Tahay, Jenny Tahay, Santiago Can, Elena Ixmata Xum, Enrique Guarchaj, Sergio Manuel Guarchaj Can, Catarina Marcela Tambriz Cotiy, Telma Can, Tara Kingsley, Charlotte Hayes, Christopher J. Walker, María Angelina Ixmatá Sohom, Jacob Sandler, Silveria Guarchaj Ixmatá, Manuela Petronila Tahay, and Susan Smythe Kung. 2018. Chqeta'maj le qach'ab'al K'iche'! https://tzij.coerll.utexas.edu/.

Frauke Sachse and Michael Dürr. 2016. Morphological glossing of Mayan languages under XML: Preliminary results. Working Paper 4, Nordrhein-Westfälische Akademie der Wissenschaften und der Künste.

Kevin Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Proceedings of the 3rd Web as Corpus Workshop*, pages 5–15.

M. Straka, J. Hajič, and J. Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France. European Language Resources Association (ELRA).

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Peter Svenonius. 2006. The emergence of axial parts. Working Paper 33.1, Universitetet i Tromsø.

Wikimedia Incubator. 2017. Wp/quc/tripanosomiasis africana — Wikimedia Incubator. https://incubator.wikimedia.org/w/index.php?title=Wp/quc/Tripanosomiasis_africana.

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

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadova, Esha

Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.