A survey of part-of-speech tagging approaches applied to K'iche'

Francis Tyers[†♢]
† Department of Linguistics
Indiana University
Bloomington, IN
ftyers@iu.edu

Nick Howell[♢]
♢ School of Linguistics
HSE University
Moscow
nhowell@hse.ru

Abstract

We study the performance of several popular neural part-of-speech taggers from the Universal Dependencies ecosystem on Mayan languages using a small corpus of 1435 annotated K'iche' sentences consisting of approximately 10,000 tokens, with encouraging results: F_1 scores 93%+ on lemmatisation, part-of-speech and morphological feature assignment. The high performance motivates a cross-language part-of-speech tagging study, where K'iche'-trained models are evaluated on two other Mayan languages, Kaqchikel and Uspanteko: performance on Kaqchikel is good, 63-85%, and on Uspanteko modest, 60-71%. Supporting experiments lead us to conclude the relative diversity of morphological features as a plausible explanation for the limiting factors in cross-language tagging performance, providing some direction for future sentence annotation and collection work to support these and other Mayan languages.

1 Introduction

This paper presents a survey of approaches to part-of-speech tagging for K'iche', a Mayan language spoken principally in Guatemala. The Mayan languages are a group of related languages spoken throughout Mesoamerica. K'iche' belongs to the Eastern branch, which contains 14 other languages, including Kaqchikel in the Quichean subgroup and Uspanteko which belongs to its own subgroup.

Part-of-speech tagging has wide usage in corpus and computational linguistics and natural language processing, and is often considered part of a toolkit for basic natural language processing.

In the definition of part-of-speech tagging we subsume the tasks of determining the part of speech, morphological analysis and lemmatisation. That is, given a sentence such as in (1) part-of-speech tagging would return both the sequence of part-of-speech tags [VERB, DET, NOUN] but also the lemmata [q'ojomaj, le, q'ojom] and the set of feature value pairs for each of the forms.[1]

(1) Kinq'ojomaj le q'ojom.
k-∅-in-q'ojomaj le q'ojom.
IMP-B3SG-A1SG-play the marimba.

'I play the marimba.'

A brief reading guide: prior work, on Mayan and other languages of the Americas and on cross-language part-of-speech tagging, is reviewed in section 2. Our experimental design including the mathematical model used for analysing performance are given section 3. Universal dependencies annotation for K'iche' and the systems tested are described in section 4, and results are presented and analysed in section 5.

2 Prior work

Palmer et al. (2010) explore morphological segmentation and analysis for the purpose of generating interlinearly glossed texts. They work with Uspanteko, a language of the Greater Quichean branch, and the closest language to K'iche' we were able to identify with published studies of computational morphology. They explore several different systems: inducing morphology from parallel texts, an unsupervised segmentation+clustering strategy, and an interactive training strategy with a linguist.

In Sachse and Dürr (2016), a set of preliminary annotation conventions for Mayan languages in general, and K'iche' in particular, are proposed.

A maximum-entropy part-of-speech tagger is presented in Kuhn and Mateo-Toledo (2004) for Q'anjob'al, which, like K'iche', is a Mayan language of Guatemala. They work with a custom selection

[1] For example for the VERB it would return Aspect=Imp, Number[obj]=Sing, Number[subj]=Sing, Person[obj]=3, Person[subj]=1, Subcat=Tran, VerbForm=Fin.

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 44–52

of 60 tags, and trained on an annotated corpus of 4100 words (no lemmatisation is performed). In contrast to the systems we will study, Kuhn and Mateo-Toledo (2004) perform feature engineering and end up with F_1 scores between 63% and 78%, depending on the features chosen.

There is much work on part-of-speech tagging for languages of the Americas outside of the Mayan family: statistical lemmatisation and part-of-speech tagging systems are described by Pereira-Noriega et al. (2017) and a finite-state morphological analyser by Cardenas and Zeman (2018) for Shipibo-Konibo, a Panoan language of the Amazonian region of Peru.

In Rios (2010) and Rios (2015), respectively, finite-state morphology and support vector machine-based tagging+parsing systems are described for Quechua. The latter uses a corpus that comprises $2k$ sentences.

Cross-language part-of-speech tagging through parallel corpora, sometimes called annotation projection, is well-studied; in Mayan languages, Palmer et al. (2010) use a parallel corpus as a bridge to a higher-resourced language for which a part-of-speech tagger already exists.

In the absence of such a corpus, so-called "zero-shot" methods are created from other (presumably higher-resourced) languages and applied to the target language. The main balance to strike is between specificity of resources (how closely-related are the other languages) and quantity of resources (how much linguistic data is accessible). UDify of Kondratyuk and Straka (2019) is an example of preferring the latter: a deep neural architecture is trained on all of the Universal Dependencies treebanks. The former strategy can be seen in Huck et al. (2019), where in addition to annotation projection, authors attempt zero-shot tagging of Ukrainian with a model trained on Russian.

3 Methodology

We used a corpus of K'iche'[2] annotated with part-of-speech tags and morphological features (Tyers and Henderson, 2021). The corpus consisted of 1,435 sentences comprising approximately 10,000 tokens from a variety of text types and was annotated according to the guidelines of the Universal Dependencies (UD) project (Nivre et al., 2020). An example of a sentence from the corpus can be seen in Table 1.

We studied the performance of several popular part-of-speech taggers within the Universal Dependencies ecosystem; these are reviewed in section 4. Performance was computed as F_1 scores for lemmatisation, universal part-of-speech (UPOS), and universal morphological features (UFeats). We performed 10-fold cross validation to obtain mean and standard deviation of F_1. We also recorded training time and model size to compare the resource consumption of the models in the training process.

We selected the best-performing system and performed a convergence study (see section 5.3 for results). We decimated the training data of one of the test-train splits from the cross-validation, and plotted the performance of models trained on the decimations.

We make the following assumption about the performance: additional training data provides exponentially decreasing performance improvement. Under this assumption, we obtain the formula:

$$F_1(n) = F_1(\infty) - \Delta F_1 \cdot e^{-n/k}. \quad (1)$$

Here $F_1(n)$ is the performance of a model trained on n tokens, $F_1(\infty)$ is the asymptotic performance, and ΔF_1 is the gap between $F_1(\infty)$ (estimated maximum performance) and $F_1(0)$ (zero-shot performance).

The parameter k is the *characteristic number of tokens*; each additional k tokens of training data causes the gap $\Delta F_1 = F_1(\infty) - F_1(n)$ to shrink by a factor of $1/e \approx 36\%$. This can be used to estimate the training data n required to meet a given performance target F_1^{target}:

$$n = k \cdot \log \frac{\Delta F_1}{F_1(\infty) - F_1^{\text{target}}} \quad (2)$$

We fit this curve against our convergence data and estimate peak performance and characteristic number. Error propagation is used with the error in parameter estimation to compute the error bands in the graph:

$$(\delta F_1)^2 = \sum \left(\frac{\partial F_1}{\partial x} \delta x \right)^2 \quad (3)$$

Here x runs over the parameters of $F_1(n)$: $F_1(\infty)$, ΔF_1 and k.

We also studied the best-performer in cross-language tagging on the related Kaqchikel and Uspanteko languages. The 10 models trained in

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

cross-validation were all evaluated on small part-of-speech-tagged corpora of 157 (Kaqchikel) and 160 (Uspanteko) sentences. For results and overviews of the languages, see section 6.

4 Systems

We tested morphological analysis on three systems designed for Universal Dependencies treebanks: UDPipe (Straka et al., 2016), UDPipe 2 (Straka, 2018), and UDify (Kondratyuk and Straka, 2019). Of these, only UDPipe had a working tokeniser. For other taggers we trained, we trained the UDPipe tokeniser and other tagger together. We thus present combined tokeniser-tagger systems.

UDPipe (Straka et al., 2016) is a language-independent trainable tokeniser, lemmatiser, POS tagger, and dependency parser designed to train on and produce Universal Dependencies-format treebanks. It uses gated linear units for tokenisation, averaged perceptrons for part-of-speech tagging, and a neural network classifier for dependency parsing. It is the least resource-hungry model in our study by an order of magnitude or more, and we trained it from-scratch using the K'iche' corpus in section 3.

UDPipe 2 (Straka, 2018) is a Python prototype for a Tensorflow-based deep neural network POS-tagger, lemmatiser, and dependency parser. It won high rankings in the CoNLL 2018 shared task on multilingual parsing (Zeman et al., 2018), taking first place by one metric. Deep neural methods have achieved impressive performance results in recent years, but take considerable computational resources to train. We used UDPipe 2 without pretrained embeddings, and trained it from-scratch using the K'iche' corpus in section 3.

UDify (Kondratyuk and Straka, 2019) is a AllenNLP-based multilingual model using BERT pretrained embeddings and trained on the combined Universal Dependencies treebank collection; we fine-tuned this pretrained model on our K'iche' data. This was our most resource-intensive model, even though we only fine-tuned on K'iche'; our initialisation was the UDify-distributed BERT+UD model.

5 Results

5.1 Energy efficiency

Resource utilisation for the three systems is summarised in Table 2. Model production is reported in kilojoules for each of our systems; these were estimated by taking the reported runtime and multiplying it by the thermal design power (TDP) of the reported hardware. Error could be introduced into these estimates from many sources: only the reported device is considered, ignoring many other components of the machine; devices are assumed to run at their TDP the entire runtime; the UDify numbers as reported by Kondratyuk and Straka (2019) are approximate.

5.2 Task performance

We evaluated the performance of the models on five tasks: tokenisation (Tokens), word segmentation (Words), lemmatisation (Lemmas), part-of-speech tagging (UPOS) and morphological tagging (Features). The difference between tokenisation and word segmentation can be explained with reference to Table 1. The word *chqawach* 'to us' counts as a single token, but two syntactic words. So the performance of tokenisation is recovering the tokens, and the performance of word segmentation is recovering the words.

We performed 10-fold cross validation on the 1435 analysed sentences, with F_1 scores for lemmatisation, part-of-speech tagging, and morphological features computed using the evaluation scripts from Zeman et al. (2018), modified to not ignore language-specific morphological features. Results are summarised in Table 3; the winner is UDPipe2.

While both UDPipe 2 and UDify have deep neural architectures, it seems UDify is unable to overcome non-K'iche' biases from the BERT embeddings and initial training on Universal Dependencies releases; neither of these components incorporate Mayan languages. We speculate that training on data with a better representation of languages of the Americas would enable UDify to surpass UDPipe 2.

The original UDPipe makes an impressively resource-efficient performance: it obtains 95%, 97%, and 96% the performance of UDPipe 2 on lemmatisation, part-of-speech tagging, and feature assignment, all with 3.5% of the training time and 3.6% of the model size.

5.3 Convergence

We performed a convergence study on the best system, UDPipe 2. Results are shown in Figure 1. Asymptotic F_1 scores are $95.4 \pm 1.9\%$, $97.4 \pm 2.2\%$, and $95.7 \pm 2.1\%$ for lemmatisation, part-of-speech tagging, and feature assignment, respectively. Gaps at full use of the 1292 sentence-, 9559-token training set are 2.5%, 2.9%, and 3.8%, respectively, and

# sent_id = utexas:123.2									
# text = Xuk'ut le K'iche' ch'ab'al le al Nela chqawach.									
# text[spa] = Manuela nos enseñó el idioma k'iche'									
# labels = tijonik-17 complete									
1	Xuk'ut	k'ut	VERB	_	[...][1]	_	_	_	_
2	le	le	DET	_	_	_	_	_	_
3	K'iche'	k'iche'	ADJ	_	_	_	_	_	_
4	ch'ab'al	ch'ab'al	NOUN	_	_	_	_	_	_
5	le	le	DET	_	_	_	_	_	_
6	al	ali	NOUN	_	Gender=Fem|NounType=Clf	_	_	_	_
7	Nela	Nela	PROPN	_	Gender=Fem	_	_	_	_
8-9	chqawach	_	_	_	_	_	_	_	_
8	ch	chi	ADP	_	_	_	_	_	_
9	qawach	wach	NOUN	_	[...][2]	_	_	_	_
10	.	.	PUNCT	_	_	_	_	_	_

[1] Aspect=Perf|Number[obj]=Sing|Number[subj]=Sing|Person[obj]=3|Person[subj]=3|Valency=2|VerbForm=Fin
[2] NounType=Relat|Number[psor]=Plur|Person[psor]=1

Table 1: An example sentence from Romero et al. (2018) that has been included in the corpus. Here it is displayed annotated in 10-column CoNLL-U format. The sentence is *Xuk'ut le K'iche' ch'ab'al le al Nela chqawach.* "Manuela taught us the K'iche' language". This demonstrates: the treatment of contractions, e.g. *chqawach* 'to us' → *chi* + *qawach*, the lemmatisation and parts of speech and the morphological features.

Model	**Energy** (kJ)	
	UD	K'iche'
UDPipe	0	50
UDPipe 2	0	1400
UDify	540000	1300

Table 2: Energy cost expended, per-source. K'iche' training costs are estimated as runtime × TDP of the processor, while UD training costs are runtime × TDP of the graphics card used in training.

characteristic numbers are 4700, 4800 and 4700 tokens. Using (2), we can use this to compute how much more training data would be required to close this gap; for example, to bring F_1 to within 1% of its maximum, we would need to annotate an additional 4400, 4500, and 5900 tokens, respectively.

6 Cross-language tagging

There are around 32 Mayan languages spoken in Mesoamerica, in the countries of Guatemala, Mexico, Honduras, El Salvador and Belize. Given the impressive performance of the best-performing system on K'iche' data, we decided to test it on two related languages spoken in Guatemala: Kaqchikel and Uspanteko. UDify is also reported as being suited to zero-shot inference, so we include two UDify-based models: fine-tuned on K'iche' (referred to as "UDify-FT") and the original UDify model (simply "UDify").

6.1 Kaqchikel

Kaqchikel (ISO-639: `cak`; previously Cakchiquel) is a Mayan language of the Quichean branch. It is spoken in Guatemala, to the south and east of the K'iche'-speaking area (see Figure 2) and has around 450,000 speakers. Some notable differences between Kaqchikel and K'iche' are the lack of status suffixes on verbs, no pied-piping inversion (Broadwell, 2005), and SVO order in declarative sentences (Watanabe, 2017).

For the Kaqchikel corpus, we extracted glossed example sentences from a number of published sources, including papers discussing topics in morphology and syntax (Henderson, 2007; Broadwell and Duncan, 2002; Broadwell, 2000) and grammar books (Garcia Matzar et al., 1999; Guaján, 2016). These sentences were then analysed with a morphological analyser (Richardson and Tyers, 2021) and manually disambiguated using the provided glosses.

6.2 Uspanteko

Uspanteko (ISO-639: `usp`; also referred to as *Uspantek*, or *Uspanteco*) is a Mayan language of the Greater Quichean branch. The language is spoken

	UDPipe	UDPipe 2	UDify
Training time	12.5 ± 0.1	356 ± 4	323 ± 2
Model size	2.3M	64M	760M
Tokens	99.7 ± 0.4	—	—
Words	98.6 ± 0.5	—	—
Lemmas	88.3 ± 1.1	**93.2 ± 0.6**	88.3 ± 0.9
UPOS	91.4 ± 1.4	**94.5 ± 0.8**	94.2 ± 1.1
Features	88.8 ± 1.1	**92.9 ± 0.8**	89.2 ± 1.2

Table 3: Results on tasks from tokenisation to morphological analysis. 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; **Features** morphological features. Model size is in megabytes, training time is in mm:ss, as run on a machine with AMD Ryzen 7 1700 8-core CPU and 32GiB of memory.

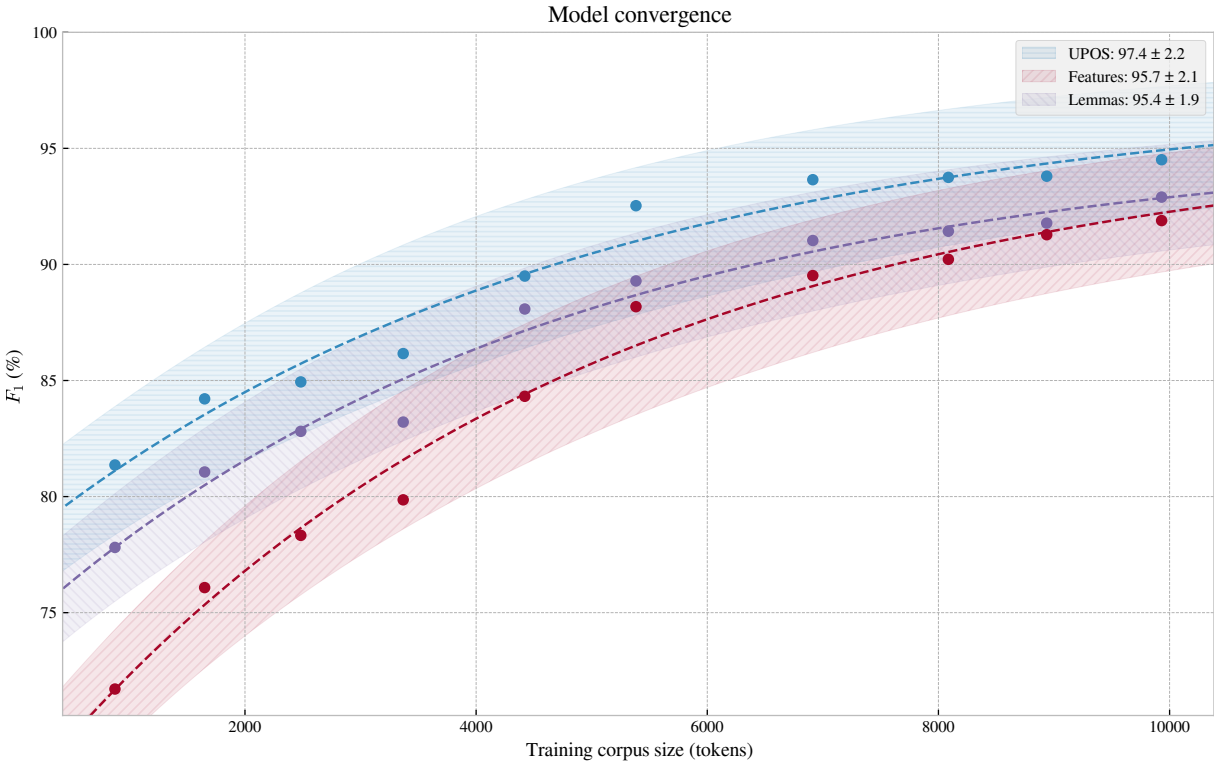


Figure 1: Convergence of the F_1 scores of the UDPipe 2 combined system for lemmas, universal part-of-speech, and universal feature tags, as a function of total number of tokens in training. The plotted points (p, s) are the decimation data: measurements of F_1 score p when given a training corpus of s tokens. Curves are obtained by constrained least-squares fitting of this data against (1). The shaded regions represent the propagation of the standard error (3) in the fit parameters through the curve; under hypothesis of the normal distribution, $\approx 68\%$ of observations are expected to lie within this region. The numbers in the legend are the asymptotic performance given by the fitting procedure; as more training data is supplied, model performance should converge to the asymptotic performance.

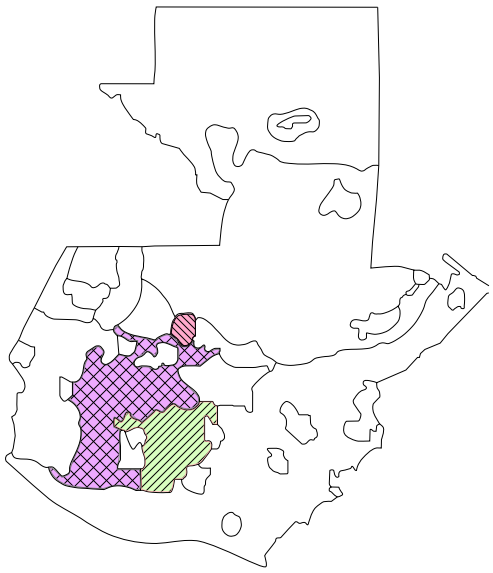


Figure 2: A map of Guatemala with approximate locations of speaker areas of Mayan languages. K'iche', Kaqchikel and Uspanteko are highlighted in purple (grid-hatched), green (forward slash-hatched), and red (backward slash-hatched), respectively.

Kaqchikel			
Sentences			157
Tokens			1091
	UDPipe 2	UDify-FT	UDify
UPOS	84.9 ± 0.4	**90.0** ± 0.4	34.3
Features	**63.4** ± 0.7	**63.4** ± 0.7	46.1
Lemmas	72.5 ± 0.5	**75.4** ± 0.5	3.2
Uspanteko			
Sentences			160
Tokens			1171
	UDPipe 2	UDify-FT	UDify
UPOS	60.8 ± 0.6	**64.7** ± 0.5	40.2
Features	**60.3** ± 0.9	59.1 ± 1.0	55.3
Lemmas	71.2 ± 0.5	**71.4** ± 0.4	6.3

Table 4: Results for cross-lingual tagging on Kaqchikel and Uspanteko, using our UDPipe 2, UDify, and UDify-FT systems for part-of-speech tagging. We evaluated on our corpora lemmatised and annotated for part-of-speech, morphological features. Performance for the K'iche'-trained systems are quoted as the average and standard deviation over the same 10 trained models used in cross-validation for K'iche' (see section 3).

in an area adjacent to the K'iche'-speaking area in Guatemala. It has around 2,000 speakers and is one of the few Mayan languages to have developed contrastive tone.

Palmer et al. (2010) present a large interlinearly-glossed corpus of Uspantek with approximately 3400 sentences and 27000 tokens. We selected 160 sentences from this corpus, totalling 1003 tokens and annotated them with part of speech, lemmas and morphological features. The lemmas were given by a morphological analyser[3] created from a lexicon provided by OKMA.

6.3 Results

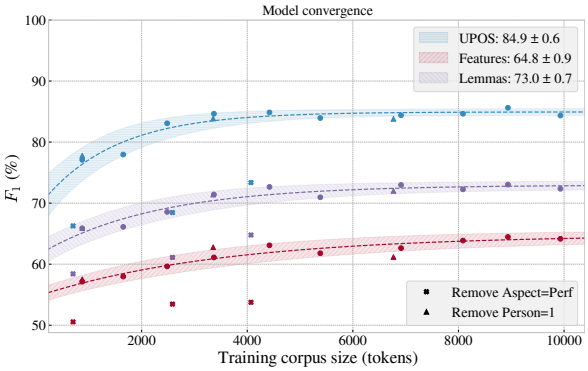
The results of our cross-language tagging study are shown in Table 4; in general the winner is UDify-K'iche'; the original UDify model itself performs very poorly. UDPipe 2 manages nearly as good performance as UDify-FT, especially impressive considering its three orders of magnitude less energy consumption. For UDPipe 2 and UDify-FT, we used the ten models trained to provide the K'iche' tagging performance and confidence. The original UDify system is a single model, thus we are unable to provide confidence intervals.

We also studied convergence for the cross-language tagging task using our UDPipe 2 decimated K'iche' models; see figures 3a and 3b. We observe that for the given set of labels our models essentially have converged, with the exception of part-of-speech tagging for Uspanteko, which might benefit from additional examples of features already present in our K'iche' corpus.

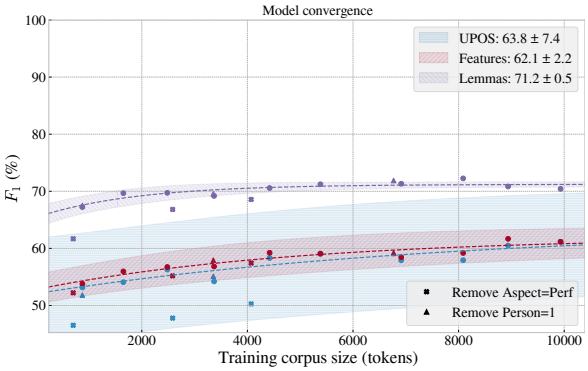
In order to understand whether our K'iche' corpus covers a sufficient variety of labels (parts of speech, features, lemmatisation patterns), we selected two labels, one of high frequency and one of low frequency (see Table 5a), from our corpus with which to disable our model. For each label, new convergence runs where made using the 10%, 40%, and 70% subsets, omitting all sentences featuring the chosen label.

If our cross-language tagging models could not be improved by a more diverse K'iche' training corpus, we would expect these disabled datapoints to fall within error of the convergence trendlines. This is the case with the low-frequency label, "first-person". On the other hand, we see that the loss of the high-frequency label, perfective aspect, has

[3] https://github.com/apertium/apertium-usp



(a) Kaqchikel. Characteristic number of tokens of annotated K'iche' for these was 2200 (lemmatisation), 1400 (part-of-speech), and 3500 (features). At nearly 10000 tokens, all are essentially converged.



(b) Uspanteko. Characteristic number of tokens for these was 1900 (lemmatisation), 7900 (part-of-speech), and 4900 (features); part-of-speech tagging might see improvement from increased annotation of K'iche' data, but with such high uncertainty (over 10% in asymptotic performance) it is difficult to be sure.

Figure 3: Convergence of our UDPipe 2 on Kaqchikel (3a) and Uspanteko (3b). The legends show projected asymptotic performance for each of universal part-of-speech tagging, universal feature assignment, and lemmatisation.

a disproportionate impact on cross-tagging performance: removing this training data has caused the convergence curve to change parameters, lowering asymptotic performance.

This raises the possibility that we might improve the asymptotic performance of our cross-tagging models by locating labels which are high-frequency in our target language (Kaqchikel or Uspanteko) and extending our K'iche' corpus with sentences featuring those labels. See Table 5b for a sample of high-frequency labels which appear in our K'iche' corpus but not our cross-tagging evaluation corpus.

These all indicate that the small test corpora of Kaqchikel and Uspanteko we annotated are not as diverse in terms of text type as the K'iche' corpus. For example, the test corpora contain no infinitive forms (for example the morpheme *-ik* in K'iche'), although these certainly exist in both Kaqchikel — see §2.7.2.6 in Garcia Matzar et al. (1999) — and Uspanteko. Additionally they contain no examples of the imperative mood, relative clauses introduced by relative pronouns, the formal second person, or reflexives. All of these features certainly exist in the languages, but not in the selection of sentences we annotated.

7 Concluding remarks

We used an annotated corpus of 1435 part-of-speech tagged K'iche' sentences to to survey a number of neural part-of-speech tagging systems from that ecosystem. We found the best performance was generally with UDPipe 2, a deep neural system integrating lemmatisation, part-of-speech and morphological feature assignment. Our UDPipe 2-trained system achieved F_1 of 93% or better on all tasks, very encouraging results for a relatively small corpus.

Convergence studies showed that on corpora of similar morphological composition even better performance is attainable, but to close the gap to within 1% of projected optimal performance requires roughly half again the amount of training data.

The high performance on K'iche' led us to experiment using our model to perform cross-language tagging on the related languages of Kaqchikel and Uspanteko. Performance on the more closely-related language, Kaqchikel, was still respectable, with F_1 ranging from 63 to 85% on the tasks; on Uspanteko performance we observed more modest performance $60-71\%$. The K'iche' fine-tuned UDify model does show noticably better performance, but possibly not worth the energy expenditure.

Our results after disabling our cross-language tagger by withholding some labels during training imply that cross-language performance could be improved by annotating more data with similar features to the Kaqchikel and Uspanteko evaluation corpora, and suggest that cross-language tagging is a path forward to greater availability of part-of-speech annotation for Mayan languages.

Label	Frequency		Discrep.
	quc	evaluation	
Person=1	3%	1%	0.12σ
Aspect=Perf	49%	62%	$-3.5\ \sigma$

(a) The two labels chosen for the label diversity study for our cross-language taggers. We studied convergence of two additional models: training data alternately lacked first-person (Person=1), or perfective aspect (Aspect=Perf). Frequency is percentage of sentences in the corpus with the feature. We give the median discrepancy, computed as the performance gap between the disabled model and the prediction for a model trained on the same number of tokens, normalised by the uncertainty in that prediction σ. For the first-person label, we see a similar distribution with a very slight bias towards higher performance; perfective aspect seems to have an outsized effect, increasing the median discrepancy to 3.5σ.

Label	Frequency (% sents.)
VerbForm=Inf	6
Mood=Imp	3
Reflex=Yes	2
PronType=Rel	2
Polite=Form	2

(b) The results of our label diversity study. The top 20 labels for our K'iche' training corpus which do not appear in our Kaqchikel and Uspanteko evaluation corpora, along with their frequencies in the K'iche' corpus. See Table 5a for the impact missing high-frequency labels can have on cross-tagging performance.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This article is an output of a research project implemented as part of the Basic Research Programme at the National Research University Higher School of Economics (HSE University).

References

George Aaron Broadwell. 2000. Word order and markedness in Kaqchikel. In *Proceedings of the LFG00 Conference*.

George Aaron Broadwell. 2005. Pied-piping and optimal order in Kiche (K'iche').

George Aaron Broadwell and Lachlan Duncan. 2002. A new passive in Kaqchikel. *Linguistic Discovery*, 1:26–43.

Ronald Cardenas and Daniel Zeman. 2018. A morphological analyzer for Shipibo-konibo. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 131–139, Brussels, Belgium. Association for Computational Linguistics.

Pedro Oscar Garcia Matzar, Valerio Toj Cotzajay, and Domingo Coc Tuiz. 1999. *Gramática del idioma Kaqchikel*. PLFM.

Pakal B'alam Rodriguez Guaján. 2016. *Rutz'ib'axik ri Kaqchikel — Manual de Redacción Kaqchikel*. Editorial Maya' Wuj.

Robert Henderson. 2007. Observations on the syntax of adjunct extraction in Kaqchikel. In *Proceedings of the CILLA III Conference*.

Matthias Huck, Diana Dutka, and Alexander Fraser. 2019. Cross-lingual annotation projection is effective for neural part-of-speech tagging. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 223–233, Ann Arbor, Michigan. Association for Computational Linguistics.

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.

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 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisboa, Portugal.

J. Nivre, 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.

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.

José Pereira-Noriega, Rodolfo Mercado-Gonzales, Andrés Melgar, Marco Sobrevilla-Cabezudo, and Arturo Oncevay-Marcos. 2017. Ship-LemmaTagger: Building an NLP toolkit for a Peruvian native language. In *International Conference on Text, Speech, and Dialogue*, pages 473–481. Springer.

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

Annette Rios. 2010. Applying finite-state techniques to a native American language: Quechua. Lizentiatsarbeit, Institut für Computerlinguistik, Universität Zürich.

Annette Rios. 2015. *A basic language technology toolkit for Quechua*. Ph.D. thesis, Institut für Computerlinguistik, Universität Zürich.

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.

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.

Francis M. Tyers and Robert Henderson. 2021. A corpus of K'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

Akira Watanabe. 2017. The division of labor between syntax and morphology in the Kichean agent-focus construction. *Morphology*, 27:685–720.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.