Morphological Tagging in Bribri Using Universal Dependency Features

Jessica Acton Karson
Dartmouth College
jess.a.karson.23@dartmouth.edu

Rolando Coto-Solano
Dartmouth College
rolando.a.coto.solano@dartmouth.edu

Abstract

This paper outlines the Universal Features tagging of a dependency treebank for Bribri, an Indigenous language of Costa Rica. Universal Features are a morphosyntactic tagging component of Universal Dependencies, which is a framework that aims to provide an annotation system inclusive of all languages and their diverse structures (Nivre et al., 2016; de Marneffe et al., 2021). We used a rule-based system to do a first-pass tagging of a treebank of 1572 words. After manual corrections, the treebank contained 3051 morphological features. We then used this morphologically-tagged treebank to train a UDPipe 2 parsing and tagging model. This model has a UFEATS precision of 80.5 ± 3.6, which is a statistically significant improvement upon the previously available FOMA-based morphological tagger for Bribri. An error analysis suggests that missing TAM and case markers are the most common problem for the model. We hope to use this model to expand upon existing treebanks and facilitate the construction of linguistically-annotated corpora for the language.

Resumen

Etiquetado morfológico del Bribri usando rasgos de Universal Dependencies. Este artículo presenta un experimento para el etiquetado automático de la morfología de las palabras en una colección de árboles sintácticos de dependencia en bribri, un idioma indígena de Costa Rica. El esquema *Universal Features* es un componente de etiquetado morfológico de *Universal Dependencies*, un estándar para el análisis sintáctico de oraciones. Este esquema busca poder etiquetar cualquier lengua del mundo y sus diversas estructuras (Nivre et al., 2016; de Marneffe et al., 2021). Empezamos el proyecto usando un sistema basado en reglas para etiquetar automáticamente una colección de árboles sintácticos con 1572 palabras. Después de una corrección manual, la colección tenía un total de 3051 etiquetas morfológicas. Esta nueva colección de árboles se usó para entrenar un modelo de UDPipe 2 que pudiera hacer etiquetado y análisis sintáctico automáticamente. Este modelo tiene una precisión de UFEATS de 80.5 ± 3.6, lo cual es una mejora estadísticamente significativa con respecto a los etiquetadores basados en FOMA disponibles para el bribri. Un análisis de errores sugiere que el principal problema para el modelo fue el no poder producir algunas etiquetas de TAM y de caso en la salida. Esperamos usar este modelo para expandir las colecciones de árboles ya existentes, y así facilitar la construcción de corpus anotados lingüísticamente para esta lengua.

1 Introduction

It is essential that the fields of linguistics and Natural Language Processing dedicate time and resources towards smaller, Indigenous, and minority languages. Building annotated and tagged corpora for smaller languages supports the expansion of NLP capabilities in processing them, and could potentially expand the languages' domain of usage and help create tools that aid in language revitalization and normalization. In this paper we worked on one small building block of future NLP tools: the morphological tagging of corpora in the Bribri language, an Indigenous language from Costa Rica. In section 1 we review the process of morphological tagging and describe the Bribri language's vitality and context. In section 2, we describe an algorithm for rule-based tagging, and how we used this for our first attempt at automatic tagging. After correcting any resulting errors, we trained a deep-learning based model to perform future tagging. Section 3 describes the tags applied to the treebank, compares the model's performance to a previously available tagger, and describes the errors that the model is making in its tagging output. Finally, section 4 describes some limitations of the tagging scheme when describing Bribri data, as well as directions of future work.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 56–66
June 21, 2024 ©2024 Association for Computational Linguistics

1.1 Morphological Analysis and Tagging

Morphological analysis is the systematic breakdown of words into smaller pieces that reflect units of meaning (i.e. morphemes). For example, the input `cats` would return the output `cat-s`. In the context of natural language processing, morphological analysis can be paired with the task of morphological tagging. In morphological tagging, a word like `cats` would produce an output like `cat+[N;PL]`, `NN2` or `Number=Plur`. These three examples, which use different standards, indicate in different ways that the word is plural. This tagging can support the building of annotated corpora, which in turn allows for more advanced linguistic research, but also for more advanced NLP tasks such as lemmatization and disambiguation tasks.

Morphological analysis is undertaken using different standards and can use language-specific or language family-specific differentiations. The UCREL CLAWS7 tagset (UCREL, 2011), for example, is made for English and uses a one-tag-per-word system which labels both the part of speech and some related morphological characteristics (e.g. `cats` → `NN2` 'common noun plural'). The UniMorph standard (McCarthy et al., 2020) attempts to describe all languages using the same tags, and it uses a one-to-many system where one word can have several tags depending on its part of speech and its morphemes (e.g. `cats` → `cat+[N;PL]`). The Universal Dependencies' (Nivre et al., 2020) Universal Features schema (UFEATS) also attempts to offer coverage for the morphology of every language. It uses its own set of tags, leaving out the part of speech but including one or more morphological tags per word (e.g. `cats` → `Number=Plur`). This standard is used to annotate numerous treebanks in Universal Dependencies, including an existing one for Bribri (see section 2.1 below). Because of this, and because it would provide an additional way to query the existing treebank for specific morphemes, the UFEATS schema will be used in this work.

1.2 Automatic Morphological Analysis and Indigenous Languages

Morphological analysis for under-resourced Indigenous languages presents unique challenges for several reasons. The limited availability of data complicates progress when determining meaningful connections between words or units within words. Additionally, the input of the language data can have inconsistencies due to lack of standardization in orthography[1] and unaccounted-for variation in data collection.

Despite these challenges, there has been work for Universal Features tagging in languages of the Americas. There are Universal Features tagged datasets for Tupí languages (Rodríguez et al., 2022), K'iche (Tyers and Howell, 2021; Tyers and Henderson, 2021) and Yupik (Park et al., 2021). There are also languages tagged using UniMorph, such as Kanien'kéha (Kazantseva et al., 2024), Plains Cree, Gitksan, Asháninka, Aymara, Seneca, Dakota, Otomí, Mixtec, Chatino, Zapotec and Tohono O'odham (Batsuren et al., 2022). There is also work on using finite state transducers to do morphological tagging and segmentation. Some languages where such taggers exist are Haida (Lachler et al., 2018), Michif (Davis et al., 2021), Cree (Snoek et al., 2014), Lushootseed (Rueter et al., 2023), Wixarika (Mager et al., 2018a), Nahuatl (Pugh and Tyers, 2021) and Guaraní (Kuznetsova and Tyers, 2021). Languages where custom methods have been used for morphological tagging and segmentation include Inuktitut (Khandagale et al., 2022; Le and Sadat, 2021), Seneca (Liu et al., 2021), Quechua (Llitjós et al., 2005), Shipibo-Konibo (Mercado-Gonzales et al., 2018) and Mapugundun (Molineaux, 2023). In addition to these papers, Mager et al. (2018b) document additional efforts to work on morphological analysis and tagging of Indigenous languages in the Americas.

1.3 Bribri Language

Bribri is a language of the Chibchan family with approximately 7000 speakers in Costa Rica (INEC, 2011). Bribri is closely related to the Cabécar language also spoken in Costa Rica, and it is more distantly related to other Chibchan languages like Malecu and Ngäbe (Quesada, 2007). Bribri is a vulnerable language (Sánchez Avendaño, 2013), which means that there are children in the community who only speak Spanish. As Bribri is a low-resource language, documentation and natural

[1] We do not necessarily advocate standardization here. This is a decision that needs to be taken by the community. Moreover, very valuable materials are being published in Bribri using orthographic conventions unique to each author (MEP, 2013; García Segura, 2016; Jara Murillo and García Segura, 2022). This is a relatively common situation in writing within Indigenous communities, and pursuing a single standard might be detrimental to language revitalization (De Korne and Weinberg, 2021). We believe that this input issue is the engineers' problem to solve, not necessarily the communities.

language processing applications for the language are limited and difficult to make. Moreover, the particularities of Bribri morphosyntax make transfer learning from large-resource languages difficult. For example, the language is morphologically ergative, it has numerical classifiers and a complex deictic system, and it has a verbal system where "now" is not the locus of division between tenses, but rather "the night before". Examples of these phenomenona are presented in section 4.

2 Methodology

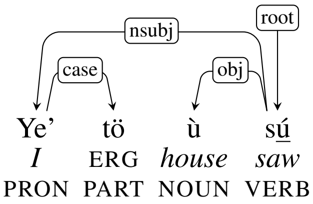
Our overall goal is to improve morphological tagging for Bribri. In this section we will explain how we used a rule-based algorithm to tag the existing treebank using Universal Dependencies Features. After manual correction, we tested these features by using them to train parsing models. When those models were trained we compared their performance to that of a pre-existing morphology analysis system for Bribri.

2.1 Bribri Data and Pre-existing Algorithms

There is relatively little unlabeled data for Bribri. The main data source is the oral corpus by Flores-Solórzano (2017a,b), which contains both text and audio for Bribri conversations. There are some printed materials which could provide written data, such as textbooks (Constenla et al., 2004; Jara Murillo and García Segura, 2013), a grammar book (Jara, 2018), two dictionaries (Margery, 2005; Krohn, 2021) and several educational books (Sánchez Avendaño et al., 2021a,b). Using this data there has been progress in NLP, in subfields such as speech recognition (Coto-Solano, 2021), forced alignment (Coto-Solano and Solórzano, 2016; Solórzano and Coto-Solano, 2017; Coto-Solano et al., 2022), machine translation (Feldman and Coto-Solano, 2020; Kann et al., 2022; Jones et al., 2023) and the study of semantics through embeddings (Coto-Solano, 2022). The work also includes the development of tools to extend the usage of the language, such as keyboards (Solórzano, 2010) and digital dictionaries (Krohn, 2020).

There are a few labelled datasets for the language (e.g. Ebrahimi et al. 2021), and one of them is a dependency treebank (Coto-Solano et al., 2021) tagged with Universal Dependencies v2 (Nivre et al., 2020) and stored in the CoNLL-U format. This treebank contains 315 sentences (1572 tokens) from some of the unlabeled sources above, and it includes information on part-of-speech and dependency arcs and labels. Figure 1 shows an example parse from this treebank.

(1) *Ye' tö ù sú̱* 'I saw the house'



Morphological analysis is one of the areas where there has been previous NLP research for Bribri. The state-of-the-art tagger is the Flores-Solórzano (2017b) FOMA-based tagger, which was built to tag the oral corpus (Flores-Solórzano, 2017a). It uses a finite state transducer (FST) which takes one word at a time, processes its characters one at a time, and follows a path that will ultimately lead the FST to an end node with a list of possible morphological features. Table 1 shows the morphological features for example sentence 1. The first word, *ye'* 'I' is correctly predicted as a first person singular pronoun. The second word *tö*, has three possible predictions: it could be a verb, a conjunction, or the ergative postposition. Here the third option is the correct one, but the FST does not output probabilities, so knowing this would require a human determination or an additional module. The third word, *ù* 'house' is correctly predicted as a noun (*sustantivo* in Spanish), but the tag does not specify the absolutive case that the noun is in. Finally, the fourth word *sú̱* 'saw' only has +? as its morphological tag. This means that the FST could not find the word amongst its states, and therefore cannot provide any morphological information.

Word	Features
ye'	`+1PSg`
tö	`te+V+Imp1Intran`
	`+Conj[subordinada]`
	`+Posp[Erg]`
ù	`u+Sust`
sú̱	`+?`

Table 1: FOMA-based morphological features for the Bribri sentence *Ye' tö ù sú̱* 'I saw the house'.

2.2 Assignment of Universal Features

The challenge we are trying to solve is to improve existing morphological tagging so that it can tag any word in the Bribri language, not just those in the existing FST. In order to do this, we wrote a series of rules (regular expressions) to tag the sentences using Universal Dependency Features. These regular expressions were created based off orthological patterns determined by the researchers with support from previously established character patterns noted in resources such as Flores-Solórzano's work with verbal conjugation (Flores-Solórzano, 2017b). In this paper we will focus our tagging on some of the major parts of speech: verbs, nouns, pronouns, adjectives and copulas. Here we detail the rules for the parts of speech we selected for this work.

2.2.1 First Pass of the Verbs

The first step in processing this data is to compile a list of the verbs present in the text, and specify which were transitive and which were intransitive. Given a transitive or intransitive verb in the surrounding context, we can then determine the case of nouns and pronouns, such as the ergative and absolutive cases which are vital to the structure of Bribri.

After this verb list is compiled, we used a series of regular expressions to assign Universal Dependency morphological features to each of the verbs. Once a `VERB` label is found in the part of speech column of the CoNLL-U, a regular expression would be used to find the conjugation of the verb and find its morphological features. For example, the regular expression `r".*r$"` is used to find the imperfective middle voice verbs. These tags would then be inserted into the 5th position of a new CoNLL-U file. For example, the verb *tkër* 'to be sitting' matches the previous regular expression, and so it would receive the tags `Aspect=Imp|Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Mid`.

2.2.2 Nouns

After the first pass of the verbs, nouns were analyzed for their plurality using the regular expression `r".*pa$"`, which triggers the tagging of that `NOUN` with `Number=Plur`. Then the cases of the nouns were determined by the presence of transitive or intransitive verbs either directly after or two words after the noun. `NOUN` subjects near verbs in the transitive verb list would receive the `Case=ERG` tag. `NOUN` subjects near verbs in the intransitive verb list would receive the `Case=ABS` tag. Finally, if the noun was an object, it would receive the `Case=ABS` tag as well.

2.2.3 Adjectives

For the most part, Bribri adjectives do not show number agreement with their nouns. However, there are a few adjectives which have irregular plural forms. For example, the word *tsîr* 'small' has the plural form *tsítsi*. We manually assessed the adjectives in the treebank and tagged the irregular plurals as `Number=Plur`.

2.2.4 Pronouns

Pronouns were analyzed in the same way as the nouns and checked for `Case` and `Number`. However, unlike nouns, the pronouns were also tagged for `Person` and for `Type`, such as personal and reciprocal pronouns. The first person plural pronouns were also tagged for `Clusivity` (i.e. *se'* 'inclusive we' and *sa'* 'exclusive we').

Possessive pronouns were also tagged. They are phonologically the same as the personal pronouns (compare *ye'* 'I' with *ye' ù* 'my house') so their `Poss=Yes` status was determined their by position directly preceding `NOUN` tokens.

2.2.5 Second Verb Pass and Copulas

A second round of verb analysis was then completed so that the newly tagged `Person` features of nouns and pronouns could be used to determined the `Person` value for verbs that appeared in the immediate context of those nominals. If the `VERB` has a `NOUN` subject, then the tag `Person=3` is assigned to the `VERB` automatically. If a `PRON` is the subject, then the `Person` of the `VERB` is directly copied from that of the `PRON`.

The copula *dör* is a special part of speech. Copulas do not behave morphologically like verbs: They don't have TAM suffixes like most action verbs, and they don't have plural forms like most positional verbs. Copulas, however, are obligatory in equative sentences, and pronoun subjects can present weak forms next to both verbs and copulas. Because of this similarity to verbs, copulas were tagged for `Person` in the same way as verbs, associating their `Person` to the that of the surrounding nouns or pronouns.

2.3 Training and Statistical Comparisons

The procedure described above was used to tag the treebank automatically. After the first and second passes, a manual revision was carried out by the researchers to correct the errors of the rule-based predictions. Approximately 24% of the 330 verbs were not recognized by the regular expressions, and so they were corrected manually by the authors, using the Flores-Solórzano (2017b) verbal description, the Jara (2018) grammar, the Constenla et al. (2004) textbook and the Krohn (2021) dictionary as our main references. The surrounding context of the verb was also referenced to support this manual correction process. All of the nouns and copulas were tagged correctly as predicted by the rules, but some of the possessive pronouns needed manual correction and this was undertaken in the same fashion as the aforementioned manual correction of some verbs. The irregular plural adjectives *tsítsi* ‘small.PL’ and *tsîrala’ralar* ‘tiny.PL’ were tagged for number manually because fitting regular expressions were not developed for these forms.

We used this new, morphologically-tagged CoNLL-U file to train twenty parsing models using UDPipe 2 (Straka, 2018). We trained these using a cloud-based system with a V100 GPU. Each model took approximately 1.5 hours to train and test, for a total of 30 hours of processing. The hyperparameters can be found in Appendix A. Once these models were trained, we calculated the precision of the feature tagging for each of them and used this information to compare the system’s performance with that of the FOMA-based tagging.

3 Results

At the end of the tagging process, a word would have its Universal Dependencies’ morphological features in the corresponding CoNLL-U column. Table 2 shows an example of a sentence and its features.

Word	POS	Features
ye’	PRON	Case=ERG|Number=Sing|Person=1|PronType=Prs
tö	PART	_
ù	NOUN	Case=ABS|Number=Sing
sú̱	VERB	Aspect=Perf|Mood=Ind|Tense=Past|VerbForm=Fin|Voice=Act

Table 2: Universal Dependency Features morphological features (UFEATS) for *Ye’ tö ù sú̱* ‘I saw the house’

3.1 Tags after Correction

After the manual corrections, there was a total of 3051 morphological features in the annotated treebank. Table 3 shows the total of features for each part of speech in the annotated dataset. The majority of the tags were dedicated to the verbs (n=1504, 49%), in particular the tense-mood-aspect (TAM) markers. There are also numerous tags for the distinction between active and middle voice, which is crucial in the description of Bribri grammar.

Part-of-Speech	Morphological Feature	n
Verb	Aspect=Imp	138
	Aspect=Prosp	45
	Aspect=Perf	65
	Mood=Des	1
	Mood=Imp	3
	Mood=Ind	245
	Person=1	47
	Person=2	16
	Person=3	32
	Polarity=Neg	3
	Tense=Pres	152
	Tense=Past	97
	VerbForm=Inf	63
	VerbForm=Fin	267
	Voice=Mid	62
	Voice=Act	268
Noun	Case=ABS	69
	Case=ERG	5
	Number=Plur	10
	Number=Sing	246
Adjective	Number=Plur	3
Pronoun	Case=ABS	121
	Case=ERG	19
	Clusivity=Ex	8
	Clusivity=In	11
	Number=Sing	274
	Number=Plur	51
	Person=1	136
	Person=2	38
	Person=3	131
	Poss=Yes	39
	PronType=Dem	16
	PronType=Int	12
	PronType=Prs	307
	PronType=Rcp	4
	Reflex=Yes	7
Copula	Person=1	14
	Person=2	6
	Person=3	20

Table 3: Part-of-Speech and Tagged Universal Features

Error	n	
TAM missing	10	28%
Case missing	9	18%
Hallucinated features	7	14%
Person missing	6	12%
Number missing	5	10%
Others	12	25%

Table 4: Types of errors in the output for morphological features in an example UDPipe 2 model (total n=49)

Pronouns were the next category in importance (n=1174, 38%). Most of the tags were for person and number, followed by case tags for those pronouns that were either the syntactic ergative or absolutive in the sentence. Importantly, the 1st person plural pronouns were also marked for clusivity (i.e. exclusive or inclusive), and the non-personal pronouns were marked for their function (e.g. demonstrative, interrogatives, reciprocals and reflexives). Nouns had the third most features (n=330, 11%). Like in the case of the pronouns, they were marked for number, and for case if they occupied a core argument (ergative or absolutive) position in the sentence.

Copula features (n=40, 1%) only have tags for the person that the copula refers to. Finally, the three irregular plural adjectives in the corpus were tagged with the corresponding plural feature.

3.2 Parsing Model Tests

Once the dataset was tagged, we used it to train a series of UDPipe 2 models in order to test whether this relatively small dataset could be used to expand our morphological tagging capabilities. We used the 315 sentences in the annotated treebank to create twenty random train/dev/test partitions (80%, 10%, 10%) and train the models. The average precision for the Universal Features (UFEATS) was 80.5 ± 3.6.

After this we randomly selected one of the models and analyzed the errors it produced. The test set contained 304 features, and 49 of these were predicted inaccurately (16%). Table 4 shows a summary of the errors produced by the model in the output hypotheses for the test set.

The most frequent errors are missing features that the model couldn't predict. Out of all of the errors, 28% were those where the TAM features was missing. 18% of the errors were the result of a missing absolutive or ergative case marking in the output, and a further 22% were because either the person (first, second or third) or the number of a word were missing.

There are also a few errors in the system where the word is assigned the correct part of speech but a wrong feature. For example, a verb could be assigned the past tense when it was actually in the present, or the verb could be tagged as having a 1st person subject, even though the gold standard had it as a 3rd person.

Finally, it is worth noting that the model does have a tendency to hallucinate morphological features. There were some parts of speech, like numerical classifiers and determiners, whose morphological tagging was not included in this work (more on this in the discussion below). However, the system would produce features for them. In the case of a numerical classifier like *bö̀k* 'two [round things]' the model treated this as a VERB and gave it features for finiteness and active voice. In the case of the determiner *i'* 'this one', the system misclassified the part of speech as a pronoun and then gave it features for singular number and tagged it as a demonstrative, probably because of its phonetic (but not syntactic) similarity with the demonstrative pronoun *i'* 'this one'.

3.3 Comparison with FOMA Parsing

While a direct comparison between the FOMA tags and the UFEATS is not possible due to the difference in their tagging conventions, we can estimate their difference in providing a tentative tag for unseen Bribri data.

In order to calculate an error for the FOMA, we devised the following test. We took the test sets from each of the twenty random samplings of the treebank. We took the words in those test sets and tagged them individually using Flores-Solórzano's (2019) FOMA tagger. This can only be done word by word because the system is based on an FST, and cannot get information from preceding or subsequent words. Then, we classified FOMA's responses into three possibilities. First, if FOMA produces no output (+?), then we consider this an error. Second, if FOMA produces more than one output (e.g. saying that the word *dör* is both the ergative marker and a copula), we consider this an error. This is because the system has no probabilistic information in its output, and it would be impossible to determine which of the two tags is correct without an additional module that considered context. The third condition is if the FOMA provides only one answer (e.g. labeling the word *ye'* 'I' as `+1PSg`). We assume this is a correct answer because of the FST nature of the FOMA system: it identifies a word directly and then it has a pre-programmed set of morphological outputs for it. Importantly, we calculate this for all the tokens in the UDPipe2 predictions, including those that are tagged with an empty response _, which is a correct gold-standard answer for words that don't have tags yet (e.g. postpositions and numerical classifiers).

When we calculate the results according to these three conditions, we get that, for the twenty runs, the average precision of the FOMA system is 59.5 ± 4.2. This is lower than the 80.5 ± 3.6 result for the UDPipe 2 model; this comparison is shown in figure 1. In fact, a paired t-student test revealed that the deep learning system performs significantly better with the same test sets (t(19)=16.5, p<0.00001)[2].

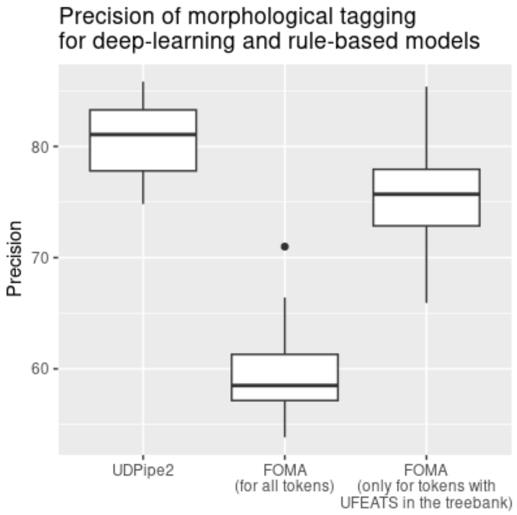


Figure 1: Precision for morphological tagging for a deep learning model (UDPipe 2), a rule-based FST model (FOMA) and a rule-based model that only looks at words with existing UFEAT tags in the gold standard.

We conducted a second, more rigorous test. In this test, we only considered words that actually have features in the gold standard treebank. As mentioned above, there are many words, for example postpositions, particles and numerical classifiers, which only have the marker _ in their feature column. In this second test we will only include tokens if the original treebank had actual UFEATS in it. After this modification, the precision of the FOMA system increases to 75.7 ± 4.7. Figure 1 shows the distribution of the twenty samples, under the condition *FOMA (only for tokens with UFEATS in the treebank)*. While this precision is higher than the FOMA for all the tokens, it is still significantly lower than the precision of the UDPipe 2 model (t(19)=3.4, p<0.005).

These results confirm that the deep learning model trained from our tagged treebank shows improvement in the state of the art for morphological tagging in the Bribri language.

4 Discussion

Overall, the rule-based tagging of the verbs was difficult due to their morphological complexity, and numerous manual corrections were needed. We had specific regular expressions for over 80 verbs, and so the rules described in section 2.2 would not be easily transferable to larger segments of written Bribri. However, our objective in using these rules was to create a new system which could accept forms it hasn't seen before as its input. The morphological feature tags we have introduced to the treebank produce acceptable results during inference. Our future work is to take this new treebank and use it to make morphological and syntactic parsings of unseen sentences of Bribri in order to expand existing corpora.

The most immediate item of future work is to expand the tags for the remaining parts of speech. For example, Bribri's deictic system includes pronouns that refer to distance from the speaker (near, far) and vertical position from the reference point (above, even and below). For example, the word *aí* means 'that one (above, near)', and the word *dià* means 'that one (below, far)'. It also includes deictics which need the feature `Deixis=NVis` (not visible), like the word *ñe̱'* 'that one (that can be heard but not seen)'. These are tags that already exist in the Universal Features, and should be easy to expand upon.

There are also places where the parts of speech treated here could be expanded. For example, Bribri has several diminutive morphemes for nouns and adjectives (e.g. *a̱mì̱* 'mother' versus *a̱mí̱la* 'mommy'). These would take the feature `Degree=Dim`, but this was not included in the present work. These morphemes are important

[2] A Shapiro-test confirms that all distributions meet the normality assumption: $p_{UDPipe2}$=0.25, $p_{FomaAll}$=0.06, $p_{FomaUfeats}$=0.99.

for the studying of Bribri discourse, and so their tagging is necessary in the future.

More complex to tag are numerical classifiers. These classifiers contain the number, but also semantic information about the geometry of the object. Some examples are: (i) *buà bö̀töm* ‘two[long] iguanas’, (ii) *apë’ bö́l* ‘two people’, (iii) *àshali bö̀k* ‘two[round] oranges’, and (iv) *kua’kua bö̀t* ‘two[flat] butterflies’. There are at least 8 of these classes, and their tagging cannot be described with the features in Universal Features. That additional information would have to be included separately.

Finally, there is additional information about the verbs that also needs to be saved separately. For example, Bribri verbs distinguish between "recent" and "remote" past perfect tenses. For example, the sentence *ye’ shka’* means ‘I walked (sometime yesterday, before I went to bed last night, or further back in the past)’. On the other hand, the sentence *ye’ shké* is also perfect, but it covers both the immediate present perfect (e.g. ‘I will walk’), and a perfect aspect action that has occurred in the recent past, after the last time one went to bed (e.g. ‘I walked (sometime today, in the recent past)’. This recent tense has also been called the *hodiernal* tense in literature (Dahl, 1983). This distinction cannot be described in the Universal Features, and would have to be stored separately as well.

One piece of future work is to make a system that performs automatic morphological segmentation. Such a system would get the input *Shkàne̲* ‘There was walking’, and would be able to produce the output `shk-àn-e̲`, with the root `shk` ‘walk’, the middle voice suffix `-àn` and the remote past tense perfect aspect suffix `-e̲`. We hope that the feature tagging described in this paper will be helpful in making such a segmentation system, which would further contribute to the creation of annotated corpora.

5 Conclusions

In this paper we have presented a new morphological tagger for the Bribri language. We automatically tagged an existing treebank with Universal Dependencies’ Universal Features. We hand-corrected any errors during the tagging process, and then used this new treebank to train a parsing model. This model has significantly better performance than the previous FST-based analyzer. We will continue to expand upon this work, using these tools to aid in the annotation of corpora for the language.

Limitations

The system is limited to written Bribri, which might hinder its usability for other applications, as most speakers of Bribri do not write the language. and much of the data we ultimately want to tag is oral narratives. Moreover, the writing system represented in the dataset is only one of the orthographies currently in use for the language, and so an input system that can easily accept all orthographies would need to be deployed alongside this tagger in the future.

Ethics Statement

The work was done using openly available materials published by Costa Rican institutions (e.g. University of Costa Rica). The models will be used to work on corpora construction, in collaboration with Bribri community members who work on the linguistics of the language.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, et al. 2022. UniMorph 4.0: universal morphology. *arXiv preprint arXiv:2205.03608*.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Rolando Coto-Solano. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A case study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467.

Rolando Coto-Solano, Sharid Loáiciga, and Sofía Flores-Solórzano. 2021. Towards Universal Dependencies for Bribri. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 16–29.

Rolando Coto-Solano, Sally Akevai Nicholas, Brittany Hoback, and Gregorio Tiburcio Cano. 2022. Managing data workflows for untrained forced alignment: examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu. *The Open Handbook of Linguistic Data Management*, 35.

Rolando Coto-Solano and Sofía Flores Solórzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de Costa Rica. *Kánina*, 40(4):175–199.

Östen Dahl. 1983. Temporal distance: Remoteness distinctions in tense-aspect systems. *Linguistics*, 21(1).

Fineen Davis, Eddie Antonio Santos, and Heather Souter. 2021. On the computational modelling of Michif verbal morphology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2631–2636.

Haley De Korne and Miranda Weinberg. 2021. “I Learned That My Name Is Spelled Wrong": Lessons from Mexico and Nepal on Teaching Literacy for Indigenous Language Reclamation. *Comparative Education Review*, 65(2):288–309.

Marie-Catherine de Marneffe, Christopher D Manning, Joachim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 255–308.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, et al. 2021. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976.

Sofía Flores-Solórzano. 2017a. Corpus oral pandialectal de la lengua bribri. `http://bribri.net`.

Sofía Flores-Solórzano. 2017b. *Un primer corpus pandialectal oral de la lengua bribri y su anotación morfológica con base en el modelo de estados finitos*. Ph.D. thesis, Universidad Autónoma de Madrid.

Sofía Flores-Solórzano. 2019. La modelización de la morfología verbal bribri - Modeling the Verbal Morphology of Bribri. *Revista de Procesamiento del Lenguaje Natural*, 62:85–92.

Alí García Segura. 2016. *Ditsö̀ rukuö̀ - Identity of the seeds: Learning from Nature*. IUCN.

INEC. 2011. X Censo Nacional de Población y VI de Vivienda 2011 - Territorios Indígenas - Principales Indicadores Demográficos y Socioeconómicos.

Carla Victoria Jara. 2018. *Gramática de la lengua bribri*. E-Digital ED.

Carla Jara Murillo and Alí García Segura. 2022. Së́bliwak Francisco García ttö̀. `https://www.lenguabribri.com/las-palabras-de-francisco`.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttö́ bribri ie Hablemos en bribri*. E Digital.

Alex Jones, Rolando Coto-Solano, and Guillermo González Campos. 2023. TalaMT: Multilingual Machine Translation for Cabécar-Bribri-Spanish. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 106–117.

Katharina Kann, Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, John E Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo A Giménez-Lugo, et al. 2022. AmericasNLI: Machine translation and natural language inference systems for Indigenous languages of the Americas. *Frontiers in Artificial Intelligence*, 5:995667.

Anna Kazantseva, Karin Michelson, Jean-Pierre Koenig, et al. 2024. Fitting a Square Peg into a Round Hole: Creating a UniMorph dataset of Kanien'kéha Verbs. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–51.

Sujay Khandagale, Yoann Léveillé, Samuel Miller, Derek Pham, Ramy Eskander, Cass Lowry, Richard Compton, Judith L Klavans, Maria Polinsky, and Smaranda Muresan. 2022. Towards unsupervised morphological analysis of polysynthetic languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 334–340.

Haakon Krohn. 2020. Elaboración de una base de datos en XML para un diccionario bribri–español español–bribri en la web. *Porto das Letras*, 6(3):38–58.

Haakon S. Krohn. 2021. Diccionario digital bilingüe bribri. `http://www.haakonkrohn.com/bribri`.

Anastasia Kuznetsova and Francis Tyers. 2021. A finite-state morphological analyser for Paraguayan Guaraní. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 81–89.

Jordan Lachler, Lene Antonsen, Trond Trosterud, Sjur Moshagen, and Antti Arppe. 2018. Modeling Northern Haida Verb Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ngoc Tan Le and Fatiha Sadat. 2021. Towards a low-resource neural machine translation for indigenous languages in Canada. *Traitement Automatique des Langues*, 62(3):39–63.

Zoey Liu, Robert Jimerson, and Emily Prud'Hommeaux. 2021. Morphological segmentation for Seneca. In *First Workshop on Natural Language Processing for Indigenous Languages of the Americas*.

Ariadna Font Llitjós, Roberto Aranovich, and Lori Levin. 2005. Building Machine Translation Systems for Indigenous Languages. In *Second Conference on the Indigenous Languages of Latin America (CILLA II), Texas, USA*.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018a. Probabilistic finite-state morphological segmenter for Wixarika (Huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018b. Challenges of language technologies for the indigenous languages of the Americas. *arXiv preprint arXiv:1806.04291*.

Enrique Margery. 2005. *Diccionario fraseológico bribri-español español-bribri*, second edition. Editorial de la Universidad de Costa Rica.

Arya D McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of The 12th language resources and evaluation conference*, pages 3922–3931. European Language Resources Association.

MEP. 2013. *Los Bribris y Cabécares de Sulá - Tomo 1 - Minienciclopedia de los Territorios Indígenas de Costa Rica*. Ministerio de Educación Pública de Costa Rica.

Rodolfo Mercado-Gonzales, José Pereira-Noriega, Marco Antonio Sobrevilla Cabezudo, and Arturo Oncevay. 2018. ChAnot: An intelligent annotation tool for indigenous and highly agglutinative languages in Peru. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Benjamin Molineaux. 2023. The Corpus of Historical Mapudungun: Morpho-phonological parsing and the history of a Native American language. *Corpora*, 18(2):175–191.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Hyunji Park, Lane Schwartz, and Francis Tyers. 2021. Expanding universal dependencies for polysynthetic languages: A case of St. Lawrence island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*.

Robert Pugh and Francis Tyers. 2021. Towards an open source finite-state morphological analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 80–85.

Juan Diego Quesada. 2007. *The Chibchan Languages*. Editorial Tecnológica de Costa Rica.

Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon, and Fabrício F Gerardi. 2022. Tupían language ressources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58.

Jack Rueter, Mika Hämäläinen, and Khalid Alnajjar. 2023. Modelling the Reduplicating Lushootseed Morphology with an FST and LSTM. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 40–46.

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en Costa Rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42.

Sofía Flores Solórzano. 2010. Teclado Chibcha: Un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, pages 155–161.

Sofía Flores Solórzano and Rolando Coto-Solano. 2017. Comparison of Two Forced Alignments Systems for Aligning Bribri Speech. *CLEI Electronic Journal*, 20(1):2–1.

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.

Carlos Sánchez Avendaño, Alí García Segura, et al. 2021a. *Se' Dalì Diccionario y Enciclopedia de la Agricultura Tradicional Bribri*. Editorial de la Universidad de Costa Rica.

Carlos Sánchez Avendaño, Alí García Segura, et al. 2021b. *Se' Má Diccionario-Recetario de la Alimentación Tradicional Bribri*. Editorial de la Universidad de Costa Rica.

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

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

UCREL. 2011. UCREL CLAWS7 Tagset.

A UDPipe2 Hyperparameters

```
batch_size: 32
beta_2: 0.99
char_dropout: 0
cle_dim: 256
clip_gradient: 2.0
dropout: 0.5
epochs: [(3, 0.001), (3, 0.0001)]
exp: None
label_smoothing: 0.03
max_sentence_len: 120
min_epoch_batches: 300
parse: 1
parser_deprel_dim: 128
parser_layers: 1
predict: False
predict_input: None
predict_output: None
rnn_cell: LSTM
rnn_cell_dim: 512
rnn_layers: 2
rnn_layers_parser: 1
rnn_layers_tagger: 0
seed: 42
single_root: 1
tag_layers: 1
tags:['UPOS','XPOS','FEATS','LEMMAS']
threads: 4
variant_dim: 128
we_dim: 512
wembedding_model: bert-base-multilingual
    -uncased-last4
word_dropout: 0.2
```