Towards Universal Dependencies for Bribri

Rolando Coto-Solano[1] **Sharid Loáiciga**[2] **Sofía Flores-Solórzano**[3]
[1]Department of Linguistics, Dartmouth College
[2]CLASP, Department of Philosophy, Linguistics & Theory of Science, University of Gothenburg
[3]Ministry of Public Education, Costa Rica
rolando.a.coto.solano@dartmouth.edu sharid.loaiciga@gu.se
sofia.flores.solorzano@mep.go.cr

Abstract

This paper presents a first attempt to apply Universal Dependencies (Nivre et al., 2016; de Marneffe et al., 2021) to Bribri, an Indigenous language from Costa Rica belonging to the Chibchan family. There is limited previous work on Bribri NLP, so we also present a proposal for a dependency parser, as well as a listing of structures that were challenging to parse (e.g. flexible word order, verbal sequences, arguments of intransitive verbs and mismatches between the tense systems of Bribri and UD). We also list some of the challenges in performing NLP with an extremely low-resource Indigenous language, including issues with tokenization, data normalization and the training of tools like POS taggers which are necessary for the parsing. In total we collected 150 sentences (760 words) from publicly available sources like grammar books and corpora. We then used a context-free grammar for the initial parse, and then applied the head-floating algorithm in Xia and Palmer (2001) to automatically generate dependency parses. This work is a first step towards building a UD treebank for Bribri, and we hope to use this tool to improve the documentation of the language and develop language-learning materials and NLP tools like chatbots and question answering-systems.

Resumen

Este artículo presenta un primer intento de aplicar Dependencias Universales (Nivre et al., 2016; de Marneffe et al., 2021) al bribri, una lengua indígena chibchense de Costa Rica. Dado el limitado trabajo existente en procesamiento de lenguaje natural (PLN) en bribri incluimos también una propuesta para un analizador sintáctico de dependencias, así como una lista de estructuras difíciles de analizar (e.g. palabras con orden flexible, secuencias verbales, argumentos de verbos intransitivos y diferencias entre el sistema verbal del bribri y los rasgos morfológicos de UD). También mencionamos algunos retos del PLN en lenguas indígenas extremadamente bajas en recursos, como la tokenización, la normalización de los datos y el entrenamiento de herramientas como el etiquetado gramatical, necesario para el análisis sintáctico. Se recolectaron 150 oraciones (760 palabras) de fuentes públicas como gramáticas y corpus y se usó una gramática libre de contexto para el análisis inicial. Luego se aplicó el algoritmo de flotación de cabezas de Xia y Palmer (2001) para generar automáticamente los análisis sintácticos de dependencias. Este es el primer paso hacia la construcción de un treebank de dependencias en bribri. Esperamos usar esta herramienta para mejorar la documentación de la lengua y desarrollar materiales de aprendizaje de la lengua y herramientas de PLN como chatbots y sistemas de pregunta-respuesta.

1 Introduction

This paper presents a first attempt to conduct dependency parsing in Bribri, an Indigenous language spoken in southern Costa Rica (Glottolog `brib1243`). There is an increasing number of Universal Dependency treebanks (Nivre et al., 2016; de Marneffe et al., 2021) available for Indigenous languages of the Americas: e.g. Yupik (Chen et al., 2020; Park et al., 2021), Arapaho (Wagner et al., 2016),

Hupa (Spence et al., 2018), Maya K'iche' (Tyers and Henderson, 2021), Shipibo-Konibo (Vasquez et al., 2018), Guaraní (Thomas, 2019), Apurinã (Rueter et al., 2021) and several Tupí languages from Brazil (Ferraz Gerardi et al., 2021).[1] However, there is no previous work on any member of the Chibchan family, a language family spoken in lower Central America, Colombia and Venezuela, so this paper seeks to address this gap and contribute to the automated syntactic analysis of these languages.

Bribri is a Chibchan language spoken by approximately 7000 people (INEC, 2011). It is a vulnerable language (Moseley, 2010; Sánchez Avendaño, 2013), still spoken by many adults and some children but mostly restricted to settings inside the home. Bribri is an morphologically ergative language (McGregor, 2009; Quesada, 1999; Pacchiarotti and Kulikov, 2021), with SOV word ordering, head-internal relative clauses and numeral classifiers. There has been some previous work on Bribri NLP: The first was the keyboard of Flores-Solórzano (2010), which allowed the language to be typed easily into computers and cellphones. The language also has an electronic Bribri-Spanish dictionary (Krohn, 2020; Krohn, 2021) and a morphological analyzer (Flores-Solórzano, 2019; Flores-Solórzano, 2017b), and there have been experiments in speech recognition (Coto-Solano, 2021), forced alignment (Coto-Solano and Flores-Solórzano, 2016; Coto-Solano and Flores-Solórzano, 2017), neural machine translation (Feldman and Coto-Solano, 2020; Mager et al., 2021) and natural language inference (Ebrahimi et al., 2021). This paper seeks to expand the work of Bribri NLP into the area of syntax and automated parsing, in the hopes of generating tools that help in the documentation and ultimately the revitalization of the language.

2 Methodology

In this section we will present the workflow that we followed for this first experiment. We collected sentences from various data sources (grammar books and oral corpora). We then tokenized the sentences and extracted the POS tag for each word. After that we designed a constituent grammar to perform the first automatic parse, and an algorithm to convert those constituent parses into dependency parses.

2.1 Data sources

For this first attempt we selected 150 sentences, containing 760 words. These ranged in complexity from simple structures (e.g. *Shkèna* 'Hello', lit. 'to have woken up') to entire conversations. For example, the longest sentence comes from an oral narration and contains 58 words. The sentences come from either published or Creative Commons licensed sources, specifically the textbook of Constenla et al. (2004), the grammar of Jara (2018) and the spoken Bribri corpus of Flores-Solórzano (2017a), and they included examples from the Amubri, Coroma and Salitre dialects. Most sentences were isolated examples, originally intended to illustrate the grammar of Bribri and chosen for the variety of their syntantic structures. However, the dataset also includes two short stories; one of them is in conversational style and it includes speech phenomena such as *reparanda* disfluencies (Universal Dependencies Contributors, 2021).

One major challenge is the normalization of the written data. As is the case with many Indigenous languages, where the orthography is of recent creation and created by outsiders to the community, there is considerable variation in how Bribri is represented in writing. There are four main sources of variation: (a) Different authors use different writing systems. For example, Constenla et al. (2004) uses a line underneath the vowel to indicate nasality, whereas Jara (2018) uses a tilde diacritic and Margery (2005) uses a Polish hook. Therefore, the word 'pot' can be found as *ù*, *ũ̀* or *ų̀* (all of them pronounced [ũ˥])[2]. (b) Phonological variation is not represented consistently. For example, the word *ami* [ã˧ˈmĩ˥] 'mother' can also be written *mì* because the unstressed vowel in the first syllable can be deleted. (c) There is variation across dialects. The word 'road', for example, is *ñalà* [ɲã˧ˈɽã˥] in the Amubri dialect and *ñolõ̀* [ɲõ˧ˈɽõ˥] in the Coroma dialect. Finally, (d) there is considerable idiosyncratic variation in and between documents, as would be expected of any language where the writing system has been recently adopted. During this work, the word 'much' has been found as *taî* (Constenla et al., 2004), *tãì*, *tãĩ̀*, *tã́ĩ* (Jara, 2018), *tã̂ĩ* (Pacchiarotti and Kulikov, 2021), *tai*, *tái*, *taí*, *tái*, *taí*, *táin*, *táin*, *taín* and *táìn* (MEP, 2017).

[1] There are some non-UD treebanks for languages like Quechua (Rios et al., 2008) and Karuk (Garrett et al., 2013).

[2] Bribri is tonal: The high tone is indicated by an grave diacritic (ù), the falling tone by the acute diacritic (ú), the low rising tone by a circumflex (û), and the low/neutral tone (Coto-Solano, 2015) is indicated by the lack of a diacritic (u).

The two NLP tools publicly available for Bribri, the keyboard layouts and the morphological analyzer (section 2.2 below) use the Constenla orthography. It is also used by the Ministry of Education of Costa Rica in school classes. Therefore, we will use that system here. However, when the Bribri treebanks are released, they will be made available in the two main orthographies, the ones in Constenla et al. (2004) and Jara (2018), and some orthographic variation might have to be standardized. When the automated dependency parser is released, it will have to be made resilient to the variation exemplified above, so that it can effectively tag and parse text that deviates from spelling norms. This is particularly important because, given Bribri's status as a vulnerable language, the main role of researchers at this stage should be to incentivize the creation of Bribri written materials, not to strictly enforce orthographic standards.

2.2 Tokenization and POS Tagging

The oral corpus in Bribri.net (Flores-Solórzano, 2017a) includes a unigram-based morphological analyzer (Flores-Solórzano, 2019). This program uses the finite-state analyzer FOMA (Hulden, 2009) to analyze each word. Example (1) shows Bribri words and their FOMA output. The FOMA was then used to extract the lemma and to extrapolate the part-of-speech for each word. For example, the word *ù* 'house' has the FOMA `ù+Sust` 'noun', so this word would be tagged as a noun with the lemma `ù`.

(1) Bribri Ye' tö ù sú
FOMA `+1PSg +Posp[Erg] ù+Sust su+V+PerfImp`
Gloss *I* ERG *house* *saw*
'I saw the house' (Constenla et al., 2004, 52)

Because the program was unigram based, it is not sensitive to context and its output can include several possibilities for the morphological analysis. For example, the word *tö* is the ergative marker in sentence (1). When this word is entered as input to the FOMA, it produces three different outputs. These were used in combination with the surrounding words to decide the most appropriate POS for a given word.

One important issue for future work is tokenization. There are a few forms, like the reduced ergative marker *r* and the clitic pronouns, that can be attached to other words. The examples in (2) show the 3rd person absolutive clitic. Different authors deal with the clitic in different ways: they attach it directly to the verb, as in (2a), they separate it with a dash, as in (2b), or sometimes they write it separately, as in (2c). In this first experiment the clitics and the ergative markers were separated manually and stored separate from other words, but the parser needs to be made more resilient to these variations.

(2) a. E'kuék és **i**kíe dör
because like.this **3SG.ABS**=to.be.called COP
'That's why they call it like this' (García Segura, 2016, 11)

b. Ie' mìne **i**-mauk
3SG went **3SG.ABS**=tie.INF
'She went to tie it up' (Constenla et al., 2004, 47)

c. Ma se' tö **i** kiè ema dlásháwö
well we ERG **3SG.ABS** call well ginger
'Well, we call it, ah, ginger' (Valengana and Flores-Solórzano, 2017)

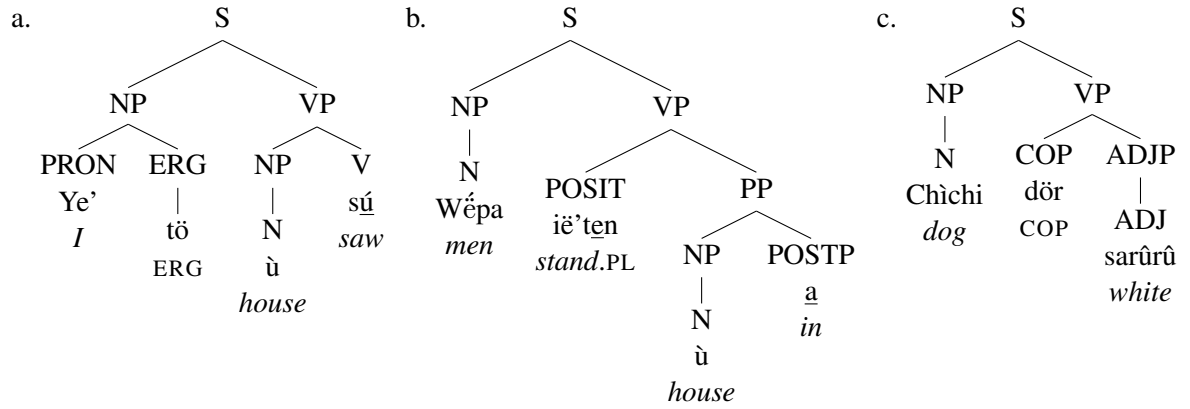
2.3 Constituency Parsing

The next step was the parsing of Bribri. We created an n-ary context-free grammar (CFG) (Chomsky, 1956; Hopcroft and Ullman, 1979) to model Bribri syntax[3], implemented using NLTK in Python (Bird et al., 2009). The grammar contains 122 rules: 10 for sentences, 14 for NPs, 52 for VPs, 23 for terminals and 23 for other non-terminal structures. Example (3) shows example sentences parsed with this grammar[4]. The grammar had to be complemented with filters to reject invalid parsings. For example, the parser rejects sentences where the main verbal phrase doesn't contain a finite verb.

[3]There were early attempts to make transformational grammars of Chibchan languages like Bribri and Cabécar (Bourland, 1976; Wilson, 1986), but most work in Bribri syntax has taken place within the functionalist tradition. There are some works, like Coto-Solano (2009), Coto-Solano et al. (2015) and Pacchiarotti (2016) which have elements of generative theories like Government and Binding and Minimalism.

[4]The current version of the CFG grammar is available at http://github.com/rolandocoto/bribri-cfg.

(3) CFG parses for transitive, intransitive and copular sentences: (a) *Ye' tö ù sú* 'I saw the house' (Constenla et al., 2004, 52), (b) *Wɛ́pa ië'ten ù a* 'The men are in the house (standing)' (Constenla et al., 2004, 67) and (c) *Chìchi dör sarûrû* 'The dog is white' (Constenla et al., 2004, 60).



This grammar can parse most simple sentences and some complex sentences, such as adverbial clauses and verbal complements. However, there are some complex structures, such as relative clauses, that cannot be parsed by the current iteration of the parser[5]. These sentences were decomposed into simpler structures and then linked together manually into a single CFG tree.

2.4 Dependency Parsing

We used the method of Xia and Palmer (2001) to raise the heads of the CFG subtrees and establish the dependencies between words. We then wrote a series of rules to establish the relations between dependencies; the relations were drawn from version 2.8 of Universal Dependencies, henceforth UD. After this first pass, some parses had to be automatically corrected to match the UD standards. For examples, copular sentences needed to be corrected to make the attribute the head. After setting the relations we converted the Bribri-specific parts of speech to Universal POS tags (UPOS). Several parts of speech were merged into a single UPOS (e.g. verbs and positional verbs were merged into UPOS VERB). Finally, the parser extracted the features of verbs and adverbs with negative polarity. The features of nouns, pronouns and determiners are pending in the current iteration of the parser.

3 Results: Common Structures in Bribri

The methodology described above was used to automatically generate dependency parses for 150 Bribri sentences. Table 1 shows the percentage of UPOS tags in the dataset. The four most common parts of speech, PRON, VERB, NOUN and ADP, account for 73% of the words in the corpus.

PRON	183 (24%)	ADP	68 (9%)	PUNCT	28 (4%)	ADJ	14 (2%)	NUM	5
VERB	163 (21%)	PART	44 (6%)	AUX	26 (4%)	DET	10 (1%)	CCONJ	3
NOUN	146 (19%)	ADV	39 (5%)	PROPN	21 (3%)	SCONJ	7 (1%)	INTJ	3

Table 1: UPOS tags in the Bribri sentences. Counts without percentages are less than 1% of the total.

Table 2 shows the relations found in the corpus. The most common relations, *root* and *nsubj*, account for 40% of the total. There are some relations, like *reparandum*, that are found infrequently, but could become more frequent as the corpus is expanded with conversational data from oral narrations.

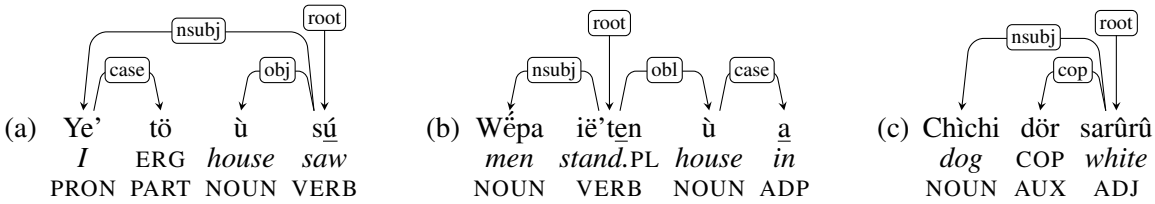
Example (4) shows dependency parses for transitive, intransitive and copular sentences. These are the same sentences that were shown as CFG parses in example (3) above. They show three different objects as roots: a verb (*sú* 'saw'), a positional verb (*ië'ten* 'to be in a place, standing') and an adjective as the attribute of a copula (*sarûrû* 'white'). They also show basic relationships such as *nsubj* for ergative and absolutive subjects, *obj* for an absolutive direct object, and *obl* for an oblique argument.

[5]Out of the 150 sentences, 104 (70%) were parsed completely automatically. For 23 of the sentences (15%), the correct POS was provided manually and the CFG and DepParses were generated automatically. For another 23 of the sentences (15%), both the POS tag and the CFG parse were provided manually and the DepParse was generated automatically.

nsubj	154 (20%)	cop	26 (3%)	advcl	8 (1%)	intj	3
root	150 (20%)	nmod	18 (2%)	amod	6	appos	2
case	89 (12%)	nmod:poss	18 (2%)	nummod	5	ccomp	2
obl	78 (10%)	xcomp	16 (2%)	compound	4	fixed	1
advmod	63 (8%)	conj	12 (2%)	acl:recl	3	reparandum	1
obj	46 (6%)	mark	11 (1%)	cc	3		
punct	28 (4%)	det	10 (1%)	flat	3		

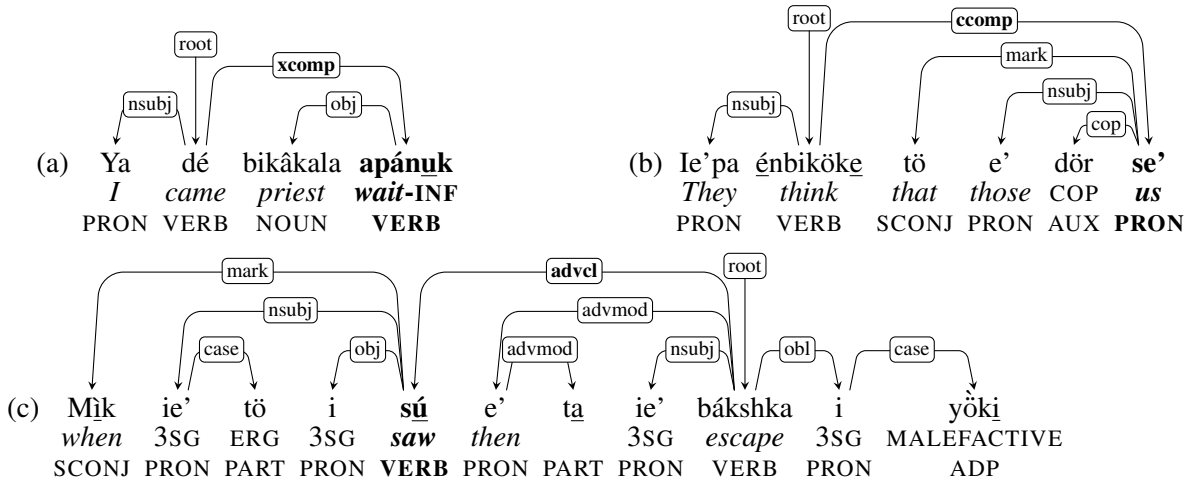
Table 2: Relations in the Bribri sentences. Counts without percentages are less than 1% of the total.

(4) Dependency parse for transitives, intransitives and copulas: (a) *Ye' tö ù sú* 'I saw the house' (Constenla et al., 2004, 52), (b) *Wẽ́pa ië'tẽn ù a* 'The men are in the house (standing)' (Constenla et al., 2004, 67) and (c) *Chìchi dör sarûrû* 'The dog is white' (Constenla et al., 2004, 60).



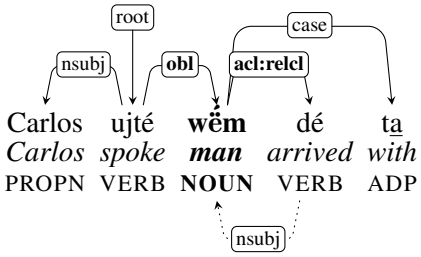
The examples in (5) show more complex sentences. The sentence in (5a) has a clausal complement marked with *xcomp*, the phrase *bikâkala apánuk* 'to wait for the master of ceremonies' (a type of priest). The sentence in (5b) has a copular clause as a direct object, and so it is marked with the *ccomp* relation. The sentence in (5c) includes an adverbial clause that precedes the main clause. Therefore, the head of the subclause is connected to the root using the *advcl* relation.

(5) Dependency parse for (a) *Ya dé bikâkala apánuk* 'I came to wait for the master of ceremonies (priest)' (Constenla et al., 2004, 47), (b) *Ie'pa énbikökẽ tö e' dör se'* 'They think that those [spirits] are one of us' (Constenla et al., 2004, 114) and (c) *Mìk ie' tö i sú e' ta ie' bákshka i yökì* 'When he saw him, he ran away from him' (Constenla et al., 2004, 112).



All of the previous examples were parsed automatically by the CFG grammar and then converted automatically into a dependency parse. However, example (6) shows a complex clause that cannot yet be parsed. This is a head-internal relative clause, the main type of relative clause in Bribri (Coto-Solano et al., 2015). The sentence *Carlos ujté* ***wẽ́m*** *dé ta* 'Carlos spoke with the man that arrived' has the main verb *ujté* 'spoke' and the relativized verb *dé* 'arrived'. (Bribri does not have an attributive conjugation, so the main and subordinate verbs have the same morphological forms). The head of the relative clause is *wẽ́m* 'man', which is an oblique argument to the main verb and the subject of the relativized verb.

(6) Dependency parse for *Carlos ujté wëm dé ta* 'Carlos spoke with the man that arrived' (Constenla et al., 2004, 54). It includes an enhanced dependency for the subject of the relative clause.



Because this structure cannot be parsed by the CFG it can't be converted to UD automatically. We parsed it separately as two clauses, and then joined them manually as a single constituency parse, which was then converted to a dependency parse using the procedure described above. This clause is also noteworthy in that an enhanced dependency was included to mark the relation between the relativized verb and the head of the relative clause. Further research needs to be conduct in order to parse these in a fully automated fashion.

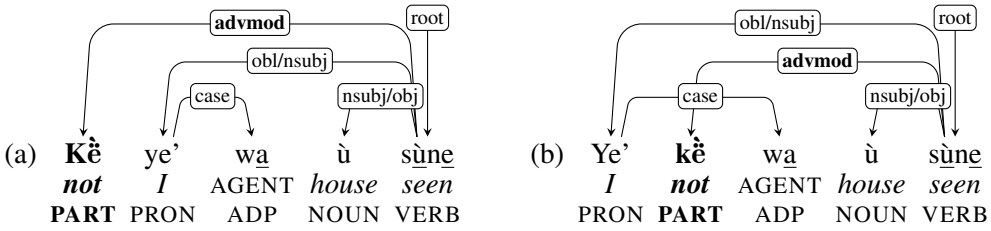
4 Challenging Bribri Structures

There were numerous challenges during the process of dependency parsing. Here we will focus on four of them: (a) structures with flexible order, (b) the treatment of sequences of verbs and positional verbs, (c) the relations of arguments in sentences with middle voice verbs, intransitives of motion and possession, and (d) the differences between UD tense features and the Bribri tense system.

4.1 Flexible word ordering

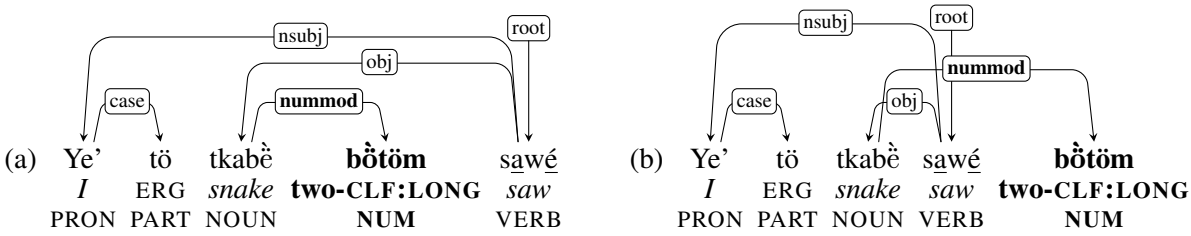
Bribri has several elements that admit flexible word-ordering, which can lead to non-projective parses. One such element is the negative adverb *kë̀* 'not'. In sentence (7a), the negative is at the edge of the sentence, without interfering with other relations. However, in sentence (7b), the negative particle is between the pronoun *ye'* 'I' and its case marker *wa*. (For whether the clause with *wa* should be labeled as *obl* or *nsubj*, see section 4.3 below).

(7) Dependency parse for (a) *Kë̀ ye' wa ù sùne* 'I didn't see the house' and (b) *Ye' kë̀ wa ù sùne* 'I didn't see the house' (Constenla et al., 2004, 53). The parse in (b) is non-projective.



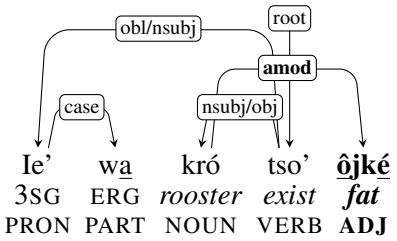
Another flexible structure is found when an absolutive noun is modified by a numeral or an adjective. In example (8a) 'I saw **the two** snakes', the noun is immediately followed by the numeral. However, in example (8b), 'I saw **two** snakes', the numeral is placed at the end of the sentence, and there is a verb between the noun and its numeral.

(8) Dependency parse for (a) *Ye' tö tkabë̀ bö̀töm sawé* 'I saw the two snakes' and (b) *Ye' tö tkabë̀ sawé bö̀töm* 'I saw two snakes' (Constenla et al., 2004, 70). The parse in (b) is non-projective.



Adjectives and participles can also show this behavior. Example (9) shows the adjective *ôjkẹ́* 'fat', which describes the noun *kró* 'rooster'. However, the noun-adjective connection crosses the connections of the root verb with its constituents. The current CFG parser can parse negatives and numerals, but the correct parsing of adjectives separate from their nouns remains for future work.

(9) Non-projective parse for *Ie' wạ kró tso' ôjkẹ́* 'She has a fat rooster' (Pacchiarotti, 2020, 254).



4.2 Positional Verbs as Auxiliaries

Example (10a) shows a sentence with the positional verb *dur* 'to be in a place, standing'. This positional would be the root of the dependency parse. Example (10b) has a sentence with the negative verb *kụ* 'not to be in a place'; this would also be the root of its sentence. However, example (10c) shows a sentence where both of these verbs are in a sequence. Which of the two should be the root?

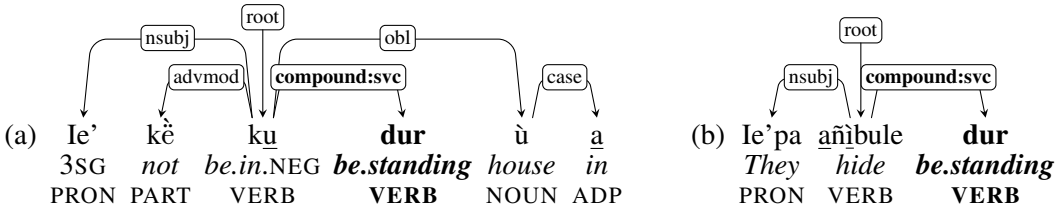
(10) a. Ie' **dur** ù ạ
3SG **ROOT:*be.standing*** *house in*
'He is (standing) in the house' (Constenla et al., 2004, 67)

b. Ie' kə̀ **kụ** ù ạ
3SG *not* **ROOT:*be.in*.NEG.IPFV** *house in*
'He is not in the house' (Constenla et al., 2004, 67)

c. Ie' kə̀ **kụ** **dur** ù ạ
3SG *not* ***be.in*.NEG.IPFV** ***be.standing*** *house in*
'He is not (standing) in the house' (Constenla et al., 2004, 67)

This second verb in this construction is not a light verb because both verbs contribute semantic content to the sentence. It is also not an auxiliary because it contains little or no information about tense, aspect, mood, voice or evidentiality. (These positional verbs do not take the set of TAM suffixes that other verbs do). Therefore, we will treat this sequence as an *asymmetrical serial verb* (Aikhenvald, 2006), where the first verb carries the TAM marking and the second verb contributes motion information to the sentence. We will also follow the analysis of Jara Murillo (2013), Pacchiarotti (2015) and Krohn (2017) and treat the first element of the verb chain as the root of the structure, and the positional verb as the secondary verb. Two examples of these serial structures are shown in (11).

(11) Dependency parse for (a) *Ie' kə̀ kụ dur ù ạ* 'He is not (standing) in the house' (Constenla et al., 2004, 47) and (b) *Ie'pa a̱ñi̱bule dur* 'They are hiding (standing)' (Jara, 2018, 203).



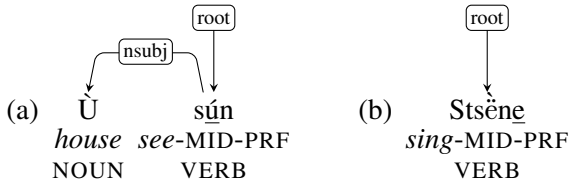
4.3 Core Arguments in Middle Voice and Intransitive Verbs

The marking of the core arguments of verbs is straightforward in most cases. As shown above, the ergative marker can be used to find the *nsubj*, and its presence or absence can be used to determine

whether the absolutive is an *nsubj* or *obj*. However, there are structures, like middle voice verbs and some intransitives, where this decision is more complex.

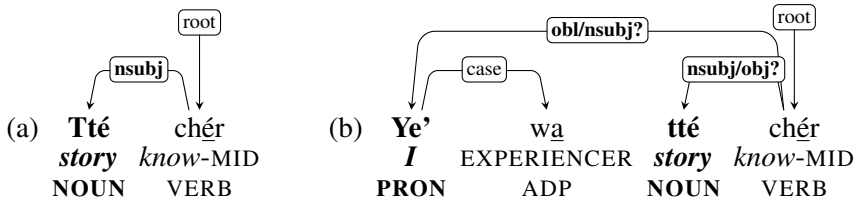
In Bribri middle voice verbs, the subject is usually the patient of the action, and the agent of the action is understood as an unspecified "general" agent. In (12a), *ù sún* 'houses are visible', the houses could be "seen" by anyone passing by. In sentence (12b), *stsë̀ne* 'there was singing', there is no specific person doing the singing. This would be similar to *on chante* or *ça chante* in French, or *man singt* in German.

(12) Dependency parse for (a) *Ù sún* 'Houses are visible' (lit: 'houses are seen') (Constenla et al., 2004, 84) and (b) *Stsë̀ne* 'There was singing' (Constenla et al., 2004, 26).



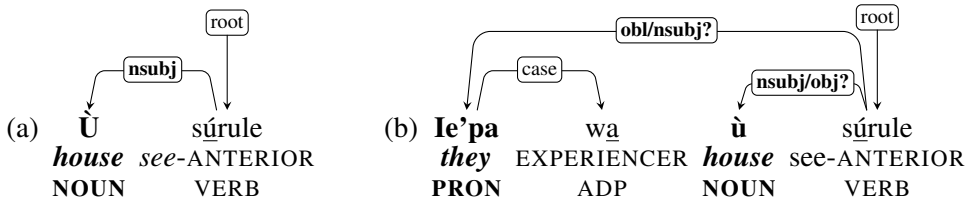
From a morphological point of view, middle verbs are not transitive, and should not be able to take agents. However, middle verbs can add an argument using the postposition *wa*. Sentence (13a), 'The story is known' is a typical middle voice structure. But sentence (13b) 'I know the story' has an additional argument to indicate who is experiencing the knowing of the story. This argument could be described as an oblique, and the noun 'story' could be the subject in both sentences. This is a consistent way to describe two verbs with identical morphology. However, there is a second alternative: The phrase *ye' wa* in (13b) could also be described as the ergative of the sentence, which would turn the noun 'story' into the direct object (Pacchiarotti and Kulikov, 2021, 4).

(13) Dependency parse for middle voice sentences: (a) *Tté chér* 'The story is known' and (b) *Ye' wa tté chér* 'I know the story' (lit: 'the story is known by me') (Pacchiarotti, 2016, 6).



This type of structure, where an argument is added using *wa*, is relatively frequent in Bribri. For example, *anterior* verbs (Constenla et al., 2004, 91), also called *antepresent* verbs (Jara, 2018, 72), are a type of pluperfect which are morphologically middle and can be used for middle voice meanings, as in (14a). But anterior verbs can take an additional argument which resembles an ergative, as in (14b).

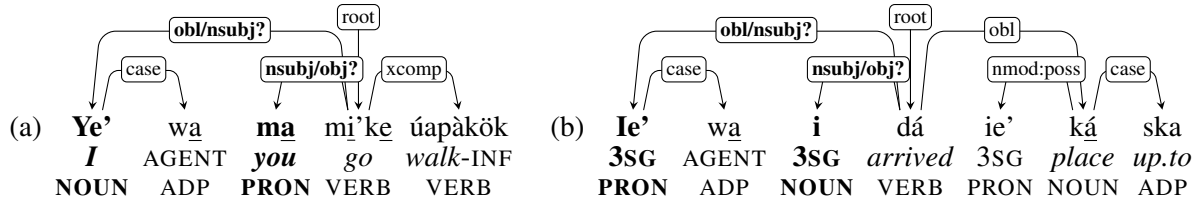
(14) Dependency parse for anterior verbs, derived from middle voice: (a) *Ù súrule* 'The house has been seen' and (b) *Ie'pa wa ù súrule* 'They have seen the house' (Constenla et al., 2004, 91).



What are the relations between the verb and the arguments in the sentences with *wa*? Following a morphological versus a semantic criterion would lead to different decisions. The sentences in (15) show motion verbs which are usually used as intransitives, but that here have an added argument for the person who causes the motion. Here the *wa* marks the causer of the movement, and the absolutive indicates the patient that is actually moved. Morphologically these verbs are intransitive, so it would make sense to label the *wa*-phrase as an oblique. On the other hand, the arguments are an agent and a patient, so they

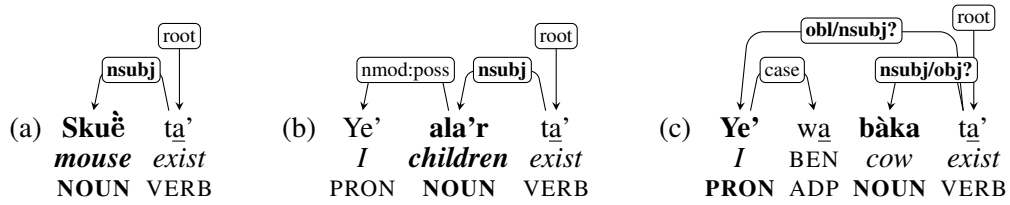
would resemble a regular ergative phrase, which would call for *nsubj*/*obj* relations coming out of the root.

(15) Dependency parse for sentences of motion: (a) *Ye' wa ma mi'ke úapàkök* 'I'll take you for a walk' and (b) *Ie' wa i dá ie' ká ska* 'She took her to her place' (Constenla et al., 2004, 117-118).



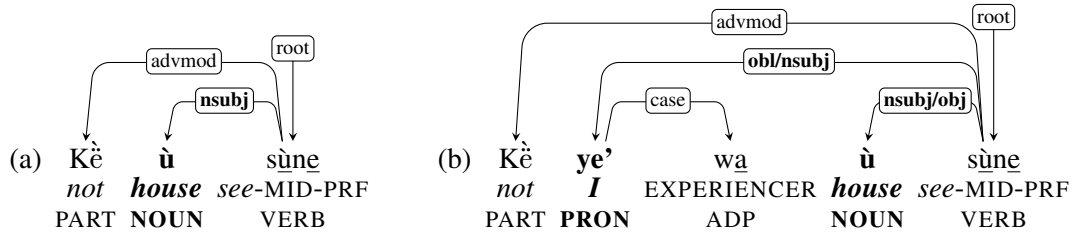
This question about how to tag the arguments of intransitives can also be seen when the verb *ta'* 'to exist' is used with alienable possessives. The sentence (16b), *Ye' ala'r ta'* 'I have children' has an inalienable possessive as its absolutive subject. Here, the possessor *ye'* 'I' is expressed as a modifier to the absolutive noun *ala'r* 'children'. On the other hand, sentence (16c) has the alienable possessor marked with *wa*. The argumentation here would be similar to that of the motion verbs: The verb *ta'* 'to exist' is morphologically intransitive and should therefore have only one core argument (the thing possessed), marked with *nsubj*. Moreover, this structure is similar to possessives in languages like Russian, where the possessor is marked with a preposition and the genitive case. On the other hand, the absolutive argument is a theme, so this would again make it a candidate for the *obj* relation.

(16) Dependency parses for existence and possession: (a) *Skuë ta'* 'There are mice', (b) *Ye' ala'r ta'* 'I have children' and (c) *Ye' wa bàka ta'* 'I have cows' (Constenla et al., 2004, 53, 74, 105).



The structures that constitute the strongest argument for labelling *wa*-phrases as *nsubj* are the transitive negatives. These are constructed using middle verbs, and the agent/experiencer does not receive its usual ergative marker. In sentence (17b), the experiencer is marked with *wa* and the theme is marked with the absolutive. In the corresponding positive version of the sentence, *Ye' tö ù sú* 'I saw the house' shown in (4a), the experiencer is marked with the ergative *tö* and the theme is again marked with the absolutive. Given the parallels between the two, it could be conceivable to mark the *wa*-structure with the *nsubj* relation and the absolutive with *obj* (Margery, 2005; Cruz Volio, 2010; Pacchiarotti, 2016). However, it would be equally useful to consistently mark the absolutive as the subject of the morphologically middle verb, so that both (17a) and (17b) have the word *ù* 'house' as their subject (Constenla et al., 2004; Jara, 1995; Barguigue, 2016).

(17) Dependency parse for (a) *Kë ù sùne* 'The house isn't seen' and (b) *Ye' kë wa ù sùne* 'I didn't see the house' (Constenla et al., 2004, 53).



So, which criterion to use, the morphological or the semantic? In the current version of our dependency parser we have chosen the relations to be consistent with the morphology of the verbs, and so

the absolutives of intransitive and middle voice verbs are marked as *nsubj*, and the other arguments are marked as *obl*. Further investigation into other syntactic properties of subjects is needed, and therefore the exact relations of these verbs could change in future iterations of the parser. One potential solution would be to mark the arguments as *obl/nsubj* in the dependencies and to use enhanced dependencies to further mark them as semantic *nsubj/obj* (Przepiórkowski and Patejuk, 2020).

4.4 Bribri Tenses and Universal Features

The Universal Feature system in UD includes the values {Past, Pres, Fut}[6] for the Tense feature. However, Bribri morphology does not match these categories, which makes the automatic extraction of features complex. The main verbal distinction in Bribri is aspect. It has perfect and imperfect verbs, and this does match the feature system. However, tense splits verbs in different ways. The temporal point of split between tenses is "the sunset of the night before" (Constenla et al., 2004, 15). This splits time into two tenses: the *remote* tense and the *recent* tense. The remote tense refers to actions that take place before yesterday's sunset, while the recent tense includes actions done in the recent past (e.g. today's morning), in the present (right now) and in the near future (e.g. "soon"). Table 3 shows examples of how these tenses interact with the aspect system. The *remote* tense is not problematic for automatic parsing, given that their UD tense will always be Past and their aspect can be determined from their morphology.

<table>
<tr><td rowspan="2">Aspect</td><td>UF Tense</td><td>Past</td><td>Past</td><td colspan="2">Present</td><td>Future</td><td>Future</td></tr>
<tr><td>Bribri tense</td><td>Remote</td><td>(today)</td><td colspan="2">Present</td><td>(near future)</td><td>Future</td></tr>
<tr><td>Perfect</td><td></td><td>Perfect remote
ya'
'drank'</td><td colspan="4">Perfect recent
yé
'drinks', 'drank'</td><td></td></tr>
<tr><td rowspan="3">Imperfect</td><td></td><td rowspan="3"></td><td colspan="3">Imperfect recent
yè
'drinks', 'was drinking'</td><td></td><td>Certain future
yèrâ
'will drink'</td></tr>
<tr><td></td><td colspan="5">Durative
yèke
'drinks', 'used to drink', 'shall drink'</td><td></td></tr>
<tr><td></td><td></td><td colspan="4">Future potential
yèmi̲
'can drink', 'shall drink'</td></tr>
</table>

Table 3: Examples of interaction between Bribri and the current version of Universal Features (UF) tenses in active voice verbs

The main issue comes with the verbs in the *recent* tense. This Bribri tense is similar to the *hodiernal* tense in Mwotlap (François, 2003), Haya, Luganda and Ancash Quechua (Comrie, 1985), in that the recent tense includes actions that have happened "today", regardless of whether they are in the past or in the near future. Depending on the context, the verbs in the recent tense could overlap with several of the time categories in Universal Features. For example, the imperfect recent form *yè* includes events that have happened before the present moment and simultaneous with the present moment, so this could be translated as 'drinks' or 'was drinking'. A sentence like *Ye' yè* could be translated as 'I drink it' or 'I was drinking it'. Without any contextual cues, it wouldn't be possible to automatically determine the appropriate tense in the Universal Feature system. There are other verbal forms, such as the *future potential* (Jara, 2018, 73), also called the *imperfect potential* (Constenla et al., 2004, 111), that also spread across two tenses of Universal Features. The sentence *Yì ki̲ be' kiàrmi̲?* (Jara, 2018, 73) can be translated as either a potential in the present tense, 'Who can love you?', or an imperfect future tense, 'Who shall love you?'. In this sentence there are no cues to aid the automatic parsing in selecting between the Tense=Pres and the Tense=Fut features.

[6]There are more Bribri verb forms that those mentioned here, and they include verbs in other tense categories of Universal Features. For example, the perfect antepresent form *yéule* 'to have drunk' would be marked with the feature Tense=Pqp.

The problem is even more pronounced with verbs in the *durative/habitual* form (Jara, 2018, 74), also called the *habitual imperfect* (Flores-Solórzano, 2017b, 34) and the *second imperfective* (Constenla et al., 2004, 90). The sentence *Ye' kanè̲blö̀ke* (Jara, 2018, 74) has an imperfect aspect, but it is spread across the recent tense. It can be translated as 'I used to work' in the recent past, 'I regularly work' in the habitual/present and 'I shall work soon' in the near future. In this sentence the tense feature could take three different values (Past, Pres, Fut), without a way to automatically distinguish between them using only the words in the sentence. One potential solution would be to leave the tense feature out of the description of these verbs, and add an annotation of their tense in the MISC field of the CONLL-U file. Another solution would be to add a feature such as Tense=Hod to the Universal Feature system, which would allow for a richer and more cross-linguistically faithful analysis of the UD database as a whole.

5 Conclusions and Future Work

This paper presents a first attempt to parse Bribri sentences using context-free grammars and dependency grammars, and it presents an adaptation of Universal Dependencies to Bribri. This preliminary effort illustrates the possibility of applying UD to Chibchan languages, but also the numerous challenges involved in implementing the task of automated parsing in Indigenous languages. In many ways these languages test the "U" in UD, and we hope that, by embracing languages where there aren't yet optimal solutions or linguistic consensus about their structures, this will help push the endeavor of Universal Dependencies forward. In future work we will expand to corpus to create a first treebank for Bribri and improve the parsers with the ultimate goal of releasing them for public use. We also seek to gather enough Bribri data in CONLL-U format so that we can train deep-learning based parsing methods like *UDPipe 2* (Straka, 2018), which might further accelerate the development of the treebank.

The parsing process presented here is done in the hope of developing tools that might be useful for the documentation and revitalization of Bribri. These should include NLP tools like chatbots and question answering systems, as well as linguistic tools like learning materials, exercises for students of Bribri, and more detailed documentation of the grammar of the language. One major challenge is to expand the process of annotation to include native speakers of Bribri. This would entail expanding the annotation process to non-automated tools, such as the manual annotation interfaces *UD Annotatrix* (Tyers et al., 2018) and *TrED* (Pajas and Fabian, 2000). Finally, we acknowledge the issues of data sovereignty with this work (i.e. non-Bribri researchers working on Bribri data). We have limited ourselves to data that is already publicly available, and in the future, we hope to expand the conversation with Bribri partners to ensure that the creation of NLP tools provides tangible benefits to Bribri partners and to the Bribri community in general.

References

Alexandra Y Aikhenvald. 2006. Serial Verb Constructions in Typological Perspective. *Serial Verb Constructions: A Cross-Linguistic Typology*, pages 1–68.

Saïd Barguigue. 2016. Predicados Afectivos en el Bribri: Un Acercamiento Tipológico-Funcional. Master's thesis, Universidad de Sonora.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

David Hawley Bourland. 1976. Una gramática generativa-transformacional del cabécar. *Revista de Filología y Lingüística de la Universidad de Costa Rica Vol. 2 Núm. 3*, pages 49–100.

Emily Chen, Hyunji Hayley Park, and Lane Schwartz. 2020. Improved Finite-State Morphological Analysis for St. Lawrence Island Yupik using Paradigm Function Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2676–2684.

Noam Chomsky. 1956. Three Models for the Description of Language. *IRE Transactions on information theory*, 2(3):113–124.

Bernard Comrie. 1985. *Tense*, volume 17. Cambridge University Press.

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

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.

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

Rolando Coto-Solano, Adriana Molina-Muñoz, and Alí García Segura. 2015. Correlative Structures in Bribri. *University of British Columbia Working Papers in Linguistics*, 43:27–41.

Rolando Coto-Solano. 2009. Reanálisis de las cláusulas relativas en la lengua bribri como un caso de linearización en la teoría minimalista. Memoria del II Congreso Internacional de Lingüística Aplicada (CILAP).

Rolando Coto-Solano. 2015. The Phonetics, Phonology and Phonotactics of the Bribri Language. In *2nd International Conference on Mesoamerican Linguistics. California State University, Los Angeles*.

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, Online, June. Association for Computational Linguistics.

Gabriela Cruz Volio. 2010. El sistema de transitividad en las cláusulas materiales del bribri según la gramática sistémico-funcional. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, pages 133–154.

Marie-Catherine de Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):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.

Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. 2021. Tudet: Tupían Dependency Treebank (version v0.2).

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. 2017a. Corpus oral pandialectal de la lengua bribri. `http://bribri.net`.

Sofía Margarita 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.

Alexandre François. 2003. *La sémantique du prédicat en Mwotlap, Vanuatu*, volume 84. Peeters Publishers.

Alí García Segura. 2016. *Ditsö Rukuö Identity of the Seeds: Learning from Nature*. IUCN.

Andrew Garrett, Clare Sandy, Erik Maier, Line Mikkelsen, and Patrick Davidson. 2013. Developing the Karuk Treebank. In *Fieldwork Forum, Department of Linguistics, UC Berkeley*.

John E Hopcroft and Jeffrey D Ullman. 1979. Introduction to Automata Theory, Languages and Computation. *Adison-Wesley*.

Mans Hulden. 2009. Foma: A Finite-State Compiler and Library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32.

INEC. 2011. Población total en territorios indígenas por autoidentificación a la etnia indígena y habla de alguna lengua indígena, según pueblo y territorio indígena. In Instituto Nacional de Estadística y Censos, editor, *Censo 2011*.

Carla Victoria Jara Murillo. 2013. Morfología verbal de la lengua bribri. *Estudios de Lingüística Chibcha*.

Carla Victoria Jara. 1995. Transitividad en el discurso bribri. *Revista de filología y lingüística de la Universidad de Costa Rica*, 21(2):93–105.

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

Haakon S Krohn. 2017. Semántica de los posicionales del bribri. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, 43(1):117–136.

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`.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez Lugo, Ricardo Ramos, et al. 2021. Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217.

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

William B McGregor. 2009. Typology of Ergativity. *Language and Linguistics Compass*, 3(1):480–508.

MEP. 2017. *Los Bribri y Cabécares de Sulá, Tomo 1 - Minienciclopedia de los Territorios Indígenas de Costa Rica*. Dirección de Desarrollo Curricular, Educación Intercultural. Ministerio de Educación Pública.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 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.

Sara Pacchiarotti and Leonid Kulikov. 2021. Bribri Media Tantum Verbs and the Rise of Labile Syntax. *Linguistics*.

Sara Pacchiarotti. 2015. The Argument Structure of some Caused Motion Constructions in Bribri: A Possible Explanation. In *18th Workshop on American Indigenous Languages (WAILS)*.

Sara Pacchiarotti. 2016. Verbal Deponency in the Chibchan Family. In *49th Annual Meeting of the Societas Linguistica Europaea*.

Sara Pacchiarotti. 2020. On the Origins of the Ergative Marker wã in the Viceitic Languages of the Chibchan Family. In *Reconstructing Syntax*, pages 241–288. Brill.

Petr Pajas and P Fabian. 2000. Tree Editor TrED, Prague Dependency Treebank, Charles University, Prague. *See URL http://ufal. mff. cuni. cz/˜ pajas/tred*.

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*, pages 131–142.

Adam Przepiórkowski and Agnieszka Patejuk. 2020. From Lexical Functional Grammar to Enhanced Universal Dependencies. *Language Resources and Evaluation*, 54(1):185–221.

J Diego Quesada. 1999. Ergativity in Chibchan. *STUF-Language Typology and Universals*, 52(1):22–51.

Annette Rios, Anne Göhring, and Martin Volk. 2008. A Quechua-Spanish Parallel Treebank. *LOT Occasional Series*, 12:53–64.

Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney da Silva Facundes, Mika Hämäläinen, and Niko Partanen. 2021. Apurinã Universal Dependencies Treebank. *arXiv preprint arXiv:2106.03391*.

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

Justin Spence, Zoey Liu, Kayla Palakurthy, and Tyler Lee-Wynant. 2018. Syntactic Annotation of a Hupa Text Corpus. Technical report, Working Papers in Athabaskan Languages: Alaska Native Language Center ....

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, October. Association for Computational Linguistics.

Guillaume Thomas. 2019. Universal Dependencies for Mbyá Guaraní. In *Proceedings of the third workshop on universal dependencies (udw, syntaxfest 2019)*, pages 70–77.

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 M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2018. Ud annotatrix: An annotation tool for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 10–17.

Universal Dependencies Contributors. 2021. reparandum: overridden disfluency.

Petronila Valengana Valengana and Sofía Flores-Solórzano. 2017. Wès sa' tsiru' chkà alèke - Cómo se prepara el cacao dulce. `https://bribri.net/B09h22m53s05sep2012.html`.

Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay. 2018. Toward Universal Dependencies for Shipibo-Konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161.

Irina Wagner, Andrew Cowell, and Jena D Hwang. 2016. Applying Universal Dependency to the Arapaho Language. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 171–179.

Jack L Wilson. 1986. Sobre la definición lingüística: el sujeto y el ergativo. *Estudios de Lingüística Chibcha*, pages 59–84.

Fei Xia and Martha Palmer. 2001. Converting Dependency Structures to Phrase Structures. Technical report, Pennsylvania Univ. Philadelphia.