# What conditions tone paradigms in Yukuna: Phonological and machine learning approaches

Magdalena Lemus-Serrano, Marc Allassonnière-Tang, Dan Dediu

Glossa
a journal of general linguistics

# What conditions tone paradigms in Yukuna: Phonological and machine learning approaches

RESEARCH

MAGDALENA LEMUS-SERRANO
MARC ALLASSONNIÈRE-TANG
DAN DEDIU
*Author affiliations can be found in the back matter of this article*

]u[ ubiquity press

## Abstract

Yukuna is an understudied Arawak language of North-West Amazonia with a privative tonal system. In this system, roots are underlyingly specified for tone, whilst affixes are toneless. However, affixation interacts with tone, leading to many variations in surface tonal patterns. This paper puts forth a qualitative analysis of Yukuna's tonal system, and provides data-driven evidence in favor of this analysis using machine learning methods. More precisely, we use decision trees and random forests to assess quantitatively the predictions of the phonological analysis. A manually annotated corpus of verbal paradigms was split into a training and a testing set. We trained the computational classifiers on the first and tested their predictions on the second. We found that they predict the majority of the patterns and support the qualitative analysis. Additionally, they suggest avenues for enhancing the phonological analysis, by providing a ranking of the variables that highlight statistical tendencies within tonal patterns. Besides its contribution to understanding tonal systems in general and of that of Yukuna in particular, our work also suggests that such machine learning approaches might become part of the complex theoretical and methodological toolkit needed for language description and linguistic theory development.

CORRESPONDING AUTHOR:
**Marc Allassonnière-Tang**
Dynamique Du Langage UMR 5596 CNRS, Université Lumière Lyon 2, FR
*marc.tang@univ-lyon2.fr*

KEYWORDS:
Yukuna; tone; machine learning; decision tree

TO CITE THIS ARTICLE:
Lemus-Serrano, Magdalena, Marc Allassonnière-Tang and Dan Dediu. 2021. What conditions tone paradigms in Yukuna: Phonological and machine learning approaches. *Glossa: a journal of general linguistics* 6(1): 60. 1–22. DOI: *https://doi.org/10.5334/gjgl.1276*

## 1 An overview of the Yukuna language

Yukuna is an Arawak language (ISO 693-3:ycn, Glottocode: yucu1253) spoken by under 1,000 speakers in various communities along the Miriti-Parana river in South Eastern Colombia (***Figure 1***). The Yukuna language is spoken by the Yukuna and Matapí ethnic groups, who are in intense, long-term contact with the Tukanoan speaking groups Tanimuka and Letuama (Fontaine 2001: 57). Despite the overall small number of speakers, the language continues to be transmitted to new generations within the Miriti Parana communities, so most ethnic Yukuna and Matapi of all ages speak their language. The relative stability of the language in the Miriti Parana contrasts with the sharp decline in vitality of the language when speakers move to nearby towns and cities, where Spanish and Portuguese are the dominant languages (Lemus Serrano 2016: 24).

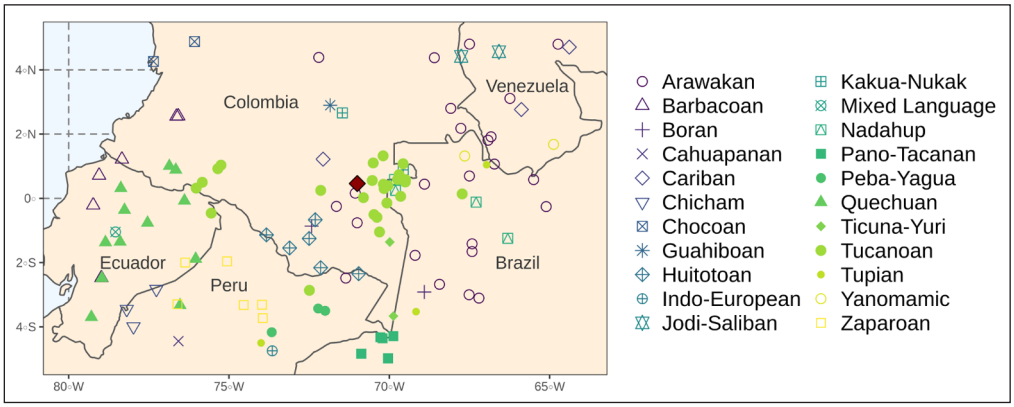


**Figure 1** Languages of Colombian Amazonia with Yukuna highlighted by a diamond shape (Hammarström et al. 2019).

In terms of morphosyntax, Yukuna is a nominative-accusative language with a canonical SVO constituent order. Core arguments are not case-marked, and obliques are marked with postpositions. The nominative argument of verbs (S) is obligatorily encoded in finite clauses either with a bound person index on the verb, or by an overt NP placed immediately before the verb. In contrast, the accusative argument of verbs (P) is neither obligatorily encoded, nor indexed on the verb. Verbs in main finite clauses can be marked for person (S), negation, tense, aspect and mood, but the only obligatory category is person. Lexical word classes in Yukuna are nouns, verbs, adverbs, adjectives, and postpositions. Non-verbal word classes can be used predicatively, in non-verbal clauses, with or without a copula (Schauer et al. 2005; Lemus Serrano 2020).

In terms of segmental phonology, Yukuna's consonant inventory is characterized by the absence of voiced stops, a near absence of fricatives (with the exception of /h/), one affricate (/tʃ/), a three way nasal distinction (/m, n, ɲ/), and a lateral/tap distinction (/l/, /ɾ/). The most important feature of this consonant inventory is the presence of an aspiration contrast in stops and sonorants, whereby all stops have an aspirated counterpart, and all sonorants have a voiceless counterpart. However, studies differ in terms of their analysis of the phonological status of aspirated consonants, either as consonant sequences involving /h/ (Ramirez 2001: 356), or as phonemes (Lemus Serrano 2020: 6).

The vowel inventory of Yukuna consists of five different vowel qualities (/a, e, i, o, u/), and two vowel features (creakiness, nasality). All vowels can be either plain (modal, oral), creaky (V̰), or nasal (Ṽ), with additionally a single vowel combining both features; creaky nasal /ã̰/. Note, however, that the analysis of the glottalization feature is also a matter of disagreement, with some studies opting to posit a phonemic /ʔ/ consonant (Schauer & Schauer 1972: 9).

The syllable structure of Yukuna is strictly (C)V, with no complex onsets and no codas. Vowel length is not phonemically distinctive, and diphthongs are analyzed as sequences of two syllable nuclei. There are thus no heavy syllables in Yukuna, and no syllable versus mora distinction. Roots and bound markers are typically mono- or disyllabic, although some trisyllabic roots are attested (Lemus Serrano 2016: 90).

Most existing studies on Yukuna are devoted to the description of the morphosyntax and segmental phonology of the language. The first and only study to specifically focus on the

word-prosodic system of Yukuna, and rightly describe it as a tonal language, is Robayo Romero (2018). In this paper, we focus precisely on this latter issue. Our aims are three-fold: first, we propose a phonological analysis of the tonal system of the language that addresses the attested variations in tonal paradigms. Second, we assess the statistical validity of the proposed analysis by using machine learning techniques (*classifiers*) to detect the factors that predict (i.e., statistically condition) with high accuracy the placement of H tones in a data set of verbal paradigms. Third, we observe tendencies in the statistical study that could suggest cues for future research on the tonal system of the language.

The structure of the paper is summarized as follows. Section 2 provides the qualitative analysis of the Yukuna tonal system, and discusses its contribution to previous literature on the language. Section 3 describes the corpus collection, annotation and dataset used for the quantitative analysis. Section 4 explains the functioning of the computational classifiers used for the analysis and show their results. Section 5 provides a discussion of the results generated by the computational classifiers and their implication for the qualitative analysis. The conclusion is given in Section 6.

## 2 The phonology of Yukuna's tonal system

Yukuna's word-prosodic system can be described as a restricted tonal system of the privative type, where H tone contrasts with the absence of tone. Such system corresponds to what is commonly labeled as 'pitch accent'. However, we avoid the latter term, following the property-driven typological phonology approach (Hyman 2009). In addition to having a limited tonal inventory (H vs. absence of H), Yukuna also displays low tonal density, in terms of the number and distribution of surface H syllables. When roots are produced without affixes, the following distributional rules apply: i) each root has at least one H syllable (e.g. *LLL), ii) each root has at most two H syllables (e.g. *HHH), iii) if there are two H tones, they are obligatorily adjacent (e.g. *HLH).

Phonologically, the distribution of surface H syllables can be captured by positing two distinct tones at the underlying level: a spreading H tone, represented as /H/, and a non-spreading H tone, represented as /HL/. Tone Bearing Units, which in Yukuna correspond to the syllable, are either specified for tone or toneless, leading to a /H/ vs. /HL/ vs. zero system. The /H/ (spreading) tone is associated to one underlying TBU, and spreads one position rightward to the next underlyingly toneless TBU. In this analysis, words with two surface H syllables only have one underlying /H/ tone, in agreement with the *Obligatory Contour Principle* (OCP) (McCarthy 1986), which disprefers adjacent identical features within a specific prosodic domain. The /HL/ (non-spreading) tone associates to one underlying TBU, and is phonetically produced as a single level H tone. The contrast between these two underlying tones is neutralized in word-final position, as they are both phonetically produced as level H tones on the final syllable. The behavior of underlying /H/ (transcribed with H tone diacritic) and /HL/ (transcribed with a falling diacritic) tones in word-initial versus word-final position is illustrated in ***Table 1***.

**Table 1** /H/ vs. /HL/.

| Tone | input | output | gloss |
|---|---|---|---|
| /H/ | /hápa/ | [hápá] | 'work' |
| | /paḷá/ | [paḷá] | 'lie' |
| /HL/ | /wâta/ | [wáta] | 'want' |
| | /wẹpî/ | [wẹpí] | 'know' |

Underlying tonal specification is a feature of roots. Each root has – at least and at most – one syllable specified for tone, either /H/ or /HL/. The specific tonal pattern of a root is entirely lexical; that is, there are no phonological, morphological or semantic features of a root that could predict its underlying tonal specification. Affixes and bound morphemes in general (clitics) are toneless. Despite being underlyingly toneless, suffixes vary in their surface

tonal value (sometimes H, sometimes L), depending on the root they combine with.[1] The surface tonal patterns brought about by suffixation reveal two additional interesting features of the Yukuna tonal system. First, that underlying /H/ tones may spread across morpheme boundaries. Second, that roots may be specified either with bound tones (as the roots in ***Table 1***) or with floating tones. Toneless suffixes are produced with a surface H tone either when they immediately follow a root with a root-final /H/ tone which spreads onto the suffix, or when they immediately follow a root with a floating tone. In all other cases, toneless suffixes are produced with a surface L tone. The behavior of underlying tones with suffixation is illustrated in ***Table 2***, which complements ***Table 1***.

**Table 2** Tone and suffixation.

| Tone | Type | root | root-suffix | gloss |
|---|---|---|---|---|
| /H/ | Bound | /hápa/ | [hápá-kahe] | 'work-NMLZ' |
| | | /paḷá/ | [paḷá-káhe] | 'lie-NMLZ' |
| /HL/ | | /wâta/ | [wáta-kahe] | 'want-NMLZ' |
| | | /wẹpî/ | [wẹpí-kahe] | 'know-NMLZ' |
| /H/ | Floating | /i̱n̥a$^{H}$/ | [i̱n̥a-káhé] | 'go-NMLZ' |
| /HL/ | | /hema̱$^{HL}$/ | [hema̱-káhe] | 'hear-NMLZ' |

As shown in ***Table 2***, floating tones may also be either /H/ or /HL/. The placement of floating tones follows the tendency of tone shift (Kenstowicz 1993; Kisseberth & Odden 2003). In Yukuna, a floating /H/ tone links to the first toneless TBU after the root, and then spreads one syllable rightward to the next toneless TBU (e.g. /i̱n̥a$^{H}$-kahe/ [i̱n̥a-káhé]). Floating /HL/ tones simply link to the first toneless TBU after the root, and surface as a single H syllable (e.g. /hema̱$^{HL}$-kahe/ [hema̱-káhe]). When no suffix is present, floating tones link to the final syllable of the root, and surface similarly to root-final bound /H/ and /HL/. This reveals that all underlying tones in Yukuna must surface, given the surface constraint for all phonological words to have at least one H syllable.

Because of these rules of floating tone association and H tone obligatoriness, four different underlying tonal patterns are produced identically -as a word-final H- in words without suffixes: root-final bound /H/ (e.g. /paḷá/ [paḷá] 'lie'), root final bound /HL/ (e.g. /wẹpî/ [wẹpí] 'know'), floating /H/ (e.g. /i̱n̥a$^{H}$/ [i̱n̥á] 'go'), and floating /HL/ (e.g. /hema̱$^{HL}$/ [hema?á] 'hear').[2] While this could be interpreted as a word-final default H tone in a language with H-tone obligatoriness, the fact that each of these roots has a different underlying tone becomes clear with suffixation.

The phonological analysis described thus far accounts for a major part of surface tonal patterns in root-plus-suffix combinations. However, irregular surface tonal patterns that cannot be explained with these rules are aplenty. Tonal irregularities mostly occur with /H/ tones at the morpheme boundary between roots and suffixes: root-final /H/ tones, as well as floating /H/ tones. Indeed, while the root-final underlying /H/ of the root *paḷá* 'tell lies' spreads onto the first syllable of the toneless suffix *-kaje* (NMLZ) in [paḷá-káhe] 'to lie', no tonal spreading occurs when the same root combines with other toneless suffixes, such as *-ri* (M) as in [paḷá-ri] 'he lies'. Likewise, while the underlying floating /H/ of /i̱n̥a$^{H}$/ surfaces as two H tones on toneless suffix *-kaje* in [i̱n̥a-káhé], it surfaces as only one H tone with suffix *-kare* (NMLZ) in [i̱n̥a-káre]. These irregularities cannot be accounted for in terms of suffix type, form nor function. Here, *-kahe* and *-kare* are both synchronically unsegmentable, disyllabic nominalizers.

In yet other cases, the surface tonal patterns in root-plus-affix combinations are so drastic that the underlying tonal pattern of the root is unclear. A case in point concerns the root *kema* 'say', which surfaces with two H tones on the suffix *-kaje* in [kema-káhé], but in other cases,

1 Note that surface tonal values are simply noted as H and L throughout this paper for practical purposes. However, the phonetic production of toneless syllables varies according to the position with respect to surface H tones.

2 Creaky vowels in word-final position require the insertion of an epenthetic copy vowel. This epenthesis leads to a resyllabification of the word, as the vowel forms an additional syllable able to carry H tone.

with two H tones on the root as in [kémá-hĩka] 'say-FAR.PST'. Irregular roots are not only problematic with respect to regular roots, but also, with respect to one another, as they do not display the same irregularities in the same way.

Lastly, irregular tonal patterns also arise depending on features of the suffix. In particular, there are two suffixes in the language that are underlyingly specified for tone: purposive subordinators /-ʧí/and /-ré/. These suffixes always surface with a H tone, irrespective of the root they combine with. This leads to rare tonal patterns, such as HH-H and HL-H sequences, as illustrated in ***Table 3*** with the suffix /-ʧí/.[3]

**Table 3** Tone and suffix /-ʧí/.

| Tone | TBU | root | root-ʧí | gloss |
|---|---|---|---|---|
| /H/ | Bound | /hápa/ | [hápá-ʧí] | 'work-PURP' |
| | | /paḻá/ | [paḻá-ʧí] | 'lie-PURP' |
| /HL/ | | /wâta/ | [wáta-ʧí] | 'want-PURP' |
| | | /wę̣pî/ | [wę̣pí-ʧí] | 'know-PURP' |
| /H/ | Floating | /i̱n̥a$^{H}$/ | [i̱n̥á-ʧí] | 'go-PURP' |
| /HL/ | | /hema̱$^{HL}$/ | [hema̱-ʧí] | 'hear-PURP' |

Yukuna's word-prosodic system is understudied. However, a few studies are worth mentioning here. The phonological sketch by Schauer & Schauer (1972: 71) explicitly mentions the phenomenon of shifting prosodic patterns caused by affixation, and provides an analysis of this phenomenon in terms of stress. Similar analyses are found in Lemus Serrano (2016: 131) and Ramirez (2001: 359). Robayo Romero (2018) is the first to analyze the word-prosodic system of the language as tonal, proving that pitch is the main correlate of prosodic prominence, and showing that surface HH sequences are attested in the language. Robayo Romero (2018) also observes the phenomenon of shifting tonal patterns with affixation, and provides an analysis in terms of tonal polarity, arguing that the first suffix after a root surfaces as L after a final H on the root, and as H after a root-final L.

The analysis presented here clearly aligns with previous studies in some respects, while also providing new insights into the prosodic system of the language. Our phonological proposal is based on two main empirical observations from our database. The first observation concerns the case of previously unmentioned tonal patterns, which cannot be accounted for in terms of tonal polarity, but which are present in our database. This is the case of surface H sequences across root-suffix boundary (as in [paḻá-káhe] 'lie-NMLZ'), or exclusively on the suffixes (as in [i̱n̥a-káhé] 'go-NMLZ'). Our phonological proposal accounts for these instances in terms of the tonal spreading of a single underlying /H/ tone (either bound or floating). This proposal has the advantage of grouping together all surface H sequences, whether in cases of roots with toneless suffixes, or root-internally (/hápa/ [hápá] 'work'), following the OCP. The second observation concerns the case of surface patterns which would be predicted by a process of tonal polarity (such as HL-H), but which are absent from our database (except in instances involving underlyingly H suffixes). For instance, /wâta/ 'want' has a root-final surface L tone, but toneless suffixes are not produced as H as in [wáta-kahe] 'want-NMLZ', see ***Table 2***). This is also captured by our phonological proposal, as the /HL/ tone in /wâta/ is simply produced as a single surface H on its corresponding TBU. Our phonological proposal thus accounts for the majority of patterns in the language, although some irregular patterns remain unexplained. We discuss these unexplained patterns in more detail in Section 5.

## 3 Materials

This section describes the corpus and dataset used as the basis for this study. First, we discuss the corpus collection and annotation process (3.1), then we introduce the variables of the dataset used for the statistical analysis (3.2).

3 Despite this singularity, both /-ʧí/and /-ré/ are suffixes, and not clitics. There is both segmental and tonal evidence in favor of this analysis, in particular, the fact that these suffixes carry in some cases the single surface H tone in a root + suffix combination, as with the verb 'hear' in Table 3.

### 3.1 Corpus collection and annotation

This study is based on a corpus containing the paradigms for a total of 80 verbal roots and stems. The verbal forms were obtained from two sources. The paradigms for 60 verbal roots and stems were obtained through direct elicitation with one consultant, a native speaker of Yukuna (anonymized in the current study), in Leticia, Colombia, between October 2017 and January 2018. During the elicitation sessions, the target words (individual root-plus-affix combinations per root) were recorded three times in non-final position within a frame sentence. The equipment used for the recording was a Zoom H4N audio recorder, with a Shure Beta 53 omnidirectional condenser head worn Microphone. The remaining 20 verbal paradigms were extracted from a corpus of Yukuna narratives collected by the first author, containing data from some 30 native Yukuna speakers. Finally, the collected paradigms were further validated during follow-up fieldwork sessions by different speakers in order to avoid overgeneralizing the potential biases and idiosyncrasies of a particular speaker or groups of speakers.

Each verbal paradigm includes a set of 30 different root-plus-affix combinations, including all of the most commonly used forms in the language corpora: person indexes (prefixes), inflectional suffixes (negation, tense), deranking suffixes (nominalizers, subordinators and gender markers).[4] The paradigm columns are the same for all verbal roots, but not all root-plus-affix combinations were possible for all verbs, nor attested in the corpus. Phonologically, the affixes in this dataset are monosyllabic and disyllabic, and they are inherently toneless, except for one suffix (the purposive subordinator /-tʃí/). This suffix was included due to its unique behavior among affixes, as discussed previously in Section 2.

Each entry from the paradigm is transcribed on two lines. One line with the alphabetic transcription of the target word, and one line with the tonal pattern and morpheme breaks. The alphabetic line used a slightly modified version of the practical Yukuna alphabet, based on the Spanish alphabet.[5] The Yukuna alphabet was a practical choice of annotation that allowed for a more surface-true transcription of H tones, which were simply marked with an acute accent on the syllable on which they were produced. Henceforth, Yukuna examples transcribed alphabetically are presented in italics. The tonal pattern line removed the segmental information from the alphabetic line, and annotated the tonal pattern of each entry (H for surface H tones, L for the rest) as well as the morphological boundaries (with dashes). This is illustrated in ***Table 4*** with a few tokens extracted from the paradigm of the verb /paḭ́á/ *pajlá* 'lie'.

**Table 4** An example of the verb paradigm data in Yukuna.

| | PR-V | V-KA | PR-V-KA |
|---|---|---|---|
| Alphabetic | *pipajlá* | *pajlákà* | *ripajlákà* |
| Tonal pattern | L-LH | LH-H | L-LH-H |

The annotation of surface H tones was done based on auditory impression for most tokens, but Praat was used to verify the pitch contours for the transcription of less common and/or problematic patterns. ***Figure 2*** provides the pitch contours for tokens illustrating six different types of tonal patterns attested in the dataset.

The six tonal patterns are coded as VH (one H tone on root, no suffix), VH_S (one H tone on root, toneless suffix), V_SH (toneless root, one H tone on suffix), VH_SH (one H tone on root, one H tone on suffix), VHH_S (two H tones on root, toneless suffix), and V_SHH (toneless root, two H tones on suffix).[6] As ***Figure 2*** indicates, the pitch contours show a marked fall in pitch after the H syllables, unless of course the H syllables are in word-final position (as in the VH and V_SHH patterns).

4 The verbal paradigms in this dataset do not include affixless roots, as these forms are less natural and harder to elicit. Additionally, affixless roots show the same tonal pattern as roots with prefixes but no suffixes.

5 The following alphabetic conventions are used: <j> /h/, <ñ> /ɲ/, <y> /j/, <V'> /V̲/ (creaky vowel), <Ch> /C$^{h}$/ (aspirated plosive), <jC> /C̥/ (voiceless sonorant).

6 Instances of root internal surface H tone sequences (VHH) are scarce in the dataset. These instances include two roots with underlying /H/ in root-medial position (/at$^{h}$úpa/ *athúpá* 'spit' and /aṃ̥úra/ *ajmúrá* 'sink'), as well as the irregular root *kémá* ~ *ímá* 'tell'. Other roots have been identified with this pattern in the language (such as /hápa/ *jápá* 'work') but they were not included in the dataset used in this study.

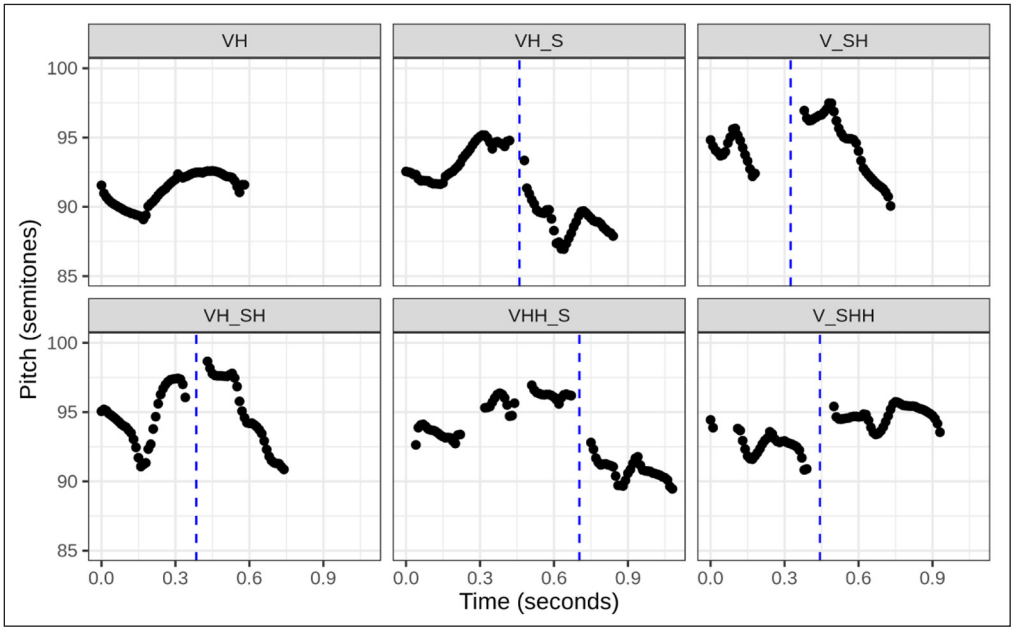


**Figure 2** The pitch contour of the main tonal patterns between the root and the suffix. The blue line indicates the limit between the root and the suffix. The patterns are extracted from the following examples: V_SH = *jo'káje* 'poke-NMLZ', V_SHH = *i'jna-kájé* 'go-NMLZ', VH_S = *amá-kaje* 'see-NMLZ', VH_SH = *ajá-káje* 'fly-NMLZ', VH =*pamá* 'you see', VHH_S = *athúpá-kaje* 'split-NMLZ'.

## 3.2 Dataset and variables

In addition to the corpus of verbal paradigms, the dataset includes seven variables shown in ***Table 5*** which were automatically extracted from the corpus using R (R-Core-Team 2020) and the *tidyverse* framework (Wickham 2017).

**Table 5** An overview of the features included in the data.

| Feature | Variables |
|---|---|
| RootTone | Absent, Initial, Final, Penultimate, Penultimate-Final |
| RootType | 1,2,3,4,5,6 |
| RootGlottal | Absent, Initial, Final |
| KajeTone | Absent, Initial, Final, Both |
| SuffixType | Je, Ri, Ka, Kana, Kare, La, Ta, Chaje, Kaje, Chi, Cha, Jika |
| PrefixMergeRoot | Yes, No |
| StructureType | Vx, Vxx, xV, xVx, xVxx |

Most of the variables chosen are purely descriptive (i.e. automatically extracted from the corpus) and concern factors that are believed to be directly or indirectly relevant to the tonal system of the language. Tonal variables simply list the positions of surface tones in root-plus-affix combinations: RootTone states whether there is a surface H tone on the root when the root carries a suffix, and if so, in which position. KajeTone indicates the position of the H tones on the suffix *-kaje*, i.e., initial, final, both syllables, or absent. Morpho-phonological variables pertain to the segmental and morphological make-up of the words. The variable PrefixMergeRoot states whether the prefix merges with the root, given that this process alters the resulting number of syllables in the word.[7] SuffixType lists the specific suffix in the root-plus-suffix combination, as it was previously shown that irregularities occur with suffixation in particular. StructureType refers to how many prefixes and suffixes are added in the word, i.e., only one suffix (Vx), two suffixes (Vxx), only one prefix (xV), one prefix and one suffix (xVx), one prefix and two suffixes (xVxx). Lastly, the variable RootGlottal refers to the position of the creaky vowels on the root (i.e., initial, final, or absent).[8]

7 For three irregular, monosyllabic roots (*la'* 'do', *a'* 'give', *ña'* 'weave'), the placement of surface H tones depends on the number of syllables in the word.

8 This variable is based on the observation that all roots with floating /HL/ tones have a root-final creaky vowel. Note however, that the reverse is not true, as not all roots with a final creaky vowel are underlyingly specified with a /HL/ floating tone.

The RootType variable is the only one which was not automatically extracted, as it pertains to the phonological analysis of tones proposed in Section 2. This variable groups together roots into six clusters according to their postulated underlying tonal specification. Type 1 contains highly regular roots that display no tonal interaction with suffixes. This includes roots with bound tones in root-initial position (e.g. /hápa/ *jápá* 'work', /wâta/ *wáta* 'want'), or a /HL/ tone in root-final position (e.g. /wẹpî/ *we'pí* 'know'). Affixes that combine with these roots always surface as L. Type 2 contains roots with a root-final /H/ tone, which spreads rightward with some toneless suffixes, but not all (e.g.*ajñá-ká* vs. *ajñá-ri*). Type 3 contains roots with a floating /HL/ tone, which are systematically produced with a single surface H tone on the first suffix after the root. Type 4 contains roots with a floating /H/ tone, which display multiple irregularities in the position and number of surface H syllables (e.g. *i'jná-ri* vs. *i'jna-kájé*). Type 5 is a small group of only three identified roots with an irregular surface tonal pattern affected by the number of syllables in the word. And lastly, Type 6 contains all highly irregular roots for which no underlying tonal pattern has been established. Unlike other root types, Type 6 is the only one that does not display any internal coherence, as roots in this type display tonal irregularities in different ways. The mapping of the underlying lexical tones presented previously, into the categories of the RootType variable used in the database is summarized in ***Table 6***.

**Table 6** Underlying tones and root types.

| type | Tone | TBU | input | output | gloss |
|---|---|---|---|---|---|
| 1 | /HL/ | bound | wâta-kaje | wáta-kaje | want |
| 1 | /HL/ | bound | amâ-kaje | amá-kaje | see |
| 1 | /H/ | bound | jápa-kaje | jápá-kaje | work |
| 2 | /H/ | bound | ajá-kaje | ajá-káje | fly |
| 3 | /HL/ | floating | jema'$^{HL}$-kaje | jema'-káje | hear |
| 4 | /H/ | floating | i'jna$^{H}$-kaje | i'jna-kájé | go |
| 5 | Irregular | | la'-kaje | la'-kajé | do |
| 6 | Irregular | | kema-kaje | kema-kájé | say |

We expect the RootType variable to be significant for the statistical analysis of the dataset in order to assess the validity of the phonological proposal. Since the phonological analysis itself is based on data observation, we also expect this variable to correlate with descriptive variables (such as RootTone). However, the variable RootType does not perfectly correlate with descriptive variables, precisely because many roots in the language display different tonal patterns throughout their paradigm. For instance, roots with floating tones (root types 3 to 6) display a VH tonal pattern (root-final H tone) when they are not carrying any suffixes due to H tone obligatoriness, but then, display V_SH, V_SHH or VH_SH when carrying suffixes. We also expect the RootType variable to be insufficient to account for all the surface patterns, and to be complemented by additional variables, given the diverse instances of irregularities in the dataset. For instance, roots included in Type 4 (floating /H/ tone) display both VH_SH as well as V_SHH patterns when combined with different suffixes.

An example of the variables and the coding conventions used in the dataset is shown in ***Table 7*** with the root /paḷá/ *pajlá* 'lie'.

**Table 7** A simplified example of the final dataset used for the analysis.

| Token | Pattern | RootType | StructureType | RootTone |
|---|---|---|---|---|
| pipajlá | VH | 2 | xV | Final |
| pajláká | VH_SH | 2 | Vx | Final |
| ripajláká | VH_SH | 2 | xVx | Final |

Lastly, a general look at the distribution of variables in the dataset reveals important tendencies. First, we note that the most frequent structures are Vx (root with one suffix) and xVxx (prefix, root and two suffixes), regardless of root type (***Figure 3***). This is due to the fact that most forms in the paradigm are morphologically complex. Indeed, while it is possible for verbal roots to occur without any affixes as phonological words in Yukuna, the preferred phonological word contains at least a root and a person prefix.

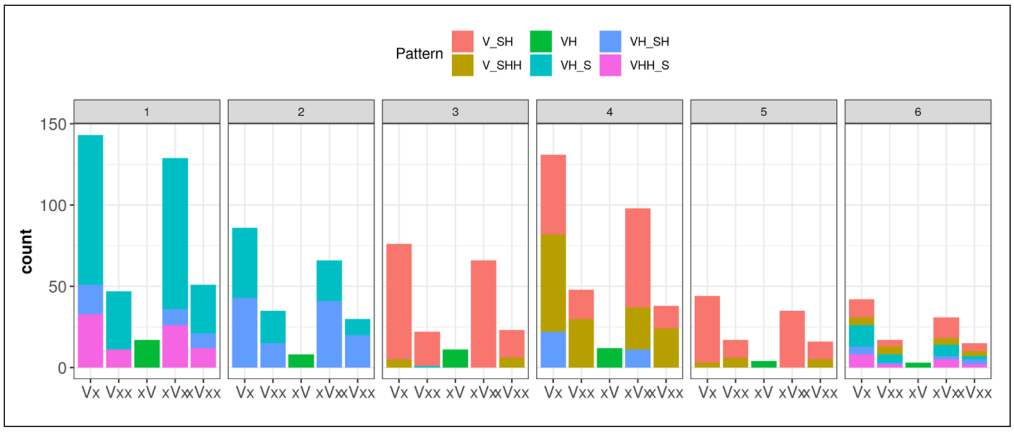


**Figure 3** An overview of the main variables in the data. The facets refer to the six main root types.

Second, we note that in terms of tonal patterns, the shares of the six main patterns are V_SH = 32.9% (448), VH_S = 27.7% (377), VH_SH = 14.7% (200), V_SHH = 13.4% (182), VHH_S = 7.3% (99), VH = 4% (55).[9] This is directly related to the number of roots per root type in the dataset. For instance, the V_SH and VH_S patterns are mostly associated with roots type 3 and 1 respectively, which together make up more than half of the roots in the dataset (root types 4, 5, and 6 are smaller). However, and importantly, these patterns also reveal a preference for only one surface H syllable per word in the language (roots with /HL/ tones, either bound or floating). Detailed information on all the variables can be found in the Supplementary Materials.

## 4 Extracting tonal rules in verb paradigms

Two computational classifiers based on *decision trees* are used to predict the tonal patterns of verb paradigms. That is to say, the classifiers generate decision trees to interpret how the variables of the data set interact with each other when making a prediction. The generated trees can then be used to identify which variables are relevant for predicting tonal patterns. The following `R` packages are used for conducting the analysis: `parsnip` (Kuhn & Vaughan 2019), `randomForest` (Liaw & Wiener 2002), `randomForestExplainer` (Paluszynska & Biecek 2017), `recipes` (Kuhn & Wickham 2019), `rpart` (Therneau & Atkinson 2019), `rpart.plot` (Milborrow 2019), and `rsample` (Kuhn et al. 2019).

To evaluate how "robust" the classification model is (i.e., how well it *generalizes*), it is first *trained* with 70% of the data, which generates a decision tree. Then, this tree is *tested* on the other 30% of the data. In other words, the decision tree is used to predict the tonal patterns of the tokens in the test set, tokens which were not used in the training phase, being, thus, new to the model. Moreover, the training and testing sets were generated such that the ratio of each tonal pattern is consistently reflected in both sets. For instance, the pattern V_SH has 448 tokens in the entire data of 1361 tokens, which equals to 32.9%; therefore, a similar ratio of 32.9% was kept within the training and test sets. Finally, to avoid the risk of biasing the results by using any specific training + test sets, ten different pairs of training and test sets were generated by randomly sampling the whole data set (respecting the constraints described

9 Only four data points were annotated with the VHH pattern. This category has thus been removed from the quantitative analysis.

above), and used to evaluate the classifiers. Due to the stability of the results across the ten sampling processes (more details in subsection 4.1), the results from the tenth sampling are reported in the current paper. For the tenth sampling, the training and the test sets have and 954 and 407 tokens respectively.

We used three different standard metrics to evaluate the performance of the classifiers on the test set: first, we estimated the *accuracy* of the classification (equal to the proportion of the correctly classified cases) of the entire data set. Second, we estimated the *f-score*, defined as the combination of two other measures: *precision* and *recall*: precision measures how many cases are correctly classified within all the predictions on a category, while recall evaluates how many cases are correctly retrieved among all the expected correct output. These measures are used in a similar way as the measures of *suppliance in obligatory context* and *target-like use* in language acquisition (Pica 1983; Tang 2017: p. 41–42). As an example, if the classifier predicts that all the tokens have the pattern VH_SH, the recall on the category VH_SH will be high since all the tokens from that category are found by the classifier. However, the precision will be low since most of the tokens classified as VH_SH actually belong to other patterns. The f-score is equal to the harmonic mean of the precision and recall, i.e. 2(recall × precision)/(recall + precision) (Ting 2010). In other words, the f-score is used as an average representation of precision and recall. Third, we compared the accuracy and the f-score of our classifiers, with the so-called *majority baseline*, which refers to a classifier which would consistently label every token with the largest category in the data set. Here, since the most frequent tonal pattern is V_SH, the majority baseline is 448/1361 = 32.9%, and the performance of our classifiers should exceed this value to be considered as good.

## 4.1 Binary recursive partitioning

The first classifier generates a decision tree through the process of binary recursive partitioning (Breiman et al. 1984). During the classification task, the data is recursively split in a binary manner to create groups that are as homogeneous as possible. At each split, each variable is tried and the variable that can create the most homogeneous split is used. This process is repeated until the data cannot be split further. The decision tree generated based on the training set of the tenth sampling is shown in ***Figure 4***.

The tree is interpreted as follows. The color of the buckets at the bottom of the tree indicate the predicted tonal pattern. The numbers included in each bucket relate to the amount of tokens identified correctly by each prediction. The prediction for each token starts from the top node and ends in a bucket at the bottom of the tree. For instance, starting from the top node 1, if the root type is either three or five (node 1 to node 2 and then to node 4), if a suffix or many suffixes are added (node 4 to 8), the verb does not carry a H tone but the suffix does. This groups together roots with a /HL/ floating tone ([hema̰-káhe] hear-NMLZ) and irregular roots in type 5 ([la̰-kahé] do-NMLZ). This flow affiliates 217 tokens to the 'V_SH' category. Within these tokens, 198 are identified correctly, which results in an accuracy of 198/217 = 91.2% for this prediction. As another example, if the root type is either one or two (node 1 to node 3) and if the verb root has two tones in the penultimate and final positions (node 3 to node 7), the two H tones stay on the verb root and the suffixes do not carry H tones. This flow affiliates 60 tokens to the 'VHH_S' category (e.g. roots with a bound /H/ tone such as /hápa/ [hápá] 'work', /at$^{h}$úpa/ [at$^{h}$úpá] 'spit', classed as type 1), and all of these tokens are assigned correctly, which results in an accuracy of 100% for this prediction. The same logic applies for the other branches of the tree.

Thus, by and large, the classifier considers that 'root types', 'suffix types', and 'the position of H tones on the root' are the most relevant variables for predicting the tonal patterns of verb paradigms. The other variables that are not shown in the tree are considered as not relevant by the classifier (for a further analysis from a linguistic point of view of this tree, see Section 5). In terms of performance, the tree generated based on the tenth

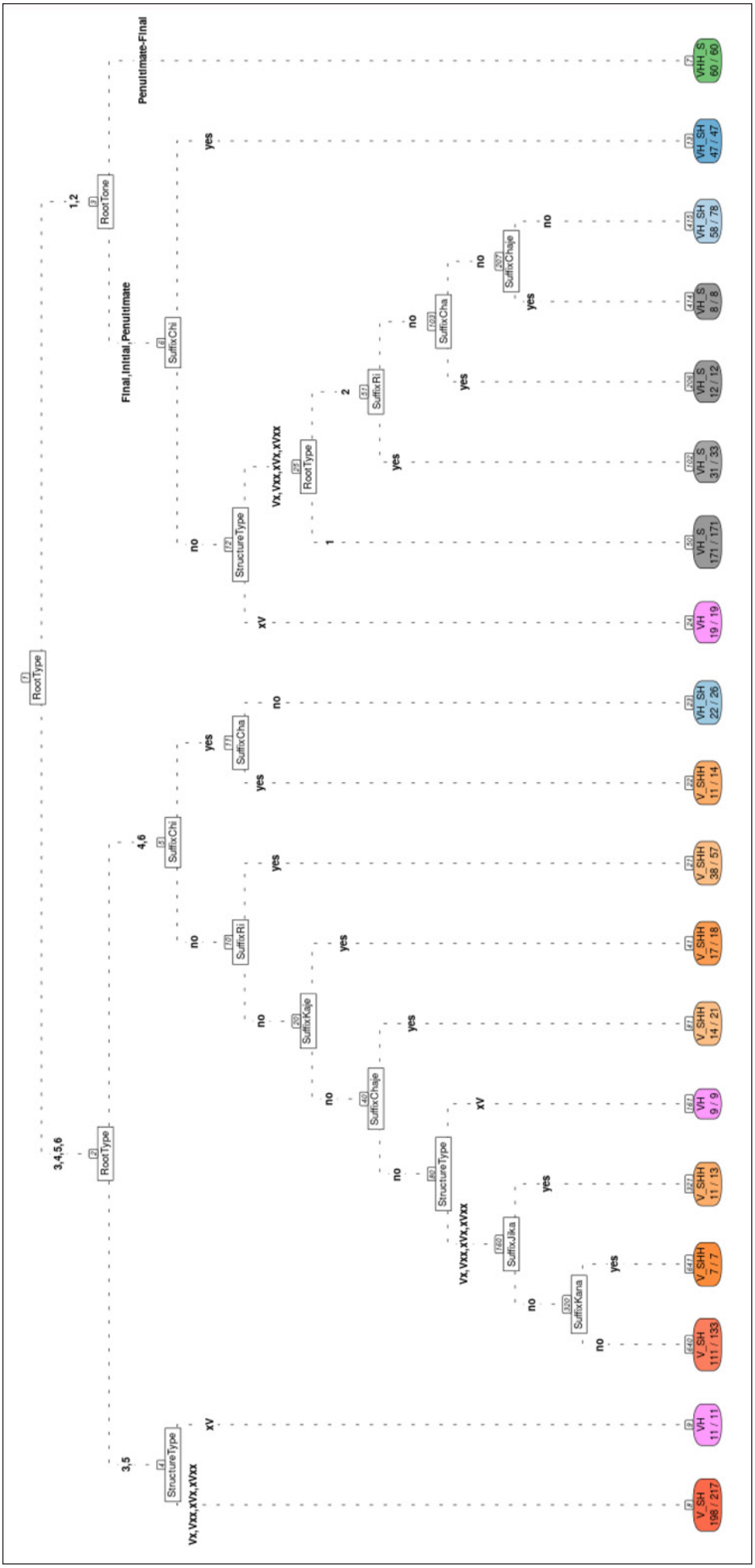


**Figure 4** The decision tree for predicting tonal patterns. See a larger version in the Supplementary Materials.

training set is evaluated with the tenth test set. The confusion matrix resulting from this evaluation is shown in ***Table 8***. The columns refer to the predicted values, the rows are the actual values, the diagonal represents the correct predictions and the off-diagonal are wrong predictions.

| | V_SH | V_SHH | VH | VH_S | VH_SH | VHH_S |
|---|---|---|---|---|---|---|
| V_SH | 135 | 0 | 0 | 0 | 0 | 0 |
| V_SHH | 16 | 36 | 0 | 0 | 0 | 0 |
| VH | 0 | 0 | 16 | 0 | 0 | 0 |
| VH_S | 4 | 5 | 0 | 101 | 6 | 0 |
| VH_SH | 1 | 8 | 0 | 0 | 52 | 0 |
| VHH_S | 2 | 2 | 0 | 0 | 1 | 22 |

**Table 8** The confusion matrix of the decision tree on the test set. The columns are the predicted values and the rows are the actual values.

Within the 407 tokens of the test set, the category V_SH has the biggest ratio of 135/407 = 33.1%, which is the majority baseline. The overall accuracy of the model is (135 + 36 + 16 + 101 + 52 + 22) / 407 = 88.9%. This number indicates that the decision tree generated by the classifier can predict correctly 88.9% of the data. Such accuracy is quite high and much higher than the majority baseline. The performance of the classifier is further analyzed by the measures of precision, recall, and f-score in ***Figure 5***. We can observe that this decision tree predicts extremely well on the VH category, with 100% of precision and recall. On the other hand, this tree does not work well on the V_SHH category, for which the performance is generally lower than the other categories. This is in line with the phonological observations on the language presented in Section 2, which identified roots involving the spreading of /H/ tones across and beyond root boundary (root Types 4,6) as the most irregular. When comparing the precision and recall, we also observe that the classifier has a higher recall for VHH_S and VH_S, which implies that the classifier *overgeneralizes* for these two categories.

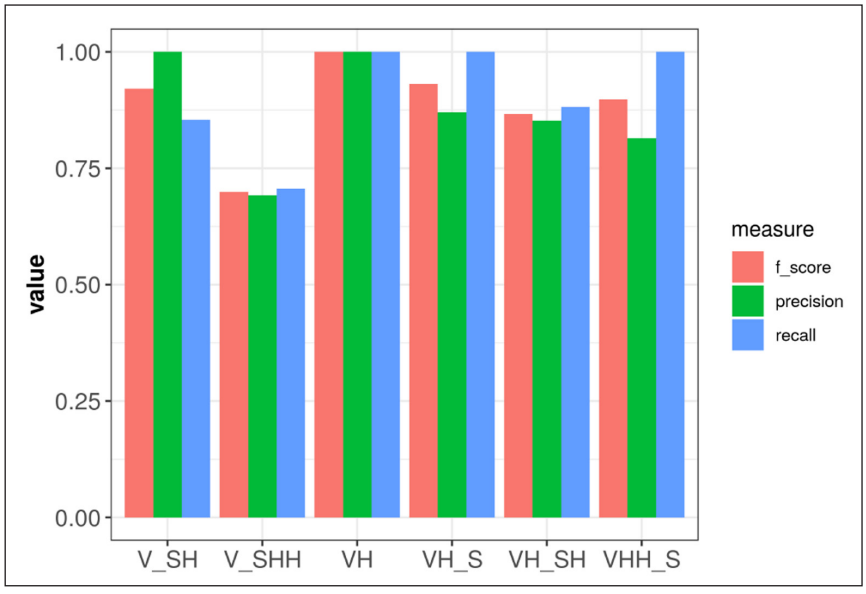


**Figure 5** The precision, recall, and f-score of each category based on the performance of the decision tree on the test set.

As a final assessment, ten different versions of training and test sets are also generated to evaluate the stability of the classifier. That is to say, the good performance of the classifier could be due to an accidental arrangement of the tokens between the training and test sets that favors the classifier. To avoid such bias, the classifier is trained and evaluated ten times with ten different versions of training and test sets. The mean accuracy of the process is 88.8%, with a standard deviation of 1.9%. There is no big variance across the ten different sets, we can thus consider that the performance of the classifier is stable and the results are reliable.

## 4.2 Random forests

The second computational classifier is also based on decision trees from binary recursive partitioning, but it generates a series of multiple decision trees instead of only one tree, justifying its name of *random forests* (Breiman 2001). For each tree, the classifier uses a *bootstrap sample* of the entire data and a random subset of the variables. In other words, if the data is visualized as a table in which the rows represent the tokens and the columns the variables, each partitioning selects a random sample of rows and columns. The trees are then analyzed as a whole and used to assess the importance of each variable in predicting the tonal patterns. A variable is considered as important if it is consistently used across the considered samples. This process of random sampling is also the main strength of random forests, as it allows the analysis of small-scale data and considers the possible auto-correlation of the variables (Tagliamonte & Baayen 2012). The second main strength of random forests in comparison with a single decision tree is the avoidance of overfitting. Single decision trees are likely to be biased depending on the data set used in the experiments. The random sampling process of random forests includes a large amount of trees that operate with different sub-samples of the data, which ensures that the observed tendencies can be generalized to the investigated subject and are not the artefact of a coincidental data set.

It is instructive to consider the decision trees discussed above and the random forests discussed here. The two are used in parallel due to the different information they convey. The ensemble of trees is used to assess the ranking of the variables in terms of their relevance for predicting tonal patterns, whereas the single tree is used to have a simplified visualization of the interaction between the variables. For random forests, the larger the importance of a variable, the more predictive it is. As an example, if the accuracy of the random forests drops the most when it does not take into account a specific variable, this variable is considered to have the highest ranking among all the variables. If the results are consistent between the two classifiers, the conclusion of the hypothesis testing can be strengthened. As an example, if both classifiers indicate that the position of H tones on the verb root is important, it is more likely to be true.

In our experiment, the sample size of trees was set to 500, since it is the sample size that reaches stabilization of the model. After 500 trees, the accuracy of the classifier plateaus and does not improve nor drop anymore. Additional information about this testing is available in the supplementary materials. The confusion matrix generated from the random forests is shown in ***Table 9***. The columns are the predicted values and the rows are the actual values. The overall accuracy of the random forests is (128 + 46 + 15 + 102 + 56 + 27) / 407 = 91.9%. This performance is considerably above the majority baseline.

| | V_SH | V_SHH | VH | VH_S | VH_SH | VHH_S |
|---|---|---|---|---|---|---|
| V_SH | 128 | 1 | 0 | 0 | 5 | 1 |
| V_SHH | 8 | 46 | 0 | 4 | 0 | 0 |
| VH | 0 | 0 | 15 | 0 | 0 | 0 |
| VH_S | 1 | 1 | 0 | 102 | 0 | 0 |
| VH_SH | 0 | 5 | 0 | 1 | 56 | 1 |
| VHH_S | 0 | 1 | 0 | 4 | 0 | 27 |

**Table 9** The confusion matrix of the random forests on the test set. The columns are the predicted values and the rows are the actual values.

The performance of random forests is shown in ***Figure 6***: we observe that the performance on each category improved compared with the single decision tree. The classifier still has 100 % precision and recall on the VH category. The performance for the V_SHH improved also, but is still the lowest in comparison with the other categories. This is in line with the description of the Yukuna phonology in Section 2, as the V_SHH category includes the most irregular roots in the language (mostly types 4 and 6). Finally, the overall balance between precision and recall also improved. The random forests are already performing random sampling of the variables and cases when generating the sample of 500 trees, therefore, there is no need to run additional randomizations of the data.

DOI: 10.5334/gjgl.1276

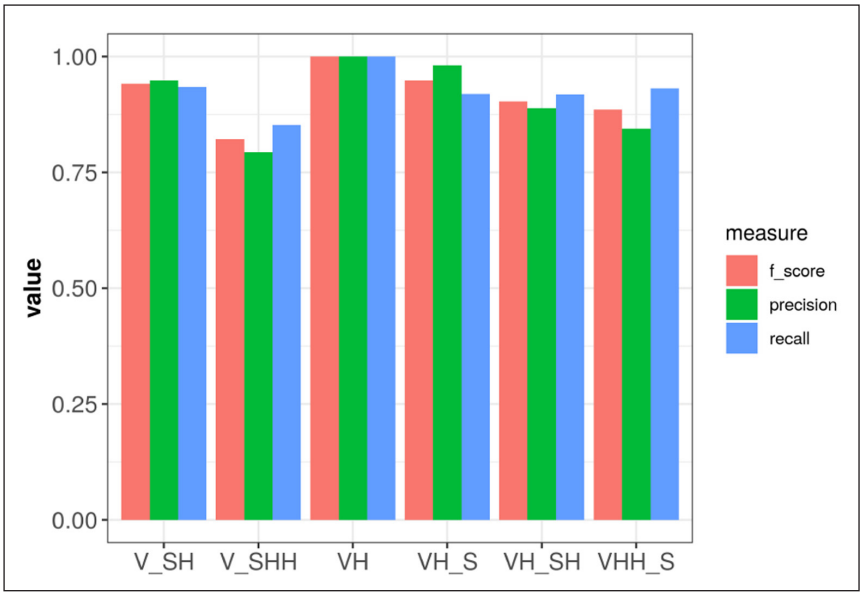


**Figure 6** The precision, recall, and f-score of each category based on the performance of the random forests on the test set.

Thus, the performance of the random forests is very good, and we want next to estimate the *importance* of each variable. We used three measures: first, we visualize the *minimal depth* of each variable for each tree within the forest. Minimal depth indicates how far is the node with a specific variable from the root node. As an example from ***Figure 4***, `RootType` is the root node, which equals to a minimal depth of zero; if a variable is frequently close to the root node, it is considered to have a high importance. The minimal depth of the top ten most important variables is shown in ***Figure 7***.

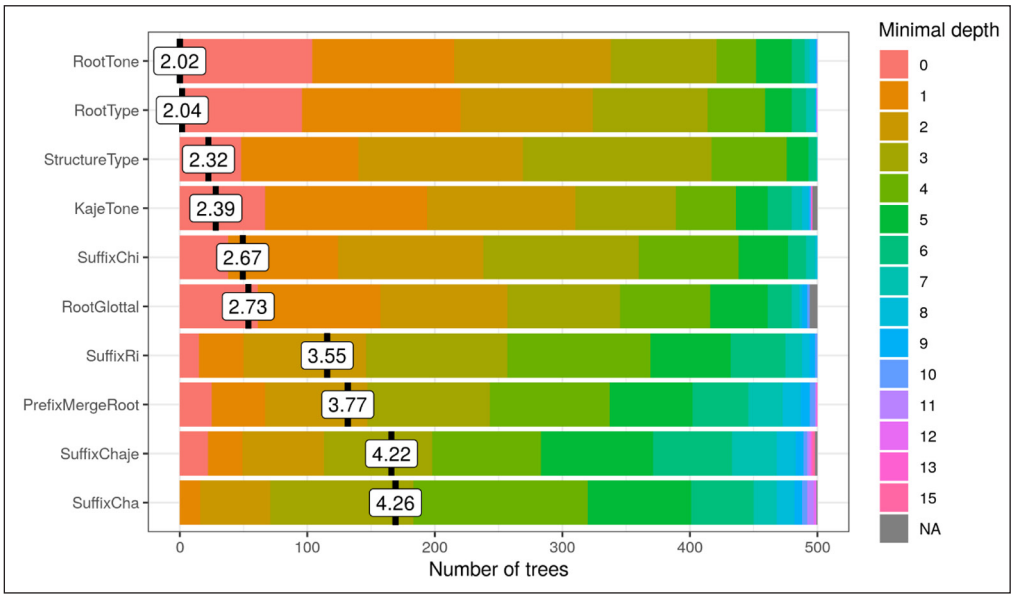


**Figure 7** The distribution of the minimal depth and its mean in random forests.

An 'elbow' can be observed in the curve, which suggests that the first six variables have a strong effect compared to the following ones. These six variables are `RootTone`, `RootType`, `StructureType`, `KajeTone`, and `SuffixChi`; except for `KajeTone`, the variables are also the ones identified by the decision tree.

Similar results are found when using other measures. In ***Figure 8***, the variables are ranked according to their effect on the *accuracy* and the *purity* of the nodes. On the one hand, the mean decrease of accuracy indicates how much worse the model performs without each variable. A high decrease suggests that the variable has a strong predictive power. On the other hand, the mean decrease of the *Gini coefficient* indicates how each variable contributes to the homogeneity of the nodes at the end of the tree. A high decrease of Gini coefficient when removing a variable suggests that this variable has a strong predictive power and therefore a high importance. For both measures, an 'elbow' is also observed around the sixth variable.

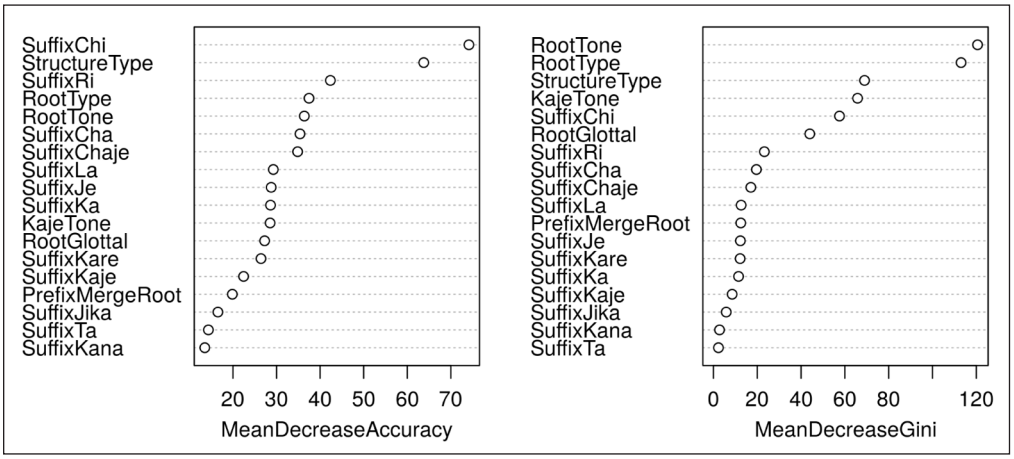


**Figure 8** The mean decrease of accuracy and purity in random forests.

To estimate the convergence of the tree measures, the top six variables that are considered important by random forests are listed in *Table 10*. Each column indicates one of the three measures, i.e., mean minimal depth, mean decrease of accuracy, and mean decrease of Gini coefficient. While different measures result in a slightly different ranking, their results are mostly consistent with each other, as four out of the six variables are found in all three measures: `RootTone`, `RootType`, `StructureType`, and `SuffixChi`.

**Table 10** The top six variables of each measure of importance. The variables in bold are the ones that are found in all three rankings.

| | Min_Depth | Decrease_Acc | Decrease_Gini |
|---|---|---|---|
| 1 | **RootTone** | **SuffixChi** | **RootTone** |
| 2 | **RootType** | **StructureType** | **RootType** |
| 3 | **StructureType** | SuffixRi | **StructureType** |
| 4 | KajeTone | **RootType** | Kaje Tone |
| 5 | **SuffixChi** | **RootTone** | **SuffixChi** |
| 6 | RootGlottal | SuffixChaje | RootGlottal |

Finally, we also visualize how 'confident' the model is when making decisions (*Figure 9*). With random forests, the confidence level of the decisions is extracted by the probability of votes across all the trees. For instance, if 450 of the trees assign a token to the VH category, then the confidence of the decision is $450/500 = 90\%$. We see that the model generally has a confidence level over 75% for correct decisions and has a confidence level of near 30% when wrong decisions are made. This indicates that the model is confident when making correct predictions and in doubt when making wrong decisions. This distribution reflects that the model is going in the right direction: it is sure about decisions that turn out to be correct while it also knows that the decision is likely to be wrong when it actually get guesses wrong.

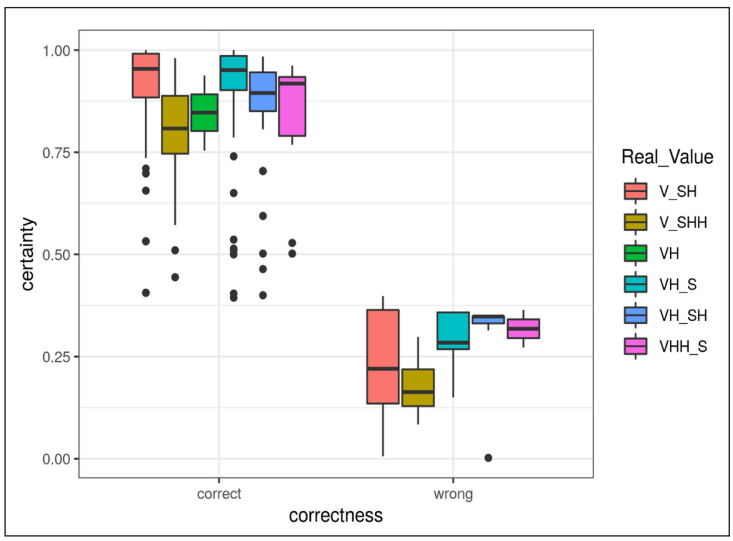


**Figure 9** The confidence of the decision from random forests.

Thus, the results from one decision tree and a forest of 500 trees converge in identifying a set of variables considered important for predicting the tonal patterns of verb paradigms in Yukuna. The performance of the two classifiers exceeds by far the majority baseline, suggesting that the results are robust across methods. Section 5 provides a phonological discussion of these results.

## 5 Discussion

The main aims of this paper were threefold: first, to provide a qualitative analysis of the language's word-prosodic system, second, to test whether computational models would accurately capture the regularities in the system as described by the qualitative analysis, and third, to test whether these models could additionally help identify relevant variables for irregular tonal patterns, and thus, potentially help refine the qualitative analysis.

First, as discussed in Section 2, previous studies on Yukuna had rightly identified the phenomenon of affixation-based shifts in the prosodic pattern of words (Robayo Romero 2018), our study addresses hitherto ignored tonal patterns in the language such as VH_SH and V_SHH, and proposes a novel qualitative analysis in terms of tonal spreading and floating tones.

Second, in terms of the assessment of the proposed qualitative analysis, the results from the two computational classifiers implemented suggest that this aim has been achieved. Indeed, we note that the decision tree and random forests converged in terms of the selection of relevant variables for the prediction of tonal patterns. Unsurprisingly, the following variables were selected by the three different measure tests applied with the random forest: `RootTone`, `RootType`, `StructureType`, and `SuffixChi`. The variables correspond to features of the qualitative analysis of Yukuna tones presented in Section 2. Indeed, the `StructureType` variable captures the *Obligatory H* rule in Yukuna, whereby all words must have at least one H tone. Because of this rule, all suffix-less verbal roots show at least one surface H tone, regardless of their underlying tonal specification. the `RootType` variable supports the clustering of roots based on their underlying tonal pattern. The decision tree in ***Figure 4*** places at its top node a split into two groups of root types: Type 1 and Type 2 (all roots with underlying bound tones) and Type 3-6 (all roots with floating tones as well as irregular roots). In terms of surface tonal patterns, the first group contains roots that always have at least one H tone syllable when suffixed corresponding to the underlying TBU. The second group contains roots that most often surface without any H tone when suffixed, as H tones are placed on the suffixes.

The `RootTone` variable captures the fact that, statistically speaking, the presence of H tone on the root often corresponds to the absence of H tone on the suffixes and vice-versa. This variable is clearly related to *culminativity* restrictions in the language, that is to say, there is a clear tendency for having only one H tone syllable per word. This effect is commonly found in tonal languages around the world, such as in the Bantu languages (Downing 2010: p. 411). Lastly, the relevance of the variable `SuffixChi` pertains to the particular features of the suffix *-chí*, the use of which leads to infrequent but allowed tonal patterns such as HHH and HLH sequences.

In addition to the identification of relevant variables, the performance of the two models implemented matched our expectations: they accurately predicted the tonal patterns of regular roots, and accurately filled parts of the verb paradigms absent from the database. The accurate predictions suggest that the qualitative analysis proposed is robust, as it systematically accounts for the majority of the data, and confidently predicts the surface tonal pattern of forms not attested in the corpus. As an example, the decision tree has been used to fill parts of the verb paradigms that are less likely to be found in corpora given the preference for morphologically simple forms in discourse (verbs rarely show more than two suffixes), and the relative low frequency of some suffix combinations. For instance, although the verb */takha*$^{*HL}$/ 'die' is relatively frequent in the corpus of narratives, it is found in only eight different forms out of the (at least) 30 possible combinations in the verbal paradigm. Thus, the decision tree created during the analysis was used to generate the less commonly used forms missing from the database (***Table 11***). These predictions were then tested based on the qualitative phonological analysis and on the intuition of native speakers. So far, the predictions of the tree match with both the qualitative phonological analysis and the intuition of the native speakers, suggesting that the rules detected by our machine learning approaches might have linguistic reality.

| Verb root | Gloss | Structure | Prediction |
|---|---|---|---|
| i'ka$^{HL}$ | throw | pr-V-ka | V_SH |
| wapa'$^{HL}$ | slice | pr-V | VH |
| wapa'$^{HL}$ | slice | pr-V-ka | V_SH |
| wapa'$^{HL}$ | slice | V-ka | V_SH |
| takha'$^{HL}$ | die | pr-V-chi | V_SH |
| takha'$^{HL}$ | die | pr-V-chaje | V_SH |
| kulá | search | pr-V-la | VH_S |
| kulá | search | V-ri | VH_S |
| kulá | search | pr-V-chi | VH_SH |
| kulá | search | pr-V-chaje | VH_S |

**Table 11** A sample of the paradigms generated based on decision tree and tested with native speakers.

However, such method should be used with caution, as elicitation with a correctness judgment is likely to induce biases in the decisions of native speakers. Nevertheless, the support of computational methods does facilitate the process, as it can narrow down the probable options and enhance the efficiency of working sessions with native speakers. The estimation of the level of confidence (as shown in ***Figure 9***) is an additional cue for the researcher useful for identifying potentially irregular paradigms.

Errors in predictions also support the qualitative analysis. An overview of the errors made by our models indicates that once more, the variable `RootType` is relevant, as illustrated in ***Figure 10***. Indeed, the model does not generate any errors for root Type 1 (bound tones unaffected by suffixation) and Type 3 (floating /HL/), as most errors concern verbs of root Type 2, 4, and 6. This is unsurprising given that root Type 2, 4 and 6 phonologically include roots with underlying /H/ tones (bound or floating) that require cross-morpheme tone spreading or tone association. These roots are not only the most irregular in behavior, but the least numerous in total number of roots, in opposition with root Type 1 and 3, which make up the largest group in the corpus and which are the most regular in behavior.

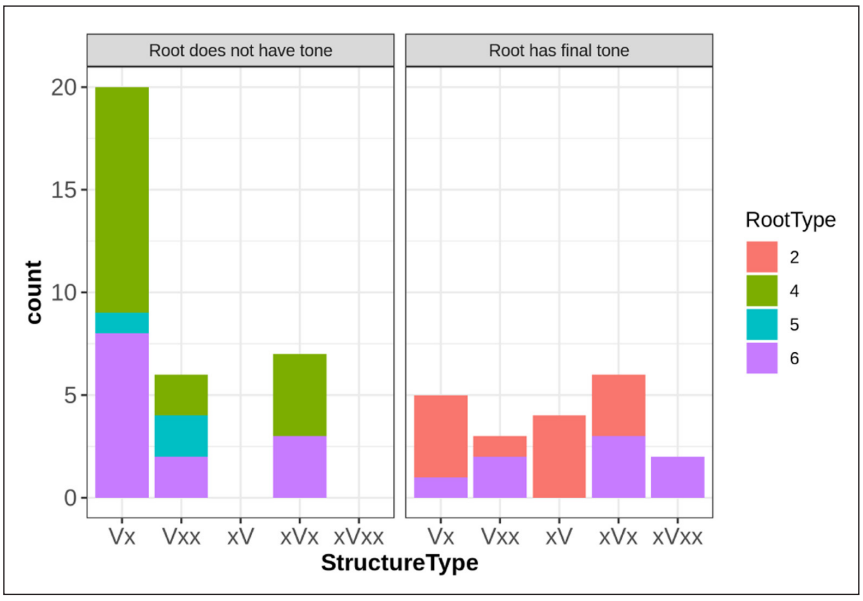


**Figure 10** The distribution of errors from the decision tree.

As expected, the most common wrong predictions are i) V_SHH predicted as V_SH in root type 4 (e.g., *a'pi-chá-chí* 'for him to yawn'), ii) VH_S and VH_SH predicted as V_SHH with root type 4 and 6 (e.g., *richí-ya-chí* 'for him to close'), iii) VH, VH_S, and V_SH predicted as VH_SH in root type 2 and 6 (e.g., *janapí-ri* 'he carries'). All errors concern irregularities in /H/ tone spreading and /H/ tone association not accounted for in the current phonological analysis. ***Figure 11*** illustrates the most common wrong predictions per root type.

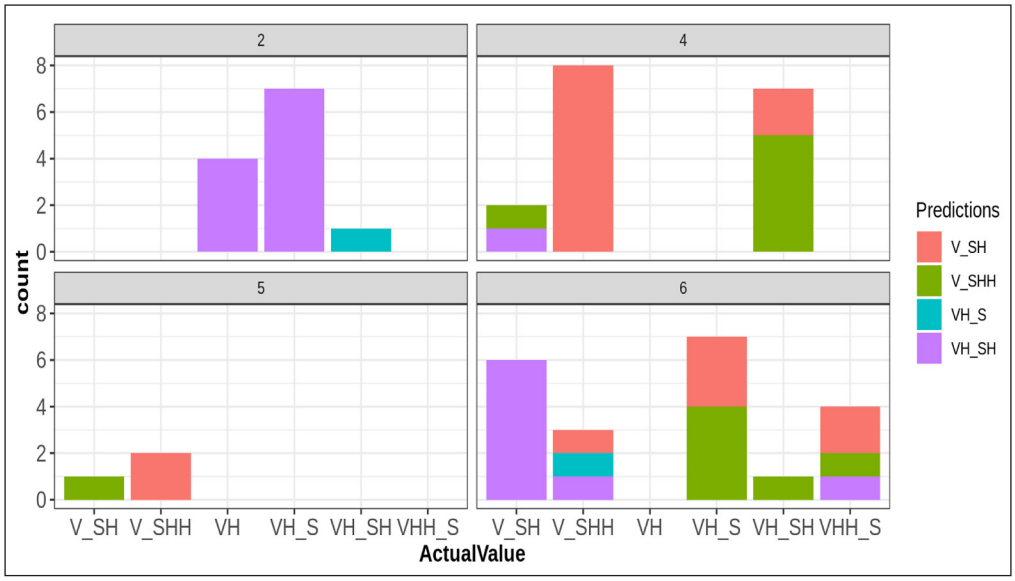


**Figure 11** Comparing the predictions and errors from the decision tree. The facets refer to different root types.

Third, in terms of the identification of tendencies among irregular patterns unexplained by the qualitative analysis, the results from the two computational classifiers implemented suggest that this aim too has been achieved. We note in particular two broad types of tendencies worth exploring in future research: tendencies concerning types of irregular roots, and tendencies concerning types of toneless suffixes.

Concerning irregular root types, the results from the decision tree in ***Figure 4*** identify two subgroupings, with root types 3 and 5 on one hand (leaf node 4), and root types 4 and 6 on the other (leaf node 5). This highlights the tendency for irregular roots in type 5 (such as *a'* 'give') to have a V_SH pattern when suffixed (Vx, Vxx, xVx, and xVxx structures), similarly to type 3 roots (roots with regular /HL/ floating tones). Likewise, there is a tendency for type 6 roots (highly irregular roots, such as *kema-kájé* 'say-NMLZ') to display a V_SHH pattern, similarly to roots in class 4 (such as *i'jna-kájé* 'go-NMLZ'). Please note that these tendencies are not absolute, resulting in the occasional wrong predictions. Nevertheless, these findings suggest that roots in types 5 and 6, for which no underlying tonal pattern was proposed, could be *tentatively* analyzed as highly irregular subtypes within types 3 (floating /HL/) and 4 (floating /HH/) respectively.

As for types of toneless suffixes, two variables are worthy of mention here: `KajeTone` and `SuffixRi`, both concerning the presence of suffixes analyzed as underlyingly toneless. Concerning suffix *-kaje* (NMLZ), we note that the variable `KajeTone` was selected by two measures of relevance with random forests, suggesting that the surface tonal pattern of this suffix is globally a reliable indicator of what the underlying tone of a given root is, more-so than other suffixes. Interestingly, this variable was selected as pertinent by the decision tree for the most irregular root types (4, 6), and its presence is associated with the presence of two adjacent surface H tones for these root types.

Concerning suffix *-ri* (M), the variable `SuffixRi` was selected by the decision tree at various branching points, with different groups of roots: Type 4 and 6 on one hand (floating /H/), and Type 2 (root-final /H/) on the other hand. It appears that the presence of *-ri* affects the surface tonal pattern of these roots, but in different ways. In the case of root Type 4 and Type 6, the presence of the suffix *-ri* affects the placement of floating /H/, and leads to a surface VH_SH pattern (one root final H, one H on the suffix) instead of the expected V_SH (no tone on root, H on suffix) as in /ḭna$^{H}$/ *i'jná-rí* vs. *i'jna-ká*. In contrast, in the case of root type 2, the presence of *-ri* affects the spreading of the bound root-final /H/ and leads to a surface VH_S pattern (root final H, no tone on suffix), instead of the expected VH_SH pattern (root final H plus H on suffix), as in /paá̰/*pajlá-ri* vs. *pajlá-ká*.

The importance of variables concerning toneless suffixes in the determination of irregular surface tonal patterns hints at the possibility that toneless suffixes may differ in terms of their behavior with tonal processes, in particular tonal spreading and floating tone association. In other words, some toneless suffixes could attract or reject H tones in specific contexts. This tendency identified by the computational model matches researchers' intuitions on certain suffixes in Yukuna, in particular, in the nominal domain. For instance, in their very first

study on Yukuna phonology, Schauer & Schauer (1972: 73) bring up the interesting case of alienability suffix *-te* on nouns, which often (but not systematically) carries the H tone, even when combined with roots that appear to have bound tones (e.g. [héma] 'tapir', [héma-na] 'tapir-PL', [hema-té] 'tapir-ALIEN').

These preliminary observations on irregular tonal patterns in Yukuna, while exciting, are to be analyzed with caution, given the limitations in size of the data set used in this study. Edge cases for which there is too little information available, or hint at very complex relationships that might require larger data sets and more powerful (but also harder to interpret) approaches such as *Deep Learning*. As an example, the recent developments of zero-shot learning (Xian et al. 2019) and/or transfer learning (Ruder et al. 2019) enables machine learning systems to make predictions with relatively small (or no) data, which could be used to accelerate the NLP pipeline for low resource languages.

## 6 Conclusion

This paper combined qualitative and quantitative approaches to the description of the Yukuna word-prosodic system. In combining these two methods, this paper provides important new insights into the tonal system of the language, and contributes to our knowledge on the typology of tonal systems in Amazonia. Indeed, the qualitative approach provided a phonological analysis that accurately predicts the majority of patterns in the data. The quantitative approach complemented the qualitative analysis by providing a ranking of relevant variables that highlight statistical tendencies in the irregular patterns. Although these irregular patterns remain unexplained in terms of phonological rules, the quantitative approach improves their *probability* of being accurately predicted.

Additionally, the results of this paper provide clues for future research questions concerning the tonal system of Yukuna. With respect to the synchronic qualitative analysis of the system, the question that arises is whether some of the exceptional tonal patterns could be integrated into rule-based analysis, by positing a phonological distinction between various subtypes of toneless suffixes, for instance, through the *co-phonologies* approach (Inkelas 1998). With respect to the diachrony of this system, a question of interest concerns the historical source of floating tones, and tonogenesis in the language. Given the predominance of irregularities concerning /H/ spreading and surface sequences of H tones, the issue of tone attrition is also of interest, as the system appears to display a preference for a single surface H tone at the level of the word.

More generally speaking, this paper argues that computational methods are a valuable tool that can be exploited for supporting linguistic analysis, circumscribed not only to phonetics/phonology as in this case, but also applicable to morphosyntax, semantics and pragmatics, and, why not, also to their interactions. Quantitative methods inspired from modern statistics, data science and machine learning (among others), are increasingly being applied to issues in the language sciences ranging from the genealogical relationships between languages (e.g., phylogenetic inferences of language families, Dunn et al. 2011) to the understanding of patterns of linguistic diversity and universals (e.g., Blasi et al. 2019). In this particular case, we chose to focus on the tone system of Yukuna for two main reasons: First, we have access to a rich database of exemplars and first-hand experience in the description of the language. Second, Yukuna's tone system is simple enough (but not too simple) and well-understood enough for hypotheses to be generated and tested, and for probable influences to be identified. A question for future research is whether such models would be as successful for accurately identifying relevant patterns in less well understood systems. Of course, while a powerful tool, computational methods need to be implemented with care. For each question of research, we need to find the most appropriate techniques (here we explored but one broad class of machine learning methods among many), the best quantification, presentation and interpretation of their results (including the degree of confidence and generalisability), and guidelines for the identification of the variables potentially relevant to the question at hand. It is of the utmost importance to highlight that quantitative approaches such as used here are not intended to replace careful linguistic analyses backed by a thorough understanding of the language and its socio-cultural context, but simply to supplement and help in the generation and testing of hypotheses, on the one hand, and the estimation of "confidence" in such generalisations, on the other.

## Abbreviations

ALIEN = alienable, COP = copula, FUT = future, H = high tone, L = low tone, M = masculine, NEG = negation, NMLZ = nominalizer, PURP = purposive, SG = singular.

## Additional files

The additional files for this article can be found as follows:

- **Supplementary file 1.** Data used for the analysis. DOI: *https://doi.org/10.5334/gjgl.1276.s1*
- **Supplementary file 2.** ***Figure 3*** reproduced in full size. DOI: *https://doi.org/10.5334/gjgl.1276.s2*

## Ethics and consent

All Yukuna speakers who participated in this research consented to being recorded, and received a salary for their participation according to the Colombian government standards.

## Acknowledgements

We address our deepest thanks to all the Yukuna, Matapi and Tanimuka collaborators who participated in the Yukuna documentation and description project led by the first author. We also thank our Dynamique du Langage colleagues who provided insightful comments on the phonology of Yukuna, in particular, Prof. Denis Creissels, Prof. Gérard Philippson, Dr. Denis Bertet, and Dr. Françoise Rose. Lastly, we thank the reviewers for their constructive remarks. All remaining errors are our own.

## Funding information

This project was funded by a LabEx Aslan doctoral grant "A grammar of Yukuna, an Arawak language of Colombian Amazonia" (2016-2019) to the first author. The second and the third authors are thankful for the support of grants from the Université de Lyon (ANR-10-LABX-0081, NSCO ED 476), the IDEXLYON Fellowship (2018-2021, 16-IDEX-0005), and the French National Research Agency (ANR-11-IDEX-0007).

## Competing interests

The authors have no competing interests to declare.

## Author Affiliations

**Magdalena Lemus-Serrano** *orcid.org/0000-0001-8312-6466*
Dynamique Du Langage UMR 5596 CNRS, Aix-Marseille Université, FR

**Marc Allassonnière-Tang** *orcid.org/0000-0002-9057-642X*
Dynamique Du Langage UMR 5596 CNRS, Université Lumière Lyon 2, FR

**Dan Dediu** *orcid.org/0000-0002-0704-6365*
Dynamique Du Langage UMR 5596 CNRS, Université Lumière Lyon 2, FR

## References

Blasi, Damián, S. Moran, S. R. Moisik, P. Widmer, D. Dediu & B. Bickel. 2019. Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363(6432). eaav3218. DOI: *https://doi.org/10.1126/science.aav3218*

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5–32. DOI: *https://doi.org/10.1023/A:1010933404324*

Breiman, Leo, Jerome Friedman, Charles J. Stone & Richard Olshen. 1984. *Classification and regression trees*. New York: Taylor & Francis.

Downing, Laura. 2010. Accent in Africa languages. In Harry van der Hulst, Rob Goedemans & Ellen van Zanten (eds.), *A survey of word accentual patterns in the languages of the world*, 381–427. De Gruyter Mouton.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345). 79–82. DOI: *https://doi.org/10.1038/nature09923*

Fontaine, Laurent. 2001. *Paroles d'échange et règles sociales chez les Indiens yucuna d'Amazonie colombienne*: Thèse de doctorat Université de la Sorbonne nouvelle Paris III. *https://tel.archives-ouvertes.fr/tel-00596637/document*.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019. *Glottolog 4.1*. Jena: Max Planck Institute for the Science of Human History. *https://glottolog.org/*.

Hyman, Larry M. 2009. How (not) to do phonological typology: the case of pitch-accent. *Language Sciences* 31(2–3). 213–238. DOI: *https://doi.org/10.1016/j.langsci.2008.12.007*

Inkelas, Sharon. 1998. The theoretical status of morphologically conditioned phonology: a case study of dominance effects. In Geert Booij & Jaap Van Marle (eds.), *Yearbook of morphology 1997*, 121–155. Springer Netherlands. DOI: *https://doi.org/10.1007/978-94-011-4998-3_5*

Kenstowicz, Michael. 1993. Evidence for metrical constituency. In Kenneth Hale & Samuel Jay Keyser (eds.), *The view from building 20: Essays in Linguistics in honor of Sylvain Bromberger*, 257–273. Cambridge: The MIT Press.

Kisseberth, Charles & David Odden. 2003. Tone. In Derek Nurse & Gerard Philippson (eds.), *The Bantu languages*, 59–70. London: Routledge.

Kuhn, Matt & Davis Vaughan. 2019. parsnip: A common API to modeling and analysis functions. *R package version* 0.0.3.1. *https://CRAN.Rproject.org/package=parsnip*.

Kuhn, Max, Fanny Chow & Hadley Wickham. 2019. rsample: General resampling infrastructure. *R package version* 0.0.5. *https://CRAN.R-project.org/package=rsample*.

Kuhn, Max & Hadley Wickham. 2019. recipes: Preprocessing tools to create design matrices. *R package version* 0.1.6. *https://CRAN.R-project.org/package=recipes*.

Lemus Serrano, Magdalena. 2016. Observations sociolinguistiques et description phonologique du yukuna: Langue arawak de l'Amazonie colombienne. Mémoire de Master 2 Université Lumière Lyon 2.

Lemus Serrano, Magdalena. 2020. *Pervasive nominalization in Yukuna, an Arawak language of Colombian Amazonia*: Ph.D dissertation Université Lumière Lyon 2.

Liaw, Andy & Matthew Wiener. 2002. Classification and regression by randomForest. *R News* 2(3). 18–22.

McCarthy, John. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry* 17(2). 207–263.

Milborrow, Stephen. 2019. rpart.plot: Plot rpart models: An enhanced version of plot.rpart. *R package version* 3.0.8. *https://CRAN.R-project.org/package=rpart.plot*.

Paluszynska, Aleksandra & Przemyslaw Biecek. 2017. randomForestExplainer: Explaining and visualizing random forests in terms of variable importance. *R package version* 0.9. *https://CRAN.R-project.org/package=randomForestExplainer*.

Pica, Teresa. 1983. Adult acquisition of English as a second language under different conditions of exposure. *Language Learning* 33(4). 465–497. DOI: *https://doi.org/10.1111/j.1467-1770.1983.tb00945.x*

R-Core-Team. 2020. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. *https://www.Rproject.org/*.

Ramirez, Henri. 2001. *Línguas Arawak da Amazônia Setentrional: Comparação e Descrição*. Manaus: Editora de Universidade do Amazonas.

Robayo Romero, Camilo Alberto. 2018. Introducción a la prosodia del verbo yukuna. *Forma y Función* 31. 33–63. DOI: *https://doi.org/10.15446/fyf.v31n1.70442*

Ruder, Sebastian, Matthew E. Peters, Swabha Swayamdipta & Thomas Wolf. 2019. Transfer Learning in Natural Language Processing. In Sarkar, Anoop and Strube, Michael (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 15–18. Minneapolis, Minnesota: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/N19-5004*

Schauer, Junia & Stanley Schauer. 1972. Fonología del yucuna. In Waterhouse, Viola G. (Ed.), *Sistemas fonológicos de idiomas colombianos 1*, 65–76. Lomalinda: Townsend. *http://www-01.sil.org/americas/colombia/pubs/12252.pdf*.

Schauer, Junia, Stanley Schauer, Eladio Yukuna & Walter Yukuna. 2005. *Meke kemakánaka puráka'aloji: wapura'akó chu, eyá karíwana chu (Diccionario bilingüe: Yukuna – Español; Español – Yukuna)*. Bogotá: Editorial Fundación para el Desarollo de los Pueblos Marginados. *http://www01.sil.org/americas/colombia/pubs/abstract.asp?id=928474518977*.

Tagliamonte, Sali A. & Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24. 135–178. DOI: *https://doi.org/10.1017/S0954394512000129*

Tang, Marc. 2017. Explaining the acquisition order of classifiers and measure words via their mathematical complexity. *Journal of Child Language Acquisition and Development* 5(1). 31–52.

Therneau, Terry & Beth Atkinson. 2019. rpart: Recursive partitioning and regression trees. *R package version* 4.1-15. *https://CRAN.R-project.org/package=rpart*.

Ting, Kai Ming. 2010. Precision and Recall. In Claude Sammut & Geoffrey I. Webb (eds.), *Encyclopedia of Machine Learning*, 781–781. Boston, MA, US: Springer. DOI: *https://doi.org/10.1007/978-0-387-30164-8_652*

Wickham, Hadley. 2017. tidyverse: Easily install and load the Tidyverse. *R package version* 1.2.1. *https://CRAN.R-project.org/package=tidyverse*.

Xian, Yongqin, Christoph H. Lampert, Bernt Schiele & Zeynep Akata. 2019. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(9). 2251–2265. *https://ieeexplore.ieee.org/document/8413121/*. DOI: *https://doi.org/10.1109/TPAMI.2018.2857768*

TO CITE THIS ARTICLE:
Lemus-Serrano, Magdalena, Marc Allassonnière-Tang and Dan Dediu. 2021. What conditions tone paradigms in Yukuna: Phonological and machine learning approaches. *Glossa: a journal of general linguistics* 6(1): 60. 1–22. DOI: *https://doi.org/10.5334/gjgl.1276*

Submitted: 08 April 2020
Accepted: 07 April 2021
Published: 05 May 2021



*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.