Measuring the similarity between languages: The case of creoles and non-creoles

Lara Rüter & Ingo Plag

December 13, 2024

Abstract

In typology, statistical methods have been successfully used to assess similarities and differences between languages. In creole studies, the use of quantitative methods has been discussed controversially. In the debate many methodological aspects of the statistical models used have been criticized (e.g. Meakins 2022; Bakker 2023). This paper presents an investigation of two methodological problems that have not been critically looked at so far: the question of which statistical models produce which results, and the question of how the amount of missing values in data sets influences the results. We present a study in which we tested different statistical models on 21 features from two arbitrarily chosen domains ('Word Order' and 'Nominal Categories') from the WALS (Dryer & Haspelmath, 2013) and APiCS (Michaelis et al., 2013) data bases. It is demonstrated that different statistical methods yield similar results, and that different sample sizes do not dramatically influence the model outcomes.

1 Introduction

Like other linguistic sub-disciplines, typological and comparative research has seen the successful rise of quantitative methods. Statistical and computational methods have been developed and successfully used over the past two decades to assess the similarities and differences between languages on a large scale, investigating mainly questions of areal distribution or genetic dependencies (e.g. Dunn et al. 2005; Jaeger et al. 2011; Bickel 2007; Bouckaert et al. 2012; List 2021.

There is also a substantial body of quantitative work in creole studies that makes use of the methods applied in typology (e.g. Bakker et al. 2011, 2017; Blasi et al. 2017; Daval-Markussen 2019). In fact, according to Google Scholar (accessed August 21, 2024), the most cited article of the present journal is one of these studies (Bakker et al. (2011)).[1] In these quantitative studies mainly three questions have been addressed: whether creoles are structurally different from non-creole languages, whether creoles are less complex than non-creoles, and whether creoles are more similar to their substrate languages than to their superstrate languages. All three questions had featured prominently in the literature on pidgin and creole languages before the advent of quantitative methods,

[1] Bakker et al. (2011) is listed with 288 citations, followed by Baker (1990) with 193 and Plag (2008) with 154.

and these questions had been controversially discussed already based on traditional structural-linguistic or theoretical-linguistic evidence and reasoning (e.g. Muysken (1988); McWhorter (1998); Lefebvre (1998); DeGraff (2003), to mention only very few pertinent publications).

Especially the paper by Bakker and colleagues (2011) has triggered a new controversy in which methodological aspects have been the center of attention. Another focus has been on the political and theoretical implications of the quantitative empirical results, but in this paper we disregard the latter aspects and concentrate on methodological problems raised by the data bases and statistical models. The main problems pointed out in the literature concern the selection of features, the sampling of languages, and the coding of linguistic features (e.g. Meakins 2022; Bakker 2023).

There are, however, two aspects that have not been critically looked at so far, and which are interesting for both proponents and critics of a quantitative approach. The first aspect concerns the question which kinds of statistical models produce which kinds of results. So far, phylogenetic networks of a particular kind have dominated the discussion, and it is unclear and unexplored whether other statistical models yield the same results. [2]

The second aspect concerns the problem of missing values. Typological data bases often provide incomplete information such that not all features or all feature specifications are available for all languages in the sample. Existing studies either work with a rather arbitrarily chosen threshold (e.g. no more than 20 percent missing values per language, as for one of the studies in (Daval-Markussen, 2019, 82, 115), or 14 percent in Blasi et al. 2017), or impute missing values (as, for example done in Blasi et al. 2017).

In this paper we present a study in which we tested different statistical models on 21 features from two arbitrarily chosen domains ('Word Order' and 'Nominal Categories') as found in two widely-used typological data bases, WALS (Dryer & Haspelmath, 2013) and APiCS (Michaelis et al., 2013).

The models were designed to investigate the question whether creoles and non-creoles are structurally different from each other. The reader should note, however, that the present paper does not aim to give a convincing answer to this research question, or to replicate or extend the findings of other studies. Rather, we want to shed light on methodological issues that are crucial for any debate on the typology of creole languages. We do this by empirically testing the consequences of competing methodological decisions and comparing the respective results with a strict focus on the separation of the two types of languages in the model. What is interesting for us in this paper is not the answer to the research question, but whether, and if so, how, this answer varies across different statistical models and across different samples of languages. An arbitrarily chosen, rather small set of features like the one we use here can not be expected to provide conclusive answers to such a big research question. But it can be used to investigate the two methodological problems of model choice and missing values. As we do not focus on the results themselves but on the differences in results across models and samples, we also do not explore the models further concerning the question of which individual languages are similar to which other individual languages. Neither do we discuss possible

[2] Murawaki (2016) and Blasi et al. (2017) are notable exceptions, as they have used other statistical tools. We will return to these papers below.

causes for the similarities and dissimilarities that the statistical models discern, e.g. by looking at which individual languages with which particular histories (or lexifiers, or substrates) cluster in which way.

We implemented two kinds of phylogenetic networks and compared their results with each other. We also included in our comparison the results yielded by other clustering and classification methods: cluster analysis, classification and regression trees, and random forests. To investigate the problem of incomplete data sets we explored the models' sensitivity to different thresholds of missing values (cf. Daval-Markussen 2019 for a similar approach, restricted, however, to phylogenetic networks). We did not optimize the models (e.g. by bootstrapping, cross-validation or cost-complexity pruning) but fitted all models with their default parameter settings in the software we were using (R, R Core Team 2021), to ensure comparability.

The evaluation of the models indicates that, independent of sample and method, the results do not change much. Creoles and non-creoles indeed clearly differ from each other, based on the features selected for this exercise. When looked at in more detail, these differences play out as rather complex constellations of particular features. The statistical models are capable of making out observable patterns which may raise new questions about the mechanisms in language contact situations that bring about certain features, but not others.

The paper is structured as follows. In the next section we discuss the merits and pitfalls of doing quantitative typology. This is followed by a description of our methodology. Section 4 presents the results, which are summarized and discussed in section 5.

2 Quantitative typology: Merits and pitfalls

2.1 Data

According to Bickel (2007), the aim of typological research is to explain linguistic diversity, i.e. to answer the question 'what's where why?'. This means that typology is chiefly concerned with areal and historical distributions of linguistic features. An analysis of such distributions leads to "probabilistic theories stated over properly sampled distributions" (p. 239).

This statement already hints at the first key problem, the sampling of languages and linguistic features. Which languages and which features are suitable for a specific comparison, given a certain aim? This problem is widely discussed and the justification of a particular sample is typically part of the methodology section of pertinent publications. Quite often, the sampling decisions are heavily constrained simply by the availability of data.

Typologists standardly make use of data gleaned from data bases that provide measurement points for a large variety of language properties. These data bases are usually compiled with no specific research question in mind, which is good and bad at the same time. It is good because the researcher who uses the data was not able to introduce any personal biases. It is bad because the data may not be fully suitable for the investigation of the research question at hand, or may reflect the biases of the data base compilers.

It should be noted that the sampling problem is neither specific nor restricted to quantitative approaches but is also prevalent in traditional theoretical-linguistic

research even though it has not received as much attention in that part of the literature, probably because in that tradition, convenience samples have been largely taken for granted.

Another key problem in comparing languages is the discernment (or development) of variables ('features') that can be used to measure the similarities and differences between languages. Although it is generally accepted that comparisons should be based on concepts that are comparable across languages (cf. Haspelmath 2010), there are still often debates about the nature of the categories (and their proper definitions) that are used in the comparisons. Again, the problem of defining the relevant features is not restricted to quantitative research, but is only rarely discussed in non-quantitative work (but see, for example, Plag 2011, 95f, for an exception). Typological data bases have taken certain decisions, resulting in advantages and disadvantages that are analogous to those concerning their sampling decisions. Like in qualitatively oriented research, the investigator is of course free to redefine the feature values for their own purposes, as we do below to make data bases compatible. A complete reconceptualization or redefinition of features, accompanied by a recoding of the feature space is, however, rarely done.

In creole studies, a number of different data bases have been used, but over the past decade two large data bases have become standards: APiCS (Michaelis et al., 2013) and WALS (Dryer & Haspelmath, 2013).[3]

APiCS provides codings for 130 linguistic features of 76 contact languages, documenting the distribution of 130 lexical, phonological, morphological and syntactic features.

The choice of features in APiCS was inspired by WALS and the features dealt with in Holm & Patrick (2007). The features "define abstract structural features that make reference to structural properties that can be identified in any language. These can be general concepts of language form such as 'precedes/follows', 'overt/zero', 'identical/different', or semantic-pragmatic concepts like 'negation', 'question', 'focus', or more complex comparative concepts defined on the basis of such elementary formal concepts and semantic-pragmatic concepts (e.g. 'subject', 'pronoun')." (Michaelis et al., 2013, xxxvii).

APiCS was compiled in such a way that for each language an expert (or a team of experts) was supposed to code each feature. Nevertheless, APiCS still contains cases of missing values. Overall, only 273 of 9880 data points (i.e. 2.8 percent) are missing. In the subset of 21 features we investigated for this paper, we counted 11 percent empty cells, i.e. missing values.

According to (Daval-Markussen, 2019, 114) the 76 APiCS languages comprise 54 creole languages, and 22 languages that may be labeled as expanded pidgins, semicreoles or mixed languages. For the present paper, we will use only the 54 creole languages.[4]

[3]Other data bases are also available, such as Grambank (Skirgård et al., 2023; Skirgård et al., 2023). We decided to use APiCS and WALS for mainly two reasons: First, they have been widely used in creole studies before, and second, due to the close relationship between the two, the features are easily comparable.

[4]As part of our methodological exercise, we also implemented analyses that cast the net as wide as possible by using all APiCS languages and comparing them to the WALS languages. Including all languages from APiCS biases the results against the hypothesis that creoles and non-creoles are different, and makes it harder to find differences between creoles and non-creole languages. The results of these analyses are documented in the supplementary material (see `https://osf.io/4vesz/?view_only=7fc25d277f9649cd9daf317c439a943d`). The nature

WALS has 192 features, with data from 2662 languages. The values of a given feature have been provided by an expert of this feature, which has the unfortunate consequence that for each feature there is a different sample of languages. About 83.5 percent of the feature-language cells are empty (e.g. Cysouw 2008) because, due to this setup, most languages are included with only very few features. There is thus a massive problem of missing data. For our 21 features, 1863 languages are represented in WALS, with 50 percent of missing values.

It is unclear how many of the WALS languages are contact languages, but the WALS data base is generally assumed to represent non-contact languages, i.e. non-creole languages by implication (e.g. Parkvall 2008; Bakker et al. 2011; Daval-Markussen 2019).

APiCS and WALS share 48 features from a wide range of linguistic domains as defined by the data base compilers (word order; nominal categories; nominal syntax; verbal categories; argument marking; clausal syntax; complex sentences; negation, questions, and focusing; lexicon; phonology). This makes the two data bases an ideal data set for quantitative work, and almost all of the quantitative studies of creole structures since 2013 have made use of them.

2.2 Statistical tools

As already mentioned, the construction of phylogenetic networks is the method of choice that has been employed in most previous quantitative studies in creolistics. This method has essentially been developed in evolutionary biology to study the historical development of species and the distribution of genes. Phylogenetic networks come in two flavors and have been used in biology and linguistics in two ways. First, they are employed to construct rooted evolutionary networks in which the internal nodes represent historical ancestors of the leaf nodes. The second type of network is unrooted and the edges in the network represent affinities or similarities between the leaf nodes. The latter type of network is therefore well suited for the exploration of synchronic similarities between languages or for testing pertinent hypotheses in typological research.

Phylogenetic networks result from certain types of clustering technique, but other clustering models are also available. As an alternative to clustering, it is also possible to use classifying algorithms that predict the membership of an item in a particular class (here: creole or non-creole) on the basis of this item's properties.

Blasi et al. (2017) use random forests to investigate the question whether the similarities (and differences) between creoles can be explained by genealogical and contact processes, i.e. by processes of transmission similar to those found in non-creole languages. Murawaki (2016) employ a linear support vector machine classifier in combination with principal component analysis to categorize languages into creoles and non-creoles, and Bayesian generative models to quantify the contribution of three different sources in creole genesis (the lexifier, the substrate and a 'global restructurer').

In summary, different studies have used different statistical models to investigate different research questions. The choice of a particular method in

of the overall results of these parallel analyses is basically the same as the ones given in this paper.

these publications is often motivated by convenience (e.g. there is a software package available), or by tradition (i.e. related studies have used this method before). Systematic comparative investigations of the methodological problems or choices are not available, and it is the aim of the present paper to provide insights into these issues through a systematic comparison of methods and of the consequences of particular methodological choices.

2.3 The present study

As has become clear from the discussion, there are plenty of methodological problems for any researcher who wants to compare languages (be it quantitatively or qualitatively). Many of these problems have been critically discussed and more or less satisfactorily solved in the literature. In this paper we want to explore two methodological problems that have not been systematically investigated in the context of creole studies: the choice of statistical models and the treatment of missing values. Do different statistical models yield different results? Are there models that should be preferred, and others that are not suitable? And can we trust data sets that have up to 30 percent missing values per language? Or only those that have only up to 10 percent? Or none at all?

For exploration of these questions we use a showcase of 21 features from APiCS and WALS, different samples of languages from these two data bases, and diverse statistical models. The details of our approach are presented in the next section. The data and the scripts for our statistical analyses are available at `https://osf.io/4vesz/?view_only=7fc25d277f9649cd9daf317c439a943d`.

3 Methodology

3.1 Data

In line with previous quantitative analyses on creole structures, we used data from the shared domains of APiCS and WALS. Of the total of 10 domains, we arbitrarily selected the first two given on the APiCS website, 'Word Order' and 'Nominal Categories', with 22 features listed therein (one feature was eliminated later, see below). Nine features are listed under the domain 'Word Order' (Table 1), and 13 features are subsumed under the domain 'Nominal Categories' (Table 2).

Table 1: The nine features of the domain 'Word Order'.

Word Order
Order of subject, object, and verb
Order of possessor and possessum
Order of adjective and noun
Order of adposition and noun phrase
Order of demonstrative and noun
Order of cardinal numeral and noun
Order of relative clause and noun
Order of degree word and adjective
Position of interrogative phrases in content questions

Table 2: The 13 features of the domain 'Nominal Categories'.

Nominal Categories
Gender distinctions in personal pronouns
Inclusive/exclusive distinction in independent personal pronouns
Politeness distinctions in second-person pronouns
Indefinite pronouns
Occurrence of nominal plural markers
Expression of nominal plural meaning
Definite articles
Indefinite articles
Pronominal and adnominal demonstratives
Distance contrasts in demonstratives
Adnominal distributive numerals
Ordinal numerals
Sortal numeral classifiers

3.2 Coding

Although the two data bases share the same 22 features, the coding of these features is not identical across the data bases. Sometimes more distinctions are made in APiCS than in WALS for a given feature, sometimes APiCS does not code values that we find in WALS, sometimes the names for feature values are not identical across the two sources.

In order to streamline the two data sets we have applied the following methods:

1. **Fixing names of feature values**.
   In those cases where different value names designated the same phenomenon, we gave the same name to the feature value.

2. **Addition of feature values (`Mixed` or `NDO`)**.

   The languages found in the two data bases can express different values for one feature, which is why there are several coding options for each feature. Every language is coded for at least one feature value, but can also exhibit multiple feature values. In the latter cases, we decided to code the respective language with `NDO` (short for 'No dominant order') or `Mixed`. The choice of using either `NDO` or `Mixed` depended on which of the two labels was already present for the WALS languages. For instance, the languages Batavia Creole, Cavite Chabacano, and Eskimo Pidgin each express two values for the feature ORDER OF ADJECTIVE AND NOUN of the domain 'Word Order'. In these cases, it is not directly possible to discern which of the two manifestations appears the most. For this reason, we coded all those languages that express various values for this feature with `NDO` ('no dominant order of adjective and noun'), using the same label as in WALS.

3. **Collapsing of feature values**.

Table 3: Original values of WALS and APiCS for the feature POLITENESS DISTINCTIONS IN SECOND-PERSON PRONOUNS. Non-identical feature values are given in italics.

WALS	APiCS
No politeness distinction	No pronominal politeness distinction
Binary politeness distinction	Binary pronominal politeness distinction
Multiple politeness distinction	Multiple pronominal politeness distinction
Pronouns avoided for politeness	*Titles used as second person forms*

Table 4: Normalized values of WALS and APiCS for the feature POLITENESS DISTINCTIONS IN SECOND-PERSON PRONOUNS.

WALS and APiCS
No politeness distinction
Binary politeness distinction
Multiple politeness distinction
Pronouns avoided for politeness

In some cases, one of the data bases makes more distinctions within one feature than the other. This leads to an unequal number of value options for that feature, which impedes the compatibility of the data. We resolved this issue by collapsing those values of one data base that were overdifferentiated compared to the corresponding value in the other data base. In such cases, the more general meaning was taken as a reference point. For instance, in the feature POLITENESS DISTINCTIONS IN SECOND-PERSON PRONOUNS of the domain 'Nominal Categories', APiCS shows a feature value that can be interpreted as a special case of the more general feature value in WALS (Table 3). Hence the more general feature value was taken as a reference point for streamlining the value names, i.e. 'Pronouns avoided for politeness' (Table 4).

4. **Omission of feature values**.

   Those features that have incompatible values due to the different definitions of values in the data bases were eliminated from our data set. This was the case with the feature ORDINAL NUMERALS of the domain 'Nominal Categories'. As a consequence, our data set includes 21 features of the 22 features as found in the domains 'Word Order' and 'Nominal Categories'.

3.3 Sampling

As already mentioned, our subset of 21 features contains 50 percent of missing values for the WALS languages and 11 percent empty cells for the APiCS languages. This is due to the fact that not all languages are coded for all features, which is attributable to the data collection procedures. In addition to testing the various statistical models that have been used to date, the aim of this study was to tackle the question of how many missing values (hereafter also referred to

as 'NAs', NA = 'not available') are acceptable in order to obtain reliable results. To test the effects of varying numbers of NA's on the results, we decided to create four different samples, each with a different maximum amount of missing data points per language: 0, 10, 20, and 30 percent of NAs, respectively.Figure 1 shows the number of APiCS and WALS languages in the four samples. We will refer to the four samples as the '0NA', '10NA', '20NA' and '30NA' samples.

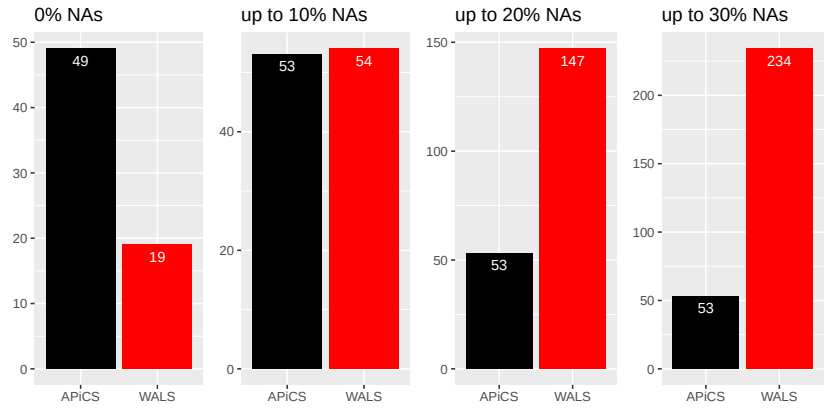

Figure 1: Number of languages from the two data bases in the different samples.

The number of APiCS languages remains largely the same across the four different data sets, whereas the number of WALS languages rises remarkably with increasing percentages of NAs. The increase in missing values, particularly for the WALS languages, is due to the fact that the number of languages coded for each feature in the WALS data base highly varies (recall the 83.5 percent of empty feature-language cells in WALS as mentioned in subsection 2.1). It should be stressed, however, that the percentages given here indicate the maximum proportion of empty cells per language to be included in the respective sample. Many of the languages included in each of the three more lenient samples have much fewer missing values than the threshold for the maximum might suggest. The overall proportion of empty cells in each data set is given in table 5. Figure 2 shows the distribution of NA's across languages in the data set with the largest amount of NA's per language (i.e. up to 30 percent, which means 7 features).

Table 5: Overall proportion of empty cells in the three data sets.

Data set	Proportion of NA's
10NA	0.02
20NA	0.07
30NA	0.12

3.4 Statistical modeling

We can quantitatively investigate the question of whether creoles are typologically distinct from non-creoles in two main ways. Firstly, we can look at measures of similarity: If creoles form a distinct typological class, they should be

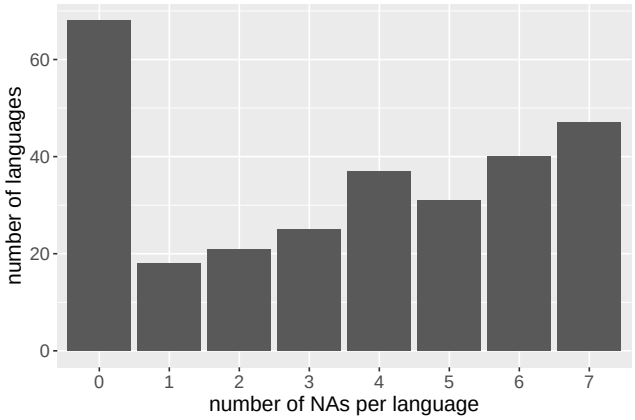

Figure 2: Distribution of NA's across languages in the 30NA data set.

more similar to each other with respect to a set of typological features than to non-creoles. There are different computational models to measure this similarity.

All of these similarity-based models employ clustering techniques that cluster languages according to their degree of similarity. If creoles are distinct from non-creoles, they should cluster together in this analysis and only join with the non-creoles late in the construction of the tree.

The second way in which we can investigate the question of the typological distinctness of creoles quantitatively is by devising models that predict whether a language is a creole or non-creole on the basis of specific feature combinations. We can then assess the predictive power of these models: if they manage to predict with great accuracy whether some language is a creole language, this supports the idea that creoles are typologically distinct.

For our analyses we used a variety of clustering algorithms and classifying models. All of these have been used and are established methods across linguistic subdisciplines such as typology, phonetics, phonology, morphology, sociolinguistics or psycholinguistics. We will explain and discuss each method in turn.

3.4.1 Clustering languages according to their features

For clustering the languages in our samples we used phylogenetic networks and hierarchical cluster analysis. To fit phylogenetic networks we employed to different models, 'neighbor-joining' and 'neighbor net'.

The neighbor-joining algorithm is a bottom-up clustering method borrowed from biology, based on a distance matrix, which specifies the distance in the feature space between taxa (in our case: languages). Based on the similarities of features a network is iteratively built up, with the edges in the network representing these similarities, and the leaf nodes representing languages.

The neighbor net model is an extension of the neighbor-joining algorithm Levy & Pachter (2011), and its computations are based on a dissimilarity map. The model allows for a direct analysis of conflicting signals within a network. We used the R packages `ape`, `phangorn` and `tanggle` (Paradis et al., 2002; Schliep,

2011; Schliep et al., 2017; Yu et al., 2017) for our analyses.

Hierarchical cluster analysis (see, for example, Baayen 2008 for an introduction) builds a hierarchy of clusters based on the similarity between the objects that are being clustered. A binary tree (a so-called dendrogram) is built bottom-up by iteratively joining the two most similar clusters, starting with each individual language as its own cluster. We used the R packages `cluster` (Maechler et al., 2023) and `colorhcplot` (Fantini, 2018) for the cluster analyses.

We will visualize the results of the clustering algorithms with the help of graphs that show the clustering of languages of the two classes. As we are, for the reasons outlined in the introduction, not interested in the exploration of the details of each individual model, and the position of individual languages in the clusters, we only label the graphs with the two data base names. In doing so, we deliberately want to draw the readers' attention away from the individual results in specific models to the differences between the samples and models. The interested reader is referred to the supplementary material for the documentation of graphs that show more detailed information.

3.4.2 Predicting class membership

As we are dealing with nominal variables as predictors and predicting the class membership of an item (in this case a language) is our goal, tree-building algorithms are the method of choice. These algorithms work through all predictors and partition the data into subsets that differ significantly in their distribution of the response variable from other subsets, with the subsets being characterized by particular constellations of the values of the predictor variables. Available tree-building algorithms differ in their underlying mathematical procedures. We used three different kinds of model: recursive partitioning and regression trees, conditional inference trees, and random forests, as implemented in the R packages `party, partykit, rpart.plot, caret, mltools` (Torsten Hothorn et al. 2009; Torsten Hothorn & Achim Zeileis 2009; Stephen Milborrow 2011; Kuhn 2007; Ben Gorman 2016).

Conditional inference trees and recursive partitioning and regression trees are binary trees, but unlike hierarchical clustering, these are built top-down. The feature that best predicts type of language splits the set up, and this is repeated as long as it improves the predictive power of the model.

Random forests combine multiple decision trees. Each tree is devised on a random subset of the data and predicts an outcome. The prediction of the random forest as a whole is determined by aggregating the predictions of its trees.

For reasons of comparability, all tree-building algorithms have been used with their default settings. For the same reasons the trees have not been pruned or cross-validated (in general, classification trees tend towards overfitting, i.e. they introduce splits that are not really meaningful (Baayen, 2008, section 5.2.1 for discussion)).

4 Results

4.1 Phylogenetic networks

We will first look at the results yielded by the two different network algorithms. For this comparison we use the 0NA data set (section 4.1.1). We then turn to the comparison of the different data sets, restricting ourselves to the neighbor-joining algorithm (section 4.1.2).

4.1.1 Neighbor-joining vs neighbor net networks

Figure 3 gives the network based on the neighbor-joining method. WALS languages are given in red, APiCS languages in black. We can see six major clusters and one branch containing a single WALS language. The two top left clusters are exclusively populated with APiCS languages. The bottom center cluster is dominated by WALS languages, with only one APiCS language branching off first. The top right cluster contains roughly the same proportion of WALS and APiCS languages, and the remaining two clusters each contain one WALS language with a few APiCS languages around it. Overall, it seems that there are clear and important differences in the feature constellations of WALS vs APiCS languages. However, some clusters tell us that there are groups of structurally similar languages that come from both data sets.

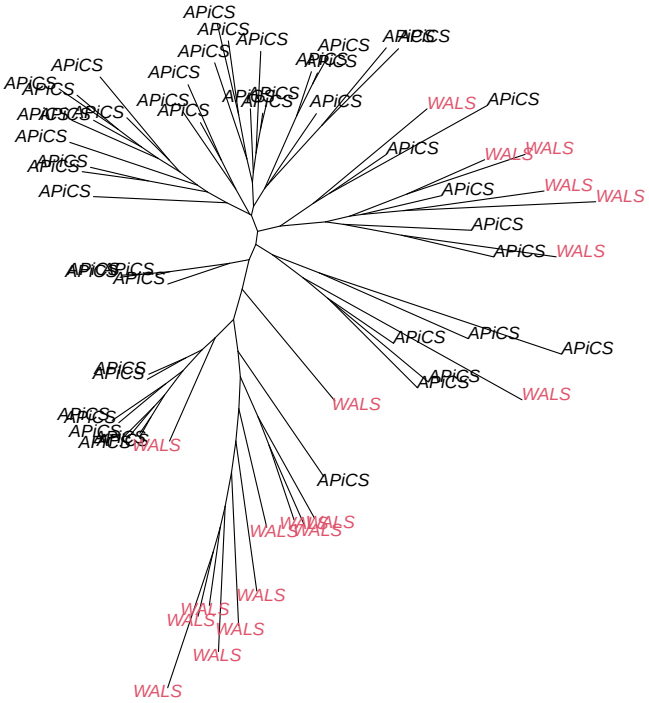

Figure 3: Neighbor-joining-based phylogenetic network of the 0NA data set. WALS languages are given in red, APiCS languages in black.

The neighbor net model, as depicted in figure 4, shows very similar results. Imagining a diagonal from the upper left to the lower right corner of the graph, we have to main clusters. The upper cluster has exclusively APiCS languages, with one WALS outlier (top center). The lower cluster predominantly contains WALS-languages, with only a small set of APiCS languages (bottom center).

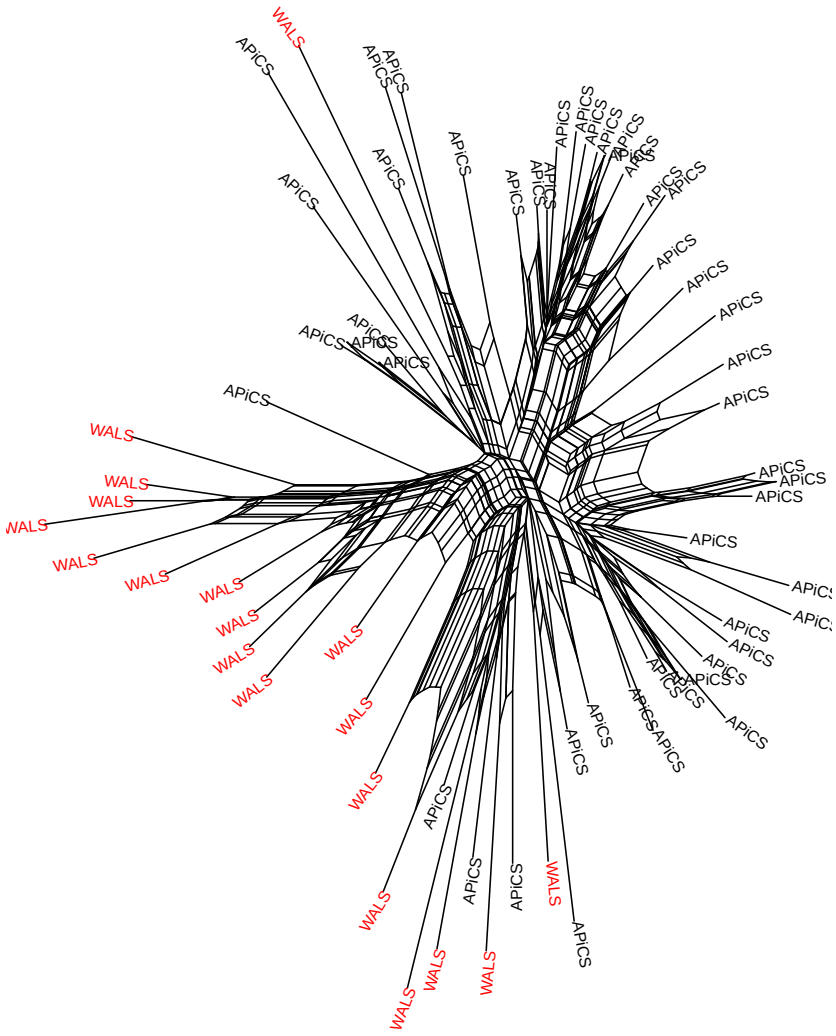

Figure 4: Neighbor net-based phylogenetic network of the 0NA data set. WALS languages are given in red, APiCS languages in black.

Overall, it seems that the networks produced by the two algorithms differ only in some details (which one could explore in more detail), but generally lead to the same conclusions concerning our research question: In both networks, we can identify clearly APiCS-dominant clusters, clearly WALS-dominant clusters, and one relatively balanced cluster (in figure 3). APiCS and WALS languages thus appear to mostly exhibit different feature constellations, with a certain overlap in the various feature values. Let us now turn to the comparison of the different samples. For this, we focus on neighbor-joining networks.

4.1.2 Neighbor-joining networks

Figure 5 shows the results of the neighbor-joining analysis of the 10NA data set. APiCS languages are indicated in black, WALS languages in red. The

network comprises seven main clusters. APiCS languages are mainly found in the three bottom left clusters. One of these clusters (bottom center) contains approximately the same number of APiCS languages as WALS languages. The two leftmost clusters are predominantly populated with APiCS languages, with only two interspersed WALS languages. The remaining four clusters are dominated by WALS languages. While a few APiCS outliers can be found in three of the WALS-dominated clusters, the rightmost cluster is exclusively made up of WALS languages. Overall, we can see a clear trend that the APiCS and WALS languages tend to form separate groups, which implies that the languages differ in their structural make-up. However, the few APiCS outliers that are found in the area predominantly populated by WALS languages and the reverse patterning (i.e. WALS outliers in APiCS-dominated clusters) illustrate that the two types of languages also share certain structural characteristics.

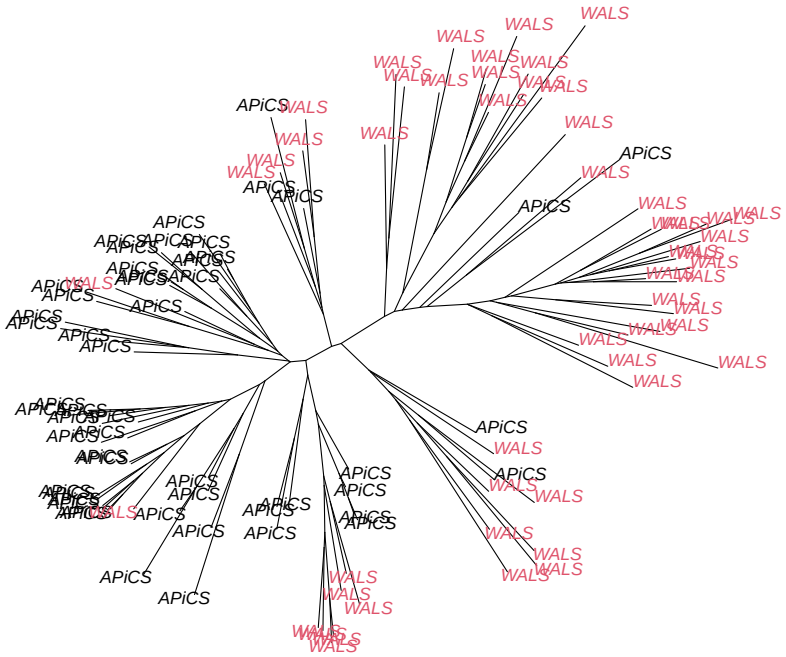

Figure 5: Neighbor net-based phylogenetic network of the 10NA data set. WALS languages are given in red, APiCS languages in black.

The neighbor net-based phylogenetic network of the 20NA data set is shown in figure 6. Similar to figure 5, which had seven major clusters, we can detect eight major clusters in figure 6. APiCS languages are predominantly found in one of the eight clusters (bottom right), in which two WALS outliers appear. Apart from a few APiCS outliers, the remaining seven clusters are populated

exclusively with WALS languages (with one cluster having only two languages). As already shown in figure 5, there seems to be a general trend that the two types of languages form distinct clusters. However, this separation is not always strict, leading to some mixed clusters.

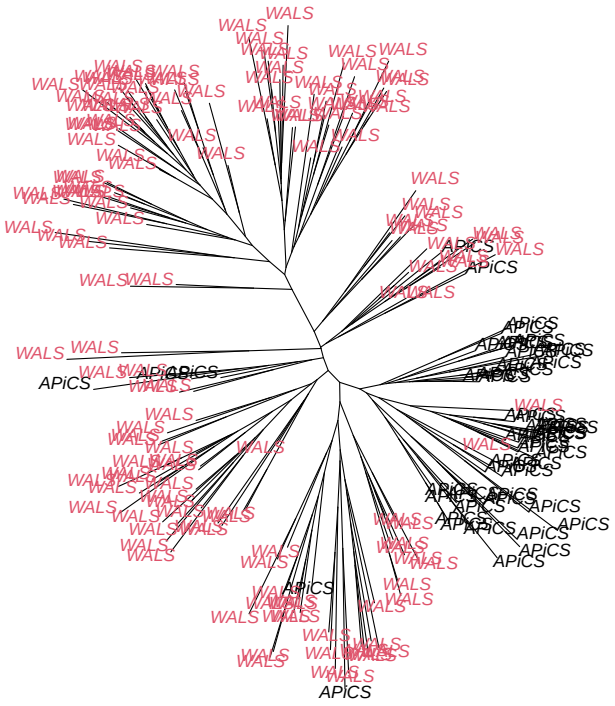

Figure 6: Neighbor net-based phylogenetic network of the 20NA data set. WALS languages are given in red, APiCS languages in black.

Figure 7 shows the results of the neighbor net-based phylogenetic tree analysis of the 30NA data set. We can see a similar pattern as observed in figures 5 and 6. The network contains a few large clusters, and almost all APiCS languages are found in only two of the clusters. All other clusters mainly contain WALS languages. Of these WALS-dominated clusters, some have no APiCS languages (e.g. the bottom left clusters), and others have a few APiCS outliers (e.g. mid left). Taken together, these results once again show that the APiCS and WALS languages seem to exhibit different feature constellations which largely results in the separate grouping in the network. On the other hand, the close grouping of some of the APiCS and WALS languages in some clusters indicates that there also seem to be certain structural features that both types of languages have in common.

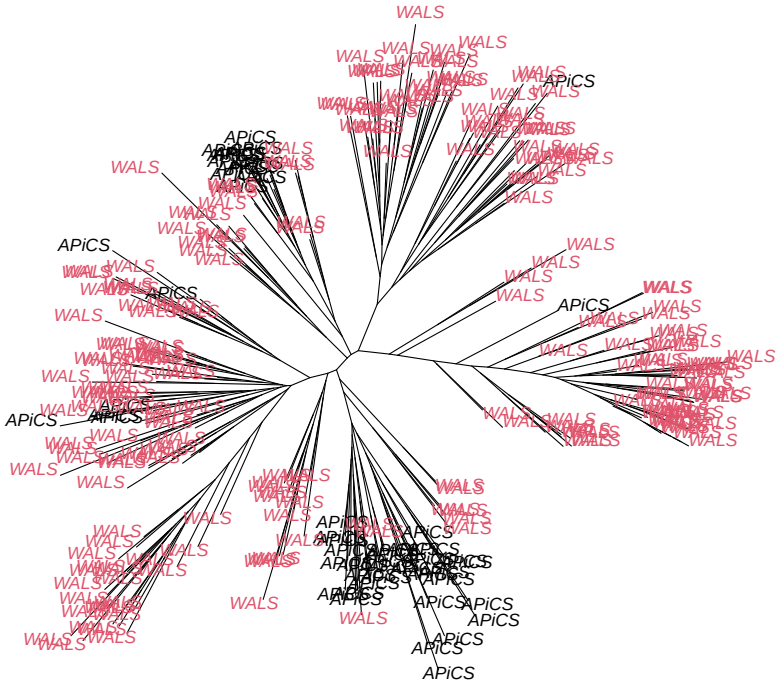

Figure 7: Neighbor net-based phylogenetic network of the 30NA data set. WALS languages are given in red, APiCS languages in black.

In short, the results of the phylogenetic network analyses remains consistent over the four data sets. In both the neighbor-joining and the neighbor net analyses, the general trend is that WALS and APiCS languages form separate groups, but there are also some mixed groups. Comparing the results of the two algorithms has additionally shown that these findings seem to be robust across two different clustering methods.

The results of a different type of cluster analysis (i.e. hierarchical cluster analysis) applied to the same four data sets, are examined in the following section.

4.2 Hierarchical cluster analysis

Figures 8 and 9 depict the results of the hierarchical cluster analysis, which arranges objects, in our case languages, into clusters according to their identified similarity. Each dendrogram represents the language data as found in one of the four data sets. The different clusters represent languages that are similar to each other based on their feature constellation. The y-axis ('height') represents the strength of the similarities between clusters or between the final leaves (i.e. the languages). Long vertical lines indicate more distinct separation between

the groups or the languages.

Languages are represented by numbers, WALS languages by red numbers, APiCS languages by black numbers. Red lines are used for clusters that exclusively contain WALS languages, black lines are reserved for clusters that only feature APiCS languages, and grey lines go down to languages that are in mixed clusters.

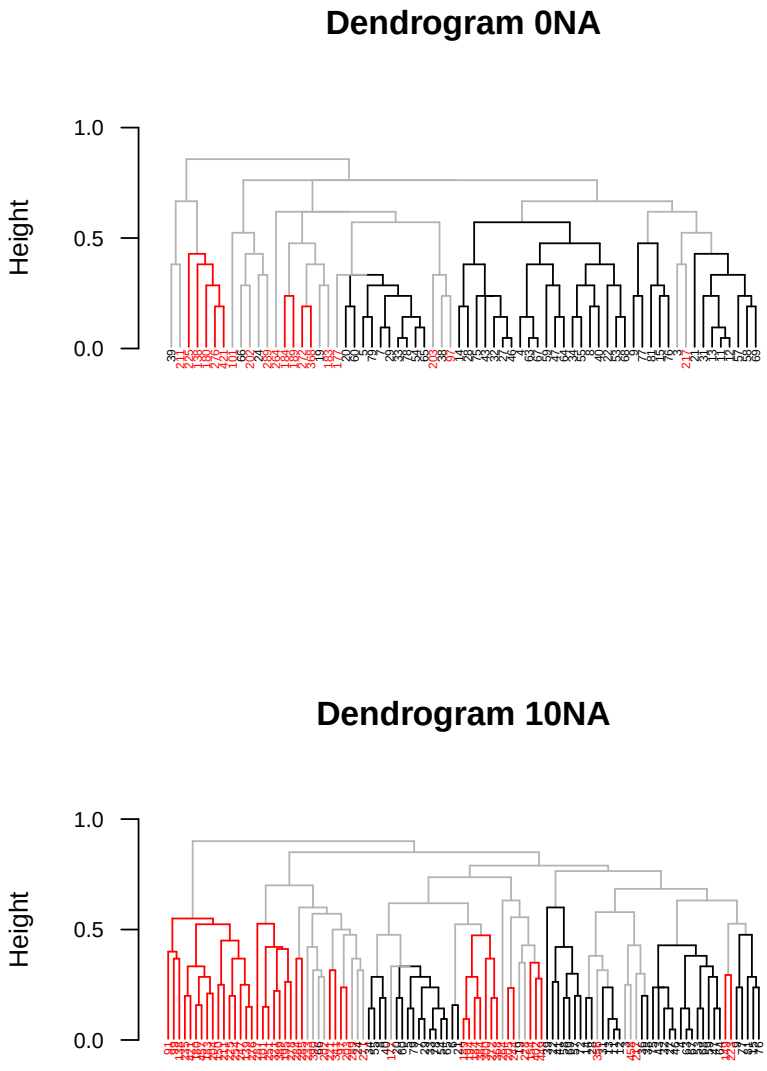

Figure 8: Dendograms of hierarchical cluster analysis for data sets 0NA and 10NA. WALS languages are given in red, APiCS languages in black.

In the dendrogam of the 0NA data set (figure 8), we can see that the first split from the top results in two main clusters which each divide further into

smaller subclusters. The left one of the top two clusters is mostly populated with WALS languages, while the right is dominated by APiCS languages. Like in the phylogenetic network for this data set, we see a general trend towards a split between creoles and non-creoles, but each group is interspersed with members of the other group.

Many subclusters are uniform, and only a few small subclusters are mixed, i.e. end in grey lines. As already observed in figure 3 we predominantely see clear differences in the grouping of APiCS and WALS languages in the dendograms, but also clusters that indicate significant similarities between the two types of languages.

The second dendrogram in figure 8 is based on the 10NA data set. In comparison to the 0NA dendrogram, it contains a larger number of subclusters and additional levels of branches. This is attributable to the increased number of WALS languages included in the sample (as seen in figure 1). Overall, we see a similar pattern as observed in the 0NA dendrogram. There are clusters only containing WALS or APiCS languages, but also mixed clusters at different levels of similarity.

The two dendrograms in figure 9 show the results of the hierarchical cluster analysis of the 20NA and the 30NA data set.

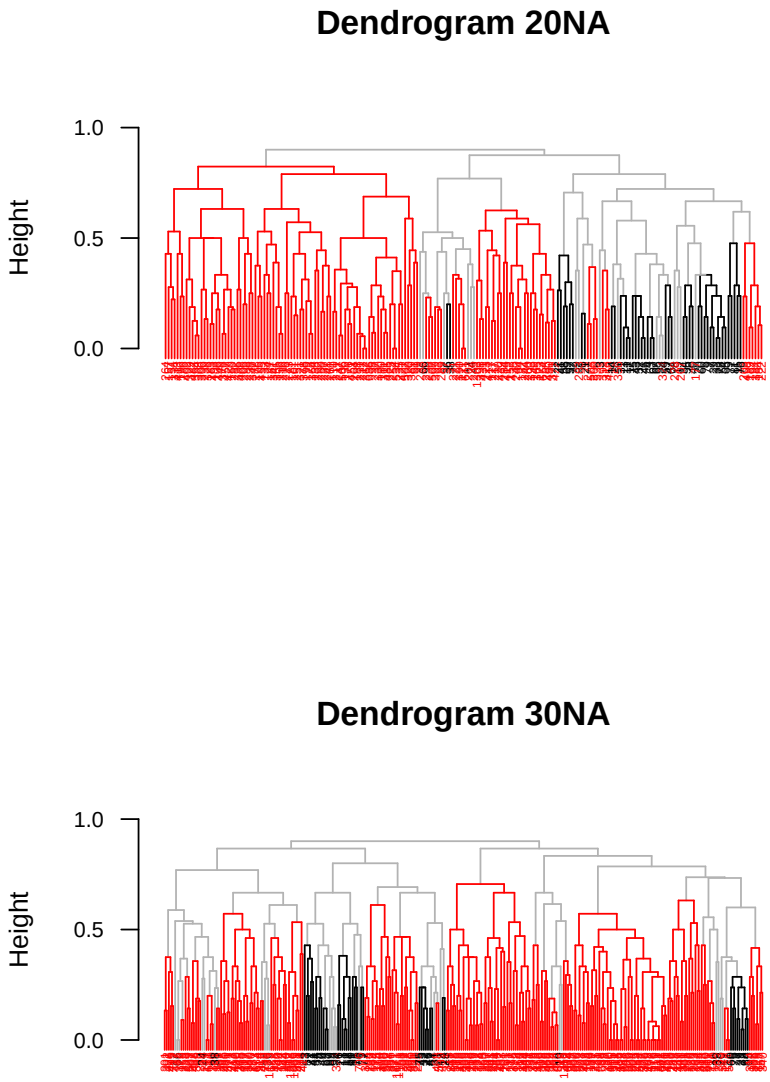

Figure 9: Dendograms of hierarchical cluster analysis for data sets 20NA and 30NA. WALS languages are given in red, APiCS languages in black.

We can see that the number of subclusters and the number of levels of branching rises as the amount of languages increases (as seen in figure 1). The

overall distribution of the APiCS and WALS languages seems to be analogous to the one found in the dendrograms of figure 8. At different similarity levels, we can differentiate between clusters that only contain either WALS or APiCS languages, and few clusters that are mixed.

Overall, these results confirm the findings of the neighbor-joining and neighbor net analyses and show that the feature constellations often differ for WALS and APiCS languages, but also share similarities in some cases.

4.3 Conditional inference trees

We now turn to algorithms that predict class membership based on the feature constellation of the item to be classified. For the assessment of the goodness of fit of the predictive models we used accuracy scores (i.e. as the percentage of correctly classified instances) and F-scores (F1). The F1 score is the harmonic mean of two measures, precision and recall. 'Recall' is the number of items for which the model correctly predicts a particular outcome divided by the number of items which have that observed outcome (i.e. the ratio of true positive cases to all cases of this class). It is thus a measure of how well the model is able to find a particular outcome. 'Precision' is the number of items for which the model correctly predicts a given outcome divided by the number of all items for which the model predicts that outcome (i.e. the ratio of true positives and all positives). Precision thus tells us how well aimed the model is in its predictions.

The logic of these analyses is that if a model can predict with high accuracy whether a language is a creole or not, it means that there are clearly identifiable features that mark creoles typologically.

The first algorithm we employ is conditional inference trees. For reasons of space we restrict the presentation to the lowest and highest sample sizes, i.e. the 0NA and 30NA data sets, as shown in figures 10 and 11, respectively. The tree can be read as follows. Each node contains the name of the feature according to which the data show a significant split. Note that not all of the splits may be interesting for our purposes because the sensitivity of the algorithm sometimes finds splits that differ only slightly in their (otherwise clear) majority choice, or finds splits that occur within the classes. The nodes are numbered for easy reference. The terminal nodes give the majority choice and the proportions of the two choices. For instance, in figure 10 node 4 has 17 languages, of which 82 percent (i.e. 14) are APiCS languages. Node 2 has 16 languages, all of which are WALS languages.

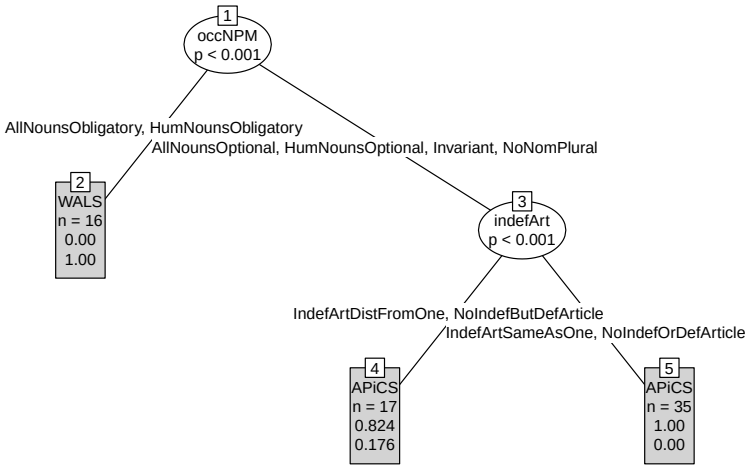

Figure 10: Conditional inference tree, data set 0NA.

The 0NA data set model has two features that have a significant influence on class membership predictions: Occurrence of nominal plural markers ('occNPM') is the most important (and only) feature differentiating between the two classes (node 2 vs. node 3). The indefinite article ('indefArt', node 3) only differentiates between different APiCS languages (nodes 4 and 5). The accuracy of the predictions based on the two significant features is 96 percent (F-score: 0.97). All 49 APiCS languages are correctly classified, and only three of the 19 WALS languages (i.e. 16 percent) are incorrectly predicted to come from the APiCS data set (all three WALS languages are in node 4).

The conditional inference tree for the largest data set shows many more influential features. Occurrence of nominal plural markers is still the most important feature (node 1). Looking at the left branch below node 1, we can see that the distance contrasts in demonstratives ('distContDem', node 3) is distinctive for a subset of the data that is also defined by the order of adjective and noun ('orderAdjNoun', node 2). Notably, these features distinguish mostly between different types of WALS languages. The right branch below node 1 brings in more features that can distinguish languages: the position of interrogative phrases in content questions ('posOfInterrogPhrasesInContentQuest', node 7), the indefinite article ('indefArt', node 8), the order of subject, verb and object ('orderSVO', node 9) and, finally, politeness distinctions in second-person pronouns ('politDist2PP', node 11).

The accuracy for this data set is 95 percent, with an F1-score of 86 percent. Of the 53 APiCS languages, 11 (i.e. 21 percent) are wrongly classified, but only 3 of the 234 WALS languages (i.e. 1 percent).

Let us summarize what we have learned from the inspection of the two inference trees: even though the results across the different samples are quite different in certain details, it is possible for both data sets to predict class membership quite successfully. A qualitative investigation of the feature constellations re-

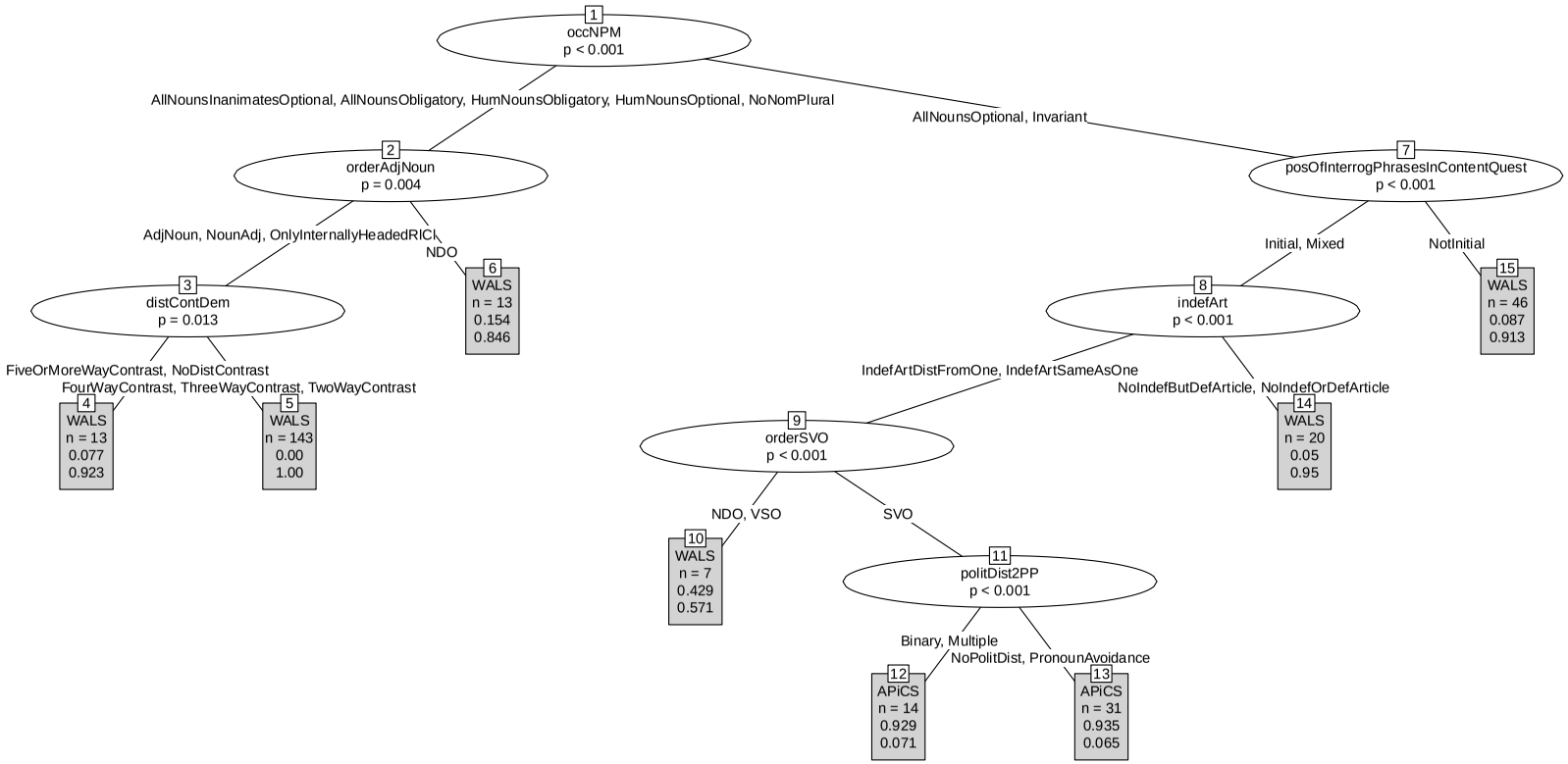

Figure 11: Conditional inference tree, data set 30NA.

veals however, that those constellations that are distinctive of the two classes vary across samples. It seems that larger samples, not unsurprisingly, can bring about a more fine-grained picture concerning the question of which features are distinctive between the classes.

4.4 Recursive partitioning trees

The recursive partitioning trees for the 0NA and 30NA data sets as given in figures 12 and 13 show results similar to the conditional inference trees. The layout of these trees is somewhat different from the conditional inference trees. Each node gives the predicted class, below which we find the predicted probability of 'WALS' as the outcome. Each node also gives the percentage of observations in this node. For instance, in figure 12 the top node predicts APiCS as the majority choice on the basis of 100 percent of the data, with 0.28 probability of WALS languages. Below each node with a split there is the name of the feature that is responsible for the split in this node, and the feature values that characterize the left branch under the node. The right node contains items with the other feature values.

The 0NA data set shows only one split, according to the feature occurrence of nominal plural markers ('occNPM'), as we also saw in the corresponding conditional inference tree. The split in terms of feature values is slightly different from the one found in the conditional inference tree, and the accuracy is 96 percent (F1=0.97).

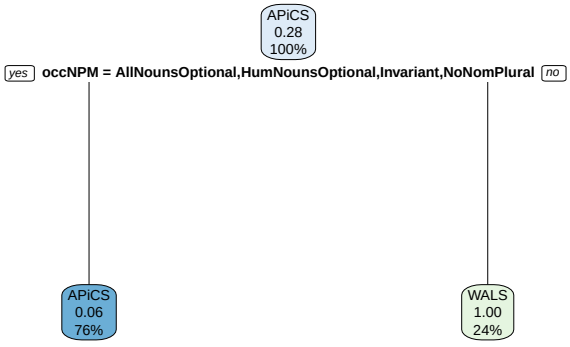

Figure 12: Recursive partitioning tree, data set 0NA.

The tree for the 30NA data set has fewer splits than the corresponding conditional inference tree. The set of influential features is, however, a subset of those of the conditional inference tree: occurrence of nominal plural markers ('occNPM'), the order of subject, verb and object ('orderSVO'), and the order of adjective and noun ('orderAdjNoun'). The accuracy of the predictions for this data set is 0.96 (F1=0.87).

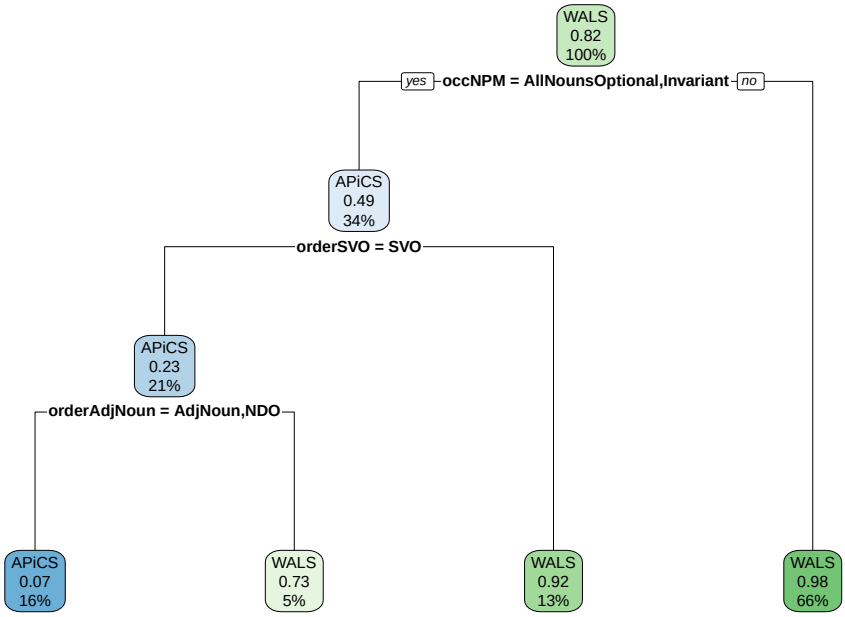

Figure 13: Recursive partitioning tree, data set 30NA.

In sum, the recursive partition trees tell a similar story as the conditional inference trees. A larger sample is more varied and quite expectedly leads to a tree in which more features are used (and necessary) to differentiate between the two classes. Even though the results across the conditional inference trees and the recursive partition trees differ somewhat concerning which constellations of features are influential, there is a large overlap in those features. All trees are highly successful in predicting class membership.

4.5 Random forests

The random forest analysis with samples that contain empty cells necessitates one-hot encoding of the features, i.e. each value of each feature gets its own variable, which is then encoded by zero or one. This in turn yields a different structure of the results (see below).

The accuracy for the 0NA data set is 93 percent (F1=0.84), that of the 30NA data set is 87 percent (F1=0.92). The random forest analysis is also informative about the importance of individual features and their respective values. Table 6 provides an overview of the most important features and their values.

Table 6: Overview of random forest models across data sets, with the three most important features listed.

Data set	feature and value	accuracy
0NA	occNPM-AllNounsObligatory	0.91
	orderAdjNoun-NDO	
	occNPM-AllNounsObligatory	
10NA	occNPM-AllNounsObligatory	0.96
	orderAdjNoun-NDO	
	adpositions-Postpositions	
20NA	occNPM-AllNounsObligatory	0.97
	adpositions-Postpositions	
	indefP-InterrogBased	
30NA	orderDemNoun-NDO	0.96
	occNPM-HumNounsObligatory	
	occNPM-AllNounsObligatory	

All models reach satisfactory to very good predictive accuracy and a large overlap in the most important feature values. Of the 21 features, we only find five represented in this list of the three most important features across data sets: The occurrence of nominal plural markers ('occNPM', in all four models), the order of adjective and noun ('orderAdjNoun', in two models), the order of adposition and noun phrase ('adpositions', in two models), the order of demonstrative and noun ('orderDemNoun', in one model) and indefinite pronouns ('indefP', in one model).

Overall, we see that across data sets the models are largely able to distinguish between APiCS and WALS languages, irrespective of which data set is selected.

5 Summary and discussion

In this paper we have addressed methodological concerns that have been raised against quantitative typological research in the field of creole studies. We investigated the influence of methodological decisions on the empirical results, testing the controversial hypothesis that creole languages are structurally different from non-creole languages. In a comparative approach, we used different samples and different statistical models. We employed the APiCS data base, taking this source as representing creole languages, and the WALS data base, taken as representing non-creole languages.

Picking up our questions from section 2.3, we can summarize our findings as follows:

- *Do different statistical models yield different results?*
  The different statistical models converge on the same general finding: In line with previous quantitative work, there are remarkable and statistically significant structural differences between creoles and non-creoles.

- *Are there models that should be preferred, and others that are not suitable?*
  Although the models support the main finding, they differ in details. Clustering techniques seem to show more fine-grained patternings of the languages than algorithms that try to predict class membership.

- *Can we trust data sets that have up to 30 percent missing values per language?*
  The differences between APiCS and WALS languages emerge irrespective of which sample is used. Whether we include languages with no, with up to 10, up to 20 or up to 30 percent of missing values, does not substantially influence the main results. However, an increase in sample size (quite expectedly) leads to more fine-grained sets or constellations of predictive features, both with clustering algorithms and regression trees.

With regard to sampling, there is a prominent debate in typology about the representativeness of samples, especially with regard to historical and areal non-independence. With some typological research questions this may be a big problem, with other research questions it may not be. Much of the discussion of historical and areal non-independence has focused on the analysis of a single feature (e.g. order of subject, verb and object). And even in those cases it has been shown that balanced samples are not necessary 'as long as the sample is a variety sample large enough to cover different areas and families' (Guzmán Naranjo & Becker, 2022, 605). We believe that this condition is fulfilled in the WALS data that we use.

However, historical and areal non-independence could still potentially be an issue for the present study if the study had the intention to arrive at a serious result. As we explicitly say in section 1, we are interested not in the results per se, but in the differences of the results across different samples. Obviously we could inspect each sample for areal or historical dependencies, and then relate these dependencies (or non-dependencies) to the nature of the results. Given the general tendencies reflected in all samples and all models, we do not think that such analyses are warranted or necessary to prove our point. The separation of interest is that between WALS and APiCS languages, and this separation is very similar independent of the samples. The relation of areal or historical dependencies to that separation is an important, but very different research question, and beyond the scope of this paper. There are studies that have focused on the issue of these dependencies (e.g. Blasi et al. 2017, or some of the papers in Bakker et al. 2017).

There are three main lessons to learn from the present study. First, it is reassuring to see that different statistical methods yield similar results, and that different sample sizes do not dramatically influence the model outcomes. Second, however, we also saw that different methods and different sample sizes do not yield exactly the same results. This also means that theoretical conclusions drawn on the basis of these results may differ somewhat. A case in point are the different features that different models might find more influential. Each model needs to be inspected critically before drawing far-reaching conclusions. Our results can also be interpreted as showing that quantitative typological research should not be restricted to a methodology that uses only a single statistical model. Third, overall, the differences between APiCS and WALS languages are not categorical but probabilistic. They concern particular features, and the constellations of their values.

In view of these points, one might also wonder whether research questions that reduce highly complex phenomena to simple yes/no questions should be replaced by more nuanced questions, which may then be more clearly answered by probabilistic models.

References

Baayen, H. 2008. *Analyzing linguistic data. A practical introduction to statistics.* Cambridge: Cambridge University Press.

Baker, P. 1990. Off target? *Journal of pidgin and creole languages* 5(1). 107–119.

Bakker, P. 2023. Empiricism against imperialism: Science, dogma and the neocolonial heritage of creole studies. Reflections on. *Journal of Pidgin and Creole Languages* .

Bakker, P., F. Borchsenius, C. Levisen & E. M. Sippola. 2017. *Creole studies–phylogenetic approaches.* John Benjamins Publishing Company.

Bakker, P., A. Daval-Markussen, M. Parkvall & I. Plag. 2011. Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole languages* 26(1). 5–42.

Ben Gorman. 2016. mltools: Machine Learning Tools. doi: 10.32614/CRAN.package.mltools. Institution: Comprehensive R Archive Network Pages: 0.3.5. URL `https://CRAN.R-project.org/package=mltools`.

Bickel, B. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology* 11. 239–251.

Blasi, D. E., S. M. Michaelis & M. Haspelmath. 2017. Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour* 1(10). 723–729.

Bouckaert, R., P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard & Q. D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097). 957–960.

Cysouw, M. 2008. Using the World Atlas of Language Structures. *Language Typology and Universals* 61(3). 181–185. doi: doi:10.1524/stuf.2008.0018. URL `https://doi.org/10.1524/stuf.2008.0018`.

Daval-Markussen, A. 2019. *Reconstructing creole.* Aarhus: Aarhus University Phd dissertation.

DeGraff, M. 2003. Against creole exceptionalism. *Language* 79(2). 391–410.

Dryer, M. S. & M. Haspelmath. 2013. WALS Online (v2020.3). *Zenodo https://doi. org/10.5281/zenodo* 7385533.

Dunn, M., A. Terrill, G. Reesink, R. A. Foley & S. C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309. 2072–2075.

Fantini, D. 2018. *colorhcplot: Colorful Hierarchical Clustering Dendrograms.* URL `https://doi.org/10.32614/CRAN.package.colorhcplot`.

Guzmán Naranjo, M. & L. Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.

Haspelmath, M. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Holm, J. & P. L. Patrick. 2007. *Comparative creole syntax*. Battlebridge.

Jaeger, T. F., P. Graff, W. Croft & D. Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15. 281–320.

Kuhn, M. 2007. caret: Classification and Regression Training. doi: 10.32614/CRAN.package.caret. Institution: Comprehensive R Archive Network Pages: 6.0-94. URL `https://CRAN.R-project.org/package=caret`.

Lefebvre, C. 1998. *Creole Genesis and the Acquisition of Grammar: The Case of Haitian Creole*. Cambridge: Cambridge University Press.

Levy, D. & L. Pachter. 2011. The neighbor-net algorithm. *Advances in Applied Mathematics* 47(2). 240–258.

List, J.-M. 2021. *Computer-assisted approaches to historical language comparison*. Jena: Friedrich-Schiller-Universität Jena Habilitation thesis.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert & K. Hornik. 2023. *cluster: Cluster Analysis Basics and Extensions*. URL `https://CRAN.R-project.org/package=cluster`. R package version 2.1.6 — For new features, see the 'NEWS' and the 'Changelog' file in the package source).

McWhorter, J. H. 1998. Identifying the creole prototype: Vindicating a typological class. *Language* 74(4). 788–818.

Meakins, F. 2022. Empiricism or imperialism: The science of Creole Exceptionalism. *Journal of Pidgin and Creole languages* 37(1). 189–203.

Michaelis, S. M., P. Maurer, M. Haspelmath & M. Huber (eds.). 2013. *The atlas of pidgin and creole language structures*. Oxford University Press, USA.

Murawaki, Y. 2016. Statistical modeling of creole genesis. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1329–1339.

Muysken, P. 1988. Are creoles a special type of language. In F. Newmeyer (ed.), *Linguistics: the Cambridge survey, vol. 2: Linguistic theory: Extensions and implications*, 285–301. Cambridge University Press.

Paradis, E., S. Blomberg, B. Bolker, J. Brown, S. Claramunt, J. Claude, H. S. Cuong, R. Desper, G. Didier, B. Durand, J. Dutheil, R. Ewing, O. Gascuel, T. Guillerme, C. Heibl, A. Ives, B. Jones, F. Krah, D. Lawson, V. Lefort, P. Legendre, J. Lemon, G. Louvel, E. Marcon, R. McCloskey, J. Nylander, R. Opgen-Rhein, A.-A. Popescu, M. Royer-Carenzi, K. Schliep, K. Strimmer & D. De Vienne. 2002. ape: Analyses of Phylogenetics and Evolution. doi: 10.32614/CRAN.package.ape. Institution: Comprehensive R Archive Network Pages: 5.8. URL `https://CRAN.R-project.org/package=ape`.

Parkvall, M. 2008. The simplicity of creoles in a cross-linguistic perspective. In M. Miestamo, F. Karlsson & K. Sinnemäki (eds.), *Language Complexity. Typology, contact change*, 265–285. John Benjamins Publishing Company.

Plag, I. 2008. Creoles as interlanguages: Inflectional morphology. *Journal of Pidgin and Creole Languages* 23(1). 114–135.

Plag, I. 2011. Creolization and admixture: Typology, feature pools, and second language acquisition. *Journal of Pidgin and Creole languages* 26(1). 89–110.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL `https://www.R-project.org/`.

Schliep, K., A. A. Potts, D. A. Morrison & G. W. Grimm. 2017. Intertwining phylogenetic trees and networks. *Methods in Ecology and Evolution* 8. 1212–1220.

Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4). 592–593.

Skirgård, H., H. J. Haynie, D. E. Blasi, H. Hammarström, J. Collins, J. J. Latarche, J. Lesage, T. Weber, A. Witzlack-Makarevich, S. Passmore, A. Chira, L. Maurits, R. Dinnage, M. Dunn, G. Reesink, R. Singer, C. Bowern, P. Epps, J. Hill, O. Vesakoski, M. Robbeets, N. K. Abbas, D. Auer, N. A. Bakker, G. Barbos, R. D. Borges, S. Danielsen, L. Dorenbusch, E. Dorn, J. Elliott, G. Falcone, J. Fischer, Y. Ghanggo Ate, H. Gibson, H.-P. Göbel, J. A. Goodall, V. Gruner, A. Harvey, R. Hayes, L. Heer, R. E. Herrera Miranda, N. Hübler, B. Huntington-Rainey, J. K. Ivani, M. Johns, E. Just, E. Kashima, C. Kipf, J. V. Klingenberg, N. König, A. Koti, R. G. A. Kowalik, O. Krasnoukhova, N. L. Lindvall, M. Lorenzen, H. Lutzenberger, T. R. Martins, C. Mata German, S. van der Meer, J. Montoya Samamé, M. Müller, S. Muradoğlu, K. Neely, J. Nickel, M. Norvik, C. A. Oluoch, J. Peacock, I. O. Pearey, N. Peck, S. Petit, S. Pieper, M. Poblete, D. Prestipino, L. Raabe, A. Raja, J. Reimringer, S. C. Rey, J. Rizaew, E. Ruppert, K. K. Salmon, J. Sammet, R. Schembri, L. Schlabbach, F. W. Schmidt, A. Skilton, W. D. Smith, H. de Sousa, K. Sverredal, D. Valle, J. Vera, J. Voß, T. Witte, H. Wu, S. Yam, J. Ye , M. Yong, T. Yuditha, R. Zariquiey, R. Forkel, N. Evans, S. C. Levinson, M. Haspelmath, S. J. Greenhill, Q. D. Atkinson & R. D. Gray. 2023. Grambank Reveals Global Patterns in the Structural Diversity of the World's Languages. *Science Advances* 9. doi: 10.1126/sciadv.adg6175.

Skirgård, H., H. J. Haynie, H. Hammarström, D. E. Blasi, J. Collins, J. Latarche, J. Lesage, T. Weber, A. Witzlack-Makarevich, M. Dunn, G. Reesink, R. Singer, C. Bowern, P. Epps, J. Hill, O. Vesakoski, N. K. Abbas, S. Ananth, D. Auer, N. A. Bakker, G. Barbos, A. Bolls, R. D. Borges, M. Browen, L. Chevallier, S. Danielsen, S. Dohlen, L. Dorenbusch, E. Dorn, M. Duhamel, F. E. H. Ali, J. Elliott, G. Falcone, A.-M. Fehn, J. Fischer, Y. G. Ate, H. Gibson, H.-P. Göbel, J. A. Goodall, V. Gruner, A. Harvey, R. Hayes, L. Heer, R. E. H. Miranda, N. Hübler, B. H. Huntington-Rainey, G. Inglese, J. K. Ivani, M. Johns, E. Just, I. Kapitonov, E. Kashima, C. Kipf, J. V. Klingenberg, N. König, A. Koti, R. G. A. Kowalik, O. Krasnoukhova, K. L. Lindsey,

N. L. M. Lindvall, M. Lorenzen, H. Lutzenberger, A. Marley, T. R. A. Martins, C. M. German, S. van der Meer, J. Montoya, M. Müller, S. Muradoğlu, HunterGatherer, D. Nash, K. Neely, J. Nickel, M. Norvik, B. Olsson, C. A. Oluoch, D. Osgarby, J. Peacock, I. O. Pearey, N. Peck, J. Peter, S. Petit, S. Pieper, M. Poblete, D. Prestipino, L. Raabe, A. Raja, J. Reimringer, S. C. Rey, J. Rizaew, E. Ruppert, K. K. Salmon, J. Sammet, R. Schembri, L. Schlabbach, F. W. P. Schmidt, D. Schokkin, J. Siegel, A. Skilton, H. de Sousa, K. Sverredal, D. Valle, J. Vera, J. Voß, D. W. Smith, T. Witte, H. Wu, S. Yam, J. Y. , M. Yong, T. Yuditha, R. Zariquiey, R. Forkel, N. Evans, S. C. Levinson, M. Haspelmath, S. J. Greenhill, Q. D. Atkinson & R. D. Gray. 2023. Grambank v1.0. doi: 10.5281/zenodo.7740140. Dataset. URL `https://doi.org/10.5281/zenodo.7740140`.

Stephen Milborrow. 2011. rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. doi: 10.32614/CRAN.package.rpart.plot. Institution: Comprehensive R Archive Network Pages: 3.1.2. URL `https://CRAN.R-project.org/package=rpart.plot`.

Torsten Hothorn & Achim Zeileis. 2009. partykit: A Toolkit for Recursive Partytioning. URL `http://R-Forge.R-project.org/projects/partykit/`.

Torsten Hothorn, Kurt Hornik, Carolin Strobl & Achim Zeileis. 2009. party: A Laboratory for Recursive Partytioning. URL `http://CRAN.R-project.org/package=party`.

Yu, G., D. K. Smith, H. Zhu, Y. Guan & T. T. Lam. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8. 28–36. doi: 10.1111/2041-210X.12628.