The relevance of valence orientation for clustering three Niger-Congo language families

Marc Tang, Sylvie Voisin, Stéphane Robert

Submitted to *Journal of Quantitative Linguistics*, 25 November 2019

The relevance of valence orientation for clustering three Niger-Congo language families

Marc Tang, Sylvie Voisin, Stéphane Robert
Université Lumière Lyon 2, DDL, CNRS
Aix Marseille University, DDL, CNRS
LLACAN, CNRS, ~~Sorbonne Paris Cité~~ & INALCO

Abstract
Valence orientation is one of the typological parameters used in linguistics to differentiate the phylogenetic affiliation of languages. Within previous studies, a sample of 18 verb pairs is commonly used to compare the valence orientation across languages. However, recent studies suggest that verb pairs 10-18 are more relevant than verb pairs 1-9 to identify the phylogenetic affiliation of languages. This study further develops these hypotheses by assessing quantitatively the predictive power of the 18 verb pairs with regard to 38 languages from the Atlantic, Mande, and Mel families. The results from clustering and classification analyses show that verb pairs 10-18 are indeed more relevant to identify the phylogenetic affiliation of languages from the current sample. Moreover, the analysis pinpoints which specific features play a major role in the classification task.

1. Introduction

The aim of this paper is to assess statistically the regularities of valence orientation in three families of languages spoken in West Africa, and to explore the role of this linguistic feature as a phylogenetic marker. Valence orientation refers to the overall tendency of a language to treat members of causal-noncausal verb alternations in a particular way (Nichols et al. 2004), that is to code the alternation by specific morphological means. The causal (C)/non-causal (nC) alternation is defined here as a semantic distinction based on the presence/absence of a causer in a pair of verbs referring to the same core event or state-of-affairs, e.g., *kill* vs. *die*, *raise* vs. *rise* in English. Five possible strategies for the coding of causal-noncausal pairs are attested crosslinguistically. They are listed and respectively symbolized by “>”, “≠”, “<”, “~”, and “=” in Table 1.

Table 1. The different coding strategies exemplified in Wolof (Atlantic)

Type	*Abbreviation*	*Example*
Causativization	nC > C	*réer* ‘to be lost’ > *réer-al* ‘to lose’
Decausativization	nC < C	*sakk-u* ‘to be sealed’ < *sakk* ‘to seal’
Lability	nC = C	*lakk* ‘to burn (intr.)’ = *lakk* ‘to burn’(tr.)
Suppletivism	nC ≠ C	*dee* ‘to die’ ≠ *rey* ‘to kill’
Equipollence	nC ~ C	*daan-u* ‘to fall’ ~ *daan-al* ‘to let fall’

Taking Wolof (Atlantic) as an example, all strategies are found in this language. Causativization refers to pairs in which the causative meaning is generated by expanding the non-causal form of the verb, as in *réer* (be_lost) 'to be lost' and *réer-al* (be_lost-CAUS) 'to lose'. The causal meaning 'to lose' is obtained here by adding a causative marker on the non-causal verb form 'to be lost'. Decausativization refers to the reverse configuration whereby the noncausal form is obtained by adding a decausative marker on the causal (base) form, namely a middle suffix in *sakk-u* (seal-MID) 'to be sealed' out of *sakk* 'to seal'. Lability applies to a causal-noncausal pair involving no formal change, like *lakk* 'to burn' (tr.) and *lakk* 'to burn (intr.)'. Suppletivism involves two distinct verbal lexemes paired in a causal-noncausal alternation, like *rey* 'kill' and *dee* 'to die'. Finally, with the equipollent strategy, the causal and non-causal meanings are generated from the same verb with an equivalent morphological complexity, as in *daan-al* (knock_down-CAUS) 'to drop, to fell' vs. *daan-u* (knock_down-MID) 'to fall'. The root verb *daan* is used in traditional wrestling when one of the opponents wins by knocking down the other one. The non-causal meaning of the 'fall/fell' pair is obtained by a middle voice derivation, whereas the causal meaning is generated by a causative derivation on the same (base) verb *daan*. Typologically, the alternation with the equipollent strategy can be achieved through derivational (as in Wolof), inflectional or phonological marking.

The Atlantic, Mande, and Mel families are affiliated to the same Niger-Congo phylum but have diverging typological profiles and long-lasting historical contacts in some parts of their extension area, namely in Senegal and the surrounding areas. This situation creates an ideal case study for both testing the predictive power of valence orientation with regard to family affiliation and inferring which linguistic features or components are more resistant to language contact. By way of illustration, some words of the lexicon are more likely to undergo borrowing while others tend to be more stable, e.g., nouns tend to be easier to borrow than adjectives or verbs (Tadmor et al. 2010). Similar tendencies are observed with morphosyntactic features (Matras 2009, 2010). Among them, valence orientation is generally considered as a general typological parameter of languages and has been studied with various approaches (Nichols et al. 2004, Haspelmath et al. 2014, Bickel 2015, Robert & Voisin 2018) but seldom with statistical methods. This study aims at filling this gap.

The data are extracted in a similar way as in previous qualitative studies (Robert & Voisin 2018), to facilitate the comparability of the analyses. In total, 18 specific verb pairs (see Appendix 1) are selected based on Nichols et al. (2004). The paradigms for these verb pairs are extracted from a balanced sample of 38 languages from the Atlantic, Mande, and Mel families (see Appendix 2) and mostly retrieved from the lexical database Reflex [Reference Lexicon of the Languages of Africa] (Segerer & Flavier 2011-2018). Grammars and linguists working on the languages are also consulted when needed.

The regularities of the coding strategies for the 18 verb pairs are investigated in two ways: clustering and classification. First, a principal

component analysis combined with k-means clustering is used to cluster the languages from Atlantic, Mande, and Mel and compare the obtained clusters with the original family groupings. Then, the relevance of each individual verb pair in predicting the original families of languages is investigated by feeding the data to a random forests classifiers. Section 2 provides an overview of the existing literature and recent studies on valence orientation in the Atlantic, Mande, and Mel language families. Section 3 explains how the data are gathered and provides a short explanation of the models of clustering and classification. The results are displayed in Section 4, while Section 5 discusses the results.

2. Literature review (hypothesis)

The Atlantic, Mel and Mande language families are generally represented as sub-branches of the Niger-Congo phylum, subdivided each in various sub-groups of languages (Figure 1). In terms of size, the Mande sub-branch has around 70 languages. The Atlantic sub-branch has near 50 languages. The Mel family is the smallest sub-branch of the sample, with only a dozen of languages. The amount of languages for each sub-branch is listed with approximate numbers due to the divergence of classification in existing studies.

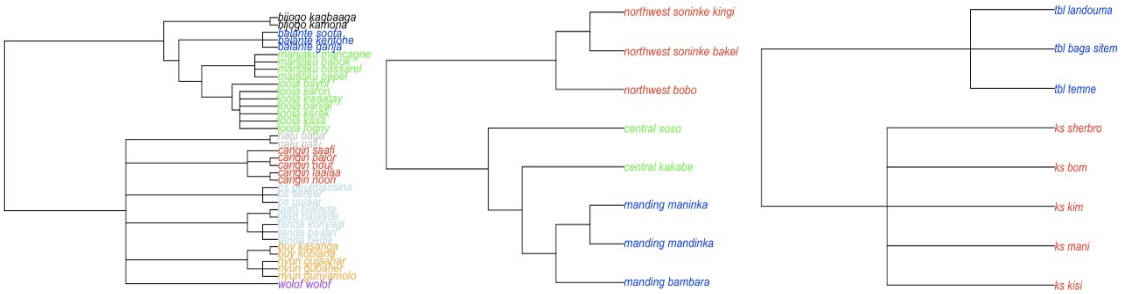


Figure 1. Overview of the languages of the families involved in this study

The three sub-branches have long lasting historical contacts in Senegal and the surrounding areas, which results in diverging opinions as to their phylogenetic affiliations (Childs 2004, 2010; Dwyer 2005; Pozdniakov et al. 2008; Vydrin and Vydrina 2010). On the one hand, the classification of the Mel languages as a sub-branch of Niger-Congo distinct from Atlantic has been recently confirmed by Pozdniakov and Segerer (to appear), after a first suggestion by Dalby (1965). On the other hand, the belonging of the Mande family to the Niger-Congo phylum is questioned (Dimmendaal 2008, 2011, Creissels 2017) and remains under debate (Vydrin 2016, Güldeman 2018[1]). Due to to the insufficient documentation of African languages, only a sample of the languages of these family are

[1]Overall, unless more robust and systematic evidence is brought forward, the long-standing but vague idea that Mande is distant from the rest of Niger-Kordofanian as one of its earliest offshoots should give way to the neutral assessment that it is a family without a proven genealogical affiliation (Güldeman 2018: 192).

investigated here, as for the Mande family but for different raisons (see section 3.1.).

As can be seen from Figure 2, in the area under study, the speakers of Mande languages are generally neighbored by speakers of Atlantic and Mel languages. The contact-induced changes among these three families have already received some attention through various studies.

Figure 2. Languages of Senegal and the surrounding areas (Pozdniakov et al. 2019)

For instance, inside the Manding subgroup of the Mande family, linguistic divergences of some languages such as Mandika and Maninka (included in this study) can be attributed to their contact with Mel and Atlantic languages spoken in the same area (Childs 2010), and viewed as a result of the historical assimilation of Atlantic or Mel speakers during the Manding domination at the time of the Manding (or Mali) Empire[2]. Various grammatical points have been investigated in light of these contacts between the three families (Creissels 2014, Robert and Voisin 2018, Voisin forthcoming). In general, the effects of these historical contacts have been mainly studied from the aspect of lexical borrowings between the Mande and Atlantic languages (Childs 2010, Pozdniakov et al. 2019). The results indicate a larger amount of lexical borrowings from the Mande to the Atlantic languages than the opposite way. However, no linguistic investigations known to the authors have been conducted on

[2]The Manding Empire is known to have lasted for several centuries during the middle-age period, from circa 1235 to 1670 but historians do not agree about the precise dates of its beginning and end.

the contact between the Mel and the Atlantic families beside those touched upon in Robert and Voisin (2018).

In terms of grammatical structures, Mande languages display a typological profile quite different from the Atlantic and Mel languages. The Mande languages do not have noun class systems, they exhibit isolating morphology, and have a strict SOV order. By contrast, Atlantic and Mel languages display a typological profile commonly found in the Niger-Congo language family with an SVO word order, a noun class system, and an agglutinative morphology characterized by a remarkably rich verbal derivation. However, Mel and Mande languages share tonal systems, which are quite rare among Atlantic languages.

Valence orientation across Atlantic, Mande, and Mel languages has recently been used to study the correlation between genetic affiliation and coding profile. For instance, Creissels (2018) investigates this question on a sample of 30 Subsaharan languages belonging mostly (though not exclusively) to the Niger-Congo phylum, including some languages from the Atlantic and Mande families. As an interesting contribution, Creissels demonstrates that several Mande languages show an extreme degree of preference for lability[3], but no language in his sample displays an extreme degree of preference for causativization. In another recent study by Robert & Voisin (2018), a sample of 36 Atlantic languages, 8 Mande languages, and 7 Mel languages is investigated to shed light on the correlation between typological profiles, valence orientation and contact-induced changes in the three Niger-Congo language families. The cross-family comparison is made by extracting the general pattern of distribution of the five main valence strategies across a set of 18 verb pairs predefined by Nichols et al. (2004) (see Appendix 1), in order to define family standard patterns for coding valence orientation and to tackle contact-induced phenomena through deviance of individual languages form their family pattern. Pairs 1 to 9 belong to a type of verbs known to universally favor the causative strategy and are considered to introduce a bias for causativization in the coding profile of languages (Haspelmath et al. 2014), whereas, in contrast, pairs 10 to 18, which belong to a type a verbs known to show the greatest cross-linguistic variation in the coding of causal-noncausal pairs, are considered to reveal the real preferences of languages for coding valence alternation (Haspelmath 1993, Creissels 2018). Concerning the correlation between genetic affiliation and coding profile, the results show that Atlantic and Mel languages share a preference for directed strategies (causativation and decausativization) whereas Mande languages combine a strong propensity for lability with a prevalence of causative coding, thus confirming the correlation when the verb sample was restricted to pairs 10-18. However, no additional quantitative analysis has been carried out to assess the predictive power of valence orientation in determining the

[3]10 languages belonging to 7 of the 15 genetic units represented in the sample show a relative prominence of lability only: Emai (Benue-Congo), Sar (Central Sudanic), Jamsay (Dogon), Minyanka (Gur), Baule (Kwa), Fon (Kwa), Bambara (Mande), Kakabe (Mande), Mano (Mande), and Gbaya (Ubangian). All of them have a very high proportion of labile pairs, of the same range as that found for example in English (between 10 and 12 out of 13).” (Creissels 2018: 5).

language affiliation of Atlantic, Mande, and Mel languages, and the differential predictive power between verb pairs 1-9 and 10-18 has not been further investigated either. This study aims at filling this gap.

3. Methodology (testing method)

In this section, explanation is provided as to how the 38 languages are selected to represent the Atlantic, Mande, and Mel families. Then, an overview about how the valence orientation strategies are associated with each verb pair is provided. Finally, the general process of clustering and classification is displayed. The analyses and visualization in this paper are produced with the packages ape (Paradis & Schliep 2018), ClusterR (Mouselimis 2019), data.table (Dowle & Srinivasan 2019), factoextra (Kassambara & Mundt 2017), ggfortify (Tang et al. 2016), party (Hothorn et al 2006), random (Eddelbuettel 2017), randomForest (Liaw & Wiener 2002), randomForestExplainer (Paluszynska & Biecek 2017), reprtree (Dasgupta 2014), tidyverse (Wickham 2017) from R (R- Core-Team 2018).

3.1 Materials

In Robert and Voisin (2018), the sample of Atlantic, Mande, and Mel families is 41 (60%), 17 (25%), and 10 (15%) languages respectively[4]. Thus, the sample of languages extracted for this study follows a similar phylogenetic distribution. In total, 26 Atlantic languages (68%), 8 Mande languages (21%), and 4 (11%) Mel languages are extracted. The same method is applied for each sub-group of the language families. By way of illustration, the Nyun sub-group accounts for 14% (6/41) of the Atlantic family, thus, a similar ratio of Nyun languages is extracted for the sample (12%, 3/26). The distribution is not exactly the same, but the similarities are considered sufficient for the current analysis.

The 18 verb pairs used in previous studies and replicated in this study are summarized in Table 2. Each pair includes a causative and non-causative meaning, e.g., *laugh* vs. *amuse*. Some verb pairs contain the same word since valence orientation is marked with lability in these English verb pairs. For instance, the verb pair 11 *burn/burn* actually refers to two different verbs with the meaning of *burn* and *make burn*. However, English does not differentiate the verb form between the two meanings so both cells are filled with the form ‘burn’.

Table 2. A simplified overview of the 18 verb pairs. The labels of the columns indicate the causative (C) and non-causative (nC) strategies

	nC	*C*		*nC*	*C*		*nC*	*C*
1	laugh	amuse	7	be angry	anger	13	open	open
2	die	kill	8	fear	scare	14	dry	make dry

[4]This sample does not reflect perfectly the distribution of the Atlantic, Mande, and Mel languages. In total, the three families are considered to have a ratio of 5:7:1 (Atlantic ~ 50 languages, Mande ~ 70 languages, and Mel ~ 10 languages). The sample in our study (as in Robert & Voisin 2018) only selects languages from the three families that are spoken in the same region, since investigating language contact is one of the main purposes of the study. For the Mel and Atlantic families, some languages had to be discarded because of insufficient documentation. For the Mande, which extends much further to the East, only the languages in contact with the two other families have been selected, and two others added for balancing the sample.

3	sit	seat	9	hide	conceal	15	be straight	straighten
4	eat	feed	10	boil	boil	16	hang	hang (up)
5	learn	teach	11	burn	burn	17	turn over	turn over
6	see	show	12	break	break	18	fall	drop

Information for each verb pair is extracted as follows: when possible, the verb pairs are retrieved from the Reflex database. If the verb pairs are not found in the database, the authors rely on linguistic resources (e.g., grammars and texts) of the language to obtain the needed information. The full list of verb pairs also includes proxies for each verb. Please refer to the supplementary materials for further details. After the causative and non-causative forms are found, the verb pair is labeled according to the valence orientation strategy used for that verb pair. A sample with strategies of valence orientation in Landuma (Mel) is shown in Table 3.

Table 3. Valence orientation in *Landuma* (Rogers & Bryant, 2012)

Verb pair (nC/C)	Strategy	nC	C
6 see/show	Causativization nC > C	*wos*	*wos-əs*
10 boil/boil	Decausativization nC < C	*wɔkəc-ʌ*	*wɔkəc*
12 break/break	Suppletivism nC =/ C	*nənk*	*mʌnk*
14 dry/make dry	Lability nC = C	*pʌc*	*pʌc*
18 fall/drop	Equipollence nC ~ C	*funp.ʌ*	*funp-əs*

In terms of data coverage, all languages included in the analysis have more than half (i.e., 9/18) of the verb pairs annotated. The ratio of missing values for each family is Atlantic: 22.6% (106/(26*18)), Mande: 14.6% (21/(8*18)), and Mel: 12.5% (9/(4*18)). A visualization of the data is provided in Figure 3. The y-axis represents the variance of the ratio of the verb pairs using a specific valence orientation strategy across all languages of the family. The colors represent the values when counting all verb pairs and only counting verb pairs 10-18. By way of illustration, no verb pairs using equipollence are found in the Mande family. Thus, the box plot of those columns indicates a ratio of 0% for equipollence. As an another example, languages in the Mande family mostly have between 10% and 30% of their verb pairs using suppletivism, which is reflected in the length of the orange boxplot for that feature.

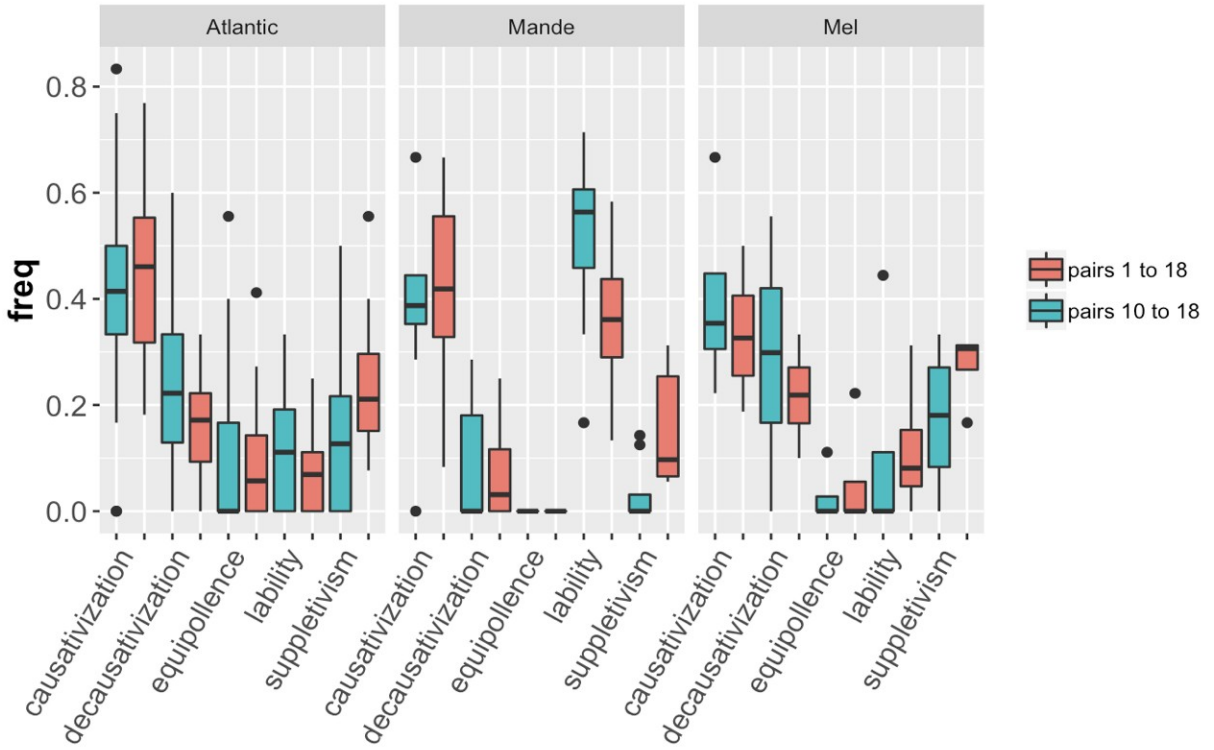


Figure 3. An overview of valence orientation in the three language families

This overview matches with the observations in Robert & Voisin (2018). On the one hand, the three languages families share the tendency to use causativization and the Mande languages use less decausativization. On the other hand, the Mande languages rely much more on lability than the Atlantic and Mel languages. With regard to the different/difference between verb pairs 1-18 and 10-18, we do observe a difference of variance when the pairs 1-9 are removed. For instance, the variance of using causativization reduces substantially when only considering the verb pairs 10-18. This difference of variance seems to support the hypothesis that verb pairs 10-18 are more resistant to language contact and thus more stable. Nevertheless, additional statistical evaluation is needed to support such claim. In terms of variance, a Kolmogorov–Smirnov test is used to assess if the two samples are from the same distribution. The results show that only suppletivism ($D = 0.4$, $p < 0.01$) has a different distribution between the two verb pairs group. When comparing the different strategy ratios across families with a t-test, a significant difference between the verb pairs 1-18 and 10-18 is found for decausativization, lability, and suppletivism ($df=37$, $p < 0.01$). These preliminary observations indicate that using different groups of verb pairs is likely to have an impact on clustering and classification. The following sub-section provides an overview of the models that are used for these tasks.

3.2 Method

The first part of the analysis is conducted using principal component analysis and k-means clustering to visualize how the surveyed languages

are clustered based on valence orientation. The second part of the analysis uses random forests classifiers to extract the predictive power of valence orientation with regard to the family affiliation of the surveyed languages. The importance of different verb pairs is also evaluated. Comparing the results of these two methods is expected to provide an insight on the robustness of the hypothesis from Robert & Voisin (2018) presented in section 2.

Principal component analysis (PCA) is a technique used for unsupervised dimension reduction (Jolliffe, 2002). High dimensional data often include variables that are correlated and/or carry similar information. If the dataset is large, it is preferable to reduce it first before feeding it to other downstream tasks, thus the need of reducing the dimensions of the data. PCA fulfills this aim by using a mathematical procedure to transform a number of correlated variables into uncorrelated variables, which are called *principal components*. The first component accounts for as much of the variance in the data as possible. The embedded variance then decreases gradually in each of the following components. If only two components can explain most of the variance, the data size is substantially reduced, which is then very helpful for further processing. This method is widely used in areas such as image processing, genomic analysis, information retrieval, among others. Figure 4 shows an example of PCA with gene data. The original three-dimensional space (left) represents the similarities and dissimilarities when comparing three gene expressions across individuals. PCA is used to identify the two-dimensional plane that captures the highest variance of the data. The extracted two-dimensional space is then rotated and displayed as a two-dimensional space (right). The x-axis relates to the first component and the y-axis indicates the second component.

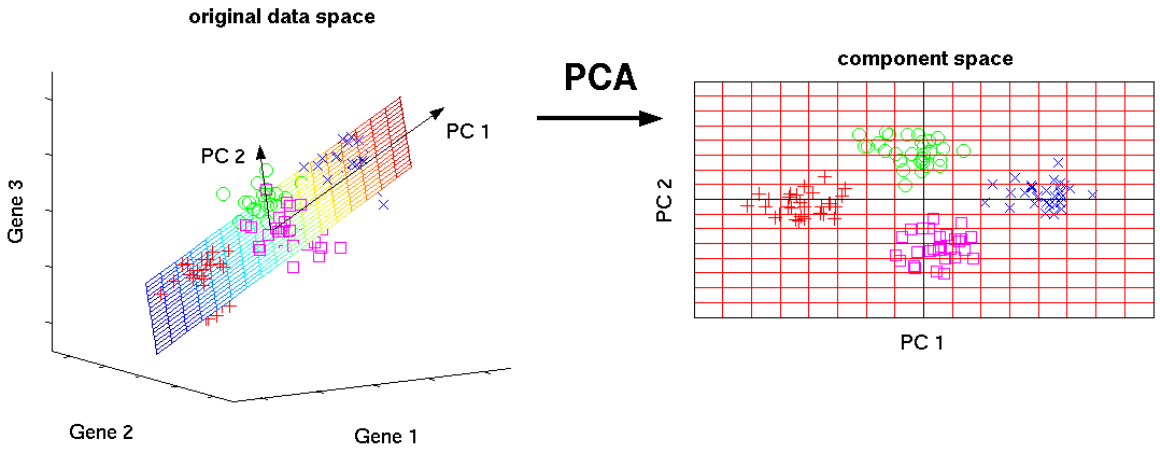


Figure 4. An example of PCA with data of gene expressions (Scholz, 2006: 16)

The extracted components can be used to cluster the data points, i.e., to find how many main groups can we find in the data. One of the most common clustering techniques is k-means clustering (Forgy, 1965; MacQueen, 1967; Hartigan & Wong, 1979; Lloyd, 1982), which is

commonly used on the output of PCA (Zha et al. 2002, Ding & He 2004). The clustering process is as follows: First, a *k* number of seed points are generated randomly within the investigated space. Second, each data point within the space is assigned to the nearest seed centroid, which represents a cluster. Third, new seed points are generated as the centroids of the current *k* clusters. Finally, the second to third step is repeated until the centroids do not change anymore, i.e., when the optimal centroids are found for each cluster.

In the first analysis, PCA is used to reduce the dimensionality of the data and compare the clusters generated by kmeans with the affiliation to the Atlantic, Mande, and Mel language families. One measure will be used for this comparison: the *Rand index*. The Rand index refers to the general accuracy of the model in clustering. It is obtained by dividing the total number of correctly retrieved tokens by the total number of retrieved tokens (Rand, 1971). This measure is used to compare the performance of the clustering when taking all verb pairs and only verb pairs 10-18. If the performance of the clustering rises when using only verb pairs 10-18, the results support the hypothesis that verb pairs 10-18 are more stable and provide more relevant information about the language families.

The second part of the analysis is based on classification tasks. A random forests classifier is used to extract the interaction of the variables and their relative importance within the data set. This classifier is based on binary recursive partitioning (Breiman, Friedman, Stone, & Olshen, 1984). Basically, the data in recursively partitioned binarily to form groups that are as homogeneous as possible. At each partitioning, the classifier uses a bootstrap sample of the original data and selects randomly a subset of the variables. A statistical test is carried out for each random sampling and the results are considered statistically significant if the proportion of the random samplings providing a test statistic greater than or equal to the one observed in the original data is smaller than the significance level. This process of random sampling is also the main strength of random forests, as it allows the analysis of small-scale data and consideration of the possible auto-correlation of variables (Tagliamonte & Baayen, 2012).

Two outputs of the random forests classifier are used: the decision tree and the conditional permutation of variable importance. On the one hand, the decision tree is expected to show the hierarchical interaction of the variables within the dataset. For instance, if both verb pairs 1 and 2 have a significant effect on distinguishing Atlantic languages from Mande and Mel languages. The decision tree will show which of the two pairs has a stronger predictive power when they are both considered. On the other hand, the relative importance of the predictors can be obtained by calculating the average difference between the estimate and the out-of-bag error without permutation. The larger the importance of a variable, the more predictive it is. By way of illustration, if the accuracy of the classifier drops the most when it does not take into account verb pair 1, this verb pair is considered to have the highest ranking within all the variables.

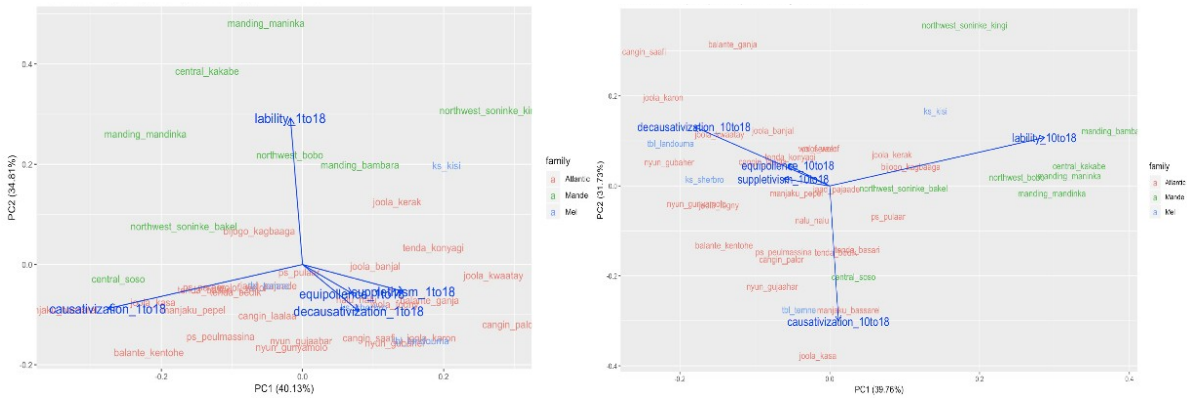
The performance of the random forests classifier is assessed with two measures, the *accuracy* and the *f-score*. One the one hand, the f-score evaluates the performance of the model in each category (i.e., language sub-branch). It is a combination of two other measures: precision and recall. Precision evaluates how many tokens are correct among all the output of the classifier, whereas recall quantifies how many tokens are correctly retrieved among all the expected correct output. The two measures evaluate the output from two different perspectives. These two measures are then combined into the f-score to interpret the overall performance of the classifier. The f-score is equal to the harmonic mean of the precision and recall, i.e. 2(recall × precision)/(recall + precision) (Ting, 2010). On the other hand, the accuracy provides an overview of the performance on the entire dataset. It is equal to all to the ratio of all the correctly retrieved tokens within the entire data. This value is expected to be used along with the *majority rule*. The majority rule relates to the biggest category in the dataset. Since most languages in our data are affiliated to the Atlantic sub-branch (68.4%, 26/38), the computational classifier could reach a precision of 68.4% just by labelling all the 38 languages as Atlantic languages. Thus, the valence orientation in the 18 verb pairs as explanatory variables should at least exceed the accuracy of 68.4% to be considered as having good discriminatory power.

4. Results (Experimental data)

This Section displays the performance of the clustering and classification tasks. The detailed comparison including the error analysis will be provided in Section 5.

4.1 Clustering

The output of the PCA with all the 18 verb pairs and only verb pairs 10-18 is visualized in Figure 5. Each point represents one of the 38 languages in the dataset. The distance between the languages reflects the similarities and dissimilarities of valence orientation strategies across the verb pairs. The more similar two languages are based on valence orientation, the closer they are in the two-dimensional space. The arrows indicate the influence of the five valence orientation strategies. For instance, tbl_landuma (in blue) is found at the extreme of the arrows of decausativization and equipollence. This is because it relies much more on the strategy of decausativization (6 on 18) and equipollence (4 on 18) than other languages. This visualization confirms what is seen in Figure 3: Atlantic (red) and Mel (blue) languages have similar strategies and they are overlapping in the PCA visualization. However, Mande languages (green) are mostly in the direction of the arrow related to lability, which means that Mande languages rely more on lability than the Atlantic and Mel languages.



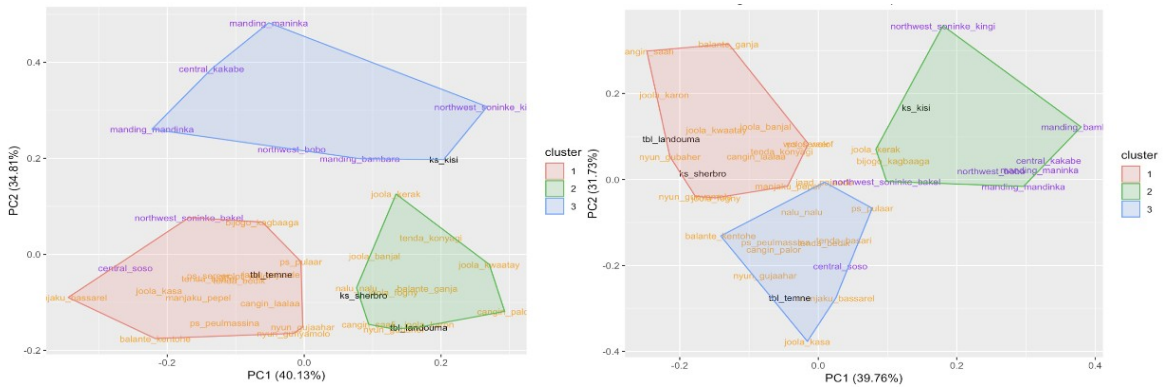
a. Verb pairs 1-18 b. Verb pairs 10-18

Figure 5. PCA visualization of valence orientation strategies in the dataset

Slightly different results are found by only taking verb pairs 10-18 (Figure 5b). On the one hand, the results show that decausativization seems to be a more Mel-type feature, but the small quantity of Mel languages (4) in the dataset limits the strength of this observation. On the other hand, Atlantic languages are divided in two main types: those that use causativization and those that mostly use decausativization combined with equipollence and suppletivism. Another benefit of using only the verb pairs 10-18 is that it allows decausativization to stand out from equipollence and suppletivism, as the three strategies has a similar effect size in Figure (5a). As a short summary, lability seems to be a Mandel-type feature, whereas Atlantic and Mel languages rely more on the other strategies, especially causativization and decausativization.

Then, the amount of components (amount of information) to keep when reducing the dimensionality is calculated. Reducing the dimensionalities means condensing information, which can result in loosing information. It is thus important to make sure that the amount of information is maintained reasonably when the dimensionality is reduced. One of the rules of thumb for selecting the number of principal components to retain in an analysis of this type is to pick the number of components that explain 85% or greater of the variation. In the current case, keeping four components for both 1-18 and 10-18 verb pairs is reasonable as 92.8% and 94.6% of the variation is kept within the first four components in both cases.

The k-means clustering can then be applied based on the extracted components. As mentioned in Section 3, $k = 3$ clusters are assumed since the languages belong to three different families (Atlantic, Mande, and Mel). The output of k-means clustering is shown in Figure 6. The results match better with the actual genealogical affiliation when only taking the verb pairs 10-18. As an example, the Mande languages (purple) are separated across clusters in 1-18, but are mostly clustered together with 10-18. In both runs, the Mel languages (black) are scattered across the three clusters, whereas the Atlantic languages are split in two clusters.



a. Verb pairs 1-18 b. Verb pairs 10-18

Figure 6. k-means clustering based on valence orientation strategies in the dataset

To evaluate statistically the performance of the two verb pair groups, the clusters generated by k-means are compared with the original genealogical clusters (Atlantic, Mande, and Mel). The measure is the Rand Index, which is used to compare clusters of the same size. The Rand index is defined as the number of pairs of objects that are either in the same group or in different groups in both partitions divided by the total number of pairs of objects. The Rand index lies between 0 and 1. When two partitions agree perfectly, the Rand index achieves the maximum value 1. A potential problem with Rand index is that the expected value of the Rand index between two random partitions is not a constant. This problem is corrected by the adjusted Rand index that assumes the generalized hyper-geometric distribution as the model of randomness. The adjusted Rand index has the maximum value 1, and its expected value is 0 in the case of random clusters. A larger adjusted Rand index means a higher agreement between two partitions. In this study, both the measures of Rand index and adjusted Rand index are shown in Table 4 to enhance the robustness of the comparison. In both the Rand-index and the adjusted-Rand-index, the score gets higher when only taking verb pairs 10 to 18. This supports the hypothesis that verb pairs 10 to 18 contain more relevant information to the genealogical affiliations of the languages we investigated.

Table 4. The performance of k-means clustering

	Verb pairs 1-18	Verb pairs 10-18
Rand index	0.634	0.744
Adjusted Rand index	0.273	0.486

As a short summary, the PCA shows that the Mande languages tend to use more lability to code causality, whereas Atlantic and Mel languages behave similarly with regard to causality-coding. Moreover, more regularities of valence orientation strategies are found in clustering when only considering verb pairs 10-18. These observations match with the hypothesis that verb pairs 10-18 encode more relevant information of valence orientation for language family identification.

4.2 Classification

Two main results are obtained via the classification task. On the one hand, the interaction of the variables in the entire dataset is visualized through the use the conditional inference trees. On the other hand, the individual importance of the variables is extracted by using the random forests classifier.

The first step provides an overview of what a classification tree looks like and how is the interaction of the variables if the entire dataset is analyzed without cross-validation. In other words, the entire dataset is used to generate the tree and assess its precision. While this has the risk of overfitting the data, the main purpose of this tree is to show a preliminary assessment of the interaction of the variables. The random forests classifier includes bootstrapping of the data. Thus, the use of conditional inference tree without cross-validation is considered appropriate. Figure 7 shows the conditional inference tree obtained via Monte Carlo simulations when including all 18 verb pairs and their value for the 38 languages as variables. The variables considered statistically significant by the classifier are displayed in the upper nodes (Node 1 and 2). These nodes divide the data into buckets (Node 3, 4, and 5). The bars in the buckets indicate the ratio of languages affiliated to each family. By way of illustration, Node 4 only includes Mel languages. In case of high performance, each bucket is expected to contain only tokens from the same category (i.e., languages from the same family). This is apparently the case: Node 3 represents Atlantic languages, Node 4 indicates Mel languages, and Node 5 mostly relates to Mande languages.

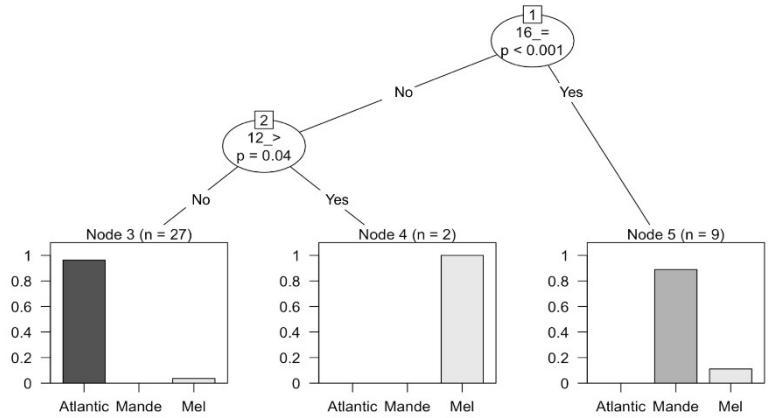


Figure 7. Conditional inference tree based on the entire dataset

The two variables included in the decision tree are 1) does verb pair 16 use lability, 2) does verb pair 12 use causativization. The tree can thus be read as follows: In a given language, if verb pair 16 uses lability as a strategy of valence orientation, it is very likely to be a Mande language. If the given language does not use lability in verb pair 16 but uses causativization with verb pair 12, then it is very likely to be a Mel language. If neither of the two preceding conditions is fulfilled, it is very likely to be an Atlantic language. However, we also observe that the effect of verb pair 16 and lability is statistically highly significant ($p <$

0.001), whereas the effect of verb pair 12 with causativization is only statistically significant ($p < 0.05$). The assessment of the predictions based on this tree is shown in Table 5. The f-score on the biggest families is high (0.981 for Atlantic and 0.941 for Mande). The overall accuracy of the model is also high. It is 0.947 (36/38), which exceeds by far the majority rule baseline of 0.684.

Table 5. The performance of the conditional inference tree

	Precision	Recall	F-score
Atlantic	0.963	1.000	0.981
Mande	0.889	1.000	0.941
Mel	1.00	0.500	0.667

Only two Mel languages are labelled incorrectly based on this tree. Kisi is wrongfully labelled as Mande, and Landuma is wrongfully labelled as Atlantic. This is not extremely surprising since these two languages were outliers in the PCA visualization. The results of the conditional inference tree thus show that the family affiliation of the 38 languages can be predicted with high accuracy based on the information of valence orientation in the 18 verb pairs. Moreover, the verb pairs 16 and 12 seem to be sufficient to predict the family affiliation of the languages, which matches with the hypothesis that verb pairs 10-18 include more relevant information. However, as mentioned at the beginning of the section, using the entire dataset makes the results subject to overfit. In other words, the generated tree may apply very well on the dataset in this study, but does not reach the same performance with other languages. This issue is covered by the random forests classifier.

In the second step, the random forest classifier is trained with the data of the 18 verb pairs and asked to predict the family affiliation of each language included in the dataset. The training and testing process are realized via 200 bootstrap samples of the dataset, which fulfills the similar function as cross-validation. The data is split randomly into subsets, on which decision trees are then generated. The individual importance of the variables (i.e., the 18 verb pairs) can thus be assessed via the conditional permutation-based variable importance. This process is expected to diminish the risk of overfit and provide a more faithful representation of the predictive power of the variables. If a variable is consistently helpful in predicting the family affiliation in most of the data subsets, it infers that this variable has a high importance for the classification task. First, the frequency and the mean of the minimal depth for each variable within all the 200 trees generated by the random forests are visualized. The minimal depth indicates how far is the node with a specific variable from the root node. As an example from Figure 7, lability in verb pair 16 is the root node, which equals to a minimal depth of zero. If a variable is frequently close to the root node, it is thus considered to have a high importance. The minimal depth of the top ten most important variable is shown in Figure 8. The majority of the variables are from verb pairs 10-18, which once more matches with the hypothesis that the valence orientation in verb pairs 10-18 are more

relevant to identify the family affiliation of the Atlantic, Mande, and Mel languages.

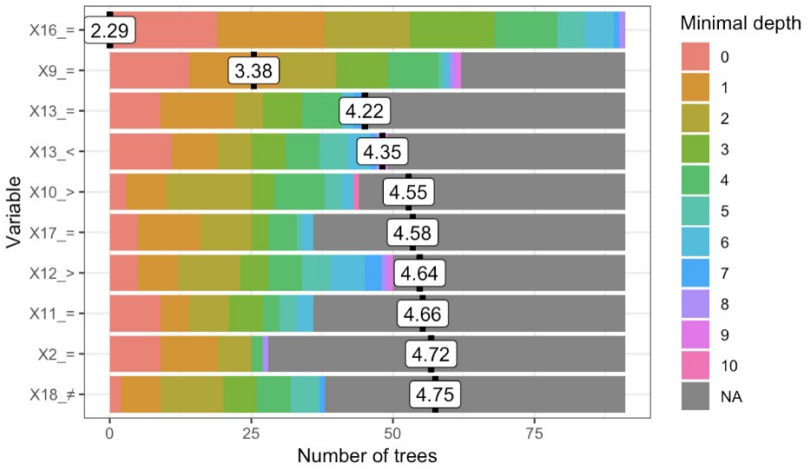


Figure 8. Distribution of minimal depth and its mean

Similar results are found when using other measures. In Figure 9, 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 worse the model performs without each variable. A high decrease infers 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 and the end of the tree. A high decrease of Gini coefficient when removing a variable infers that this variable has a strong predictive power and therefore a high importance. In both measures, the top ten ranked variables are mostly from verb pairs 10-18. However, verb pair 9 and 2 are also consistently included in the top ten rankings. A more detailed analysis of the verb pairs is provided in Section 5.

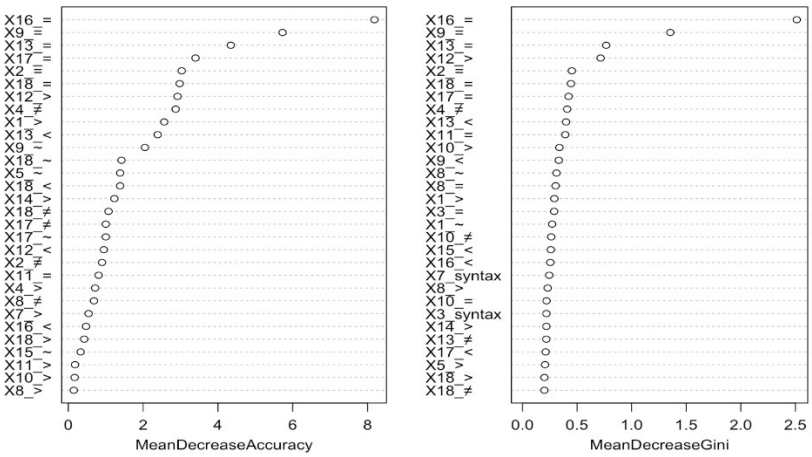


Figure 9. Accuracy and purity of the nodes

Finally, the performance of the random forest classifier is evaluated according to its f-score and accuracy. The precision, recall, and f-score of the random forests classifier are shown in Table 6. A drop in all three

measures is found, especially for the Mel family, where the model does not perform well. A look at the data indicates that the classifier tends to classify languages as either Atlantic and Mel. This may be due to the small data size of the Mel languages. This drop of performance is also found in the measure of accuracy. The accuracy of the model is 0.842 (32/38). As mentioned previously, this drop of accuracy is expected since the data was randomly split to avoid the risk of overfit. Moreover, the accuracy is still above the majority rule baseline of 0.684. Thus, the classifier is still considered performant.

Table 6. The performance of the random forests

	Precision	Recall	F-score
Atlantic	0.813	1.000	0.897
Mande	1.000	0.750	0.857
Mel	0.000	0.000	0.000

Finally, the interaction of the variables in random forests is also visualized. A representative tree is generated based on the d2 metric defined in Banerjee et al (2012). The average distance D(T) of each tree in the set of trees in computed, and trees with the lowest D(T) value are extracted and formatted to be compatible with the tree class. Due to the randomization and cross-validation process in the random forests, the representative tree generated with this method does not fully represent the variation of the data, but it allows the visualization of the interaction of the variables. The representative tree from the random forests data is shown in Figure 10. The tree can be read as follows: if the branch goes right it means TRUE, if the branch goes left it means FALSE. For instance, if starting from the root, if a language uses lability (=) for verb pair 13 and does not use equipollence (~) for verb pair 5, it is very likely to be a Mande language. In the data, six languages use lability for verb pair 13 (Kagbaaga, Soninke_Kingi, Soninke_Bakel, Kakabe, Mandinka, and Bambara). Among these six languages, only Kagbaaga uses equipollence and is correctly identified as belonging to the Atlantic family, whereas the other five languages are also correctly affiliated to the Mande family.

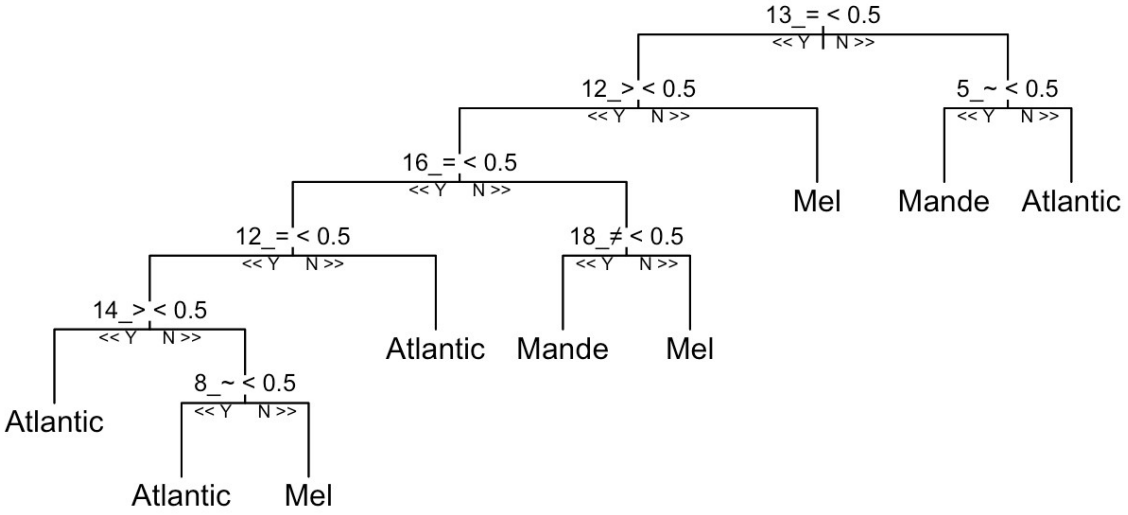


Figure 10. Representative tree of the random forests.

As a summary, the results of the conditional inference tree and the random forests classifiers both show that the family affiliation of the 38 languages could be predicted with high accuracy based on the information of valence orientation in the 18 verb pairs. Furthermore, verb pairs 10-18 tends to be highly ranked in terms of importance for this classification task. A qualitative analysis of the results is provided in the following section.

5. **Discussion**

The variables that were considered important by the random forests are listed in Table 7, which combines the top ten variables when using the three different measures of 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 seven of the ten variables are found in all three rankings: X16_=, X9_=, X13_=, X13_<, X17_=, X12_>, X2_=.

Table 7. An overview of the importance of the variables. The shaded cells represent variables that are present in all the three types of ranking.

	Minimal depth ranking	Mean Decrease Accuracy	Mean decrease Gini
1	X16_=	X16_=	X16_=
2	X9_=	X9_=	X9_=
3	X13_=	X13_=	X13_=
4	X13_<	X17_=	X12_>
5	X10_>	X2_=	X2_=
6	X17_=	X18_=	X18_=
7	X12_>	X12_>	X17_=
8	X11_=	X4_≠	X4_≠

9	X2_=	X1_>	X13_<
10	X18_≠	X13_<	X11_=

Among these seven variables, five of them belong to the 10-18 verb pairs (X16_=, X13_=, X13_<, X17_=, X12_>). Thus, verb pairs 10-18 enclose more relevant information on clustering the languages from the Atlantic, Mande, and Mel families. This observation supports the hypothesis that languages share universal tendencies between verb pairs 1-9 and only the verb pairs 10-18 display variance across language sub-groups. The occurrence of verb pairs 2 and 9 in the ranking can be explained from a linguistic point of view. Verb pair 2 (*kill vs die*) is expected to differentiate between the Mande languages using lability and the Atlantic and Mel languages using the suppletivism strategy. The lexicalization of frequent actions such as *kill* and *die* is a universal tendency, from which Mande languages diverge due to the use of lability. Verb pair 9 (*hide vs conceal*) may be used with a non-human subject in the non-causal form, which makes it more similar to verb pairs 10-18 than verb pairs 1-9.

Furthermore, five of the seven verb pairs are related to lability (X16_=, X9_=, X13_=, X17_=, X2_=). We speculate that this is due to the fact that lability helps to distinguish the Mande languages from the Atlantic and Mel languages. Except verb pair 2, the verb pairs involved must show specific-family trends. As a result, directed strategies (causativization and decausativization) should be favored in Atlantic and Mel languages. Conversely, the Mande languages should use them to a lesser extent and give priority to the strategy of lability.

6. Conclusion

The overarching theme of the analysis was to assess quantitatively the use of valence orientation as a typological paramenter of languages. The two main aims were to 1) assess statistically the regularities of valence orientation in the languages of the Atlantic, Mande, and Mel sub-branches of the Niger-Congo phylum spoken in West Africa, 2) to explore the variation of valence orientation between the verb pairs 1-9 and 10-18.

With regard to the first aim, the analysis shows that valence orientation has a strong predictive power on the affiliation of languages of the three sub-branches. Languages in the Mande family rely more on lability, while the Mel languages have a stronger tendency to use decausativization than the Atlantic languages. These results match with the hypothesis of Nichols & al. (2004) that ”high morphological complexity favors decausativization”. However, the tendency of Atlantic and Mel languages to favor directed strategies does not match with the trends proposed by Creissels (2018) for the sub-Saharan languages. The error analysis of the classifier shows that the errors can be mostly attributed to language contact (Kisi in Mande) or internal changes (Landuma) affecting the valence strategies of individual languages. Further investigation in other language groups of the area are thus required to identify areal tendencies and explore further the effect of contact. For the second aim, the results show that the verb pairs 10-18

are more relevant for differentiating the languages of the three sub-branches. However, the verb pair 2 and 9 are also ranked as important by the classifier. The results thus partially match with the hypotheses from previous studies, but for both aims, the current analysis provides additional insights by pinpointing which verb pairs are more relevant to identify the language affiliation.

Abbreviations

CAUS: causative voice marker; MID: middle voice marker

Acknowledgements

Our greatest thanks go to the following researchers for helping us generously to complement our data or analyses on various languages: Sokhna Bao Diop (Baynunk Guñaamolo), Alain-Christian Bassène (Jóola Banjal), Tucker Childs (Kisi and Sherbro), El Hadji Dièye (Laalaa), Gérard Dumestre (Bambara), Dame Ndao (Pepel), Pierre Sambou (Jóola Karon), Kirk Rogers (Landuma), Amadou Sow (Pulaar). The first author is thankful for the support of the IDEXLYON Fellowship grant (16-IDEX-0005).

References

Banerjee, M., Ding, Y., & Noone, A-M. (2012). Identifying representative trees from ensembles. *Statistics in Medicine*, *31*(15), 1601-1616.

Bickel, Balthasar. (2015). Distributional typology: statistical inquiries into the dynamics of linguistic diversity. In H. Bernd & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 901–923). Oxford: Oxford University Press.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). *Classification and regression trees*. Taylor & Francis, New York: Chapman & Hall.

Childs, George Tucker. (2004). The Atlantic and Mande groups of Niger-Congo: A study in contrasts, a study in interaction. *Journal of West African Languages* 30(2), 29-40.

Childs, G. T. (2010). The Mande and Atlantic groups of Niger-Congo: prolonged contact with asymmetrical consequences. *Journal of Language Contact*, 3(1), 15–46.

Creissels, D. (2014). Le développement d'un marqueur de déplacement centripète en mandinka. In C. de Féral (Ed.), *In and Out of Africa, Languages in Question, In Honor of Robert Nicolaï* (pp. 95–102). Louvain-la-Neuve / Walpole MA: Peeters.

Creissels, Denis. (2017). Copulas originating from the imperative of 'see/look' verbs in Mande languages. In B. Walter & A. Malchukov (Eds.), *Unity and diversity in grammaticalization scenarios* (pp. 45–66). Berlin: Language Science Press.

Creissels, Denis. (2018). The noncausal-causal alternation in the languages of Subsaharan Africa: a preliminary survey of noncausal-causal pairs involving inanimate undergoers. 51th annual meeting of the Societas Linguistica Europaea, Tallin,. Workshop Valence

Orientation in Contact. http://www.deniscreissels.fr/public/Creissels-noncausal_causal_alternation.pdf.
Dalby, D. (1965). The Mel languages: A reclassification of southern 'West Atlantic'. *African language studies 6*, 1–17.
Dasgupta, A. (2014). reprtree: Representative trees from ensembles. R package version 0.6.
Ding, C., & He, X. (2004). K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on Machine learning*, 225-232.
Dimmendaal, Gerrit J. (2008). Language Ecology and Linguistic Diversity on the African Continent. *Language and Linguistics Compass*, *2*(5), 840–858.
Dimmendaal, Gerrit Jan. (2011). *Historical Linguistics and the Comparative Study of African Languages*. Amsterdam, Philadelphia: John Benjamins.
Dowle, M., & Srinivasan, A. (2019). data.table: Extension of data.frame. R package version 1.12.2. https://CRAN.R-project.org/package=data.table.
Dwyer, D. 2005. The Mende Problem. In B. Koen & J. Maniacky (Eds.), *Studies in African Comparative Linguistics with Special Focus on Bantu and Mande* (pp. 29–42). Tervuren: Royal Museum for Central Africa.
Eddelbuettel, D. (2017). random: True random numbers using random.org. R package version 0.2.6. https://CRAN.R-project.org/package=random.
Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, *21*, 768-769.
Greenberg, J. H. (1963). *The languages of Africa*. Bloomington: Indiana University.
Güldemann, T. 2018. Historical linguistics and genealogical language classification in Africa. In T. Güldemann (Ed.), *The Languages and Linguistics of Africa* (pp. 58–444). Berlin, Boston: De Gruyter Mouton.
Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, *28*, 100-108.
Haspelmath, M. (1993). More on the typology of inchoative/causative verb alternation. In B. Comrie & M. Polinsky (Eds.), *Causatives and transitivity* (pp. 87–120). Amsterdam, Philadelphia: John Benjamins.
Haspelmath, M., Calude, A., Spagnol, M., Narrog, H., & Bamyaci, E. (2014). Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. *Journal of Linguistics* , *50*(3), 587–625.
Hothorn, T., Hornik, K., & Zeileis, Achim. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674.
Jolliffe, I. (2002). *Principal component analysis*. Boston, MA: Springer.
Kassambara, A., & Mundt, F. (2017). factoextra: Extract and visualize the results of multivariate data Analyses. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18-22.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*, 128-137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297). Berkeley, CA: University of California Press.

Matras, Y. 2009. *Language contact*. Cambridge: Cambridge University Press.

Matras, Y. 2010. Contact, convergence, and typology. In R. Hickey (Ed.), *The handbook of language contact* (pp. 66-85). Oxford: Wiley-Blackwell.

Mouselimis, L. (2019). ClusterR: Gaussian mixture models, k-means, mini-batch-kmeans, k-medoids and affinity propagation clustering. R package version 1.1.9. https://CRAN.R-project.org/package=Cluster.

Nichols, J., Peterson, D. A., & Barnes, J. (2004). Transitivizing and detransitivizing languages. *Linguistic Typology*, *8*, 149-211.

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

Paradis, E., & Schliep, K. (2018). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526-528.

Pozdniakov, K., & Segerer, G. (to appear). A Genealogical classification of Atlantic languages. In L. Friederike (Ed.), *The Oxford guide to the Atlantic languages of West Africa*. Oxford, New York: Oxford University Press.

Pozdniakov, K., Segerer, G., & Vydrine, V. (2019). Mande-Atlantic Contacts. *Oxford Research Encyclopedia of Linguistics*, Oxford University Press. Online Publication Date: May 2019 DOI:10.1093/acrefore/9780199384655.013.39.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846-850.

R-Core-Team. (2018). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Robert, S., & Voisin, S. (2018Comparing causal-noncausal alternation in three West-African families in contact: Atlantic, Mel and Mande. 51th annual meeting of the Societas Linguistica Europaea, Tallin. Workshop Valence Orientation in Contact.

Scholz, M. (2006). *Approaches to analyse and interpret biological profile data*. PhD dissertation. Postdam: University of Postdam.

Segerer, G., & Flavier, S. (2011-2018). *RefLex: Reference Lexicon of Africa*. http://reflex.cnrs.fr/.

Tadmor, U., Haspelmath, M., & Taylor, B. (2010). Borrowability and the nation of basic vocabulary. *Diachronica*, *27*(2), 226-245.

Tang, Y., Horikoshi, M., & Li, W. (2016). ggfortify: Unified interface to visualize statistical result of Popular R Packages. *The R Journal*, *8*(2), 478-489.

Tagliamonte, S. A., & Baayen, H. (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.
Ting, K. M. (2010). Precision and Recall. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 781). Boston, MA: Springer US.
Voisin, Sylvie. (forthcoming). Associated Motion and Deictic Directional in Atlantic languages. In G. Antoine & H. Koch (Eds.), *Associated motion*. Berlin, New York: De Gruyter Mouton.
Vydrin, V., & Vydrina, A. (2010). Impact of the Pular on the Kakabe language (Futa Jallon, Guinea). *Journal of Language Contact*, *3*(1), 86–106.
Vydrin, V. (2016). Toward a Proto-Mande reconstruction and an etymological dictionary. In K. Pozdniakov (Ed.), *Comparatisme et reconstruction : tendances actuelles* (pp. 109–123). Le Mans: Faits de Langues.
Wickham, H. (2017). tidyverse: Easily install and load the tidyverse. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
Zha, H., Ding, C., Gu, M., He, X., & Simon, H. (2002). Spectral relaxation for K-means clustering. *Advances in Neural Information Processing Systems 14 (NIPS'01)*, 1057-1064.

Appendix 1

The 18 pairs of verbs included in the survey and their proxies (Nichols et al. 2004: 186)

	Non-causative	Causative	Proxies
1	laugh	make laugh, amuse	cry
2	die	kill	
3	sit	seat, have sit, make sit	lie down, go to bed, put to bed
4	eat	feed, give food	drink, give to drink
5	learn, know	teach	understand, find out, grasp
6	see	show	
7	be/become angry	anger, make angry	annoy(ed)
8	fear, be afraid	frighten, scare	
9	hide, go into hiding	hide, conceal	
10	(come to) boil	(bring to) boil	cook
11	burn, catch fire	burn, set fire	be aflame; char
12	break	break	split, shatter, smash
13	open	open	close
14	dry	make dry	wet, clean; black, white
15	be/become straight	straighten, make straight	crooked, long, round, flat
16	hang	hang (up)	lean (incline), extend, project
17	turn over	turn over	turn, rotate, roll, shake
18	fall	drop, let fall	fall down, fall over, etc.; sink

Appendix 2

List of languages and an overview of their valence orientation across verb pairs. The percentages indicate the ratio of each strategy within each language. The columns indicate the five strategies for valence orientation: “>” is causativization, “≠” is suppletivism, “<” is decausativization, “~” is equipollence, and “=” is lability. The first part of the language code refers to the sub-group of the language. The second part of the language code refers to the name of the language.

Languages	>	≠	<	~	=
Wolof_Wolof	50%	11%	22%	6%	11%
Nyun_Gunyamolo	47%	18%	29%	6%	0%
Nyun_Gubaher	33%	25%	33%	8%	0%
Nyun_Gujaahar	46%	18%	18%	18%	0%
Tenda_Bedik	55%	27%	9%	0%	9%
Tenda_Basari	56%	11%	22%	0%	11%
Tenda_Konyagi	25%	33%	25%	0%	17%
Jaad_Pajaade	50%	30%	10%	0%	10%
Ps_Pulaar	44%	28%	0%	17%	11%
Ps_Sereer	56%	11%	17%	6%	11%
Ps_Peulmassina	58%	25%	0%	17%	0%
Cangin_Laalaa	50%	22%	22%	0%	6%
Cangin_Palor	22%	56%	11%	11%	0%
Cangin_Saafi	40%	40%	20%	0%	0%
Nalu_Nalu	40%	40%	13%	0%	7%
Balante_Ganja	24%	12%	18%	41%	6%
Balante_Kentohe	67%	11%	22%	0%	0%
Joola_Fogny	33%	33%	27%	0%	7%
Joola_Kasa	64%	14%	0%	14%	7%
Joola_Kerak	25%	19%	19%	13%	25%
Joola_Banjal	31%	19%	13%	25%	13%
Joola_Kwaatay	18%	36%	9%	27%	9%
Joola_Karon	29%	29%	29%	14%	0%
Manjaku_Pepel	60%	20%	13%	0%	7%
Manjaku_Bassarel	77%	8%	8%	0%	8%
Bijogo_Kagbaaga	47%	20%	7%	7%	20%
Northwest_Soninke_Kingi	8%	25%	25%	0%	42%
Northwest_Soninke_Bakel	56%	11%	11%	0%	22%
Northwest_Bobo	40%	27%	0%	0%	33%
Central_Soso	67%	7%	13%	0%	13%
Central_Kakabe	44%	6%	0%	0%	50%
Manding_Maninka	33%	8%	0%	0%	58%
Manding_Mandinka	56%	6%	0%	0%	39%
Manding_Bambara	31%	31%	6%	0%	31%
Ks_Kisi	19%	31%	19%	0%	31%
Ks_Sherbro	38%	31%	25%	0%	6%
Tbl_Temne	50%	30%	10%	0%	10%
Tbl_Landouma	28%	17%	33%	22%	0%