Universal and lineage-specific trends of valence orientation: Quantitative testing in three Niger-Congo language families

Marc Allassonnière-Tang, Stéphane Robert, Sylvie Voisin

The noncausal/causal alternation and genealogical affiliation: Quantitative testing in three Niger-Congo language families

Marc Allassonnière-Tang*, Stéphane Robert**, Sylvie Voisin***
*EA Eco-Anthropologie, MNHN/CNRS/University of Paris
** LLACAN Langage langues et cultures d'Afrique, CNRS INALCO & EPHE
*** Aix Marseille University, DDL, CNRS

Abstract
The noncausal/causal alternation is the pairing of two verb forms that refer to the same core event but differ in the absence vs. presence of a causer for this event (e.g. *rise* vs. *raise, open* (intr.) vs. *open* (tr.), *die* vs. *kill*). Languages differ in their overall preferences among the possible strategies for coding this alternation. This study uses machine-learning methods (clustering and tree-based computational classifiers) to investigate the predictive power of the noncausal/causal alternation for the genealogical affiliation of 38 languages belonging to the Atlantic, Mande and Mel families. The languages studied here belong to different contact areas in Senegal and its surroundings. The three families are all affiliated to the Niger-Congo phylum but display quite different typological profiles. The present paper elaborates on an earlier study that used a standard list of 18 verb pairs to establish the coding strategies in these languages. Apart from highlighting which coding strategies are favored in each family, our quantitative analyses show that the family affiliation of the 38 languages can be predicted with an accuracy above the majority baseline based on the information of the noncausal/causal alternation in the 18 verb pairs, but that the predictive power of verb pairs 1-9 is generally lower than of verb pairs 10-18. Our results confirm the hypothesis that the first group of verb pairs shows universal rather than lineage-specific tendencies concerning the noncausal/causal alternation. Furthermore, our analyses identify which of the 18 verb pairs (and their correlated coding strategies) have the highest predictive power. This study opens new avenues for identifying the relevant synchronic data for genealogical classification in historical linguistics. Future studies could replicate the same analysis in different language families to assess if our results are universal or specific to some language families.

Keywords
valency orientation, noncausal/causal alternation, genealogical affiliation, Atlantic, Mande, Mel, clustering, decision trees

L'alternance non causal/causal et l'affiliation généalogique : Analyses quantitatives dans trois familles de langues Niger-Congo

Résumé

L'alternance noncausal/causal désigne l'appariement de deux formes verbales référant à un même événement mais se distinguant par l'absence vs. présence d'un causateur de l'événement (e.g. *rise* vs. *raise, open* (intr.) vs. *open* (tr.) ou *die* vs. *kill*). Les langues emploient diverses stratégies pour encoder cette opposition. Cette étude utilise des méthodes d'apprentissage-machine (clustering et classificateurs computationnels fondés sur des arbres de décision) afin d'analyser le pouvoir prédictif des stratégies d'encodage sur l'affiliation généalogique de 38 langues appartenant aux familles Atlantique, Mandé et Mel. Les langues de l'enquête sont essentiellement localisées au Sénégal et dans ses environs où existent des zones de contact. Ces trois familles relèvent du même phylum Niger-Congo mais présentent des profils typologiques assez différents. Le traitement de l'alternance causale comme marqueur généalogique est traité ici à partir d'une étude antérieure utilisant une liste standardisée de 18 paires de verbes pour établir les stratégies employées dans ces langues. En plus d'indiquer quelles stratégies sont favorisées dans chaque famille, nos analyses quantitatives montrent que l'affiliation généalogique des 38 langues peut être prédite avec une grande précision à partir des stratégies de codage de l'alternance causale utilisées pour les 18 paires de verbes, mais que le pouvoir prédictif des paires de verbes 1-9 est généralement inférieur à celui des paires de verbes 10-18. Ces résultats confirment l'hypothèse selon laquelle le premier groupe de verbes montre des tendances universelles plutôt qu'une variation interlinguistique dans le marquage de l'alternance noncausal/causal. De plus, notre analyse identifie pour ces familles de langues quelles paires de verbes spécifiques parmi les 18 sélectionnées ont le pouvoir prédictif le plus élevé. Cette approche ouvre de nouvelles voies pour résoudre un problème fondamental de la linguistique historique, celui des filtres nécessaires pour trier les données synchroniques pertinentes pour la classification généalogique. Des études futures pourraient reproduire la même analyse sur différentes familles de langues pour évaluer si ces résultats sont universels ou spécifiques à certaines familles de langues.

Mots-clés

orientation de valence, alternance noncausal/causal, affiliation généalogique, atlantique, mandé, mel, clustering, arbres de décision

1. Introduction

The noncausal (nC)/causal (C) alternation is defined here as a semantic distinction based on the absence/presence of a causer in a pair of verbs referring to the same core event or state-of-affairs, e.g. *die* vs. *kill* and *rise* vs. *raise* in English (see Introduction of this issue). The overall tendency of a language to code the members of noncausal/causal verb alternations by specific morphological means has been claimed by Nichols et al. (2004) to reflect a general typological parameter defining transitivizing/detransitivizing languages and called 'valence orientation'. Haspelmath (1993) identifies five morphological[1] strategies attested cross-linguistically for the coding of this alternation. They are listed and exemplified in Table 1.

Table 1 — The different coding strategies exemplified in Wolof (Atlantic) (examples from Diouf 2003)

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

Table 1 shows that all strategies are found in Wolof (Atlantic). Causativization refers to pairs in which the causative meaning is generated by expanding the noncausal form of the verb, as in *réer* (be_lost) 'be lost' and *réer-al* (be_lost-CAUS) 'lose'. 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) 'be sealed' from *sakk* 'seal'. Lability applies to a noncausal/causal pair involving no formal change, like *lakk* 'burn (intr.)' and *lakk* 'burn (tr.)'. Suppletion involves two distinct verbal lexemes paired in a noncausal/causal alternation, like *dee* 'die' and *rey* 'kill'. Finally, for the equipollent strategy, the causal and noncausal meanings are generated from the same root with two different forms displaying an equivalent morphological complexity so that none of the two forms can be analyzed as derived from the other, as in *daan-al* (knock_down-CAUS) 'drop, fell' vs. *daan-u* (knock_down-MID) 'fall'. The noncausal 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 verb *daan*. Typologically, the equipollent strategy can be achieved through derivational (as in Wolof) marking or inflectional class alternation, but also through formal means involving non-concatenative morphology, such as a phonological alternation in the root (e.g. Ablaut as in *fall/fell*) or a tonal change. These subtypes are fused under the same label of equipollence in this study.

[1] Periphrastic (e.g. English *laugh/make laugh*) and, more largely, morphosyntactic strategies have been excluded from this study. This choice was supported by the typological characteristics of these three families in which the morphological oppositions seem particularly relevant. Only some rare cases of periphrastic causatives have been found, such as the use of a verb 'make' (verb root *kaan*) in Jóola Keeraak to form the causative of 'laugh' (verb root *ɬu)*, as in *a-kaan-ɔm-mɪ mɪn ɪ-ɬʊ* (SBJ.3SG-make-OBJ.1SG-COMPL so_that SBJ.1SG-laugh) 'he made me laugh' lit. 'he made me so that I laugh' (S. Robert's fieldwork data). For the same reason, we follow Haspelmath's (1993) labels for strategies (only "decausative" is preferred to "anticausative"): these are more suited to the languages under study.

Languages of the world differ in their coding preferences and, more generally, in the proportion of use (i.e. the relative rate of use) of each strategy. Therefore, the language coding profile for the noncausal/causal alternation is one of the many linguistic features that could be used to identify the different genealogical affiliations of languages (see Grünthal & Nichols 2016).[2] Some lexical categories are more likely to undergo borrowing while others tend to be more stable, e.g. nouns tend to be more easily borrowed than adjectives or verbs (Tadmor & Haspelmath 2010). Under such an assumption, the verbal domain should be more adequate for identifying the genealogical affiliation of languages. That is to say, languages from the same family are more likely to use the same or cognate verb forms, while the same analysis is more difficult to conduct on nouns, since these are easily borrowed across languages. Although the study of the noncausal/causal alternation is primarily concerned with morphological devices, verbal roots play an important role, on the one hand because they are directly involved in two strategies, i.e. lability and suppletion, and on the other hand because the semantics of the base verb (noncausal vs. causal) conditions the orientation of two other possible strategies, i.e. causativization and decausativization. Thus, a bare verb stem with causal meaning can make its non-causal counterpart by suppletion or by decausative derivation. Regarding morphology, it should be noted that in the sample used for this study, we did not identify any borrowing of derivational suffixes from one family to another. Moreover, in their study on contact phenomena (see below), Voisin and Robert (2018) identified only a few cases of contact-induced change for the noncausal/causal alternation.

Contact and other phenomena have been investigated from different lexical and grammatical perspectives in the Atlantic, Mande and Mel families (Creissels 2014; Robert & Voisin 2018; Voisin 2021). On the one hand, lexical borrowings are indeed frequent between the Mande and Atlantic languages (Pozdniakov, Segerer & Vydrin 2019). As an example, linguistic divergences of languages belonging to the same family, such as Mandinka and Maninka (Mande), can be attributed to different contact scenarios 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.[3] On the other hand, 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, exhibit isolating morphology, display a limited inventory of verbal affixes and have a strict SOV(X) order. By contrast, Atlantic and Mel languages have a noun class system, an SVO word order, and an agglutinative morphology characterized by a remarkably rich system of verbal derivation, which allows us to assume a wider use of derivational strategies.

Among the verbal morphosyntactic features that can be used in comparative analyses (Matras 2009; Matras 2010), valence orientation has been studied with various approaches in different areas of the world (Nichols et al. 2004; Haspelmath et al. 2014; Bickel 2015; Robert & Voisin 2018). More specifically, this domain has

[2] Actually, on a sample of language families of northern Eurasia, Grünthal & Nichols (2016) show that the coding patterns for noncausal/causal alternation combined with additional information provide NeighborNet trees which approximate well the known phylogeny of the family and also help to uncover language-family history.

[3] 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 on the precise dates of its beginning and end.

recently been studied in light of the genealogical affiliation of languages spoken in Africa. Creissels (this issue) investigates more largely the cross-linguistic variation in the coding of the noncausal/causal alternation in 30 Sub-Saharan languages belonging to 15 different genealogical units. He uses a specific list of 13 verb pairs whose noncausal member is a monovalent verb referring to a process typically undergone by concrete inanimate entities. These verbs correspond to the inchoative type we discuss below. The results indicate that several Mande languages show an extreme degree of preference for lability. For example, Bambara, Kakabe and Mano have a proportion of labile pairs as high as in languages such as English (between 10 and 12 out of 13) (Creissels this issue: Appendix1). No language in the sample displays an extreme degree of preference for causativization, as expected from previous studies (Nichols et al. 2004).

In another recent study by Robert & Voisin (2018), a comparison between 36 Atlantic languages, 8 Mande languages and 7 Mel languages was made by extracting the general pattern of distribution of the five main coding strategies across the set of 18 verb pairs (see Appendix) defined by Nichols et al. (2004). The study aimed at defining the family profiles in coding the noncausal/causal alternation (i.e. the relative rate of use of the different coding strategies in each family) and at tackling contact-induced phenomena through deviance of individual languages from their family profile. Hence the focus on the area where these three families are in contact. The results show that Atlantic and Mel languages share a preference for causativization or directed strategies (causativization and decausativization), whereas Mande languages combine a strong propensity for lability with a prevalence of causative coding. While these results contribute to the definition of the coding profile of the three families for noncausal/causal alternation and to the discussion of contact between Atlantic, Mande and Mel languages, the methods used were mostly qualitative in nature and have not been yet published. Furthermore, the analysis included all the 18 verb pairs from Nichols et al. (2004). In this list, pairs 1 to 9 actually correspond to verb types that are attested to universally favor the causative strategy (i.e. dynamic verbs using prototypically an animate subject for the noncausal member of the pair) (Haspelmath et al. 2014). This may have introduced a bias for causativization in the languages' coding profiles. On the other hand, pairs 10 to 18 roughly correspond to inchoative verbs, which are considered to reveal the actual preferences of languages for coding valence alternation (Haspelmath 1993). This factor was considered but not investigated quantitatively in this previous study. Considering these two factors, one can assume that verb pairs 10-18 have greater power for predicting genealogical membership of a language than verb pairs 1-9 which exhibit a universal bias toward causativization.

A precision must be made on the term "inchoative" that we shall retain for the sake of convenience in this article. According to Haspelmath (1993: 90), inchoative verbs generally refer to a change of state by excluding a causing agent and by presenting the situation as occurring spontaneously. For example, *The stick broke* (inchoative) vs. *The girl broke the stick* (causative). First of all, as Haspelmath himself points out (*ibid*.: 108), the term is not very felicitous, as it should not be understood here with the aspectual value that is usually attributed to it. Second, we prefer to follow Creissels' (this issue) slightly different characterization of this particular type of verbs, namely monovalent verbs which refer to a process typically undergone by concrete animate entities irrespective of their willingness.

Elaborating on Robert & Voisin (2018), this paper aims at (i) quantifying the predictive power of the noncausal/causal alternation for identifying the genealogical affiliations of the Atlantic, Mande and Mel languages, (ii) comparing the predictive power of the verb pairs 1-9 and 10-18 for predicting the genealogical affiliation of the Atlantic, Mande and Mel languages. Moreover, the results of these quantitative analyses will allow to shed light on the hypothesis of a possible correlation between the typological profiles of these three language families and their coding profiles for the noncausal/causal alternation.

This article is structured as follows. The languages, data and results of the previous investigation on the noncausal/causal alternation in the three families (using the 18 verb pairs) are presented in Section 2. Based on this material, two quantitative analyses using machine learning methods are then conducted to investigate the predictive power of the noncausal/causal alternation for genealogical affiliation (Section 3). First, a principal component analysis combined with k-means clustering is used to cluster the languages of the data according to their coding strategies for the noncausal/causal alternation and to compare the obtained clusters with the original families (Section 3.1). Second, the predictive power of the noncausal/causal alternation for predicting the genealogical affiliation of the languages is investigated by feeding the data to a decision-tree-based classifier (Section 3.2) and by using information gain (Section 3.3). The decision tree is used to find out which combinations of coding strategies and verb pairs are statistically significant for predicting the genealogical affiliation of a language in the sample. The information gain is used to extract a ranking of all verb pairs when it comes to predicting the genealogical affiliation of the languages. Finally, Section 4 discusses the overall results before the final conclusion on the noncausal/causal alternation as a genealogical marker (Section 5).

2. Data on the noncausal/causal alternation in the three families

In this section, an explanation is first provided as to how languages were selected to represent the Atlantic, Mande and Mel families, and how they were investigated. Then, a comparative overview is provided about the coding strategies associated with each verb pair across the language sample.

In terms of size, the Mande family has around 70 languages[4]. The Atlantic family has approximately 50 languages. The Mel family is the smallest family of the three, with only a dozen languages, some of which are already extinct (e.g. Bom). It is also the least documented family. The sample of languages included in our study does not perfectly reflect the distribution and diversity of the Atlantic, Mande and Mel languages. However, following Robert and Voisin (2018) and the reasons presented in Section 1, we only considered the languages from the three families that are spoken in the same region (in and around Senegal). Furthermore, among these languages, we had to discard those that are not well documented. In total, 26 Atlantic languages (68%), 8 Mande languages (21%) and 4 (11%) Mel languages have been extracted.[5] A map of

[4] These approximate numbers are due to an insufficient documentation to reliably distinguish languages and dialects or variants in some cases. The current numbers are mostly based on data from Glottolog (Hammarström et al. 2021).

[5] Based on the estimated number of languages per family mentioned at the beginning of the paragraph, the ratio of languages studied here per family is 26/~50 for Atlantic, 8/~70 for Mande and 4/~10 for Mel. The unbalanced sampling for Mande is due to the original purpose of the previous study (Robert & Voisin 2018). The statistical analysis conducted in this paper shall make it possible to check whether or

the languages of Senegal and the surrounding areas is provided in Figure 1, showing the contact areas. A geographical distribution of the languages included in our study is shown in Figure 2, where languages are reduced to dots for visual convenience, followed by the detailed list of the languages in Table 2.

Figure 1 — Languages of Senegal and the surrounding areas (Pozdniakov et al. 2019)

not the Mande family retains a distinctive profile for noncausal/causal alternation despite this uneven sampling.

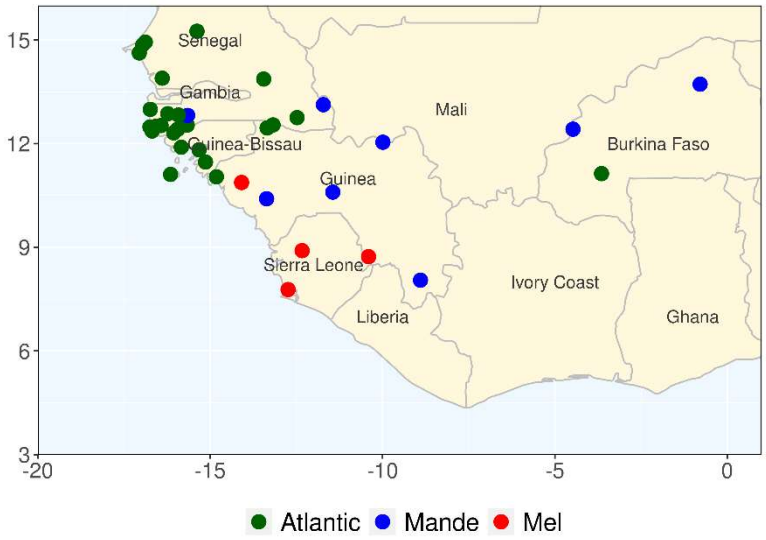

Figure 2 — The geographical distribution of the Atlantic, Mande and Mel languages covered in the study

Table 2 — The 38 languages of the study

ATLANTIC (26)				MANDE (8)	MEL (4)
NORTH branch	Wolof	BAK branch	Balant Kəntohe	Bobo	Sherbro
	Nyun Guñaamolo		Balant Ganja	Bambara	Kisi
	Nyun Gubëeher		Jóola Keeraak	Maninka of Niokolo	Temne
	Nyun Gujaher		Jóola Fóoñi	Mandinka	Landuma
	Konyagi		Jóola Banjal	Kakabe	
	Bedik		Jóola Kasa	Soso	
	Basari		Jóola Kwaatay	Soninke of Bakel	
	Pajaade		Jóola Karon	Soninke of Kingi	
	Sereer		Manjaku of Bassarel		
	Pulaar (Ful of Futa Toro)		Pepel		
	Ful of Massina		Bijogo Kagbaaga		
	Laalaa				
	Palor				
	Saafi				
	Nalu				

All the languages of the Atlantic and Mel families are in contact with the two other families. For Mande, we have added to the six languages in contact with Atlantic and Mel two other languages with no contact, Bobo and Bambara, for balancing a possible effect of contact in the coding profile emerging from our study of Mande. The list of the languages and their metadata is in Supplementary Material 1.

As in Robert & Voisin (2018), the 18 verb pairs provided by Nichols et al. (2004) were investigated. These 18 verb pairs are summarized in Table 3 and provided with more details in the Appendix (i.e. including the proxies that have been alternatively searched when some pairs could not be found, following Nichols et al. (2004) again).

Table 3 — A simplified overview of Nichols et al.'s (2004) 18 pairs of noncausal (nC) and causal (C) verbs.

n°	nC	C	n°	nC	C
1	laugh	amuse	10	boil	boil
2	die	kill	11	burn	burn
3	sit	seat	12	break	break
4	eat	feed	13	open	open
5	learn	teach	14	dry	make dry
6	see	show	15	be straight	straighten
7	be angry	anger	16	hang	hang (up)
8	fear	scare	17	turn over	turn over
9	hide	conceal	18	fall	drop

The verb pairs were mostly retrieved from the lexical database *Reflex* [*Reference Lexicon of the Languages of Africa*] (Segerer & Flavier 2018), which gathers lexical information (such as form, segmentation when available and meaning) extracted from referenced and accessible sources on (presently 789) African languages. The information retrieved from the database was also completed and substantiated (for the morphological analysis) by the available grammars and by additional inquiries with specialists of individual languages when needed and possible. After the causal and noncausal forms had been determined and analyzed, each verb pair was analyzed for its morphological structure and labeled according to the strategy used in that verb pair. A sample of the five strategies for the noncausal/causal alternation in Landuma (Mel) is shown in Table 4. The linguistic data on the noncausal/causal alternation in the 38 languages surveyed and their sources are available in Supplementary Material 2. For convenience, the bibliographical references of these sources are also provided in Supplementary Material 2.

Table 4 — The noncausal/causal alternation in Landuma (Rogers & Bryant 2012). The differences between the noncausal (nC) and causal (C) alternations are highlighted in bold.

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	Suppletion	nC ≠ C	***nənk***	***mʌnk***
14 dry/make dry	Lability	nC = C	***pʌc***	***pʌc***
18 fall/drop	Equipollence	nC ~ C	*funp.**ʌ***[6]	*funp-**əs***

Despite the extraordinary coverage of RefLex and our personal efforts to expand the data, the 18 verb pairs could not be fully completed for many of the languages in the sample. This is due to the insufficient documentation for many African languages. In terms of data coverage, all languages included in the analysis have more than half (i.e.

[6] A hyphen (-) indicates a derivational morpheme, a dot (.) indicates a morpheme analyzable as an inflectional ending or as a frozen suffix.

9/18) of the verb pairs annotated. The ratio of missing values for each family sample is: Atlantic 22.6% (106/(26*18)), Mande 14.6% (21/(8*18)) and Mel 12.5% (9/(4*18)). For example, 26 Atlantic languages are included in our sample, which results in 26 times 18 verb pairs, which is 468 pairs. Amongst these 468 pairs, 106 have missing values, which results in a missing ratio of 106/468 = 22.6%. A visualization of the data is provided in Figure 3, indicating the various strategies found for each verb pair in the languages under study. The y-axis represents the languages included in the data and the x-axis indicates the verb pairs. The correspondence of types/formal strategies are coded by the colors of the heatmap plot.

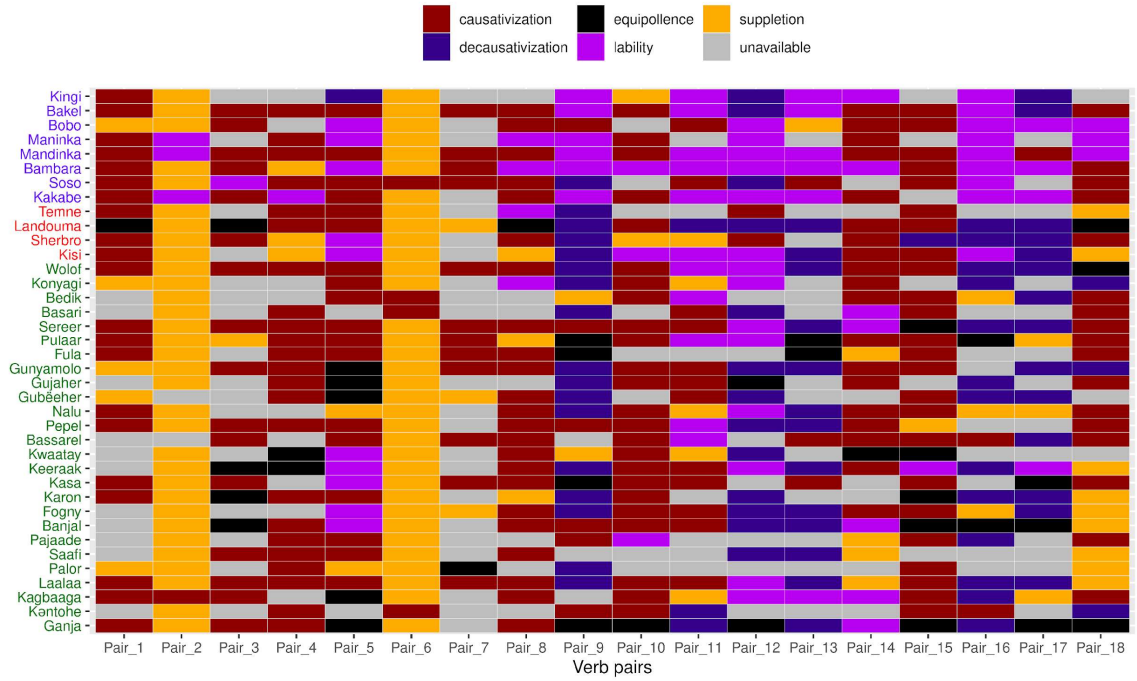

Figure 3 — An overview of the noncausal/causal alternation across the 18 verb pairs in the 38 languages (ordered as Mande, Mel, Atlantic)

An eyeballing of this overview points to different tendencies across verb pairs. Some verb pairs do not vary much across the three language groups. As an example, the great majority of the languages use suppletion for the verb pairs 2 (‘die/kill’) and 6 (‘see/show’). Furthermore, the causativization strategy is much more common in the verb pairs 1-9, while it is mostly found in verb pairs 10 and 15 within verb pairs 10-18. Tendencies across verb pairs and language families can also be found. For example, the Mande languages show the largest use of lability for coding verb pair 16 (‘hang’/‘hang up’). These observations support the claim that verb pairs 1-9 show more uniformity in their coding, cross-linguistically, than verb pairs 10-18, as expected from previous work on other languages (see Section 1). To get a more numerical overview of the noncausal/causal alternation in the languages of the sample, we also display the ratio of different coding strategies across language families in Table 5. The numbers in the cells refer to the mean number of verb pairs using a given coding strategy in each language family. The ratios thus add up to 18 in each column.

Table 5 — Coding profiles of the Atlantic, Mande and Mel languages for the noncausal/causal alternation based on the 18 verb pairs

Coding strategy		Atlantic	Mande	Mel
Causativization	>	**6.00**	**6.63**	**4.75**
Decausativization	<	2.19	0.88	**3.50**
Equipollence	~	1.27	0.00	1.00
Lability	=	1.08	**6.00**	1.75
Suppletion	≠	3.12	2.00	**4.00**
Unavailable	NA	4.35	2.50	3.00

The distribution shows that the causativization strategy is the most used strategy in the three families, whereas lability is by far more frequent in Mande than in Atlantic and Mel, and suppletion and decausativization are more frequent in Mel than in Atlantic and Mande. Nevertheless, additional quantitative analyses are required to verify the statistical significance of these tendencies when it comes to predicting the genealogical affiliation of Atlantic, Mande and Mel languages.

3. The noncausal/causal alternation and genealogical affiliation: Machine learning experiments

Three quantitative analyses were conducted to assess the distribution of the coding strategies for the noncausal/causal alternation in the language sample and its predictive power for the genealogical affiliation of the languages. First, we used principal component analysis (PCA) and k-means clustering to visualize how the surveyed languages are clustered based on the noncausal/causal alternation from verb pairs 1-9 and verb pairs 10-18. This comparison provided an overview as to (i) how different are the information encoded in the two groups of verb pairs, (ii) how well the noncausal/causal alternation in the two verb pair groups matched the language families. Second, we used a decision tree to quantify how exactly the information on the noncausal/causal alternation can help to predict the genealogical affiliation of a given language in the sample. This analysis considered all the verb pairs simultaneously and indicated which coding strategy with which verb pair had a statistically significant predictive power with regard to language family in the sample. The decision tree also showed the interaction and the hierarchy between different strategies and verb pairs. Third, to have a ranking of the importance of the variables (including those with a weak effect that are not shown in the preceding decision tree), we used information gain to quantify the amount of information captured by each pair with regard to the affiliation of languages to each family. This analysis was meant to provide a ranking of the 18 verb pairs for predicting language families in the data set. If verb pairs 10-18 contain more information for predicting language families, it is expected that most of these verb pairs will be highly ranked in terms of information gain. The detailed code and data used for the analysis are available in Supplementary Material 3.

3.1 Clustering: comparison between the two verb types (pairs 1-9 vs. 10-18)

In the first experiment, principal component analysis (PCA) was used to reduce the dimensionality of the data and then to allow for the clustering of the resulting data by

k-means. This experiment is mostly of an exploratory nature to test the role of the noncausal/causal alternation as a genealogical marker. We compared the language clustering resulting from the coding of this alternation with the established genealogical classification of languages.

PCA is a technique used for unsupervised dimension reduction (Jolliffe 2002). PCA transforms a number of correlated variables into uncorrelated variables, which are called “principal components”. To apply PCA to our data, the variation found within the 18 columns representing the verb pairs (as shown in Figure 3) was first condensed by the count of each annotated strategy. For example, when considering verb pairs 1 to 9, we extracted the sum of the verb pairs that use a certain strategy for each language. If for a given language, verb pairs 1 to 3 use causativization and verb pairs 6 to 9 use equipollence, the sum of causativization tokens is three and the sum of equipollence tokens is six for that language. The sum of other strategies is zero for that language. The five columns with the sum of each marking (causativization, decausativization, equipollence, lability and suppletion) found in the data were then compressed into two columns (i.e. two principal components), which can be visualized in a two-dimensional representation using an X-Y graph. This method does not fully capture the variance in the data, as the sum of the tokens do not take into account which pairs use which strategy. For example, two languages which are very different in their verb pairs can happen to get the same vector. Ideally, methods employing measures of distance (e.g. Gower distance) should be used. However, these measures are easily affected by the number of missing data points, which are not scarce in our data. Therefore, we used the sum of strategy count as a way to reflect the general tendencies for valence orientation in each language.

These extracted components can be used to cluster the data points, i.e. to find how many main groups exist in the data. One of the most common clustering techniques is k-means clustering (Forgy 1965; Hartigan & Wong 1979; Lloyd 1982; MacQueen 1967), 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 center points are generated randomly within the investigated space. Second, each data point within the space is assigned to the nearest center point, which represents a cluster. Third, new center points are generated as the centers of the current *k* clusters. Finally, the second to third step is repeated until the optimal center points are found for each of the clusters.

When conducting k-means clustering, three clusters are assumed to emerge since the languages of the data belong to three different language families (Atlantic, Mande and Mel). In other words, we asked the clustering method to group the languages in the sample into three clusters based on their coding strategies across the 18 verb pairs. This process was done separately for verb pairs 1-9 and verb pairs 10-18. The output of k-means clustering is shown in Figure 4 and 5. Each point represents one of the 38 languages in the dataset. The x- and y-axes represent the percentage of variance captured by the first two principal components. The distance between the languages reflects their similarities and dissimilarities in the use of coding strategies across the verb pairs. The more similar two languages are, based on the noncausal/causal alternation, the closer they are in the two-dimensional space.

Figure 4 — k-means clustering based on the PCA of the noncausal/causal alternation in the verb pairs 1-9. The color of the labels represents the language families (Atlantic = yellow, Mande = purple, Mel = black). The colors of the shapes indicate the different clusters identified by k-means clustering.

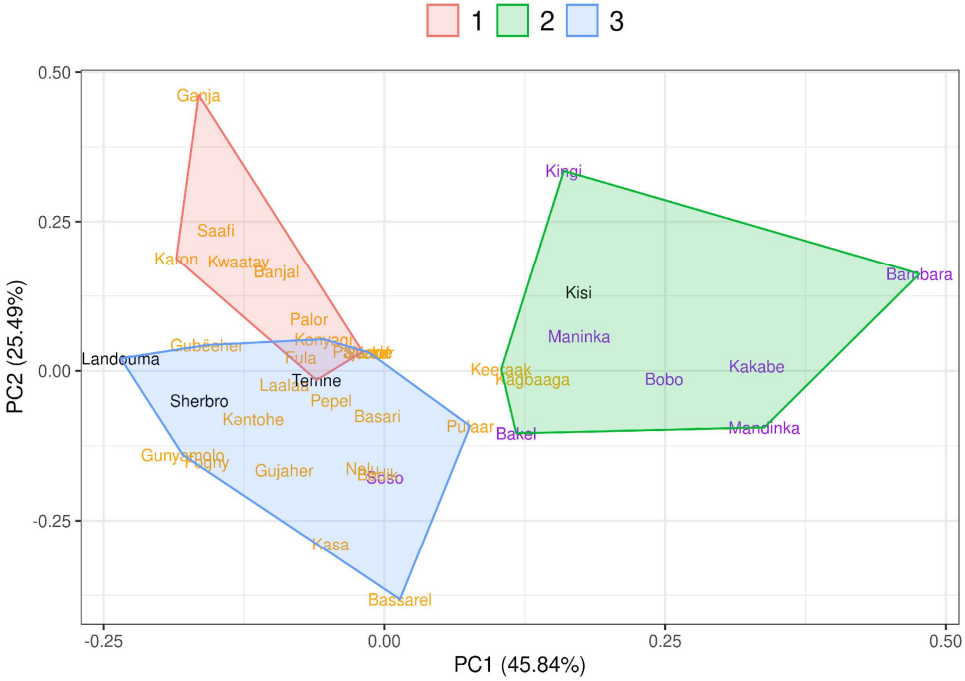

Figure 5 — k-means clustering based on the PCA of the noncausal/causal alternation in the verb pairs 10-18. The color of the labels represents the language families (Atlantic = yellow, Mande = purple, Mel = black). The colors of the shapes indicate the different clusters identified by k-means clustering.

The results match better with the actual genealogical affiliation when taking the verb pairs 10-18 than the verb pairs 1-9, confirming the role of inchoative verbs for indicating the language specific preferences for the noncausal/causal alternation. As an example, the Mande languages are scattered across clusters based on verb pairs 1-9 (Figure 4) but are mostly clustered together with results based on verb pairs 10-18 (Figure 5). In both runs, the Mel languages are scattered across two clusters, whereas the Atlantic languages are spread across three clusters with verb pairs 1-9 and are mostly found in two clusters with verb pairs 10-18. To evaluate the performance of the two verb pair groups statistically, the clusters generated by k-means were compared with the original genealogical affiliations (Atlantic, Mande and Mel). To do so, we used the Rand Index, which 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 has the value of 1. A potential problem with the Rand index is that its expected value between random partitions is not constant. This problem was 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. It is below 0 if the performance is worse than random guessing. A larger adjusted Rand index means a better match between two partitions. The measures of Rand index and adjusted Rand index are shown in Table 6 to enhance the robustness of the comparison. In both measures, 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 about the genealogical groups of the languages we investigated.

Table 6 — The performance of k-means clustering

	Verb pairs 1-9	Verb pairs 10-18
Rand index	0.489	0.744
Adjusted Rand index	-0.015	0.486

As a short summary, more regularities in the coding of the noncausal/causal alternation are found in clustering when only verb pairs 10-18 are considered. These observations match the hypothesis that verb pairs 10-18 encode more relevant information on the noncausal/causal alternation for language family identification. Nevertheless, the clustering method does not indicate explicitly which verb pairs and/or which coding strategies are more important for classifying languages into different clusters. To fill this gap, we conducted the following analyses based on decision tree and information gain.

3.2 Single decision tree

In the second experiment, a decision-tree-based computational classifier (more specifically, a conditional inference tree via Monte Carlo simulations) was used to extract the interaction of coding strategies and verb pairs when predicting language families. The decision tree shows which combinations of coding strategies and verb pairs are statistically significant to predict the language families in the sample. Furthermore, it also allows us to visualize the hierarchy of interaction between these combinations. While such a decision tree is absolutely not a representation of the

phylogenetic tree of the surveyed languages, it allows a visual interpretation of which verb pairs and noncausal/causal alternation strategies are specific to which language family.

The decision tree classifier is based on binary recursive partitioning (Breiman et al. 1984). To summarize the operating process, first, our data was one-hot encoded. That is to say, the 18 columns for the 18 verb pairs were expanded so that each combination of verb pair and valence strategy was annotated as a column filled with the binary values of 1 and 0. For example, a column will mark if a language uses the causativization strategy for verb pair 1, another column will mark if a language uses the decausativization strategy for verb pair 1, among others. This format was selected for two reasons. First, it avoids that the models consider the missing values in the data when comparing languages and verb pairs. Second, it enhances the processing speed of the algorithms. When this transformed data is fed to the decision tree, the data is repeatedly partitioned to form groups that are as homogeneous as possible. First, the model tests the null hypothesis of independence between the predictors (i.e. the columns of valence strategy in each verb pair) and the response (i.e. the language families). The strength of this association is quantified by the p-value of a permutation test. The results were considered statistically significant if the proportion of the permutations providing a test statistically greater than or equal to the one observed in the original data was smaller than the significance level. The predictor with the strongest association with the response was then used to split the data. This process of permutation is also the main strength of the classifier, as it allows the analysis of small-scale data and consideration of the possible auto-correlation of variables (Tagliamonte & Baayen 2012). This aspect was particularly relevant for us, considering the gaps in our data, both in terms of languages per family and verb pairs per language. In the experiment, we did not perform cross-validation. In other words, the entire dataset was used to generate the tree and assess its precision, since the algorithm conducts a test of statistical significance at each split. For the same reason, pruning of the tree was not required either.

Figure 6 shows the decision tree obtained when considering the 18 verb pairs and their coding strategies across the 38 languages. The verb pairs considered statistically significant by the classifier are displayed in the tree. That is to say, noncausal/causal coding strategies and verb pairs that are not helpful for distinguishing Atlantic, Mande and Mel languages are not shown in the decision tree. The current tree only has one node (Node 1), which divides the data into buckets (also named Node 2 and 3). The bars in the buckets indicate the ratio of languages affiliated to each family. 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 almost the case: Node 2 represents Atlantic languages, Node 3 mostly represents Mande languages, while Mel languages are scattered across the two buckets.

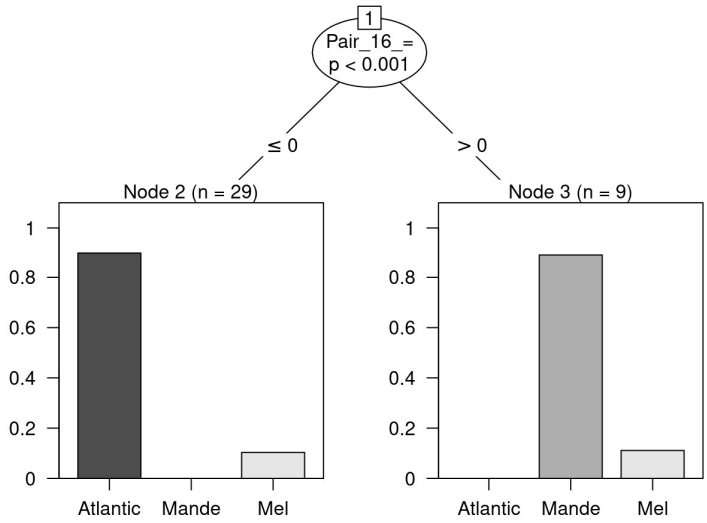

Figure 6 — Conditional inference tree based on the entire dataset

The combination of coding strategy and verb pairs showing up in the decision tree is verb pair 16 (‘hang’ vs. ‘hang up’) combined with lability. The tree can thus be read as follows: In a given language, if verb pair 16 uses lability as a coding strategy for the noncausal/causal alternation, it is very likely to be a Mande language. If the given language does not use lability in verb pair 16, then it is very likely to be an Atlantic or Mel language. The current data is not sufficient to efficiently identify Mel languages. The resulting classification can be explained by the particular features of the verb pair 16: ‘hang’ does not strictly meet the definition of an inchoative verb, i.e. a verb whose noncausal member is a monovalent verb referring to a process typically undergone by concrete *inanimate* entities, at least in the Niger-Congo languages considered here (a monkey may hang). Therefore, for this pair, Mande languages make use of lability as their preferred strategy for the pairs 10-18, whereas Mel and Atlantic languages display more variation in the coding of this pair, as they do for the whole set of pairs.

The performance of the decision tree is assessed with two measures, the accuracy and the f-score. The f-score evaluates the performance of the model in each category (i.e. in each language family). It is a combination of two other measures: precision and recall. Precision evaluates how many tokens are correct among all the outputs of the classifier. For example, if the classifier predicts that 30 languages belong to the Atlantic family and within these 30 languages, 26 languages are indeed affiliated to the Atlantic family, the precision for the Atlantic family is 26/30 = 86.7%. Recall quantifies how many tokens are correctly retrieved among all the expected correct outputs. As an example, if within the 26 Atlantic languages found in the data, 20 are correctly identified as Atlantic languages, the recall is 20/26 = 76.9%. The two measures evaluate the output from two different perspectives, as the measures of Target-Like Usage (TLU) and Suppliance in Obligatory Context (SOC) do in research on language acquisition (Pica 1983). 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). In parallel, the accuracy provides an overview of the performance for the

entire dataset. The accuracy is the ratio of the correctly retrieved tokens within the entire data. For example, if 30 of the 38 languages in the data are classified correctly, the accuracy is equal to 30/38 = 78.9%. 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 grouping (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 noncausal/causal alternation 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.

We used a confusion matrix (Table 7) to compare the predictions of the conditional inference tree with the actual genealogical affiliation of the languages in the data. The accuracy of the model was high, i.e. 0.895 (34/38), which exceeds by far the majority rule baseline of 0.684. The overall performance of the classifier was thus considered to be good. The detailed numbers reflect that the classifier is good at distinguishing Atlantic (Precision = 0.897, Recall = 1.000, f-score = 0.946) and Mande (Precision = 0.889, Recall = 1.000, f-score = 0.941) languages. However, the classifier had difficulties identifying Mel languages, as none of them were labelled correctly by the classifier. The detailed predictions of the conditional inference tree can be found in Supplementary Material 3.

Table 7 — The confusion matrix of the conditional inference tree. The rows are the predictions of the classifier and the columns indicate the actual values.

	Atlantic	Mande	Mel
Atlantic	26	0	3
Mande	0	8	1
Mel	0	0	0

All four Mel languages are labelled incorrectly based on this decision tree. For example, Kisi is wrongfully labelled as Mande and Landuma is wrongfully labelled as Atlantic (as are Sherbro and Temne). These errors most likely point to contact-induced changes since Landuma is geographically surrounded by Atlantic languages whereas Kisi is surrounded by Mande languages (see Figure 1). This study thus reveals that the coding of the noncausal/causal alternation can also be subject to contact-induced changes, in case of intense contact. Moreover, it should be noted that the Mel languages are generally under-documented (that is why only four out of the twelve Mel languages have been studied here), which probably does not allow us to define a family profile sufficiently distinct from that of the Atlantic family in which they have long been classified. The results of the conditional inference tree thus show that the family affiliation of the 38 languages can be predicted with an accuracy above baseline based on the information on the noncausal/causal alternation in the 18 verb pairs. Moreover, the verb pair 16 (‘hang’ vs. ‘hang up’) seems to be sufficient to predict the family affiliation of most languages in the data, which matches the hypothesis that verb pairs 10-18 include more relevant information.

3.3 Information gain

To analyze the contribution of all verb pairs and the noncausal/causal alternation regardless of statistical significance, we considered information gain, which represents

the quantity of information gained about a variable based on information from another variable. More specifically, the information gain is calculated by measuring the reduction in information entropy of the data when a variable is used to split the data into groups.

Entropy (Shannon 1948) represents the uncertainty of the data. The entropy ranges between 0 and 1 in the case of two groups in the data (but it can also be applied to situations with a larger number of groups, in which case the entropy can be larger than 1). A high entropy indicates a high uncertainty in the data. For example, if we are measuring the entropy of the masculine and feminine grammatical gender of nouns in a language and, if 50% of the nouns are masculine and 50% feminine, the uncertainty of the data is at its maximum, i.e. 1. The uncertainty is high since it is hard to guess the gender of a random noun. If all nouns in the given language are feminine, the uncertainty for guessing the gender of a random noun is 0, since we are sure that each randomly selected noun will be feminine. As a third example, if 95% of the nouns are masculine and 5% of the nouns are feminine, the uncertainty is equal to -((0.05*log2 (0.05)) + (0.95*log2 (0.95))) = 0.2864. The uncertainty is closer to 0 and far from 1, as it is much more likely to get a masculine noun than a feminine noun when a noun is selected randomly.

To visualize how entropy is used to calculate information gain, consider an example from verb pair 1 (‘laugh/amuse’), as shown in Table 8. The entropy of the Atlantic, Mande and Mel languages in the table is equal to -((15/27)*log2(15/27) + (8/27)*log2(8/27) + (1/27)*log2(1/27)) = 1.399208. If we use the valence strategies to split the data, we first consider the languages using the suppletion strategy. Four languages are affiliated to the Atlantic family, while one language is affiliated to the Mande family. The entropy for the languages using suppletion is equal to -((4/5)*log2(4/5) + (1/5)*log2(1/5)) = 0.72. In the same way, the entropy for languages using decausativization and equipollence is 1.42 and 0, respectively. The entropy for the languages using equipollence is 0 since all languages using equipollence for verb pair 1 are affiliated to the Mel language family. The entropy after splitting the data based on the valence strategies is then obtained by the sum of the weighted entropy for each valence encoding strategy, i.e. it is equal to (5/27*0.72) + (21/27*1.42) + (1/27*0) = 1.2366. The information gain (i.e. the diminished entropy) by splitting the data based on valence strategies in verb pairs is thus equal to 1.399208-1.2366 = 0.162608.

Table 8 — The distribution of valence strategies across the Atlantic, Mande and Mel families for verb pair 1.

	Atlantic	Mande	Mel
Causativization <	0	0	0
Decausativization >	11	7	3
Equipollence ~	0	0	1
Lability =	0	0	0
Suppletion ≠	4	1	0

Based on the same method, the information gain of each pair can be calculated. The output is displayed in Figure 7. First, we see that verb pair 16 (‘hang’ vs. ‘hang up’) has

by far the highest information gain. This result matches the output of the conditional inference tree, which uses verb pair 16 to categorize languages into different families. Second, we observe that most pairs from 10-18 are found in the top ten in terms of information gain. For example, within the top five verb pairs with the highest information gain, four are from verb pairs 10-18 (Pair 16 'hang/hang up', pair 13 'open/open', pair 17 'turn over/turn over' and pair 18 'fall/drop'). Only one of the verb pairs from 10-18 has a low information gain, i.e. verb pair 14 ('dry/make dry').

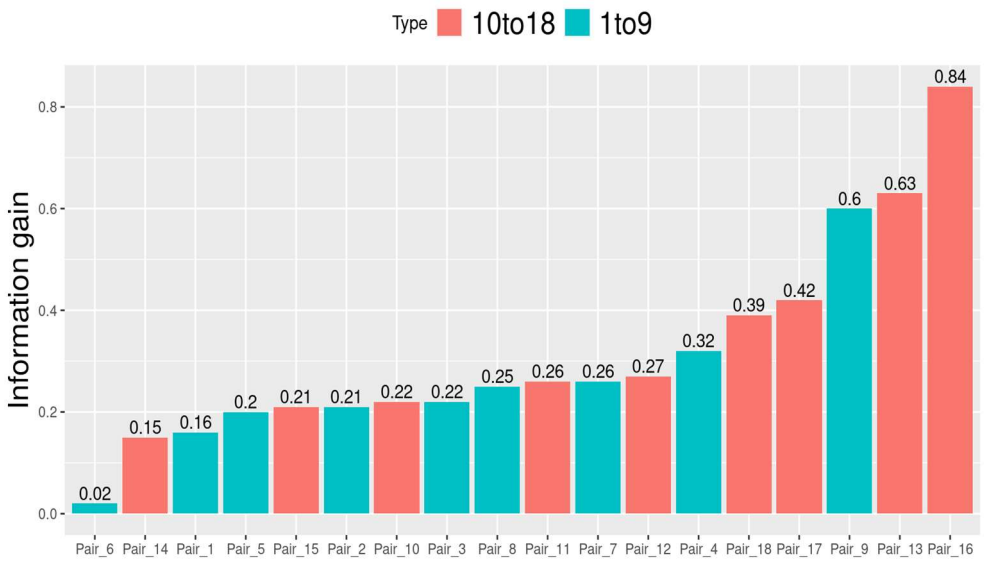

Figure 7 — The ranking of information gain for each verb pair when predicting language families. A higher information gain indicates higher information relevant for predicting language families.

As a summary, the results from the measure of information gain also show that the genealogical affiliation of languages in our sample can be better categorized based on the information of the noncausal/causal alternation from verb pairs 10-18. A more detailed analysis of these results is provided in Section 4.

4. Discussion

The main purpose of the quantitative analyses was to assess the distribution of the coding strategies in the language sample in order to quantify the predictive power of the noncausal/causal alternation for identifying the genealogical affiliations of the Atlantic, Mande and Mel languages. The experimental results show that the language-specific coding of the noncausal/causal alternation across the verb pairs is effective for predicting the genealogical affiliations of the languages. For instance, the accuracy of the conditional inference tree is 89.5%, which is far above the baseline of 68%. The lower scores for Mel are probably due to the very small number (4) of Mel languages that could be documented for this study.

The second objective was to evaluate the information relevant to genealogical classification across the different verb pairs and, in particular, to compare the predictive power of the verb pairs 1-9 and 10-18 for the genealogical affiliation of the Atlantic, Mande and Mel languages through their coding of the noncausal/causal alternation.

All three experiments show that verb pairs from the group 10-18 are more relevant for distinguishing languages from the Atlantic, Mande and Mel families. For instance, the performance of k-means clustering is much higher for this verb group, as visible from Table 6. The outputs of the conditional inference tree and the measurements of information gain converge and show that verb pair 16 is by far the most relevant to distinguish the language families. Moreover, the output of the information gain analysis also shows that verb pairs 10-18 have a higher information gain for predicting language families than verb pairs 1-9. As an example, the top ten verb pairs in terms of information gain are in descending order: pair 16 ('hang/hang (up)'), pair 13 ('open/open'), pair 9 ('hide/conceal'), pair 17 ('turn over/turn over'), pair 18 ('fall/drop'), pair 4 ('eat/feed'), pair 12 ('break/break'), pair 7 ('be angry/anger'), pair 11 ('burn/burn'), pair 8 ('fear/scare'). Among these ten pairs, six belong to the group 10-18. Furthermore, if we only consider the top five verb pairs, four pairs are from the group 10-18. Thus, in contrast to Grünthal & Nichols' (2016: 29) results for Slavic, verb pairs 10-18 ("inanimate verbs" in the authors' terminology) enclose more relevant information for clustering the languages from the Atlantic, Mande and Mel families than verb pairs 1-9. These results are consistent with the linguistic claim made by Haspelmath (1993, 2016) and Creissels (this issue) about the crosslinguistic trends in the coding of the noncausal/causal alternation. Verb pairs 10-18 actually correspond almost perfectly to the noncausal verb types for which the cross-linguistic variation in the coding of the noncausal/causal alternation is particularly important according to these authors. These verbs are characterized by Creissels (*ibid*.) as "monovalent verbs referring to a process (not a state) typically undergone by concrete inanimate entities, and easily conceived as occurring without the involvement of a clearly identified external instigator". These "inchoative" or "inanimate" verbs show the preferences of individual languages for specific coding strategies. That is why they prove to be good markers of genealogical affiliation in our experiments, in that they reveal family preferences in the coding of the noncausal/causal alternation. On this point, the deviations of individual languages from the family preferences, which were visible in their wrong affiliation by the decision tree, can be due to language contact, as explained for Landuma and Kisi (cf. Section 3.2 and Robert & Voisin (2018)).

The presence of some verb pairs from the 1-9 list at the top of the ranking can also be explained linguistically. For example, for verb pair 9 ('hide/conceal'), ranking third, the noncausal form is a monovalent verb referring to a process that may be used with a non-human subject,[7] which makes it more like verb pairs 10-18 than verb pairs 1-9. It is worth mentioning that, in spite of Nichols et al.'s (2004) recommendations, we did not systematically distinguish between human and non-human subjects when collecting the linguistic data because this information was often not available from the dictionaries.

A few additional comments on some high ranking verb pairs belonging to the 10-18 list can be made. The core event of verb pairs 12 ('break/break') and 13 ('open/open') refer to spontaneous events and that of pair 17 ('turn over/turn over') to a change in body posture. As noted by Kemmer (1993), these types of events correspond to "middle situations" which share a low degree of elaboration of events by including two participants (the Initiator and the Endpoint), which are however not fully

[7] As pointed out by an anonymous reviewer, an obvious complication for the analysis is that 'hide' used intransitively with an inanimate subject is more likely to have a passive interpretation (i.e. implying the presence of an unexpressed agent) rather than a true noncausal one.

physically and conceptually distinguishable from one another. This property most saliently concerns verbs of grooming, change in body posture and non-translational motion. “Because these [middle] properties are already intrinsically part of the meaning of the middle situation types, languages simply tend not to mark them” (Kemmer 1993: 234). This prediction gives an asset to the use of lability for coding these verb classes. Consequently, when the strategy used for these pairs is not lability, it provides a highly significant indicator of the language-specific preference in the coding of the noncausal/causal alternation for these kinds of events. That is why, in our experiments, these pairs also have a good predictive power for distinguishing Mande languages from others.

Finally, a review on the limits of our analyses is in order. Our data had missing values, which should ideally be filled in. This would require fieldwork since the information is not available in the published documentation. However, since the amount of missing information was controlled across languages and verb pairs, and the methods of permutation and bootstrapping were used to conduct sampling of the data, we consider that these missing values did not have a large effect on the output of the model. This is also mirrored in the accuracy of the computational classifiers. Also, we only considered a small sample of verb pairs and features. Additional features could be tested with the same method. Nevertheless, the features selected for this study (namely the five coding means for the noncausal/causal alternation) are the only morphological strategies used across the three families for the noncausal/causal alternation (other strategies are syntactic), which we intended to test as a genealogical marker and did successfully. Moreover, while covering almost the whole Atlantic and Mel families, our study is based on a very small sample of Mande languages. This is due to the original purpose in collecting the data for Robert & Voisin (2018), which aimed at identifying contact-induced change in the three families and, therefore, focused on the area where members of the three families are in contact. A similar study should be conducted on more languages of the Mande family for more robust results.

5. Conclusion

The results of the machine learning experiments show that, in spite of the unbalanced language sample and the missing data, the language profile for coding the noncausal/causal alternation has a strong predictive power on the genealogical affiliation of the Atlantic, Mande and, to a much lesser extent, Mel languages. However, they do not confirm the hypothesis of a correlation between the typological profiles of the families and their valence profile. The Mande languages show a massive propensity to use the labile strategy in accordance with their isolating morphology. Nevertheless, Creissels (this issue) shows that the lability strategy is also predominant in morphologically complex languages such as Basque and Avar. In our study, in spite of a comparable typological profile with large inventories of derivational suffixes, Atlantic and Mel languages do not show the same profile of use of the different strategies: Mel languages have a stronger tendency to use suppletion and decausativization than Atlantic languages. What our study shows is that each language family has rather stable trends for coding the noncausal/causal alternation and that these trends define a valency profile specific enough to predict the genealogical affiliation of a language.

As for the distinctive significance of the two subgroups of verb pairs for predicting language-specific preferences in the coding of the noncausal/causal

alternation, the results show that the verb pairs 10-18 are more relevant for differentiating the languages of the three families than verb pairs 1-9. These results generally match the hypotheses from previous studies about the two different verb types, namely the dynamic verbs (1-9) that prototypically take an animate subject for the noncausal member of the pair and which are known to show a universal trend for causative coding vs. the inchoative verbs (10-18) that were claimed to show language-specific preferences. However, a verb pair such as 9 ('hide/conceal') is also ranked as important by the classifier. This surprising result has been accounted for by some specific features of the noncausal verb of the pair that make it similar to the inchoative verbs, pointing here to less control of the verb types in Nichols et al.'s (2004) list. Moreover, the quantitative analyses have provided additional insights. First, they identified the noncausal/causal alternation as a good marker for predicting language affiliation, at least among the three families we investigated. Then, they pinpointed which verb pairs are more relevant to identify language affiliation. In addition, the wrong affiliation of Mel languages by the classifier has usefully pointed to plausible contact-induced changes.

Finally, this study provides a pipeline of methods (including the data and code in the supplementary materials) that can be tested on other geographical areas and families. Other strategies for coding noncausal/causal alternation (e.g morphosyntactic ones) could also be investigated, if relevant for the languages. Moreover, it contributes in terms of methodology to the fundamental question in historical linguistics about the relationship of quantitative and qualitative methods for the study of African languages, supplementing Grünthal & Nichols' (2016) pioneering contribution. A very fundamental issue in historical linguistics is that most synchronic data are irrelevant or misleading for genealogical classification or sub-classification and a filter is needed to only keep the relevant information (Teeter 1964: 1030; Campbell 2013: ch14). The ranking of variables used in the decision tree classifier provides a concrete application of such a filter by suggesting which verb pairs are the most relevant.

Abbreviations

CAUS	causative voice marker
COMPL	completive
MID	middle voice marker
OBJ	object
SBJ	subject
SG	singular

References

Banerjee, Mousumi, Ying Ding & AnneMichelle Noone. 2012. Identifying representative trees from ensembles. *Statistics in Medicine* 31(15). 1601–1616. https://doi.org/10.1002/sim.4492.

Bickel, Balthasar. 2015. Distributional typology: statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 901–923. 2nd edn. Oxford: Oxford University Press.

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

Campbell, Lyle. 2013. *Historical Linguistics. An Introduction.* Edinburgh: Edinburgh University Press.

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

Creissels, Denis. 2014. Le développement d'un marqueur de déplacement centripète en mandinka. In Carole de Féral (ed.), *In and Out of Africa, Languages in Question, In Honor of Robert Nicolaï*, 95–102. Walpole: Peeters.

Creissels, Denis, this issue. The noncausal-causal alternation and the limits of ambitransitivity in the languages of Subsaharan Africa.

Ding, Chris & Xiaofeng He. 2004. K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on Machine learning* 225–232.

Diouf, Jean-Léopold. 2003. *Dictionnaire wolof-français et français-wolof*. Paris: Karthala.

Forgy, Edward W. 1965. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics* 21. 768–769.

Grünthal, Riho & Nichols, Johanna. 2016. Transitivizing-detransitivizing typology and language family history. *Lingua Posnaniensis* 58(2). 11-31. doi: https://doi.org/10.1515/linpo-2016-0008.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2021. *Glottolog 4.5*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5772642 (Available online at http://glottolog.org, Accessed on 2022-05-02.)

Hartigan, John A & M Anthony Wong. 1979. Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics* 28. 100–108.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternation. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity*, 87–120. Amsterdam: John Benjamins.

Haspelmath, Martin. 2016. Universals of causative and anticausative verb formation and the spontaneity scale. *Lingua Poznaniensis* 58(2). 33-63.

Haspelmath, Martin, Andreea Calude, Michael Spagnol, Heiko Narrog & EliF Bamyaci. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics* 50(3). 587–625. https://doi.org/10.1017/S0022226714000255.

Jolliffe, Ian. 2002. *Principal component analysis*. New York: Springer. http://www.ebrary.com (19 September, 2019).

Kemmer, Suzanne. 1993. *The middle voice* (Typological Studies in Language : (TSL), 0167-7373 ; 23). Amsterdam ; John Benjamins.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2). 129–137. https://doi.org/10.1109/TIT.1982.1056489.

MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In Lucien Le Cam & Jerzy Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297. Berkeley: University of California Press.

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

Matras, Yaron. 2010. Contact, Convergence, and Typology. In Raymond Hickey (ed.), *The Handbook of Language Contact*, 66–85. Oxford, UK: Wiley-Blackwell.

https://doi.org/10.1002/9781444318159.ch3. https://onlinelibrary.wiley.com/doi/10.1002/9781444318159.ch3 (17 November, 2021).
Nichols, Johanna, David A. Peterson & Jonathan Barnes. 2004. Transitivizing and detransitivizing languages. *Linguistic Typology* 8(2). https://doi.org/10.1515/lity.2004.005. https://www.degruyter.com/document/doi/10.1515/lity.2004.005/html (17 November, 2021).
Pica, Teresa. 1983. Adult acquisition of English as a second language under different conditions of exposure. *Language Learning* 33(4). 465–497. https://doi.org/10.1111/j.1467-1770.1983.tb00945.x.
Pozdniakov, Konstantin, Guillaume Segerer & Valentin Vydrin. 2019. Mande-Atlantic Contacts. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.393. http://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-393 (17 November, 2021).
Robert, Stéphane & Sylvie Voisin. 2018. Comparing 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*. https://hal.archives-ouvertes.fr/hal-03832646
Rogers, Kirk, & Daniel Bryant. 2012. *Diksiyɔ nɛr k əlʌndma – kətabu – Dictionnaire landouma - français*. Toolbox Dictionary and pdf. Ms. Available at www.reflex.cnrs.fr.
Segerer, Guillaume & Sébastien Flavier. 2011-2022 [accessed 2018]. *RefLex: Reference Lexicon of Africa*, Version 2.0 Paris, Lyon. http://reflex.cnrs.fr/.
Shannon, Claude. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 379–423.
Tadmor, Uri & Martin Haspelmath. 2010. Borrowability and the notion of basic vocabulary. *Diachronica 27* ((2)). 226–246.
Tagliamonte, Sali A & Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24. 135–178.
Teeter, Karl V. 1964. “Anthropological Linguistics” and Linguistic Anthropology. *American Anthropologist* 66(4). 878–879. https://doi.org/10.1525/aa.1964.66.4.02a00120.
Ting, Kai Ming. 2010. Precision and Recall. In Claude Sammut & Geoffrey I. Webb (eds.), *Encyclopedia of Machine Learning*, 781–781. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_652.
Voisin, Sylvie. 2021. Associated motion and deictic directional in Atlantic languages. In Antoine Guillaume & Harold Koch (eds.), *Associated Motion*, 611–664. De Gruyter. https://doi.org/10.1515/9783110692099-016.
Zha, Hongyuan, Chris Ding, Ming Gu, Xiaofeng He & Horst Simon. 2002. Spectral relaxation for k-means clustering. *Advances in Neural Information Processing Systems 14* 1057–1064.

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

	Non-causative[8]	***Causative***	***Proxies***
1	laugh	make laugh, amuse, strike as funny	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, put into hiding	
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, protrude
17	turn over	turn over	turn, turn around, rotate, revolve, roll; shake, tremble, vibrate
18	fall	drop, let fall	fall down, fall over, etc.; sink

Supplementary materials
The following materials are available at https://doi.org/10.17605/OSF.IO/3P6MC

- Supplementary Material 1: The languages of the study and their metadata
- Supplementary Material 2: The data on noncausal/causal alternation in the surveyed languages and their sources
- Supplementary Material 3: The pdf report of the code used for the analyses.

[8] "Non-causative" and "causative" are the terms used by Nichols *et al*. (2004) for what we refer to as "noncausal" and "causal".