THE EVOLUTIONARY TRENDS OF NOUN CLASS SYSTEMS IN ATLANTIC LANGUAGES

Neige Rochant[1], Marc Allassonnière-Tang[2], and Chundra Cathcart[*3]

[*]Corresponding Author: chundra.cathcart@uzh.ch
[1]Sorbonne Nouvelle University/Lacito, CNRS, Paris, France
[2]EA, MNHN/ CNRS/ University Paris City, Paris, France
[3]Department of Comparative Language Science, University of Zurich, Zurich, Switzerland

Nominal classification systems such as grammatical gender (e.g., the masculine/feminine distinction in French) and noun classes (e.g., Bantu noun classes based on fruits, plants, liquids, among others) provide a window on how the human brain perceives and categorizes objects and experiences it encounters. While the diachronic development of grammatical gender systems is well studied, noun class systems have received less attention. We use phylogenetic comparative methods to analyze where noun classes are marked (on nouns, pronouns, demonstratives, articles, adjectives, numbers, and verbs) in thirty-six Atlantic languages and how these markers change diachronically. Our results show that noun class marking is generally preferred and more stable within the noun phrase, i.e., on nouns, demonstratives, and adjectives.

1. Introduction

Languages can rely on various strategies to categorize nouns of the lexicon (Seifart, 2010; Kemmerer, 2017). One the most common strategies is noun class systems (Corbett, 2007), in which each noun of the lexicon is assigned to a specific category (i.e., noun class), which can relate to humans, plants, fruits, liquids, among others (Corbett, 2013). For example, in Swahili, nouns are affiliated to more than ten noun classes.

Two of the most common formal criteria to define noun classes are flexibility of category assignment and grammatical agreement (Corbett, 1991). First, noun class systems typically have a rigid assignment system. Each noun of the language belongs to one of the noun classes found in the language. This assignment is not flexible and using the agreement pattern of another noun class would result in ungrammaticality. Second, noun class systems may have marking on nouns, as shown in Swahili, with prefixes on the nouns, e.g., *m-toto* (CLASS.1-child) 'child'. However, noun classes also generate grammatical agreement with words associated to the noun and its referent. The noun classes can be marked on the adjectives, verbs, demonstratives, numerals, among others. Taking Swahili again to illustrate this type of agreement: *m-toto yu-le a-li-anguka* (CLASS.1-child CLASS.1-that CLASS.1.SUBJ-past-fall) 'that child fell'. These requirements of as-

signment and agreement distinguishes noun class systems from inflection classes such as on number marking. Grammatical number systems have agreement; however, number is not a fixed inherent feature of a noun/referent. For example, count nouns can generally be singular or plural depending on the context while the noun class of a noun is fixed. It is generally held that languages may develop a nominal classification system that proceeds through different stages of grammaticalization (Grinevald, 2002; Aikhenvald, 2016). At the beginning, classification is based on lexical nouns, which then develop into classificatory morphemes, which in turn further grammaticalize and become agreement markers. For example, the Niger-Congo noun class systems may have developed from nouns, along with the nominal classification systems found in modern Amazonian languages (Grinevald & Seifart, 2004). At the same time, questions remain regarding their development that are relevant to the diachronic change of linguistic complexity (Walchli, Olsson, & Di Garbo, 2020). For example, were noun classes first marked on nouns? If so, how did they spread to pronouns or verbs? While several studies have investigated the synchronic distribution of noun class systems in languages of the world (Corbett, 2013; Allassonnière-Tang et al., 2021), most diachronic quantitative studies have focused on the evolution of agreement marking in Indo-European languages (Allassonnière-Tang & Dunn, 2020; Carling & Cathcart, 2021). Few studies have investigated the evolution of agreement marking in languages with noun class systems, and even fewer studies have approached this question from a quantitative perspective with phylogenetic methods. The current study aims to fill this gap, quantifying the diachronic preference, speed of change, and stability of noun class markers across multiple morphosyntactic domains in the Atlantic languages of West Africa, with an eye to inferring the relative chronology of their development.

2. Data

In this section, we describe how the data on noun class marking was gathered in a sample of 36 Atlantic languages. We also provide information about how the phylogeny of the languages in the sample was generated.

2.1. *Noun class systems*

The Atlantic languages were selected due to their frequent presence of noun class systems. We consider noun class marking on the noun, in the noun phrase, and the verb phrase. To be more precise, we consider noun class marking on the noun itself (via a PREFIX), ARTICLES, PRONOUNS, DEMONSTRATIVES, ADJECTIVES, CARDINAL NUMBERS, and VERBS. For instance, Segerer (2002, 85-92) shows that Bijogo (Glottocode `bidy1244`) has fourteen noun classes that are marked on the noun itself (*e-we* [class.E-goat] 'goat'), adjectives and demonstratives (*ɲo-ogo n-nɛ n:-gbon* [CLASS.M-rock CLASS.M-DEM CLASS.M-be.big] 'These rocks are big.'), pronouns (*ya-g* [CLASS.YA-PR] 'them'), cardinal numbers (*ya-to ya-nsom*

[CLASS.YA-two CLASS.YA-people] 'two people'), and verbs (*bisaw wɔ-gbe na-jɔkɔ na-kotong* [Bissau CLASS.WO-ACC.have CLASS.ÑA-house CLASS.ÑA-big] 'There are big houses in Bissau.'). A language is encoded as marking a domain as long as one instance of marking is attested. Data were extracted manually from language grammars and sketches with sufficient available information, while maintaining a balance across the sub-branches of the language family. This process resulted in a sample of 36 languages, displayed in Figure 1.

2.2. *Tree Sample*

The models used in this paper require a phylogenetic representation of the languages in our sample, in the form of a tree sample. We matched each speech variety in our data set to its closest correspondent in a data set of automatically generated lexical cognacy and sound class characters (Jäger, 2018). We inferred a phylogeny of the Atlantic languages using RevBayes (Höhna et al., 2016), using a Birth-Death tree prior (Yang & Rannala, 1997) and a General Time-Reversible model of character evolution (Tavaré, 1986). We employed clade constraints, enforcing a split between the North and Bak languages and including the seven higher-order subgroups found in Glottolog (Cangin, Central Atlantic, Fula-Sereer, Jaad, Naluic, Tenda, and Wolof-BKK), ensuring that trees that do not contain these subgroups are assigned zero posterior probability. We run 500,000 iterations of Markov chain Monte Carlo over 4 chains, discarding the first half of samples as burn-in and monitoring convergence by comparing the log posteriors of the chains.

3. Method

We explore differences in the diachronic behavior of noun class marking across the seven domains of marking in our sample via phylogenetic comparative methods. We assume that marking in each domain evolves independently according to a continuous-time Markov process parameterized by two rates, a gain rate α_d and loss rate $\beta_d : d \in \{1, ..., 7\}$. We infer the rates for all features jointly using RStan (Carpenter et al., 2017). We place Gamma(λ, λ) and Gamma(μ, μ) priors over $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively, where $\lambda, \mu \sim$ Exp(1), and incorporate phylogenetic uncertainty by inferring rates for 100 trees from our tree sample and aggregating posterior samples of rates across trees. We use posterior rate values in order to generate a number of quantities of interest to the properties mentioned above, as described below. Code employed in this paper is available at `https://github.com/chundrac/JCoLE2022-atlantic`.

Stationary probabilities of noun class marking We make use of the stationary probability of noun class marking in each marking domain in order to operationalize the LONG-TERM PREFERENCE FOR NOUN CLASS MARKING across different morphosyntactic elements. The stationary probability of a continuous-time Markov chain represents the probability that the system will be in a particular

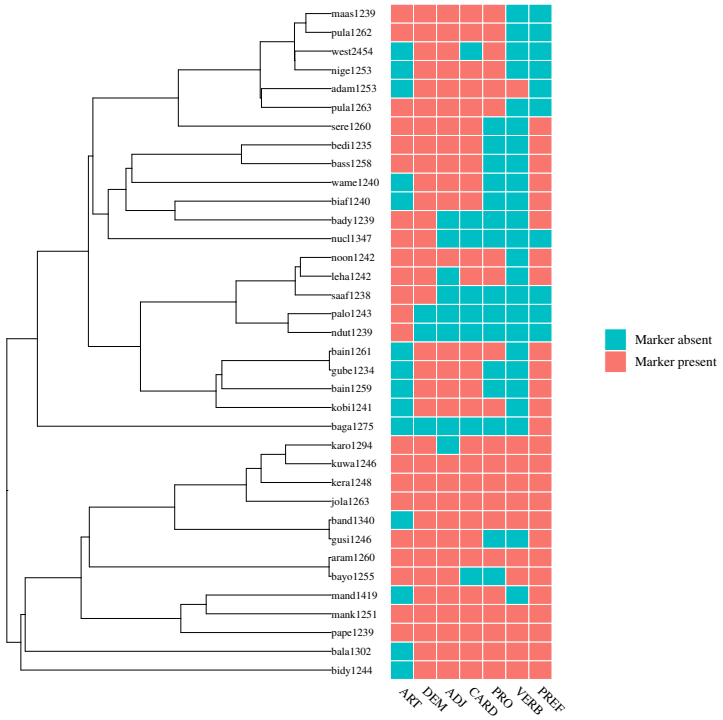

Figure 1. Maximum clade credibility tree of languages in data set, along with data

state as time approaches infinity, as well as the long-term preference of the system for a particular state. For a binary feature with a gain rate α and loss rate β, the stationary probability is equal to $\frac{\alpha}{\alpha+\beta}$. We compute posterior stationary probabilities of noun class marking in each domain from the posterior distribution of rates.

Speed of change Inferring gain and loss rates also allows us to compute the OVERALL SPEED OF CHANGE, irrespective of the direction of change for all posterior samples. Given a gain rate α and a loss rate β, the overall speed of change is $\alpha + \beta$.

Phylogenetic stability We carry out ancestral state reconstruction for the internal nodes of the tree on the basis of the inferred rates. For each pair of rates in the posterior sample, we compute the probability of noun class marking at each node in the tree and draw a Bernoulli variate indicating the presence or absence of noun

class marking. We then average these values, yielding a posterior probability of presence between 0 and 1. On the basis of these reconstruction probabilities, we infer the phylogenetic stability of marking in each domain according to the Bayesian method of Borges et al. (2019), which provides an index of a feature's relative invariance over time, regardless of whether it is preferred or dispreferred. We infer posterior values of the stability metric jointly across marking domains for all trees in the sample.

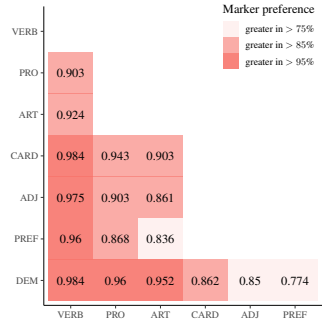

Figure 2. Cross-sample pairwise differences in long-term preference for noun class marking between marking domains

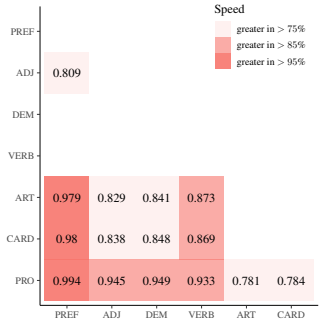

Figure 3. Cross-sample pairwise differences in speed of change between marking domains

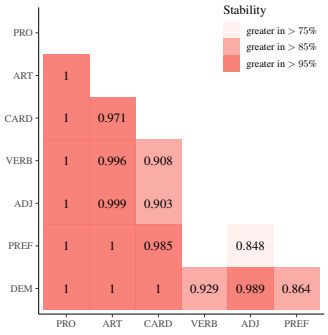

Figure 4. Cross-sample pairwise differences in stability between marking domains, across samples

4. Results

We compare posterior distributions or each metric of interest across pairs of domains. Following Gelman, Hill, and Yajima (2012), feature pairs display deci-

sively contrasting behavior if differences between values in 95% or more of samples are greater/less than zero.[1] The 85% cutoff serves as strong but not decisive evidence. Inter-feature comparisons for long-term noun class marker preference, speed, and stability can be found in Figures 2, 3 and 4, respectively. Features are organized according to their median posterior values, in increasing order.

The following cline can be found with respect to long-term noun class marking: VERB $<$ PRO, ART $<$ CARD, ADJ, PREF, DEM (where $<<$ indicates decisive and $<$ indicates strong evidence for a difference). There is additionally decisive evidence that cardinal numbers, adjectives, prefixes and demonstratives show greater preference for noun class marking than verbs and that demonstratives show greater preference for noun class marking than pronouns and articles.

Prefixes exhibit the lowest speed of change, followed by adjectives, demonstratives, verbs, articles, cardinal numbers and pronouns. Unlike preference for noun class marking, no decisive breaks can be found along this cline. There is decisive evidence that articles, cardinal numbers, and pronouns exhibit faster change than prefixes; additionally, there is strong evidence that pronouns exhibit faster change than adjectives, demonstratives, and verbs, and that articles, cardinal numbers and pronouns exhibit faster change than verbs.

In some ways, results for stability mirror those of speed, with pronouns, articles and cardinal numbers exhibiting lowest stability and highest speed, verbs exhibiting intermediate values for both metrics, and adjectives, prefixes and demonstratives exhibiting highest stability and lowest speed. Here, however, we see decisive and strong evidence for breaks along the cline of stability: PRO $<<$ ART $<<$ CARD, VERB, ADJ $<$ PREF, DEM.

5. Concluding discussion

The high preference and stability for demonstratives is in line with the literature. From a diachronic point of view, demonstratives are frequently the source of grammaticalization processes in Atlantic languages. Demonstratives inflected for class frequently grammaticalize as class pronouns, which then reinforce and extend class agreement (Creissels & Pozdniakov, 2015). The high preference and stability for marking on nouns also dovetails with received wisdom. For example, in noun class languages that lost grammatical agreement, markers are generally still found on nouns (Kießling, 2018). The lower preference for noun class marking on verbs is expected as noun class marking typically starts from within the noun phrase and then extends to the verb phrase (Tang & Her, 2019). Results show that the marking on pronouns is the least stable and most in flux, which reflects that

[1] Debate exists regarding the danger of false discoveries under multiple comparisons in a Bayesian framework. Our hierarchical model has a partial pooling index of 0.67 (Ogle et al., 2019), which corresponds to a low rate of false positives but a high rate of false negatives. A full appraisal of this issue is outside of this paper's scope but will be addressed in future work, e.g., via a mixture model.

the presence of class pronouns is likely to interact with the grammaticalization of demonstratives and other diachronic factors.

The results also indicate that marking on demonstratives, prefixes, and adjectives is the most stable and preferred in comparison to articles, cardinal numbers, pronouns, and verbs. This speaks to the development of noun class marking first in nouns and demonstratives, subsequently spreading to other elements in the noun phrase and finally to the verbs. However, our models do not quantify the complex diachronic interactions between these different markers. For example, are noun class markers most likely to be found on nouns before demonstratives, or vice-versa? In future work, models of correlated evolution can be employed to further assess the interactive dynamics between different noun class markers.

Acknowledgments

The first and second authors acknowledge the support of Sorbonne Nouvelle University and Lacito – UMR 7107, CNRS, and the French National Research Agency grant EVOGRAM (ANR-20-CE27-0021), respectively.

References

Aikhenvald, A. Y. (2016). *How gender shapes the world.* Oxford: Oxford University Press.

Allassonnière-Tang, M., & Dunn, M. (2020). The evolutionary trends of grammatical gender in Indo-Aryan languages. *Language Dynamics and Change*, *Advance articles*, 1–30.

Allassonnière-Tang, M., Lundgren, O., Robbers, M., Cronhamn, S., Larsson, F., Her, O.-S., Hammarström, H., & Carling, G. (2021). Expansion by migration and diffusion by contact is a source to the global diversity of linguistic nominal categorization systems. *Humanities and Social Sciences Communications*, *8*(1), 331.

Borges, R., Machado, J. P., Gomes, C., Rocha, A. P., & Antunes, A. (2019). Measuring phylogenetic signal between categorical traits and phylogenies. *Bioinformatics*, *35*(11), 1862–1869.

Carling, G., & Cathcart, C. (2021). Reconstructing the evolution of Indo-European grammar. *Language*, *97*(3), 561–598.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Corbett, G. G. (1991). *Gender.* Cambridge: Cambridge University Press.

Corbett, G. G. (2007). Gender and noun classes. In T. Shopen (Ed.), *Language typology and syntactic description Volume 3: Grammatical Categories and the Lexicon* (pp. 241–279). Cambridge: Cambridge University Press.

Corbett, G. G. (2013). Number of Genders. In M. S. Dryer & M. Haspel-

math (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Creissels, D., & Pozdniakov, K. I. (Eds.). (2015). *Les classes nominales dans les langues atlantiques.* Köln: Rüdiger Köppe Verlag.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of research on educational effectiveness*, *5*(2), 189–211.

Grinevald, C. (2002). Making sense of nominal classification systems: Noun classifiers and the grammaticalization variable. In I. Wischer & G. Diewald (Eds.), *New reflections on grammaticalization* (pp. 259–275). Amsterdam: John Benjamins.

Grinevald, C., & Seifart, F. (2004). Noun classes in African and Amazonian languages: Towards a comparison. *Linguistic Typology*, *8*(2), 243–285.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., & Ronquist, F. (2016). Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, *65*(4), 726–736.

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Reports*, *5*.

Kemmerer, D. (2017). Categories of object concepts across languages and brains: the relevance of nominal classification systems to cognitive neuroscience. *Language, Cognition and Neuroscience*, *32*(4), 401–424.

Kießling, R. (2018). Niger-Congo numeral classifiers in a diachronic perspective. In W. B. McGregor & S. Wichmann (Eds.), *Diachrony of classification systems.* Amsterdam: John Benjamins.

Ogle, K., Peltier, D., Fell, M., Guo, J., Kropp, H., & Barber, J. (2019). Should we be concerned about multiple comparisons in hierarchical bayesian models? *Methods in Ecology and Evolution*, *10*(4), 553–564.

Segerer, G. (2002). *La langue bijogo de bubaque.* Paris: Peeters.

Seifart, F. (2010). Nominal Classification. *Language and Linguistics Compass*, *4*(8), 719–736.

Tang, M., & Her, O.-S. (2019). Insights on the Greenberg-Sanches-Slobin generalization: Quantitative typological data on classifiers and plural markers. *Folia Linguistica*, *53*(2), 297–331.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, *17*(2), 57–86.

Walchli, B., Olsson, B., & Di Garbo, F. (2020). *Grammatical gender and linguistic complexity, Volume 1: General issues and specific studies.* Language Science Press. (OCLC: 1229599495)

Yang, Z., & Rannala, B. (1997). Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Molecular biology and evolution*, *14*(7), 717–724.