Lara Rüter & Ingo Plag
Heinrich-Heine-Universität Düsseldorf

Assessing different quantitative approaches to measuring the similarity between languages: The case of creoles and non-creoles

Like in other linguistic sub-disciplines, typological and comparative research has seen the successful rise of quantitative methods. In typology and comparative linguistics, statistical methods have been used to assess the similarities and differences between languages. In creole studies, the use of quantitative methods has provided rather solid empirical evidence concerning three questions: whether creoles are structurally different from non-creole languages, whether creoles are less complex than non-creoles, and whether creoles are more similar to their substrate languages than to their superstrate languages (e.g. Bakker et al. 2011, Bakker et al. 2017, Daval-Markussen 2019). Most studies of this kind have used phylogenetic trees and networks, a family of statistical modeling methods originally developed in evolutionary biology.

The relevance and interpretation of this quantitative evidence has led to a debate in which many methodological aspects of the statistical models used have been criticized, with the selection of features, the sampling of languages, and the coding of features being the most prominent ones (e.g. Meakins 2022, Bakker 2023). There is, however, one aspect that has not been critically looked at so far (presumably because most of the critics do not work quantitatively themselves), and which is interesting for both proponents and critics of a quantitative approach. This aspect concerns the question which kinds of statistical models produce which kinds of results. So far, phylogenetic networks have dominated the discussion, and it is unclear and unexplored whether other statistical models yield the same results. Furthermore, phylogenetic trees are not easy to interpret when it comes to the question which features are particularly important in classifying a given set of languages.

In this paper we present a study in which we tested different statistical models on 21 features from two arbitrarily chosen domains ('Word Order' and 'Nominal Categories') as found in two widely-used typological databases, WALS (Dryer & Haspelmath 2013) and APiCS (Michaelis et al. 2013). The list of features and the number of languages per data base for which a given feature is coded are given in table 1. Due to the nature of the databases the number of languages per features varies a great deal, which is a challenge for any statistical analysis.

We implemented phylogenetic networks and compared the results with those yielded by other clustering and classification methods: cluster analysis, classification and regression trees, random forests and generalized linear models. We also explored the models' sensitivity to missing data points and tested different sampling methods (cf. Daval-Markussen 2019 for a similar approach, restricted, however, to phylogenetic networks).

It turns out that the results from different models are not always the same, and that they complement each other nicely, giving new insights into the relationships of the languages under investigation. Phylogenetic trees seem to somewhat overemphasize the similarities among creole languages and play down some of the typological differences to non-creoles. Other statistical models show that the interpretation of the patterning of the observable similarities and dissimilarities can be both more intricate and more illuminating than suggested by eye-balling phylogenetic networks. The results indicate that creoles and non-creoles indeed clearly differ from each other, but, when looked at in more detail, these differences play out as rather complex constellations of particular features. This holds within and across the two domains under investigation. The observed patterns raise new questions about the mechanisms in language contact situations that bring about certain features, but not others.

Our results are theoretically important in four respects. First, we observe, in line with previous quantitative work, that there are remarkable differences between creoles and non-creoles. Second, these differences are not categorical and do not hold across the board, but concern particular features, and the constellations of their values. Third, quantitative typological research should not be restricted to a methodology that uses only a single statistical model. Finally, in view of the first two points, extreme theoretical positions concerning the question of whether creoles are different from non-creoles, or not, should probably be replaced by more nuanced positions that take into account the complexities that can be unearthed by quantitatively analyzing large data sets with different statistical tools.

Table 1: APiCS-WALS features used in this study

Feature	**Domain**	**APiCS**	**WALS**
Order of subject, object, and verb	Word Order	78	1377
Order of possessor and possessum	Word Order	77	1248
Order of adjective and noun	Word Order	76	1366
Order of adposition and noun phrase	Word Order	77	1185
Order of demonstrative and noun	Word Order	79	1223
Order of cardinal numeral and noun	Word Order	76	1154
Order of relative clause and noun	Word Order	76	825
Order of degree word and adjective	Word Order	77	481
Position of interrogative phrases in content questions	Word Order	76	901
Gender distinctions in personal pronouns	Nom. Categ.	80	378
Incl./excl. distinction in independent pers. pronouns	Nom. Categ.	76	200
Politeness distinctions in second-person pronouns	Nom. Categ.	75	207
Indefinite pronouns	Nom. Categ.	75	326
Occurrence of nominal plural markers	Nom. Categ.	79	291
Expression of nominal plural meaning	Nom. Categ.	78	1066
Definite articles	Nom. Categ.	79	620
Indefinite articles	Nom. Categ.	76	534
Pronominal and adnominal demonstratives	Nom. Categ.	78	201
Distance contrasts in demonstratives	Nom. Categ.	75	234
Adnominal distributive numerals	Nom. Categ.	72	251
Ordinal numerals	Nom. Categ.	70	321
Sortal numeral classifiers	Nom. Categ.	76	400

References

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

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

Bakker, P. (2023). Empiricism against imperialism: Science, dogma and the neocolonial heritage of creole studies. Reflections on. *Journal of Pidgin and Creole Languages. https://doi.org/10.1075/jpcl.00119.bak*

Daval-Markussen, A. (2019). *Reconstructing creole* (Doctoral dissertation, PhD Dissertation, Aarhus University).

Dryer, M. S. & Haspelmath, M. (Eds.) 2013. WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at https://wals.info, Accessed on 2024-01-14.)

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

Michaelis, S. M. & Maurer, P. & Haspelmath, M. & Huber, M. (Eds.). (2013). *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://apics-online.info)