**Experimental explorations on the typology of Amazonian languages**

Amazonia is one of the most linguistically diverse regions on earth. With approximately 300 languages linked to at least 50 genetic units (including language families and linguistic isolates) (Epps and Salanova 2013), Amazonia exhibits a unique mosaic-like linguistic face, with a salient degree of genetic diversity. One outcome of such genetic diversity is a good deal of typological diversity. Amazonian languages have indeed significantly contributed to our general understanding of what is and what is not possible for a language.

Nevertheless, one fundamental task in Amazonian linguistics is to identify typological trends that may contribute to provide a unified understanding of Amazonian languages. Various authors have attempted to identify lists of typological features that might define some sort of “Amazonian language” prototype (Derbyshire 1987, Derbyshire & Pullum 1986, Dixon & Aikhenvald 1999). These approaches to Amazonian languages based on lists of general morphosyntactic features have been criticized by more recent studies (see Campbell 2012 and Epps & Michael 2017), which claim that it is not possible indeed to find typological features that may apply to all or almost all Amazonian languages.

In this study, we conduct keyword searches (Hammarström, Her, & Tang 2021) based on the typological features generally attributed to Amazonian languages using the Gramfinder tool (Hammarström 2021), which comprises a full collection consisting of over 37,000 digitized books and articles relating to descriptive linguistics. The most important subset is made up of some 12,000 grammatical descriptions (see Virk et al. 2020). Each typological feature was defined by a list of terms and morphological and capitalization variants of such terms were counted as well. For the curation process, we checked each positive hit to determine whether the term was truly used to characterize the typology of the language under discussion.

Based on the keyword searches, we have identified 117 key terms formulated as True/False features closely related to the description of Amazonian languages. According to our collection, 203 languages of South America have a full grammatical description to conduct searches. With this, each language is simply represented by a vector of True/False values. For each key term (from the set of 117 features), we attempt to ask for (1) its average value across Amazonian (and non-Amazonian) languages in South America; and (2) the importance of the feature in order to define the “Amazonian language” prototype. To answer the first question, we select several Amazonian and non-Amazonian families. With this, we measure whether or not the considered features are attested in the two groups of languages. To answer the second question, the set of variables are used in modeling the prediction of Amazonian versus non-Amazonian languages. To deal with the complex relationships and correlations between the features, we apply Random Forests (RFs) classifiers (Breiman 2001). RFs average many noisy but approximately unbiased statistical models (trees), in order to reduce the error of the predictions. Based on the RFs predictions, we produce a ranked list of features ordered by their importance to characterize an “Amazonian language” prototype.

Our overall results suggest that features often considered as characteristic of Amazonian languages (e.g., noun incorporation, object-inicial orders or classifiers) are indeed attested in less than 50% of our sample, whereas prefixation and liquid sound are attested in more than 90% of our sample. Regarding the second question, the most important result suggests that the 3 top-ranked features are the expression of possession, prefixation, nasalized vowels, nasalisation and ergatives. Interestingly, the 2 top-ranked features (possession and prefixation) are attested in more than 90% of the Amazonian languages of our sample. Our findings also confirm that key terms characterizing Amazonian languages exhibit a large amount of variability as well, making it highly difficult to build a typological prototype for Amazonian languages.

References

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5-32.

Campbell, Lyle. 2012a. Typological characteristics of South American indigenous languages. In Lyle Campbell & Verónica Grondona (eds.), *The indigenous languages of South America: A comprehensive guide*, 259-330. Berlin: de Gruyter Mouton.

Chang, William & Lev Michael. 2014. A relaxed admixture model of language contact. *Language Dynamics and Change* 4(1). 1-26.

Derbyshire, Desmond. 1987. Morphosyntactic areal characteristics of Amazonian languages. *International Journal of American Linguistics* 53. 311-326.

Derbyshire, Desmond & Geoffrey Pullum. 1986. Introduction. In *Handbook of Amazonian Languages*, Vol.1, pp. 1-28. Berlin: Mouton de Gruyter.

Dixon, R.M.W. & Alexandra Y. Aikhenvald. 1999. Introduction. In R. M. W. Dixon & Alexandra Aikhenvald (eds.), *The Amazonian languages*, 1-22. Cambridge: Cambridge University Press.

Epps, Patience & Lev Michael. 2017. The areal linguistics of Amazonia. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 934-963. Cambridge: Cambridge University Press.

Epps, Patience, & Andrés Salanova. 2013. The languages of Amazonia. *Tipití: Journal of the Society for the Anthropology of Lowland South America* *11*(1): 1-28.

Hammarström, Harald. 2021. Gramfinder: Human and machine reading of grammatical descriptions of the languages of the world. Trier, Germany: DBLP.

Manning, Christopher, Raghavan, Prabhakar., and Schütze, Hinrich. 2008. Introduction to Information Retrieval. Cambridge University Press.

Virk, Shafqat Mumtaz, Harald Hammarström, Markus Forsberg, & Søren Wichmann. 2020. The DReaM Corpus: A multilingual annotated corpus of grammars for the world’s languages. In Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, & Stelios Piperidis (eds.), *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, 11–16 May, 871–877. Marseille, France: European Language Resources Association.