Statistical bias control in typology

Matías Guzmán Naranjo & Laura Becker

(Paris, Freiburg)

Introduction Controlling for bias in sampling has mostly focused on ways of including languages from as many genealogical groupings as possible, while ensuring that they are as unrelated as possible. Different methods have been proposed in e.g. Bickel (2008), Dahl (2008), Dryer (1989, 2011), Jaeger et al. (2011), Maslova (2008), Perkins (1989), and Rijkhoff and Bakker (1998). Controlling for areal effects often relies on similar techniques: Choosing a sample of languages which are assumed to have as little contact with each other as possible. Most studies also try to balance the number of languages selected from each macroarea in some way (cf. Jaeger et al. 2011). These methods all face the same issue: The researcher can only include a portion of her data. We present an alternative approach that, using relatively recent statistical methods, can control for the types of biases mentioned without the need to exclude data.

Materials For illustration, we focus on the relation between verb-object orders and the preference for affix position in a language. It has been argued that, while VO orders can occur with both prefixes and suffixes, OV orders show a strong preference against prefixation (e.g. Bybee, Pagliuca, and Perkins 1990; Hawkins and Gilligan 1988; Siewierska and Bakker 1996). We use the datasets from WALS chapters 26 and 83 (Dryer 2013a,b).

Method We fitted a series of Bayesian ordinal models with affix position as the dependent variable (7 levels: strongly suffixing to strongly prefixing) and with verb-object order as the predictor (3 levels: OV, no dominant order, and VO). All models were fitted with Stan (Carpenter et al. 2017) using the brms package (Bürkner 2018) in R. To control for family biases, we included a phylogenetic term (Housworth, Martins, and Lynch 2004) in our regression model. Unlike simple group-level effects, phylogenetic regression can take into account a complete phylogenetic tree, resulting in a gradient representation of genetic relations including all (known) genetic relations between languages in the sample. Assuming that closely-related languages are generally more likely to share a given linguistic feature than more distantly-related languages, the model estimates the effects to be more similar for languages closer in the phylogenetic tree, but less so for languages less close in the tree. For instance, Spanish, French, and Farsi are modeled as related, but with a much closer genetic relation between Spanish and French than between those two languages and Farsi. To control for areal bias, we include latitude and longitude information into our model using a two-dimensional Gaussian Process for each macro-area. This allows us to capture areal effects in a non-linear way across geographical areas, integrating the following issue pointed out by Cysouw, Dediu, and Moran (2012), Dryer (2018), Jaeger et al. (2011), and Rijkhoff, Bakker, et al. (1993): Two languages spoken in areas like Siberia with a distance of 100km may still share properties due to contact, while languages that are spoken 100 km apart in New Guinea are less

likely to have been in contact. Thus, our model can capture that distances across languages (a proxy for quantifying the amount of contact between them) do not have a uniform effect across the globe. Results We compared a model including the three controls described above to a hierarchical model (group-level effects for family and macroarea), and a no-controls model. Our model performed much better than the hierarchical model and the no-controls model in predicting the affix position. With regards to the association between verb-object order and affix position, our model confirmed a very mild effect of verb-object order on the preferred affix position of the language, with most of the variance being accounted for by family and areal effects. In contrast, the group-level effect model and the model without controls strongly overestimated the effect of verb-object order on affixation preferences.

Conclusion Our paper has two main points concerning sampling in typology: Firstly, we show how statistical bias control can offer an alternative to restricting a language sample and excluding otherwise available information. Secondly, our results show that areal bias, even if modeled in a very crude way, is at least as important as genetic bias and should be controlled for in any quantitative typological study exploring crosslinguistic distributions.

References

Bickel, Balthasar. 2008. A Refined Sampling Procedure for Genealogical Control. *Sprachtypologie und Universalienforschung* 61. 221–233.

Bürkner, Paul-Christian. 2018. Advanced Bayesian Multilevel Modeling with the R Package Brms. *The R Journal* 10(1). 395–411.

Bybee, Joan L., William Pagliuca & Revere Perkins. 1990. On the asymmetries in the affixation of grammatical material. In William A. Croft, Suzanne Kemmer & Keith Denning (eds.), *Studies in typology and diachrony*, 1–42. Amsterdam: Benjamins.

Carpenter, Bob et al. 2017. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 76(1). 1–32.

Cysouw, Michael, Dan Dediu, & Steven Moran. 2012. Supporting Online Material for: Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa".

Dahl, Östen. 2008. An exercise in a posteriori language sampling. *Language Typology and Universals* 61(3). 208–220.

Dryer, Matthew S. 1989. Article-Noun Order. *Chicago Linguistic Society* 25. 83–97.

Dryer, Matthew S. 2011. The Evidence for Word Order Correlations. *Linguistic Typology* 15(2). 335–380.

Dryer, Matthew S. 2013a. Order of Object and Verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthro- pology.

Dryer, Matthew S. 2013b. Prefixing vs. Suffixing in Inflectional Morphology. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, Matthew S. 2018. On the Order of Demonstrative, Numeral, Adjective and Noun. In *Languages* 94(4). 798–833.

Hawkins, John A. & Gary Gilligan. 1988. Prefixing and suffixing universals in relation to basic word order. In *Lingua* 74. 219–259.

Housworth, Elizabeth A, Emília P Martins, & Michael Lynch. 2004. The phylogenetic mixed model. *The American Naturalist* 163(1). 84–96.

Jaeger, T. Florian et al. 2011. Mixed Effect Models for Genetic and Areal Dependencies in Linguistic Typology. *Linguistic Typology* 15(2) 281–319.

Maslova, Elena. 2008. Meta-typological distributions. *Language Typology and Universals* 61(3). 199–207.

Perkins, Revere. 1989. Statistical techniques for determining language sample size. *Studies in Language* 13(2). 293–315.

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314.

Rijkhoff, Jan, Dik Bakker, et al. 1993. A method of language sampling. *Studies in Language* 17(1). 169–203.

Siewierska, Anna & Dik Bakker. 1996. The distribution of subject and object agreement and word order type. *Studies in Language* 20. 115–161.