CLUES TO LANGUAGE EVOLUTION FROM A MASSIVE DATASET WITH TYPOLOGY, PHONOLOGY, AND VOCABULARY FROM MANY LANGUAGES

SVERKER JOHANSSON[*1]

[*]Corresponding Author: sja@du.se
[1]Dalarna University, Falun, Sweden

1. Introduction

A major component in the evolution of language is the evolution of the human language capacity, whatever biological endowments humans have that make us language-ready. But the language capacity is not well understood and is difficult to study directly. Clues may come from biases displayed by humans in language acquisition and language change. Even weak underlying biases can lead to strong patterns in the resulting languages (Smith, 2011). Biases can be studied at the individual level in learning experiments (e.g. Culbertson, 2012, Tamariz et al., 2014), but they can also be inferred at the macro level from patterns in the features of natural languages (e.g. Dediu & Ladd, 2007). Biases can be seen either in the synchronic patterns of language features today, or in the diachronic patterns of transition probabilities between features as languages culturally evolve (e.g. Dunn et al, 2011).

Patterns that reveal biases may be found in any aspect of language, e.g. syntax, morphology, phonology, or lexicon, and may be subtle enough to be discernible only in large samples of languages. This work is an exploratory study across the widest possible set of languages, combining typological, phonological, lexical and phylogenetic data on a significant fraction of the languages of the world, with the goal of mapping any biases that may be present. Both synchronic and diachronic patterns are studied, with the emphasis on the latter.

2. Data set

The following data sources are used:

- **Phylogeny and geography**: Ethnologue (Simons & Fennig 2017); ~7,500 languages.
- **Phonological inventories:** PHOIBLE (Moran & McCloy & Wright 2014); ~1,800 languages.
- **Typology:** WALS (Dryer & Haspelmath 2013); ~2,500 languages.
- **Lexicon** (Swadesh lists): Rosetta Project Digital Language Archive (2009); ~1,300 languages.

All four types of data are available for ~300 languages. At least three types are available for ~1,600 languages from 132 different stocks. In order to keep the data set as homogeneous as possible, each type of data has been imported from a single source only. Languages are identified between data sources by their ISO codes.

3. Methods

The language phylogeny from Ethnologue is taken as given in the analysis. For the synchronic analysis, the phylogeny is taken into account in the character statistics by down-weighting multiple "hits" in the same family, in order to control for phylogenetic bias and lineage-specific patterns. Geographic data is also available to control for areal effects. Cross-correlations between different types of characters are analysed for possible patterns.

For the diachronic analysis, the phylogeny together with modern-day character data are used to infer both ancestral character states up the language tree for phonological and typological characters, and transitional probabilities between states (including the probability of characters appearing and disappearing), in a bootstrapping process.

4. Some preliminary results

Well-known typological patterns are reproduced. But correlations between features are observed that go beyond those normally discussed in typology, or those observed by Dunn et al (2011). Interestingly, there are also some modest cross-correlations between grammatical features and phonemes. For example, the presence of aspirated consonants and nasal vowels correlates with certain word-order features, even after controlling for phylogeny.

In the diachronic analysis, there are hints of patterns beyond the obvious one that transition probabilities into common features are larger, but much work remains to be done in the interpretation of these patterns.

References

Culbertson, J. (2012) Typological Universals as Reflections of Biased Learning: Evidence from Artificial Language Learning. *Language and Linguistics Compass 6/5*, 310–329

Dediu, D. & Ladd, D.R. (2007) Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *Proc Nat Acad Sci 104*, 10944-10949.

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2017-08-30.)

Dunn, M., Greenhill, S. J., Levinson, S. C. & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature 473*, 79–82

Moran, Steven & McCloy, Daniel & Wright, Richard (eds.) 2014. *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://phoible.org, Accessed on 2017-08-30.)

Roberts, S & Dediu, D & Levinson, S (2012) *Detecting differences between the languages of Neanderthals and modern humans.* Presented at Evolang 10.

Rosetta Project Digital Language Archive (2009) http://rosettaproject.org/ , https://archive.org/details/rosettaproject

Simons, Gary F. and Charles D. Fennig (eds.). 2017. *Ethnologue: Languages of the World, Twentieth edition*. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Smith, K. (2011) Learning Bias, Cultural Evolution of Language, and the Biological Evolution of the Language Faculty. *Human Biology 83*, 261-278.

Tamariz, M., Ellison, M., Barr, D.J., & Fay, N. (2014) Cultural selection drives the evolution of human communication systems. *Proc. R. Soc. B: Biological Sciences 281*, DOI: 10.1098/rspb.2014.0488.