Oleg Belyaev

Temporal stability of features in Jazyki Mira

September 28, 2008

Abstract

Stability of a language feature reflects its susceptibility to change, i. e., if defined in more rigorous terms, it is the probability that the feature remains unchanged during an arbitrary period of time. The idea that some features are more prone to mutate due to areal contact than others is not a new one and has been elaborated on, among others, by Edward Sapir (Sapir 1921) and, later, by Joseph Greenberg (Greenberg 1978). (Nichols 1992) was, however, the first study to propose precise mathematical metrics for calculating the parameter of stability and applying them to a relatively wide range of the world's languages.

New large typological databases like WALS in its digital form (Haspelmath et al. 2008) provide new opportunities for conducting similar quantitative research. A recent study notable in this regard is (Wichmann & Holman n. d.), where a metric similar in spirit to Nichols' metrics is applied to the data of the WALS database, making it the first such effort to use the data of thousands of different languages.

Jazyki Mira ('Languages of the World', JM) is another large database analogous to WALS which has been developed in the Institute of Linguistics of the Russian Academy of Sciences starting from the mid-1980s (cf. (Polyakov & Solovyev 2006) for a full description). It is mostly limited in scope to Eurasia and consequently contains only 318 languages. JM's advantage is, however, in its detail, both in terms of the representation of different local genera and in terms of the feature set: JM contains 3821 binary features on most aspects of grammar, and each language is supposed to be exhaustively described by them. A calculation of stability conducted on JM's data would not only be useful to double-check the results received on WALS and to verify if stability is tied to linguistic areas, but would also serve as a plausible means of comparison of the two databases' performance.

Since (Wichmann & Holman n. d.) is, to the author's knowledge, the only quantitative approach to this problem to be based on WALS or any other typological database, it seemed natural to apply the same algorithm to JM, using the same classification of languages and comparing the results for features with more-or-less reliable correspondences.

The results that have been received to this moment are mixed. The first problem was establishing corresponding features for WALS and JM, which is a complicated undertaking in its own right. Even for the relatively small set of 42 most reliable pairs absolute values for stabilities do not seem to correlate in any plausible way. On the other hand, approximate evaluation of these values based on a four-way division of the whole range of percentages (as used in

(Wichmann & Holman n. d.)) shows a much better correspondence: 23 (∼55%) fall into the same categories, 13 (∼31%) – into adjacent ones, and only 6 (∼14%) have no correlation at all. This means that the correspondences between stabilities are acceptable for 86% of all feature pairs. When comparing with statements in the literature (as it is done in the original paper), the results are even better, with only 2 pairs with uncertain correlation out of 13. This seems to bring us to the conclusion that while the databases are quite different in data structure and scope, they do often reflect the same typological realities and are suitable for similar objectives with similar final results. Among other things, this means that one of them can be used for double-checking the results gained on the other one.

Another important observation can be drawn from comparing the frequency distribution of JM stabilites with the distribution of WALS stabilities (for features decomposed into binary form). Both of them resemble Gaussian distributions and both have the mean at about 0.2. Generally speaking, the most stable features seem to constitute a relatively small 'core' of language.

A possible way of testing the plausibility of JM's results would be to conduct a simulation similar to the one conducted on the WALS results, but another possibility is using stability values as weights while calculating typological (dis)similarity of languages. In principle, more stable features should better reflect genetic relationships, while the unstable ones should show more evidence of areal contact. Such a result would demonstrate the metric's plausibility in practice.

The author would like to thank Søren Wichmann, Eric W. Holman, Vladimir Polyakov, Valery Solovyev and Dmitry Egorov for their helpful comments, suggestions and collaboration on this project.

References

[Greenberg 1978] Greenberg, Joseph H. *Diachrony, synchrony and language universals.* Universals of Human Language, Vol. III: Word Structure, ed. by Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik, 47–82. Stanford: Stanford University Press, 1978.

[Haspelmath et al. 2008] Haspelmath, Martin, Matthew S. Dryer, David Gil and Bernard Comrie (eds.) *The World Atlas of Language Structures Online.* Munich: Max Planck Digital Library, 2008. Available online at http://wals.info .

[Nichols 1992] Nichols, Johanna. *Linguistic Diversity in Space and Time.* Chicago: The University of Chicago Press, 1992.

[Polyakov & Solovyev 2006] Polyakov, Vladimir N. Valery D. Solovyev. *Компьютерные модели и методы в типологии и компаративистике* (Computational Models and Methods in Typology and Comparative Linguistics). Kazan: Kazanskij Gosudarstvennyj Universitet, 2006.

[Sapir 1921] Sapir, Edward. *Language: An introduction to the study of speech.* New York: Harcourt, Brace and company, 1921.

[Wichmann & Holman n. d.] Wichmann, Søren and Eric W. Holman: *Assessing temporal stability for linguistic typological features.*, pending publication, available online: http://email.eva.mpg.de/w̃ichmann/WichmannHolmanIniSubmit.pdf .