A language sample for a typological study of caritives

In our talk, we will present a language sample that was developed for our typological study of caritives and will describe its underlying principles and the process of its development. The typological study of caritives aims at revealing the correlations between morphological, syntactic, and semantic parameters of the different means of expressing caritive semantics. The sample should be representative of the entire set of the world languages. For example, if 90% of the world languages lack a dedicated grammatical marker for caritive, we expect it to be absent in 90% of languages in our sample and vice versa. So, we seek to make a language sample that would be representative in terms of genetic and areal diversity (while typological bias remains in the sample because we cannot predict *a priori* which linguistic parameters of the languages are important for our study).

The size of the sample is decided to be between 100 and 200 languages (the specific size is not so important as the proportion of languages from every area / family).

The surface of the Earth was divided into 6 macroareas, according to (Dryer 1992): Africa, Eurasia, Southeast Asia and Oceania, Australia and New Guinea, North America, and South America. The proportion of languages taken from each macroarea depends on the amount of separate language families in that macroarea and on the size of these families. To balance the size of different families, we used the so-called *genera* introduced in Dryer (1989). The genera are maximal subgroups of one family which are supposed to have split no earlier than 4000 years ago. Our list of (supposedly) all genera in the world was based on the list of all genera featured in WALS (https://wals.info/languoid/genealogy), with enrichments from Ethnologue (https://www.ethnologue.com/). We decided from the beginning to exclude creoles, pidgins and sign languages as well as extinct languages and some cases with doubtful status. In total, we arrived at 660 genera. Then we calculated the amount of languages that needed to be taken from each macroarea: it was made proportionate to the total number of genera in that macroarea.

Finally, we selected languages for the sample. The choice of languages is based on several principles: we could use no more than one language per genus, the amount of languages per large families was proportional to the size of the family. We also took into account such additional factors as existence of collections of glossed texts, high-quality modern grammar descriptions, comprehensive dictionaries providing sentence examples, existence of Bible translations and the areal (geographical) balance. We also took into account whether the language was included in the 200-languages sample of WALS.

References

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13. 257–292.
Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68. 81–138.