Corpora and language universals: Opportunities and challenges

Natalia Levshina

Max Planck Institute for Psycholinguistics, Nijmegen

natalia.levshina@mpi.nl

Donnerstag,
05.03.2020
11:45–12:15
ESA1 HG HS M

This paper discusses the role of corpus data in testing and explaining language universals. First of all, corpora can fill in existing gaps in grammatical descriptions and provide missing information. For example, Stolz et al. (2017) extract interrogative spatial pronouns from numerous translations of Le Petit Prince. Second, by using corpus data, one can avoid the existing bias towards a restricted set of linguistic patterns, which display cross-linguistic bimodal distributions with low language-internal variability (Wälchli 2009), and investigate universal patterns in a broad range of constructions. Third, corpora can be used to fine-tune existing universals and reformulate them with greater precision. Here, we will demonstrate how Greenberg's (1963) Universal 25 can be reformulated at the finer-grained level of intralinguistic probabilities instead of the coarse-grained categorical variables. Fourth, corpora are indispensable for providing functional explanations of language universals, which emerge due to various communicative and cognitive pressures, such as the formal asymmetries in marking of causal and non-causal events (Haspelmath et al. 2014) or the cross-linguistic preferences for particular ordering of syntactic constituents (Hawkins 1994). Fifth, some important universals are inherently usage-based, since they are formulated at the level of usage events and describe probabilistic tendencies within a language, e.g. Zipf's law of abbreviation (Zipf 1935; Bentz & Ferrer-i-Cancho 2015), or the correlation between average surprisal and word length (Piantadosi et al. 2011). Finally, one needs corpora in order to establish universals related to human interaction in contex (e.g. Dingemanse et al. 2013).

At the same time, the use of corpus data is accompanied by several challenges, such as the Indo-European bias, difficulties in extraction of semantic and pragmatic information, lack of stylistic and pragmatic diversity in most multilingual corpora, and low frequencies of some linguistic phenomena.

References: Bentz, Ch. & R. Ferrer-i-Cancho (2015). Zipf's law of abbreviation as a language universal. Capturing Phylogen etic Algorithms for Linguistics. Lorentz Center Workshop, Leiden, October 2015. Dingemanse M., F. Torreira, N.J. Enfield (2013). Is ''Huh?'' a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. PLoS ONE 8(11): e78273. Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, J (ed.), Universals of Human Language. Cambridge, Mass.: MIT Press, 73–113. Haspelmath, M., A. Calude, M. Spagnol, H. Narrog, E. Bamyacı (2014). Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. Journal of Linguistics 50(3), 587–625. Hawkins, J. A. (1994). A Performance Theory of Order and Constituency. (Cambridge Studies in Linguistics, 73.) Cambridge: Cambridge University Press. Piantadosi, Steven, H. Tily & E. Gibson (2011). Word lengths are optimized for efficient communication, Proceedings of the National Academy of Sciences 108(9), 3526. Stolz, Th. et al. (2017). Spatial Interrogatives in Europe and Beyond: Where, Whither, Whence. Berlin: De Gruyter Mouton. Wälchli, B. (2009). Data reduction typology and the bimodal distribution bias. Linguistic Typology 13, 77–94. Zipf, G. (1935). The Psychobiology of Language: An Introduction to Dynamic Philology. Cambridge, MA: M.I.T. Press.