Corpus-based typology: spoken language from a cross-linguistic perspective

Geoffrey Haig, Stefan Schnell
Universität Bamberg
geoffrey.haig@uni-bamberg.de, stefan.schnell@uni-bamberg.de

Raum: Von-Melle-Park 5 (VMP5) 2101/2105

Workshop description

Linguistic typology has traditionally taken the „language" as a unit of comparison, and compared these units on the basis of features extracted from grammatical descriptions. A complementary approach is corpus-based or token-based typology, an emergent field of comparative linguistics that involves harnessing recent developments in corpus linguistics and variationist sociolinguistics to cross-linguistic data and that deals with probabilistic generalizations drawn from observed language usage, as recorded in corpora. Its object of study is a population of utterances, rather than languages as holistic artefacts (cf. Wälchli 2009). This approach to language typology is currently undergoing a major upsurge, fueled by the growing availability of digital corpora from typologically diverse languages, and increasingly sophisticated statistical modelling (see among many others Wälchli 2009, Haig & Schnell 2016, Dingemanse et al 2015, Levshina 2019). While a growing body of research drawing on written corpus data has become increasingly influential in linguistic typology (Haspelmath et al. 2014, see esp. the cross-linguistic Universal Treebank initiative, http://universaldependencies.org/, and Levshina 2019 for recent application to classic issues in language typology), in this workshop we are interested in specific properties of spoken language as the ontologically primary type of linguistic performance, under consideration of a maximally diverse sample of languages. Topics covered include prosodic structuring and partitioning, speech rate, interactivity and intersubjectivity, universals of discourse, and corpus-based approaches process to first-language acquisition. We will also attend to methodological challenges, in particular issues of annotation and data formats.

References: Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., & Enfield, N. J. (2015). Universal Principles in the Repair of Communication Problems. PLoS One, 10(9): e0136100. doi:10.1371/journal.pone.0136100. Du Bois, John (1987). The discourse basis of ergativity. Language 63, 805–855. Haig, Geoffrey & Schnell, Stefan (2016). The discourse basis of ergativity revisited. Language 91(3), 591–618 (DOI: 10.1353/lan.2016.0049/). Himmelmann, Nikolaus P. (2014). Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. Language 90(4), 927–960 (DOI: 10.1353/lan.2014.0105/). Haspelmath, Martin, Calude, Andreea, Spagnol, Michael, Narrog, Heiko & Bamyacı, Elif (2014). Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. Journal of Linguistics 50(3), 587–625 (DOI: 10.1017/S0022226714000255). Levshina, Natalia (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. Linguistic Typology, 23(1), 533–572. Wälchli, Bernhard (2009). Data reduction typology and the bimodal distribution bias. Linguistic Typology 13, 77–94.