On potential statistical universals of grammar in discourse: evidence from Multi-CAST

Geoffrey Haig, Stefan Schnell

University of Bamberg

geoffrey.haig@uni-bamberg.de, stefan.schnell@uni-bamberg.de

In contributing to the current upsurge in corpus (or token) based typology (cf. Wälchli 2009, Levshina 2019), this presentation summarizes a set of cross-linguistically robust regularities in the grammar of discourse, drawing on natural (non-scripted, non-elicited) spoken language corpora from 12 languages (Haig & Schnell 2019). We focus on two domains: (1) Preferred Argument Structure and Referent Introduction (Du Bois 1987, Newmeyer 2003: 686). While our findings largely confirm the overall dispreference for transitive subject to be filled by lexical NP's (Haig & Schnell 2016) with new referents, we relate these findings to more general principles following from a convergence of humanness and topicality leading to low rates of new, lexical human subjects (Haig & Schnell 2016: 612–613) whether transitive or intransitive. Moreover, Referent Introduction does not pose specific challenges to discourse processing and hence do not map isomorphically onto morphosyntax, which actually appears to be solely determined by content. (2) The factors determining referential choice in discourse (cf. Ariel 1990, Torres Cacoullos & Travis 2019). Here we find that different referential expressions (zero, pronoun, NP) do not in fact line up along a continuum of decreasing accessibility (Ariel 1990): the choice between zero and pronoun, on the one hand, displays language-specific idiosyncrasies that cannot be satisfactorily accounted for in terms of accessibility or processing costs. In regards to the deployment of lexical NPs, however, we find less evidence for the claim that languages vary fundamentally in the density of information carried by discourse (e.g. Huang 2000: 262, Bickel 2003: 710); in fact we observe surprisingly little variability in the overall rate of lexical NP's across our sample (Vollmer 2019), suggesting that the undeniable differences in the rates of pronoun versus zero expression is balanced by a relatively constant rate of lexical NP's in connected discourse, providing a bedrock of referential information around which typological variation plays out.

References: Ariel, Mira. (²2014). Accessing noun-phrase antecedents. London, New York: Routledge. Bickel, Balthasar (2003). Referential Density in Discourse and Syntactic Typology. Language 79(4), 708–736. Du Bois, John (1987). The discourse basis of ergativity. Language 63, 805–855. Haig, Geoffrey & Stefan Schnell (2016). The discourse basis of ergativity revisited. Language 92(3), 591–618. Haig, G. & Stefan Schnell (eds.) (2019). Multi-CAST: Multilingual corpus of annotated spoken texts. (Online resource available at: multicast.aspra.uni-bamberg.de/) (accessed 15.11.2019). Levshina, Natalia (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. Linguistic Typology, 23(1), 533–572. Newmeyer, Frederick (2003). Grammar Is Grammar and Usage Is Usage. Language 79(4), 682–707. Vollmer, Maria (2019). How radical is pro-drop in Mandarin? A quantitative corpus study on referential choice in Mandarin Chinese. MA-thesis, University of Bamberg. Wälchli, Bernhard (2009). Data reduction typology and the bimodal distribution bias. Linguistic Typology, 13(1), 77–94. Huang, Yan (2000). Anaphora. Cambridge: CUP. Torres Cacoullos & Travis (2019). Variationist typology: shared probabilistic constraints across (non-)null subject languages. Linguistics 57(3), 653–692.

AG 10