

Language Documentation & Linguistic Theory 2

From text to typology: towards implementing quantitative typology on corpora from endangered languages

Geoffrey Haig, *University of Kiel*

Stefan Schnell, *University of Kiel*

From text to typology: towards implementing quantitative typology on corpora from endangered languages.

Typology still predominantly relies on data from grammars, that is, data that are pre-analysed and to some extent selective. However, there is a growing recognition that beyond the traditional typological generalizations, based on the mere presence or absence of abstract structures in a given languages, a cross-linguistic bedrock of commonalities also exists at the level of statistically significant distributions of grammatical features across discourse. Pioneering work in this area has been undertaken by Du Bois (1987, 2003) and more recently Bickel (2003). These authors have uncovered regularities based on the cross-linguistic comparison of corpora of naturally spoken language, for example in the way grammatical relations are distributed across pragmatic properties such as definiteness (Du Bois 1987), or the way different grammatical relations are realized according to a lexical vs. pronominal coding distinction (Bickel 2003). Similar effects have been discussed in the literature, for example, how animacy interacts with grammatical relations (cf. contributions in Lamers et al. 2008).

The current global initiatives for language documentation (e.g. DoBeS, HRELP, ELF) have, (among many other things), produced an unprecedented number of corpora of transcribed spoken language, often from small and typologically unusual language communities. These texts provide a rich resource for text-based, as opposed to grammar-based typology. In this paper, we present a one-tier model for morpho-syntactic annotation that can be applied to a large number of typologically diverse languages. The system aims at reaching a compromise between the basic annotation system of Du Bois (1987), and the more complex tagging systems currently used in tagging corpora of standardized languages (e.g. the tagging system employed by the British National Corpus). It is intended to be simple and flexible enough to accommodate languages of any type, but rich enough to facilitate cross-linguistic text-based typological research on grammatical relations, animacy, and referential density. The system has been trialed on corpora from two languages, Gorani (West Iranian) and Vera'a (Austronesian, Oceanic) and is intended to be extended to further languages from other documentation projects.

References:

- Bickel, B. 2003. Referential density in discourse and syntactic typology. *Language* 79, 708 – 736.
- Du Bois, J. 1987. The discourse basis of ergativity. *Language* 63, 805 - 855.
- Du Bois, J. . 2003. *Preferred argument structure: grammar as architecture for function*. Amsterdam.
- Lamers, M. / Lestrade, S. / P. de Swart (eds.). 2008. Animacy, Argument Structure, and Argument Encoding. Special Edition of *Lingua* 118.2, 131 – 260.