FOLK taxonomies:
Text type stratification by POS n-grams in a monolingual conversation corpus

Comparing the relative frequencies of different parts of speech is not only an interesting avenue for crosslinguistic research, but also for language-internal text typology (Biber 1988, Rayson et al. 2002). The present paper explores the potential of using differently sized POS n-grams for recognising/inducing substrata in a monolingual conversation corpus, FOLK. FOLK (Forschungs- und Lehrkorpus Gesprochenes Deutsch, Deppermann & Hartung 2011) is the Research and Teaching Corpus of Spoken German, a large, stratified collection of recordings of authentic spoken interaction and their transcriptions. The aim of FOLK is to cover a broad spectrum both in terms of regional variation and in terms of different interaction types. In its first publicly available version, FOLK comprises 70 hours of transcribed audio recordings, published via the Database of Spoken German (DGD2).
We begin by comparing simple POS distributions (unigrams) in FOLK and LIMAS (Glas 1975), a ‚balanced' corpus of written German. We then turn to the internal structure of FOLK and compare POS ratios across (possible) substrata on different taxonomic levels. In the next step, grain size is increased to POS bigrams. Gries (2010) clusters collocation strength values of all lexical bigrams in all registers and sub-registers of the BNC Baby corpus and finds that the analysis recreates the register structure of the corpus perfectly. We replicate Gries' analysis with our German data (and their much freer word order) and compare the results for POS bigrams to those obtained for both lexical bigrams and POS unigrams. In the last step, grain size is varied flexibly to extend the analysis to larger n-grams whose mean attraction (across constitutive bigrams) is maximised compared to strings with n-1 and n+1 (cf. Gries & Mukherjee 2010). We close with a discussion of the potential and the limitations of the approach for inductive stratifications of monolingual conversation corpora.

References

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Deppermann, A. & M. Hartung. 2011. Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des „Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim). In: E. Felder, M. Müller & F. Vogel, eds. *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen.* Berlin/New York: de Gruyter, pp. 414-450.

Glas, R. 1975. Das LIMAS-Korpus, ein Textkorpus für die deutsche Gegenwartssprache. *Linguistische Berichte* 40, 63–66.

Gries, S. T. 2010. Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. In *Proceedings of Corpus Linguistics 2009*, University of Liverpool.

Gries, S. T. & J. Mukherjee. 2009. Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes. *International Journal of Corpus Linguistics* 15, 4, 520–548.

Rayson, P., A. Wilson & G. Leech. 2002. Grammatical word class variation within the British National Corpus sampler. In: P. Peters, P. Collins & A. Smith, eds. *New frontiers of corpus research: Papers from the Twenty first International Conference on English Language Research on Computerized Corpora.* Amsterdam: Rodopi, pp. 295–306.