Assessing and comparing corpus composition using lexico-grammatical features

Felix Bildhauer[1], Roland Schäfer[2]
[1]IDS Mannheim, [2]Freie Universität Berlin
bildhauer@ids-mannheim.de, roland.schaefer@fu-berlin.de

Many large corpora are built not by sampling texts or documents according to a pre-defined sampling scheme (like the BNC) but by acquiring huge unstructured collections of textual data from available sources. The Corpora from the Web (COW; Schäfer and Bildhauer 2012) and the WaCky corpora (Baroni et al. 2009) are examples of such corpora. Little is known about their composition in terms of genres, registers, or even basic lexico-grammatical properties. However, many believe that distributions of linguistic phenomena depend on registers or similar categories. Assessing the composition of corpora and annotating documents with appropriate categories is thus vital.
Methods of comparing corpora and methods of assessing the homogeneity of corpora have emerged from the web corpus construction scene (e.g., Kilgarriff 2012; Ciamarita and Baroni 2006). Such methods can be used in order to assess how similar a corpus is to a corpus with a known composition, and how similar sub-corpora of a corpus are to each other. They are often based on simple word or lexeme frequencies. We present an approach for German corpora using many automatically extracted lexico-grammatical features (frequencies of function words, morphological categories, syntactic constructions, etc.). Using our COReX feature-extractor, we compare the DECOW16B web corpus with the German reference corpus DeReKo. We also measure the difference in the distribution of lexico-grammatical features between forum documents and non-forum documents in DECOW16B, as the divide between standard and non-standard written German overlaps with the forum/non-forum distinction. We also compare DECOW16B to the RandyCOW web corpus constructed using a bias-free web-crawling method (Schäfer 2016). Finally, we assess the homogeneity of all of these corpora.
The results show that there are relevant differences in the distributions of lexico-grammatical features in most of the comparisons described above. The interpretation of those differences proves to be a much more intricate problem, with the exception of a broad distinction between colloquial, less standard, dialogic texts vs. more standard narrative or reporting texts.

References: ● Baroni, M. et al. (2009): The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation 43(3), 209–226. ● Ciamarita, M. and M. Baroni (2006): Measuring web-corpus randomness: A progress report. Baroni, M. and S. Bernardini, Wacky! Working papers on the Web as Corpus. Bologna: GEDIT. ● Kilgarriff, A. (2012): Getting to know your corpus. Sojka, P. et al., Proceedings of Text, Speech and Dialogue – 15th International Conference, 3–15. Heidelberg: Springer.