The taming of the data:
Using text mining in building a corpus for diachronic analysis

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, Noam Ordan and Elke Teich

Universität des Saarlandes, Saarbrücken, Germany

Social and historical linguistic studies profit from corpora encoding contextual meta-data (time, register/genre, social group etc) and relevant structural information (e.g. document structure). While small, hand-crafted corpora control over selected contextual variables (e.g. the Brown/LOB corpora encoding variety, register and time) and are readily usable for analysis, big data (such as Google or Microsoft n-grams) are typically poorly contextualized and considered of limited value for linguistic analysis (see however Lieberman et al. (2007)). Similarly, when we compile new corpora, sources may not contain all relevant meta-data and structural data (see e.g. the Old Bailey sources vs. richly annotated corpus (Huber, 2007)).

For corpora with rich meta-data and structural data, we can draw on well-established methods of analysis, ranging from descriptive statistics to machine learning (see e.g. Kilgarriff (2004) for an overview). For analysis of corpora with few/no meta-data or structural data, we first need to learn more about our data. Relevant methods are found in data mining (Witten et al., 2011) which is concerned with detecting patterns in complex, unfamiliar and potentially noisy data sets. This is exactly the situation we have when building a corpus from uncharted material.

We are currently building a corpus from the Philosophical Transactions and Proceedings of the Royal Society of London, covering the first two centuries (1665-1870) of publication (Khamis et al., to appear). The sources (obtained from JSTOR) contain some but not all relevant meta-data (year of publication and authors but not discipline) and no structural data. We apply a combination of pattern-based techniques and text mining methods (e.g. clustering, classification, topic modeling) to explore the data. Apart from understanding our data better and (semi-)automatically enriching it with relevant contextual and structural information, we obtain positive effects regarding data quality (detection of artifacts, such as OCR errors, text duplicates, running headers/footers).

(298 words without title, authors, references, acknowledgments)

Acknowledgements: This research is funded by the German Research Foundation (*Deutsche Forschungsgemeinschaft*) under grants SFB 1102: *Information Density and Linguistic Encoding* (http://www.sfb1102.uni-saarland.de/) and EXC-MMCI: *Multimodal Computing and Interaction* (www.mmci.uni-saarland.de/). We are also indebted to Peter Fankhauser (IdS Mannheim) for his continuous support in questions of data analysis.

References

Huber, M., 2007. "The Old Bailey Proceedings, 1674-1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English". Meurman-Solin, Anneli and Nurmi, Arja (eds.) *Annotating Variation and Change* (Studies in Variation, Contacts and Change in English 1), University of Helsinki: Department of English.

Khamis, A., S. Degaetano-Ortlieb, H. Kermes, N. Ordan and E. Teich, to appear. A resource for the diachronic study of scientific English: Introducing the Royal Society Corpus. 8th International Corpus Linguistics Conference, Lancaster, UK, July 2015.

Kilgarriff, A., 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6 (1): 1-37.

Lieberman E., J.-B. Michel, J. Jackson, T. Tang and M. Nowak, 2007. Quantifying the evolutionary dynamics of language. *Nature* 449:713-716.

Witten, I. H., F. Eibe, M. A. Hall, 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 3rd edition.