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

(Saarland University, Saarbrücken, Germany)

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

Social and historical linguistic studies benefit from corpora encoding contextual metadata (e.g. time, register, genre) 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 (e.g. 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 metadata and structural data (e.g. the Old Bailey sources vs. the richly annotated corpus in Huber 2007).

For corpora with rich metadata and structural data, we can draw on well-established methods of analysis, from descriptive statistics to machine learning (see e.g. Kilgarriff 2004 for an overview). For the analysis of corpora with few/no metadata or structural information, 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 and potentially noisy datasets. This is what 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 (see e.g. Atkinson 1998; Taavitsainen et al. 2010), covering the first two centuries (1665–1869) of publication (Khamis et al. 2015). The sources (obtained from JSTOR) contain some but not all relevant metadata (year of publication and author, 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 like OCR errors, text duplicates, and running headers/footers).

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

ATKINSON, DWIGHT. 1998. *Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. New York: Routledge.

HUBER, MAGNUS. 2007. The Old Bailey Proceedings, 1674–1834: Evaluating and annotating a corpus of 18th- and 19th-century spoken English. *Annotating Variation and Change* (Studies in Variation, Contacts and Change in English 1), ed. by ANNELI MEURMAN-SOLIN and ARJA NURMI, n.p. Online: http://www.helsinki.fi/varieng/series/volumes/01/huber/. Helsinki: University of Helsinki, Department of English.

KHAMIS, ASHRAF; STEFANIA DEGAETANO-ORTLIEB; HANNAH KERMES; NOAM ORDAN; and ELKE TEICH. 2015. A resource for the diachronic study of scientific English: Introducing the Royal Society Corpus. *The 8th International Corpus Linguistics Conference (CL)*, Jul 21–24, 2015. Lancaster University, Lancaster, UK.

KILGARRIFF, ADAM. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1).1–37.

LIEBERMAN, EREZ; JEAN-BAPTISTE MICHEL; JOE JACKSON; TINA TANG; and MARTIN A. NOWAK. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449.713–6.

TAAVITSAINEN, IRMA; PETER M. JONES; PÄIVI PAHTA; TURO HILTUNEN; VILLE MARTTILA; MAURA RATIA; CARLA SUHR; and JUKKA TYRKKÖ. 2011. Medical texts in 1500–1700 and the corpus of Early Modern English Medical Texts. *Medical writing in Early Modern English*, ed. by IRMA TAAVITSAINEN and PÄIVI PAHTA, 9–29. Cambridge: Cambridge University Press.

WITTEN, IAN H.; EIBE FRANK; and MARK A. HALL. 2011. *Data mining: Practical machine learning tools and techniques*, 3rd edn. Burlington, MA: Morgan Kaufmann.