The Royal Society Corpus: Towards a high-quality corpus for studying diachronic variation in scientific writing

Big data are a potential source for quantitative research in the humanities, but typically they do not contain all relevant contextual meta-data (time, register/genre, social group, author, etc.) to be readily usable for social, historical or philological studies (cf. Schöch, 2013). Small corpora, in contrast, are typically carefully hand-crafted and provide rich meta-data as well as structural and linguistic data, but the application of data-driven analysis techniques is impeded by their small size.

We introduce a diachronic corpus of English scientific writing - the Royal Society Corpus (RSC) - adopting a middle ground between big and 'poor' and small and 'rich' data. The corpus has been built from an electronic version of the Transactions and Proceedings of the Royal Society of London and comprises c. 35 million tokens from the period 1665-1869 (see Table 1). The motivation for building a corpus from this material is to investigate the diachronic development of written scientific English.

Journal	Period	Text type				
		Book reviews	Articles	Miscellaneous	Obituaries	Total
Philosophical Transactions	1665–1678	124	641	154	–	919
Philosophical Transactions	1683–1775	154	3,903	338	–	4,395
Philosophical Transactions of the Royal Society of London (PTRSL)	1776–1869	–	2,531	283	–	2,814
Abstracts of Papers Printed in PTRSL	1800–1842	–	1,316	15	–	1,331
Abstracts of Papers Communicated to RSL	1843–1861	–	429	5	–	434
Proceedings of RSL	1862–1869	–	1,476	38	14	1,528
Total		278	10,296	833	14	11,421

Table 1: Material used for the RSC

In terms of corpus building (see Figure 1 for a schematic overview), the sources for the RSC were obtained from JSTOR and include some but not all relevant meta-data (year of publication and authors, but not disciplines), structural data is partial and erroneous (e.g. scrambled pages, text duplicates), and the base text contains OCR errors. To move towards a cleaner and richer version of the corpus, an approach is needed that allows obtaining good-quality base-text data and relevant meta-data as well as structural and linguistic data with affordable effort. For this purpose, we use a combination of pattern-based techniques (e.g. by adapting the patterns for OCR corrections made available by Underwood and Auvil)[1] and data-mining methods (e.g. topic modeling (Blei et al., 2003) to approximate disciplines; cf. McFarland et al. (2013) for an overview of types of topic models applied to capture differentiation in scientific language). Additionally, to enrich the RSC with basic linguistic annotations, we build on existing tools adapting them to the diachronic material. For normalization we use VARD (Baron & Rayson, 2008) with a model we trained on a manually normalized subset of the RSC, and for tokenization, lemmatization, segmentation and part-of-speech annotation we use TreeTagger (Schmid, 1994) on the normalized texts. Moreover, experience taught us that issues in data quality are detected at all stages of corpus handling, from pre-processing to analysis. Thus, we keep the corpus-building process as modular and automatic as possible, allowing continuous improvement in data quality whenever an issue is encountered and applying manual work before the first automatic step.

In terms of analysis, our main assumption is that due to specialization, scientific texts exhibit greater encoding density over time (Halliday & Martin, 1993), resulting in a specific discourse type characterized by high information density (Crocker et al., 2015) that is functional for expert communication (but rather inaccessible to lay persons). Linguistically, this may be reflected in lexical compression (e.g. compounding, derivation) and syntactic reduction (e.g. relativizer omission, contractions). For instance, there is evidence from the Thesaurus of the OED (Oxford English Dictionary)[2] that affixation rises considerably as a means of word formation in scientific texts in the mid-17th century. For the identification

[1] http://usesofscale.com/gritty-details/basic-ocr-correction/

[2] http://www.oed.com/thesaurus/

of further linguistic features possibly involved in denser encoding, we draw, on the one hand, on existing literature (e.g. Harris, 1991) and, on the other hand, on exploratory data-mining techniques (e.g. pattern mining as in Vreeken, 2010). In the poster, we show the corpus-building process and selected analyses of diachronic development in the RSC with dedicated visualizations (Fankhauser et al., 2014).

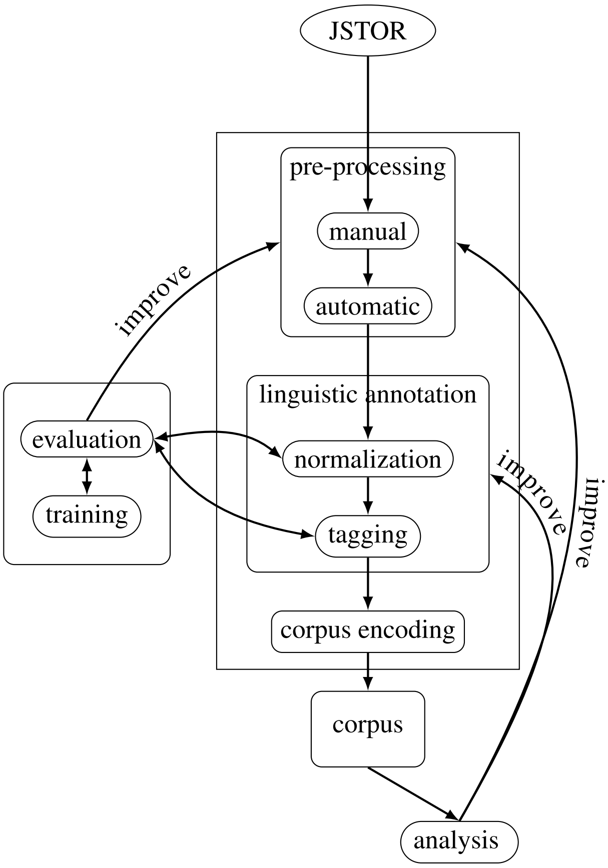

Figure 1: Corpus-building steps

References

Baron, A. & Rayson, P. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham, UK.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Crocker, M. W., Demberg, V., & Teich, E. 2015. Information density and linguistic encoding (IDeaL). *KI - Künstliche Intelligenz*, pages 1–5.

Fankhauser, P., Kermes, H., & Teich, E. 2014. Combining Macro- and Microanalysis for Exploring the Construal of Scientific Disciplinarity. In *Digital Humanities*, Lausanne, Switzerland.

Halliday, M. & Martin, J. 1993. *Writing science: literacy and discursive power*. Falmer Press, London.

Harris, Z. S. 1991. *A theory of language and information: a mathematical approach*. Oxford University Press, USA.

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. 2013. Differentiating language usage through topic models. *Poetics*, **41**(6), 607–625.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Schöch, C. 2013. Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities*, **2**(3), 2–13.

Vreeken, J. 2010. Making pattern mining useful. *ACM SIGKDD Explorations Newsletter*, **12**(1), 75–76.