

# A resource for the diachronic study of scientific English: Introducing the Royal Society Corpus

**Ashraf  
Khamis**

Saarland University  
ashraf.khamis@uni-saarland.de

**Hannah Kermes**

Saarland University  
h.kermes@mx.uni-saarland.de

**Noam Ordan**

Saarland University  
noam.ordan@uni-saarland.de

**Stefania Degaetano-Ortlieb**

Saarland University  
s.degaetano@mx.uni-saarland.de

**Jörg Knappen**

Saarland University  
j.knappen@mx.uni-saarland.de

**Elke Teich**

Saarland University  
e.teich@mx.uni-saarland.de

There is a wealth of corpus resources for the study of contemporary scientific English, ranging from written vs. spoken mode to expert vs. learner productions as well as different genres, registers and domains (e.g. MICASE (Simpson et al. 2002), BAWE (Nesi 2011) and SciTex (Degaetano-Ortlieb et al. 2013)). The multi-genre corpora of English (notably BNC and COCA) include fair amounts of scientific text too.

Diachronic resources of scientific texts are more limited in that existing corpora are typically fairly small, including only few small samples per discipline (e.g. ARCHER with approximately 258,000 words covering all scientific disciplines in British and American English texts (Biber et al. 1994) and the Coruña Corpus in which 10,000 words are taken to represent astronomy in the 18<sup>th</sup> and 19<sup>th</sup> centuries (Moskowich and Crespo 2007)) or covering one discipline only (e.g. the corpus of Early Modern English Medical Texts (Taavitsainen et al. 2011)).

To increase the pool of corpus resources for the diachronic study of scientific English, we are building a corpus from the Philosophical Transactions and Proceedings of the Royal Society of London, starting from the date of their inception (1665) to modern time. At present, we work on processing materials from the period 1776 to 1869 (2,454 articles amounting to around 23 million tokens), with other periods to follow. The materials contain texts from a variety of scientific areas ranging from biology, chemistry, physics and geography to medicine.

We describe the steps we take to get from the source materials to a usable corpus, focusing in particular on the interaction of automatic and manual

processing. The source materials are in XML format and contain metadata on journal, title, author and year of publication. Although the texts are partially structured, they need a considerable amount of preprocessing, including cleaning of OCR errors and hidden markup, ordering of scrambled pages, identification of article beginnings and endings and removal of duplicates, headers and footers. After preprocessing, we normalize the texts using VARD (Baron and Rayson 2008), annotate them for tokens, lemmas and parts-of-speech using TreeTagger (Schmid 1994) and finally encode the corpus in Corpus Query Processor (CQP) format (Evert and Hardie 2011). Furthermore, we mark up document structure as provided by the XML source as well as century, fifty-year period and decade so as to enable analyses on different temporal resolution frames.

Once a reasonable level of data quality has been reached, the Royal Society Corpus will be made available through CLARIN-D. In our own research, we use the corpus to study the diachronic development of scientific English as a distinct discourse type as well as register diversification, applying various methods of data mining.

## References

- Baron, A. and Rayson, P. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. *Postgraduate Conference in Corpus Linguistics* 2008, May 22. Birmingham, UK: Aston University.
- Biber, D., Finegan, E. and Atkinson, D. 1994. ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In U. Fries, P. Schneider and G. Tottie (eds.), *Creating and using English language corpora*, 1–14. Amsterdam/New York: Rodopi.
- Degaetano-Ortlieb, S., Kermes, H., Lapshinova-Koltunski, E. and Teich, E. 2013. SciTex - a diachronic corpus for analyzing the development of scientific registers. In P. Bennett, M. Durrell, S. Scheible and R. J. Whitt (eds.), *New methods in historical corpus linguistics: Corpus linguistics and interdisciplinary perspectives on language (CLIP)*, vol. 3. Tübingen: Narr.
- Evert, S. and Hardie, A. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK.
- Moskowich, I. and Crespo, B. 2007. Presenting the Coruña Corpus: A collection of samples for the historical study of English scientific writing. In J. Pérez-Guerra, D. González-Álvarez, J. L. Bueno Alonso and E. Rama-Martínez (eds.), *'Of varying language and opposing creed': New insights into Late Modern English*, 341–357. Bern: Peter Lang.
- Nesi, H. 2011. BAWE: An introduction to a new resource. In A. Frankenberg-Garcia, L. Flowerdew and

G. Aston (eds.), *New trends in corpora and language learning*, 213–228. London: Continuum.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, 44–49. Manchester, UK.

Simpson, R. C., Briggs, S. L., Ovens, J. and Swales, J. M. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Taavitsainen, I., Jones, P. M., Pahta, P., Hiltunen, T., Marttila, V., Ratia, M., Suhr, C. and Tyrkkö, J. 2011. Medical texts in 1500–1700 and the corpus of Early Modern English Medical Texts. In I. Taavitsainen and P. Pahta (eds.), *Medical writing in Early Modern English*, 9–29. Cambridge: Cambridge University Press.

## **SYN2015: a representative corpus of contemporary written Czech**

**Michal Křen**

Institute of the Czech National Corpus  
Charles University

`michal.kren@ff.cuni.cz`

### **1 Background**

The Czech National Corpus aims at extensive and continuous mapping of the Czech language and its varieties. This effort results in compilation, maintenance and providing access to a number of corpora (synchronic/diachronic, written/spoken etc.), including corpora of contemporary written Czech making up the SYN series.<sup>142</sup>

The SYN-series corpora can be described as traditional (as opposed to the web-crawled corpora), featuring cleared copyright issues, well-defined composition, reliability of annotation and high-quality text processing (Hnátková et al. 2014). All the corpora are also disjoint, i.e. any document can be included only into one of them.

### **2 Representative corpora of the SYN series**

Currently, the SYN series consists of three large newspaper corpora with total size exceeding 2 billion tokens and three 100-million corpora representative of written Czech (SYN2000, SYN2005, and SYN2010; the number denotes the corpus publication year).

The representative corpora cover three consecutive time periods in a regular five-year interval (i.e. SYN2010 covers the 2005–2009 period) and they contain a large variety of written genres in proportions based on language reception studies (Králík and Šulc 2005). Their design, strengths and weaknesses are described in detail in Křen (2013: 46–53) including the comparability, which is desirable to enable modern diachronic studies.

### **3 Design of SYN2015**

The aim of this paper is to introduce SYN2015, a 100-million corpus of contemporary Czech. SYN2015 will be a continuation of the series, but at the same time, it will reflect necessary methodological and technical changes outlined below.

SYN2015 is designed as a representation of the printed language of 2010–2014. Specific language of the internet (discussion forums, blogs etc.) is kept

---

<sup>142</sup> <http://ucnk.ff.cuni.cz/english/struktura.php>