Drucker, J. (2012). Humanistic theory and digital scholarship. In Gold, M. K. (ed), *Debates in the Digital Humanities*. Minneapolis: University Of Minnesota Press, pp. 85–95.

Gmiterek, G. (2014). Książka w erze nowych technologii, integracji i interaktywności mediów. In Sobczak, A., Cichocka, M. and Frąckowiak, P. (eds), *Historia 2.0 : Panta Rhei Materiały Sympozjum XIX Powszechnego Zjazdu Historyków Polskich 17 Września 2014 W Szczecinie*. Lublin: E-naukowiec, pp. 67–74. https://repozytorium.lectorium.pl/handle/item/898 (accessed 5 March 2016).

International ISBN Agency (2012). *ISBN Users' Manual*. 6th ed. London: International ISBN Agency. https://www.isbn-international.org/sites/default/files/ISBN Manual 2012 -corr.pdf (accessed 5 March 2016).

International Standard Serial Number International Center (2005). *ISSN Manual: International Standard Serial Number*. http://www.issn.org/wp-content/uploads/2013/09/ISSN-Manual_ENG2015_23-01-2015.pdf (accessed 5 March 2016).

Jessop, M. (2008). Digital visualization as a scholarly activity. *Literary and Linguistic Computing*, **23**(3): 281–93.

Kindred Britain. http://kindred.stanford.edu/# (accessed 10 May 2015).

Mapping the Republic of Letters. http://republicofletters.stanford.edu/index.html (accessed 8 May 2015).

Nahotko, M. (2010). *Komunikacja Naukowa W środowisku Cyfrowym : Globalna Biblioteka Cyfrowa W Informatycznej Infrastrukturze Nauki*. Warszawa: Wydawnictwo SBP.

ORBIS: The Stanford Geospatial Network Model of the Roman World. http://orbis.stanford.edu/ (accessed 10 May 2015).

Pikas, C. K. (2006). *The Impact of Information and Communication Technologies on Informal Scholarly Scientific Communication: A Literature Review*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.9216&rep=rep1&type=pdf (accessed 5 March 2016).

Radomski, A. (2014). *Humanistyka W świecie Informacjonalizmu*. http://e-naukowiec.eu/wp-content/uploads/2014/06/A.Radomski.pdf (accessed 5 March 2016).

Sapa, R. (2009). *Metodologia Badań Obszaru Pośredniczenia W Komunikacji Naukowej Z Perspektywy Nauki O Informacji*. Kraków: Wydawn. Uniwersytetu Jagiellońskego.

Słownik języka polskiego PWN. http://sjp.pwn.pl/sjp/wizualny;2579950.html (accessed 3 May 2015).

Terras, M. Infographic: Quantifying Digital Humanities. http://melissaterras.blogspot.co.uk/2012/01/infographic-quanitifying-digital.html (accessed 6 May 2015).

The Valley of the Shadow: Two Communities in the American Civil War. http://valley.lib.virginia.edu/ (accessed 5 March 2016).

Thompson, E. and Mahoy, S. (2013). The Roaring 'Twenties : an interactive exploration of the historical soundscape of New York City. *Vector: Journal of Culture and Technology in a Dynamic Vernacular*, **4**(1). http://vectorsdev.usc.edu/NYCsound/777b.html (accessed 5 March 2016).

Vectors : Journal of Culture and Technology in a Dynamic Vernacular. http://vectors.usc.edu/journal/index.php?page=Introduction (accessed 5 May 2015).

Virtual Pauls Cross Website: a Digital re-creation of John Donne's Gunpowder Day sermon. http://vpcp.chass.ncsu.edu/ (accessed 19 May 2015).

Wieczorek-Tomaszewska, M. (2013). Cyfrowa humanistyka jako metaforyczna współczesna Republika Listów. *23. Ogólnopolskie Sympozjum Naukowe „Człowiek - Media - Edukacja" 27-28 Września 2013*. Kraków. http://ktime.up.krakow.pl/symp2013/referaty_2013_10/wieczorek.pdf (accessed 5 March 2016).

Wilkowski, M. (2013). *Wprowadzenie Do Historii Cyfrowej*. Gdańsk: Instytut Kultury Miejskiej. https://depot.ceon.pl/handle/123456789/2001 (accessed 5 March 2016).

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

Hannah Kermes
h.kermes@mx.uni-saarland.de
Universität des Saarlandes, Germany

Stefania Degaetano-Ortlieb
s.degaetano@mx.uni-saarland.de
Universität des Saarlandes, Germany

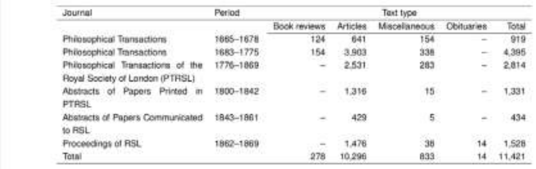
Ashraf Khamis
a.khamis@uni-saarland.de
Universität des Saarlandes, Germany

Jörg Knappen
j.knappen@mx.uni-saarland.de
Universität des Saarlandes, Germany

Elke Teich
e.teich@mx.uni-saarland.de
Universität des Saarlandes, Germany

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 modelling (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 and 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 useTreeTagger (Schmid,1994)on the normalized texts. Inspired by the idea of Agile Software Development (Cockburn, 2001), we intertwine the actual corpus building with corpus annotation and analysis, continuously building new versions of the corpus whenever we see a recurrent problem in data quality. We work with a dedicated pipeline and keep the corpus-building process as modular and automatic as possible, applying manual work before the first automatic step. In the last step, the corpus is encoded in CQP format (cf. IMS Open Corpus Workbench (CWB) (Evert and Hardie, 2011)) and can be accessed via a CQPweb interface (Hardie, 2012)[2].

In terms of analysis, our main assumption is that due to specialization, scientific texts exhibit greater encoding density over time (Halliday and 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)[3] that affixation rises considerably as a means of word formation in scientific texts in the mid-17th century. For the identification that affixation rises considerably as a means of word formation in scientific texts in the mid-17th century. For the identification 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).

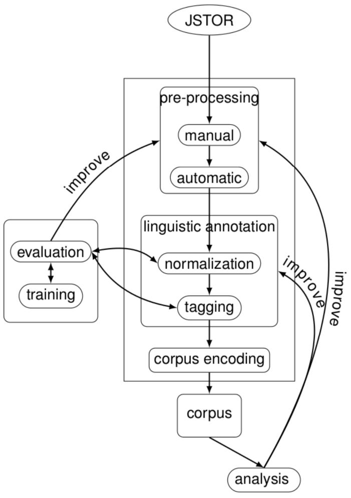


Figure 1: Corpus building steps

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).

Bibliography

Baron, A. and Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. In *ProceedingsofthePostgraduateConferenceinCorpusLinguistics*, Birmingham, UK.

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

Cockburn, A. (2001). *Agile Software Development*. Addison-Wesley Professional, Boston, USA.

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

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 Conference*, Birmingham, UK.

Fankhauser, P., Kermes, H. and 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.

Hardie, A. (2012). *International Journal of Corpus Linguistics*

17(3): 380-409.CQPweb – combining power, flexibility and usability in a corpus analysis tool.

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. and Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, **41**(6): 607–25.

Notes

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

[2] https://fedora.clarin-d.uni-saarland.de/cqpweb/

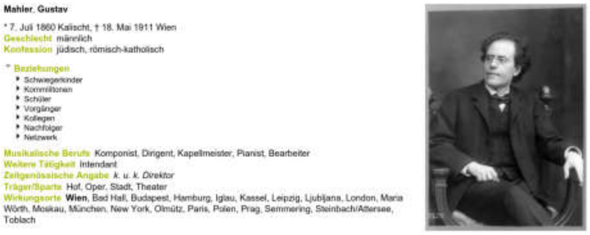
[3] http://www.oed.com/thesaurus/

On the Distant Reading of Musicians' Biographies

Richard Khulusi
richard.khulusi@web.de
Leipzig University, Germany

Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de
Leipzig University, Germany

The Bavarian Musicians Encyclopedia Online (Bayerisches Musiker Lexikon Online, BMLO) is a web-based platform that provides access to biographical information about musicians associated to Bavaria's music history (BMLO, 2016). Most of the musicians contained in the corresponding database had an active lifetime period living in Bavaria or a considerable influence on Bavaria. Initiated in 2004, the musicians database contains biographical data about nearly 28,000 musicians now. This suggests the rather global scope of the BMLO – underpinned by many musicologists worldwide using the BMLO for their daily work. The screenshot in Figure 1 shows the BMLO entry for the composer Gustav Mahler.



Mahler, Gustav

* 7. Juli 1860 Kalischt, † 18. Mai 1911 Wien
Geschlecht männlich
Konfession jüdisch, römisch-katholisch

Beziehungen
- Schwiegerkinder
- Kommilitonen
- Schüler
- Vorgänger
- Kollegen
- Nachfolger
- Netzwerk

Musikalische Berufe Komponist, Dirigent, Kapellmeister, Pianist, Bearbeiter
Weitere Tätigkeit Intendant
Zeitgenössische Angabe *k. u. k. Direktor*
Träger/Sparte Hof, Oper, Stadt, Theater
Wirkungsorte **Wien**, Bad Hall, Budapest, Hamburg, Iglau, Kassel, Leipzig, Ljubljana, London, Maria Wörth, Moskau, München, New York, Olmütz, Paris, Polen, Prag, Semmering, Steinbach/Attersee, Toblach

Figure 1: Biographical information about Gustav Mahler in the BMLO

A recently published article facilitates the profiling of musicians based on the BMLO (Jänicke et al., 2015). The profile of a musician of interest can be visualized, and according to biographical information, similar musicians are determined in a semi-automated process (MusikerProfiling, 2016) . A profiling result for Gustav Mahler is shown in Figure 2. Although the profiling system has been proven useful for the collaborating musicologists, it does not support generic research questions like “In which cities Roman Catholic conductors worked during the 18th century?” or “What are the differences and similarities among the careers of pianists and violinists?” Therefore, the musicologists desired a system that facilitates the dynamic exploration of musicians' characteristics with the help of interactive visual interfaces. The design of the resultant visualization system is outlined below.

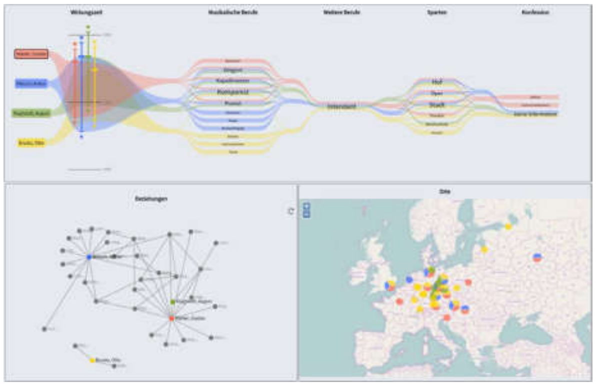


Figure 2: Interactive visual profiling of Gustav Mahler comparatively visualizes Mahler's profile to the profiles of the three most similar musicians in three views (Column Explorer, Relationship Graph, Map)

To support the dynamic exploration of musicians' biographies, we provide various views that visualize aggregate biographical information of musicians inherent in the database. For the divisions where musicians worked, we use a tag cloud (Fig. 3a). As musical (Fig. 3b) and further professions (Fig. 3c) are organized in a hierarchy, we apply a sunburst technique tailored for such structures (Stasko et al., 2000). A map plots all places of activity (Fig. 3d). Using GeoTemCo (Jänicke et al., 2013) for that purpose, occluding dots are clustered and metropolises of music history, e.g., Munich, Vienna and Berlin, are salient as large circles. To illustrate the denominations of musicians, we use again a tag cloud (Fig. 3e), and a pie chart to visualize musicians' sexes (Fig. 3f). Finally – based on the dates of birth, the first mentioned dates and the dates of death – we define an activity time for each musician. The aggregate of all activity times is shown in a timeline graph (Fig. 3g). With mouse interaction, each view can be used for filtering purposes. So, the investigation of rather generic research questions in musicology gets possible.

Figure 4 shows the filter steps required to explore in which cities Roman Catholic conductors worked during the 18th century. The first filter is applied in the denomination tag cloud (Fig. 4a) by clicking “römisch-katholisch”