The Royal Society Corpus: From Uncharted Data to Corpus

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen and Elke Teich
Universität des Saarlandes
Universität Campus A2.2, 66123 Saarbrücken, Germany
{s.degaetano, h.kermes, j.knappen, e.teich}@mx.uni-saarland.de, ashraf.khamis@uni-saarland.de

Abstract

We present the Royal Society Corpus (RSC) built from the Philosophical Transactions and Proceedings of the Royal Society of London. At present, the corpus contains articles from the first two centuries of the journal (1665–1869) and amounts to around 35 million tokens. The motivation for building the RSC is to investigate the diachronic linguistic development of scientific English. Specifically, we assume that due to specialization, linguistic encodings become more compact over time (Halliday, 1988; Halliday and Martin, 1993), thus creating a specific discourse type characterized by high information density that is functional for expert communication. When building corpora from uncharted material, typically not all relevant meta-data (e.g. author, time, genre) or linguistic data (e.g. sentence/word boundaries, words, parts of speech) is readily available. We present an approach to obtain good quality meta-data and base text data adopting the concept of Agile Software Development.

Keywords: corpus creation, corpus annotation, metadata

1. Introduction

As science developed to become an established sociocultural domain from the early modern times, it underwent a process of specialization. We assume that due to specialization, scientific texts exhibit greater encoding density over time, i.e. more compact, shorter linguistic forms are increasingly used, in order to maximize efficiency in communication. Examples of linguistic densification can be found at all linguistic levels, e.g. reductions at the syntactic level (e.g. relativizer omission), nominalizations at the morphological level or contractions at the word level.

Our assumption is that such densification effects are measurable in the linguistic signal in terms of information density, i.e. the number of bits needed to encode a given message (Shannon information), which is conventionally represented as the (log) probability of a linguistic unit given some context (Crocker et al., 2015). The more predictive a given context, the shorter the linguistic encoding (cf. e.g. variation in word length in Mahowald et al. (2013)) and the fewer the bits needed for encoding will be.

To test this assumption, we need an appropriate data set. There are a number of diachronic corpora of scientific English, but these are typically discipline-specific, cover a certain time period only (e.g. the corpus of Early Modern English Medical Texts (EMEMT) (Taavitsainen et al., 2011)) and are fairly small (e.g. the Coruña Corpus (Moskowich and Crespo, 2007) with c. 10,000 words for each discipline in the 18th and 19th centuries). Given that the Royal Society of London played a major role in shaping science from the mid-17th century (cf. Atkinson (1998)), we obtained a digitized version of the first two centuries of its publications.

When building new corpora from uncharted material, typically not all relevant meta-data or linguistic data is readily available. We describe the procedures applied to enrich the base text we use for the RSC, employing a combination of pattern-based techniques and data mining so as to obtain better-quality base text data and meta-data.

In the following, we describe in detail the corpus material (Section 2.), the processing steps taken to obtain better quality base text and richer, consistent meta-data and the basic linguistic annotation (spelling normalization, PoS tagging) (Section 3.). We conclude with a brief summary and envoi (Section 4.).

2. Corpus Material

The text sources for the Royal Society Corpus were obtained from JSTOR[1] in a well-formed XML format. The data includes different production types, such as articles, book reviews and abstracts as well as different modes of presentation, such as abstracts of printed papers and oral papers. A detailed description of the single sources is shown in Table 1.

As the material comes from scanned pages, OCR errors are present and have to be corrected.

Some meta-data is already stored in XML elements in the JSTOR version: ISSN number of the journal, abbreviation of the journal name, author(s), text type (e.g. article, abstract), page range, day, month and year of publication, first and last page numbers, head ID, volume, text ID, and title.

3. Agile Corpus Building

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. Our experience shows that such problems are often detected only in the actual work with the corpus, so our strategy is to allow as much feedback as possible from other stages of processing as well as analysis into corpus building (see Figure 1 for a graphical overview).

Corpus building is divided into three main steps: (i) preprocessing, (ii) linguistic annotation, and (iii) corpus encoding, which are described in the following subsections.

3.1. Preprocessing

Preprocessing includes the transformation of data into a standardized format, cleaning of data (e.g. OCR errors) and derivation and annotation of meta-data.

[1] http://www.jstor.org/

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

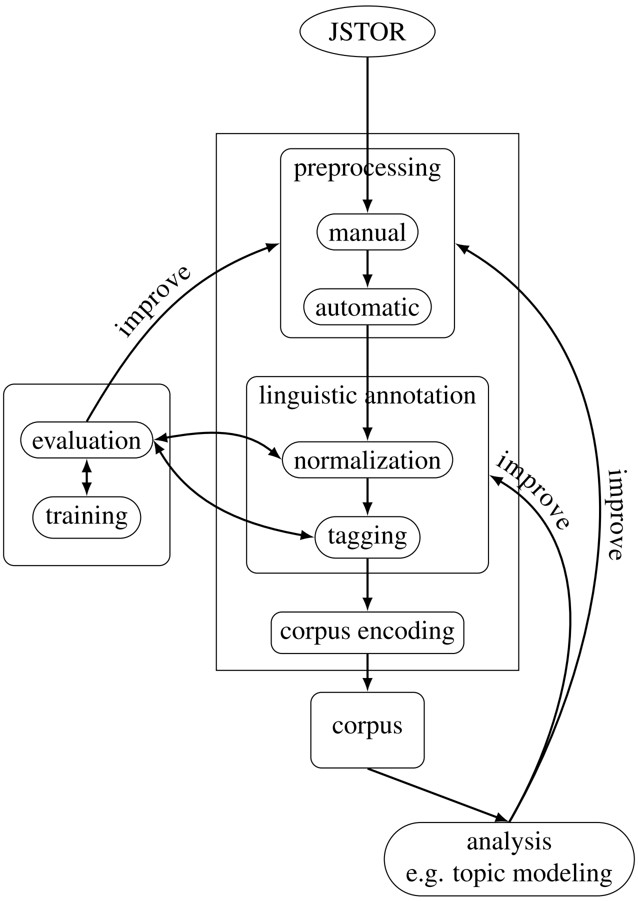

Figure 1: Corpus-building steps; interaction with annotation and analysis

3.1.1. Better Data Quality

We use dedicated scripts for preprocessing wherever possible. Manual work is invested only if automatic procedures cannot build on recurrent triggers (e.g. where the layout does not have recurrent patterns indicating article boundaries) and is applied prior to the first automatic step. In preprocessing we mainly address two types of quality issues: OCR errors and layout problems.

For dealing with OCR errors, we adapt the patterns provided by Underwood and Auvil[2]. We eliminated unused patterns, added new patterns specific to the RSC corpus, and changed some of the original patterns for a better fit to our data, e.g. *fhe* is mapped to *the* instead of *she*. Currently, we apply 1,282 correction patterns.

With respect to layout, we identified the following problems: Headers and footers are included in the running text, line breaks and paragraph boundaries are not preserved, pages may be scrambled, pages may be numbered in arabic or roman or be unnumbered, there may be gaps in the page sequence, first and last pages of articles may be duplicated, article boundaries are not explicitly marked. Also, the journals have different layout types. For example, the Philosophical Transactions (PTRSL; see again Table 1), has four different layouts (1776–1791, 1792–1827, 1828–1839, 1840–1869) that require separate scripts adapted to the individual layouts.

3.1.2. Richer and Consistent Meta-data

The procedures to obtain and annotate meta-data are similar to the procedures we apply to ensure data quality. Sources for relevant information are multiple: (i) the given meta-data, (ii) (lexical) triggers in the texts, (iii) a combination of (i) and (ii), (iv) results of pattern-based and/or data-mining techniques. In order to enrich the corpus with the meta-data, we use dedicated scripts which are incorporated in the corpus-building process as described above.

The source data from JSTOR includes meta-data such as title, author, year and journal of publication, pages and different scientific production types, such as research articles, book reviews and abstracts. All meta-data included in the source data is preserved and used to identify meta-data within the articles (title, authors, journals). However, other relevant meta-data is missing, and the given meta-data is not always consistently given.

With regard to production type, abstracts are marked if they occur in specific volumes, but not if they occur in other volumes. The latter may be identified using lexical triggers, such as the title string *Abstract*. As abstracts are stored in different volumes and files from their articles, the relation between them has to be restored. We therefore apply a matching algorithm based on matching titles. To approxi-

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

mate scientific disciplines, we apply topic modeling (Blei et al., 2003) using MALLET (McCallum, 2002). This allows us not only to identify "scientific disciplines" (e.g. *Meteorology, Astronomy, Paleontology, Optics, History*), but also articles written in languages other than English (*French, Latin*).

3.2. Linguistic Annotation

For the time being, we annotate mainly at the token level: words, lemmas, parts of speech and normalized (modernized) word forms. We build on existing and freely available tools, using VARD (Baron and Rayson, 2008) for normalization and TreeTagger (Schmid, 1994; Schmid, 1995) for tokenization, lemmatization and part-of-speech tagging. In the spirit of Agile Corpus Building, whenever errors are detected in token level annotation, a new corpus version is created. For evaluation, we created a small manually annotated subset of the RSC (~ 56.000 tokens). For training and evaluation of the normalization, we divided the subset into training and test set of equal size. For the evaluation of TreeTagger, we used the whole subset.

Normalized word forms are annotated for two reasons: (1) improvement of performance of natural language tools trained on modern texts (e.g. taggers) and (2) comparability of texts on the lexical level across time (e.g. to investigate conventionalization on the level of spelling). We use a trained model of VARD for automatic normalization based on a sample of manually normalized texts. Evaluation shows that our trained model increases the performance of VARD (see Table 2). Each time a new training model was created based on new normalized texts, the corpus was updated accordingly.

	Untrained VARD	Trained VARD
Precision	61.8%	72.8%
Recall	31.4%	57.7%

Table 2: Precision and recall of untrained and trained VARD models

For the annotation of sentence boundaries, lemmas, and parts of speech, we use TreeTagger, a PoS tagger trained on contemporary English newspaper texts. During analysis, wrong sentence boundaries were detected based on abbreviations not included in TreeTagger. Again, after including them, a new corpus version was created. The evaluation of the tagger on the RSC shows relatively good results (see Table 3 for a comparison of TreeTagger's performance on modern data, the whole RSC and its different time periods). Reasons for tagging errors are: (1) spelling variations, (2) changes in derivational morphology, and (3) grammatical/syntactic changes (e.g. word order). Evaluation on (manually) normalized tags shows an increase in tagger performance (see Table 4).

3.3. Corpus Encoding

In the last step, the corpus is encoded in CQP format (cf. IMS Open Corpus Workbench (CWB) (Evert and Hardie, 2011)) for query and analysis. The CWB requires simple XML as an input format (see Figure 2 for an example). Annotations on the token level (positional attributes, e.g. word,

	Precision
Modern English	97.0%
RSC	94.0%
1665–1715	90.8%
1816–1869	96.0%

Table 3: Comparison of TreeTagger's performance on modern and historical data

	Original	Normalized
RSC	94.5%	95.1%
1665–1715	90.7%	92.3%
1816–1869	96.0%	96.3%

Table 4: TreeTagger's performance on original and normalized word forms

pos, lemma) are represented one-word-per-line and TAB deliminated, annotations beyond token level (structural attributes, e.g. sentence boundaries, pages) as XML-tags .

```
<text id="100997" issn="03702316" title="An Extract of a
Letter Written by Dr. Edward Brown from Vienna in Austria
March 3. 1669. Concerning Two Parhelia's or Mocksuns,
Lately Seen in Hungary" fpage="953" lpage="953" year="1669"
decade="1660" period="1650" century="1600" volume="4"
journal="Philosophical Transactions (1665-1678)"
author="Edward Brown" type="fla" corpusBuild="1.17"
jstorLink="http://www.jstor.org/stable/100997">
[...]
<s no="s_0008">
One        CD    One       One
of         IN    of        of
them       PP    them      them
(          (     (         (
the        DT    the       the
lesset     NN    lesset    lesset
of         IN    of        of
the        DT    the       the
two        CD    two       two
)          )     )         )
began      VVD   begin     began
to         TO    to        to
decay      NN    decay     decay
before     IN    before    before
the        DT    the       the
other      JJ    other     other
,          ,     ,         ,
and        CC    and       and
then       RB    then      then
the        DT    the       the
other      JJ    other     other
grew       VVD   grow      grew
bigger     JJR   big       bigger
,          ,     ,         ,
and        CC    and       and
continued  VVD   continue  continued
well       RB    well      well
nigh       IN    nigh      nigh
two        CD    two       two
<normalised orig="houres" auto="true">
hours      NNS   hour      houres
</normalised>
,          ,     ,         ,
```

Figure 2: CQP input format

The CWB has a built-in encoding tool. Parameters need to be specified for positional and structural attributes. We use a dedicated script to derive the parameters for structural attributes automatically from our output files.

4. Summary and Envoi

High-quality corpora are extremely important for conducting humanities research in areas such as history, cultural studies, literary studies or linguistics. However, to build such corpora, usually high manual effort is involved. Existing corpora are therefore often fairly small. In order to build larger corpora with good quality, we have adopted the idea of Agile Software Development, which promotes a close interaction between development and application and the continuous and fast production of new versions. In corpus building, this means that after a first version of a corpus is available, users apply common annotations and analyses (e.g. PoS tagging, topic modeling) and closely monitor the quality of the output for feedback into the next corpus version. In addition, this approach to corpus building is characterized by the interaction of (few) manual steps with (largely) automatic procedures, which are kept strictly separate. If a change to the data is needed (e.g. correcting a list of OCR errors), the basic automatic processing pipeline is not affected. Furthermore, we made sure that we employ existing and open tools for processing as much as possible (such as VARD for normalization).

Finally, our approach has some interesting side-effects regarding our linguistic research. Since we monitor the output quality of the applied processing tools very closely, we effectively combine the analysis of data quality with linguistic analysis. For instance, less spelling variation and increasing accuracy of PoS tagging over time clearly indicate linguistic change. In our ongoing work, we exploit such observations in our diachronic analyses and incorporate them in modeling variation in encoding density.

Our goal is to eventually make the RSC available to humanities research at large through a CLARIN-D repository.

5. Bibliographical References

Dwight Atkinson. 1998. *Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London*. Routledge, New York.

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Alistair Cockburn. 2001. *Agile Software Development*. Addison-Wesley Professional, Boston, USA.

Matthew W. Crocker, Vera Demberg, and Elke Teich. 2015. Information density and linguistic encoding (IDeaL). *KI - Künstliche Intelligenz*, pages 1–5.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics Conference*, Birmingham, UK.

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

M.A.K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of Written English: Situational Factors and Linguistic Features*, pages 162–177. Pinter, London.

Kyle Mahowald, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Isabel Moskowich and Begoña Crespo. 2007. Presenting the Coruña Corpus: A collection of samples for the historical study of English scientific writing. In Javier Pérez-Guerra, Dolores González-Álvarez, Jorge L. Bueno-Alonso, and Esperanza Rama-Martínez, editors, *'Of varying language and opposing creed': New insights into Late Modern English*, pages 341–357. Peter Lang, Frankfurt.

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

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Irma Taavitsainen, 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. In Irma Taavitsainen and Päivi Pahta, editors, *Medical writing in Early Modern English*, page 9–29. Cambridge University Press, Cambridge, UK.

the Royal Society Corpus (Ker-. mes et al., 2016). To ensure high. embedding quality, these corpora. are divided into temporal slices of. roughly similar size, thus covering. between 10 and 50 years each.Â 2016. The Royal Society Corpus: From uncharted data to corpus. In LREC 2016, pages 1928â€“1931. Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In LT-CSS 2014 @ ACL 2014, pages 61â€“65. Evgeny Kim, Sebastian PadoÂ´, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. The royal society corpus: From uncharted data to corpus. H Kermes, S Degaetano-Ortlieb, A Khamis, J Knappen, E Teich. Proceedings of the Tenth International Conference on Language Resources and â€¦, 2016. 19. 2016. The linguistic construal of disciplinarity: A dataâ€mining approach using register features. E Teich, S Degaetanoâ€Ortlieb, P Fankhauser, H Kermes, Journal of the Association for Information Science and Technology 67 (7 â€¦, 2016. 17. 2016. An information-theoretic approach to modeling diachronic change in scientific English.Â Data Mining with Shallow vs. Linguistic Features to Study Diversification of Scientific Registers. S Degaetano-Ortlieb, P Fankhauser, H Kermes, E Lapshinova-Koltunski, LREC, 1327-1334, 2014.