Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 794–802
Marseille, 11–16 May 2020

The Royal Society Corpus 6.0
Providing 300+ Years of Scientific Writing for Humanistic Study

Stefan Fischer, Jörg Knappen, Katrin Menzel, Elke Teich
Universität des Saarlandes
Campus A2.2, 66123 Saarbrücken, Germany
stefan.fischer@uni-saarland.de
{j.knappen, k.menzel, e.teich}@mx.uni-saarland.de

Abstract

We present a new, extended version of the *Royal Society Corpus* (RSC), a diachronic corpus of scientific English now covering 300+ years of scientific writing (1665–1996). The corpus comprises 47 837 texts, primarily scientific articles, and is based on publications of the Royal Society of London, mainly its *Philosophical Transactions* and *Proceedings*. The corpus has been built on the basis of the FAIR principles and is freely available under a Creative Commons license, excluding copy-righted parts. We provide information on how the corpus can be found, the file formats available for download as well as accessibility via a web-based corpus query platform. We show a number of analytic tools that we have implemented for better usability and provide an example of use of the corpus for linguistic analysis as well as examples of subsequent, external uses of earlier releases. We place the RSC against the background of existing English diachronic/scientific corpora, elaborating on its value for linguistic and humanistic study.

Keywords: Corpus (Creation, Annotation, etc.), Digital Humanities, Language Modelling, Text Analytics

1. Introduction

We present the newest release of the *Royal Society Corpus* (RSC), a diachronic corpus of scientific English covering the period from 1665 until 1996. The corpus comprises 47 837 texts, mainly scientific articles, and is based on the *Philosophical Transactions* and *Proceedings* of the Royal Society of London. The corpus is made available under a Creative Commons license, excluding the recent decades, which are still under copyright.

We describe the origin of the data for the RSC, the processing pipeline and available annotation layers (Sections 3 & 4) and place it in the landscape of comparable corpora (Section 2). We provide several file formats for download, both plain text and XML, and we also host an installation of the corpus on our corpus query platform for easy access (Section 5). We conclude with a detailed example of application of the corpus and its accompanying infrastructure (Section 6) followed by a brief summary and outlook (Section 7).

2. Related Corpora and Own Previous Work

There are several diachronic corpora of written English that contain scientific texts but, to the best of our knowledge, there is no dedicated diachronic corpus of scientific English that comes with similarly good conditions of use as the RSC.

For example, the Corpus of Late Modern English Texts (CLMET) covers the period from 1710 to 1920 and comprises 333 texts from five different genres (De Smet et al., 2015). CLMET contains 34M words from British authors and is available under a Creative Commons license.

The Corpus of Historical American English (COHA) contains more than 400M words of historical English from the 1810s until the 2000s (Davies, 2010). COHA is balanced by genre and time with the following four genres: Fiction, Magazine, Newspaper, and Non-Fiction. Querying the corpus is free, but full-text access must be purchased.

The Scientific Text Corpus (SciTex) contains English scientific research articles from the 1970s and the early 2000s (Degaetano-Ortlieb et al., 2013). SciTex covers nine scientific disciplines and contains 34M tokens. Access to the corpus is limited due to copyright restrictions on the research articles.

ARCHER (A Representative Corpus of Historical English Registers; (Biber et al., 1994; Yáñez-Bouza, 2011)) is designed to contain 10 samples of 2k words each per 50-year period, language variety (British or American English), and genre. ARCHER-3.2 contains 12 genres, among them medicine and science, spanning the years 1600–1999. The complete ARCHER corpus has a size of about 3.3M words and is only accessible for members of the ARCHER consortium. For other users, there is restricted access to a query interface with a limited amount of downloads. The subcorpora belonging to the genres science and medicine are small compared to specialised corpora on scientific texts.

The Coruña Corpus of English Scientific Writing (Crespo-García and Moskowich, 2015) is a collection of text samples representing late Modern English scientific writing except medical texts. It covers the period between 1700 and 1900 and is comprised of eight subcorpora on different scientific disciplines (e.g. astronomy and philosophy), some of which are still under development (e.g. the subcorpora of life sciences and chemistry), each containing approximately 400k words and representing a variety of text types. Representativeness according to sociolinguistic criteria and balance within the corpus were important design criteria. The Coruña Corpus can be searched with the Coruña Corpus Tool (CCT).

The corpus of Middle English Medical Texts (MEMT) contains 86 texts about medicine with 0.5M words from ca. 1375 to 1500. It is available on CD-ROM via a commercial publisher (Taavitsainen et al., 2005). It is diachronically succeeded by the corpus of Early Modern English Medical Texts (EMEMT) with ca. 450 texts and 2M words

from 1500 to 1700, which is also published as a CD-ROM (Taavitsainen et al., 2010). A third corpus in this series[1], the LMEMT (Late Modern English Medical Texts) covering the years 1700 to 1800 is now also available.
Early English Books Online (EEBO)[2] is a large collection of Early English printed books prepared by the Text Creation Partnership[3]. EEBO Phase I is freely available and contains 750M words from more than 25k texts covering the years 1470–1699. EEBO Phase II aims at additional 45k books and extends the range of years to 1820. Access is currently restricted but a public release is announced for July 2020. It is currently available on subscription basis from Sketch Engine[4]. There is no available genre annotation in the EEBO corpus.
Against this background, the RSC fills a gap in that it provides a coherent, diachronic corpus of scientific English covering the entire period of Late Modern English (LModE, ca. 1700–1900) as well as the transition periods at the beginning and the end of LModE. It has a fair size, has been processed according to current best practices (see Sections 3 & 4) and the larger part is made available according to the FAIR principles (Section 5). Given the role of the Royal Society in scientific publications, the RSC is not only highly relevant for diachronic linguistic analysis, see e.g. (Feltgen et al., 2017; Degaetano-Ortlieb and Teich, 2018; Degaetano-Ortlieb and Teich, 2019), but also to historical and cultural analysis, e.g. (Fyfe et al., 2015; Moxham and Fyfe, 2018).
Table 1 gives an overview of earlier versions (2.0 & 4.0) of the RSC as well as the new ones.

Version	Years	# Texts	# Tokens
RSC 2.0	1665–1869	9 813	35 311 790
RSC 4.0	1665–1869	9 779	31 952 725
RSC 6.0 Open	1665–1920	17 520	78 605 737
RSC 6.0 Full	1665–1996	47 837	295 895 749

Table 1: History of RSC releases. Compared to previous releases, the current *Open* version covers 51 additional years.

3. Corpus Building

The corpus is built inspired by the principles of Agile Software Development (Cockburn, 2001; Voormann and Gut, 2008), i.e. corpus preprocessing, corpus annotation and linguistic analysis are intertwined and repeated cyclically.
There are several processing steps, which are described in detail in (Kermes et al., 2016). In the beginning, due to format inconsistencies, we had to manually edit some of the raw files which we received from JSTOR and the Royal Society. This was only done once and the remaining processing steps are fully automatised. As a first step, we filter texts written in languages other than English such as papers in Latin in the earlier Philosophical Transactions. We use `langid.py` (Lui and Baldwin, 2012) for language identification and we only keep texts that are considered English with absolute certainty.
While versions 2.0 and 4.0 used a pattern-based OCR post-correction with high precision and low recall, version 6.0 integrates the Noisy-Channel Spell Checker by (Klaus et al., 2019) for a better recall and F-score at the cost of some loss in precision.

3.1. Origin and Content of Texts

Versions 2.0 and 4.0, which both cover 1665–1869, are fully based on data obtained from JSTOR. Version 6.0 contains additional data, which we received from the Royal Society. In version 6.0, texts from 1665–1869 are still based on JSTOR data with improved processing as described above, whereas later texts are based on the new data from the Royal Society. We chose this approach in order to maintain comparability with analyses based on earlier releases of the corpus.
The texts in the corpus cover a wide range of areas from both the physical sciences and the biological sciences. During the three centuries covered by the corpus, scientific discourse formed as a discipline and underwent considerable changes. Hence, more recent articles can be classified into modern fields of study without difficulty, e.g. physics, chemistry, mathematics, engineering or biology. However, many of the early texts cannot be described by these modern categories.

3.2. Statistics

In total, the corpus contains 295 895 749 tokens in 47 837 texts. Of these, 17 520 texts and 78 605 737 tokens are part of the *open* release. Table 2 shows a detailed overview of the number of texts and tokens over time. As can be seen, the number of available texts and tokens increases exponentially.

Years	# Texts	# Tokens
1665–1699	1 325	2 582 856
1700–1749	1 686	3 414 795
1750–1799	1 819	6 342 489
1800–1849	2 774	9 112 274
1850–1899	6 754	36 993 412
1900–1949	10 011	65 431 384
1950–1996	23 468	172 018 539

Table 2: Size of the Royal Society Corpus over time.

4. Metadata and Annotations

4.1. Subcorpora and Texts

Texts in the RSC are classified by time periods of different granularity (year, decade, 50 years, century). All texts are annotated with their original metadata from the Royal Society whenever possible. For a small fraction of texts, if we could not establish a correspondence between their JSTOR ID and the DOI from the Royal Society, the original JSTOR metadata are used.

[1] `http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/CEEMcorpora.html`
[2] `https://quod.lib.umich.edu/e/eebogroup/`
[3] `https://www.textcreationpartnership.org`
[4] `https://www.sketchengine.eu/early-english-books-online-corpus/`

For example, apart from author and time of publication, we have information on text types. JSTOR provides four text types (book review, article, miscellaneous, and obituary), whereas the data from the Royal Society have a more fine-grained classification including abstract, appendix, article, bill-of-mortality, biography, book-review, lecture, report.

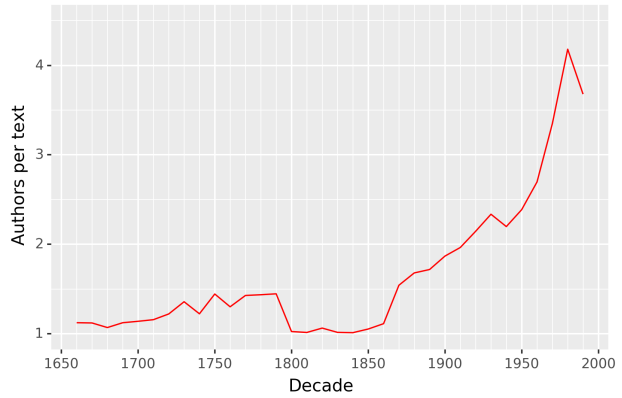


Figure 1: Average number of authors per text over time.

We also provide some statistical data on the texts, such as the number of tokens and sentences per text. For better usability, we added references to other resources, such as links to the full-text PDFs of the original articles on JSTOR and the Royal Society journal archive based on their DOIs. We also inserted links between texts and abstracts when such a relation could be determined. See Table 3 for an overview of all text attributes including proportions of coverage (some documents have missing metadata).

On the basis of the metadata, subcorpora can be built dynamically or they can be used directly in our corpus search (see Section 6). Furthermore, the metadata can be useful on their own, e.g. to explore writing/publication practice over time. See, for example, Figure 1 showing the diachronic development of multiple authorship over time.

4.2. Sentences and Tokens

As attributes of sentences we encode a running ID (within a document/text) and the number of tokens they contain. Each token is annotated as word (normalized form), original word form (historical spelling), lemma and part-of-speech. For part-of-speech tagging we use TreeTagger (Schmid, 1994; Schmid, 1995) and the Penn Treebank Tagset (Santorini, 1990) with some minor modifications. For an analysis of part-of-speech tagging performance on a previous version of the corpus see (Knappen et al., 2017). Since the newly added text material is closer to present-day language, no particular tagging problems arise.

4.3. Surprisal Annotation

As a special feature, we provide information on the (average) *surprisal* of words. Average surprisal (Kermes and Teich, 2017) is a measure of the amount of information transmitted by a linguistic unit (e.g. word or part-of-speech), averaged over all its instances (e.g. in a given time period):

Attribute	Description	Coverage
author	Author of article	96.52 %
century	Century of publication	100.00 %
corpusBuild	Version number	100.00 %
decade	Decade of publication	100.00 %
doi	DOI of article	98.37 %
doiLink	Link to original text	98.37 %
fpage	First page of article	100.00 %
hasAbstract	ID of abstract	1.64 %
id	JSTOR ID	100.00 %
isAbstractOf	ID of article	1.64 %
issn	ISSN of journal	100.00 %
journal	Journal of publication	100.00 %
jrnl	Journal abbreviation	100.00 %
jstorLink	Link to JSTOR source	20.44 %
language	Language of article	98.37 %
lpage	Last page of article	100.00 %
pages	Number of pages	100.00 %
period	Period of publication	100.00 %
sentences	Number of sentences	100.00 %
title	Title of article	100.00 %
tokens	Number of tokens	100.00 %
type	Text type	100.00 %
visualizationLink	Link to visualization	100.00 %
volume	Volume of article	100.00 %
year	Year of publication	100.00 %

Table 3: List of text attributes and percentage of texts with available metadata.

$$AvS(token) = \frac{1}{|token|} \sum_i -\log_2 p(token|context_i)$$

context here refers to an ngram context of three previous words or parts-of-speech.

Diachronically, it is interesting to observe whether certain kinds of words become more or less informative on average. For example, lexical words carry more information on average than function words. In the RSC, we find that the frequency of nouns compared to verbs increases steadily over time (see Figure 2), indicating a shift towards a more nominal style. In general language, in contrast, no such change in frequency occurs, observed on the basis of the Corpus of Late Modern English. However, inspecting surprisal on nouns vs. verbs, we find that it stays fairly stable over time (see Figure 3) for both corpora. Interestingly, both nouns and verbs show higher mean surprisal in CLMET.

Surprisal is annotated into the corpus and calculated with separate language models on the whole corpus (*srp*), individual documents (*doc*), 50-year periods (*s50*) and decades (*s10*). See Table 4 for a list of all provided attributes.

For better usability we also provide an interactive visualization of surprisal scores (Fischer et al., 2017). In the visualization, words are scaled based on their surprisal values. See Figure 4 for an example.

Another perspective provided on the corpus is relative entropy (Kullback-Leibler Divergence; KLD) across *sub-*

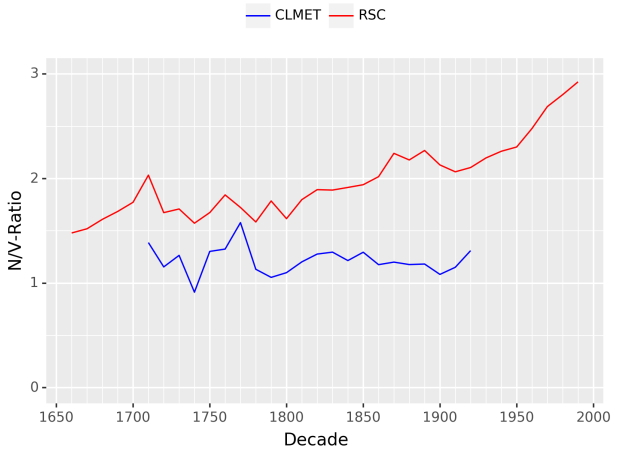


Figure 2: Ratio of nouns and verbs in the RSC and CLMET over time. The usage of nouns increases steadily in the RSC.

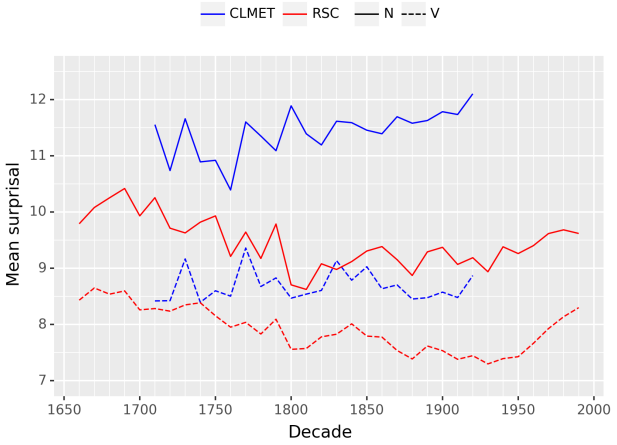


Figure 3: Surprisal of nouns and verbs over time in the RSC and CLMET.

corpora (e.g. 50-year periods). KLD is an asymmetric information-theoretic measure for the comparison of probability distributions, measuring the additional bits needed for encoding when a non-optimal code is used. Applied to diachronic analysis, KLD gives us an indication of the linguistic differences between time periods.

Again, for better usability, an interactive visualization is provided (Fankhauser et al., 2014a; Fankhauser et al., 2014b), see Figure 5 for an example. The heat map on the left shows overall KL divergence between subcorpora (green=low, red=high). The word clouds in the middle and on the right show the most typical words of a given time period where color encodes relative frequency (blue=low, red=high) and size shows the contribution to the overall divergence. Both size and color are scaled logarithmically. The visualization is interactive and the words are linked to an installation of a web-based corpus analysis tool (CQPweb) (Hardie, 2012) based on the Corpus Query Processor. See Section 6 for a detailed example of analysis using the KLD visualization together with CQPweb.

Attribute	Description
word	Normalized word form (VARD)
pos	Part-of-speech tag
lemma	Lemma, according to TreeTagger
orig	Original word form
srp	Corpus surprisal
srp_avg	AvS on corpus
srp_rnd	Corpus surprisal (rounded)
srp_avg_rnd	AvS on corpus (rounded)
doc	Document surprisal
doc_avg	AvS on document
doc_rnd	Document surprisal (rounded)
doc_avg_rnd	AvS on document (rounded)
s50	Surprisal on 50-year periods
s50_avg	AvS on 50-year periods
s50_rnd	Surprisal on 50-year periods (rounded)
s50_avg_rnd	AvS on 50-year periods (rounded)
s10	Surprisal on decades
s10_avg	AvS on decades
s10_rnd	Surprisal on decades (rounded)
s10_avg_rnd	AvS on decades (rounded)

Table 4: Positional attributes in the RSC. *word*, *pos* and *lemma* are based on TreeTagger output. *orig* is the word form before normalization. Surprisal and average surprisal (AvS) are also provided as integers (`_rnd`).

5. Access and Usage

5.1. FAIR Principles

The RSC is designed and built according to the FAIR data principles (Wilkinson et al., 2016). It is hosted at the CLARIN-D repository at Saarland University[5] and findable by a persistent and globally unique identifier, in our case a handle provided by the EPIC consortium[6]. The RSC is described by rich CMDI (Broeder et al., 2011) metadata with a link to the landing page of the corpus. The metadata are indexed and searchable by the CLARIN Virtual Language Observatory (Van Uytvanck et al., 2012).

The *Royal Society Corpus 6.0 Open* is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. We provide files in several common formats (see Section 5.2). Furthermore, there are multiple options for searching the corpus online (see Section 5.3).

5.2. Download

The RSC can be downloaded in several formats[7]. Our default format, which contains all available metadata, is

[5] `https://fedora.clarin-d.uni-saarland.de`

[6] `https://www.pidconsortium.eu/`

[7] RSC 2.0 is available at `https://hdl.handle.net/21.11119/00-246C-0000-0023-8D26-7`, RSC 4.0 is available at `https://hdl.handle.net/21.11119/0000-0001-7E8B-6` and RSC 6.0 Open is available at `https://hdl.handle.net/21.11119/0000-0004-8E37-F`. The downloadable files of the new release will be made available before the conference.

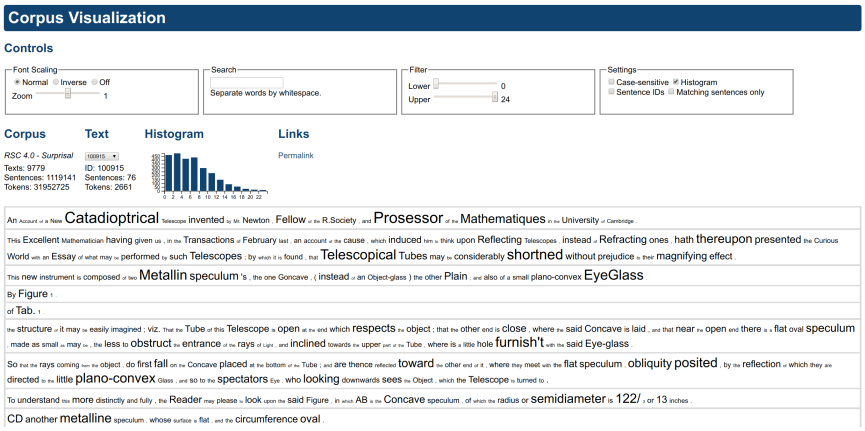


Figure 4: Visualization of surprisal at token level. Words are scaled based on their surprisal score.

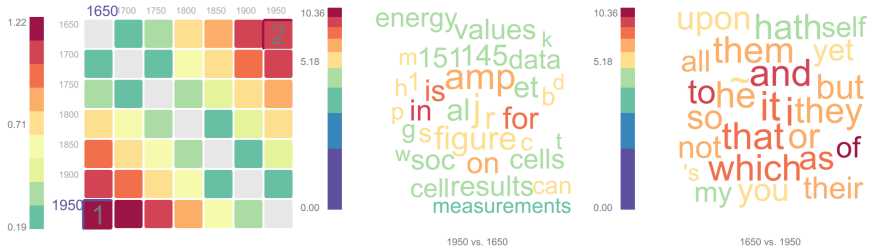


Figure 5: Visualization of RSC 6.0 based on KLD (50-year periods). The heat map shows overall KLD (left). The word clouds show relative frequency (color) and contribution to overall KLD (size) of individual words for 1950 (vs. 1650) (middle) and 1650 (vs. 1950) (right).

the vertical text format (VRT) of CQP (Evert and Hardie, 2011). Using the VRT format, users can encode the RSC on their own CWB/CQPweb servers. For those who do not need their own installation, we provide a CQPweb (Hardie, 2012) server.

We also provide plain text and two XML formats: WebLicht Text Corpus Format (TCF) (Hinrichs et al., 2010) and TEI format (TEI Consortium, 2019). Downloading the corpus as a bundle of plain text files most likely will suit the needs of researchers from disciplines such as machine learning, while the XML formats are more suitable for other kinds of users, including linguists. TCF is provided to facilitate the usage of the corpus within the CLARIN-D infrastructure, while TEI is commonly used in the field of digital humanities.

5.3. Online Access

The full text of the corpus can be queried via CLARIN Federated Content Search (CLARIN-FCS)[8]. Also, a CQPweb (Hardie, 2012) installation is made available[9], enabling a quick navigation to the original texts and further analyses making full use of the text metadata described in Section 4, including time periods of different granularity (1 year, 10, 50, 100 years).

5.4. External Subsequent Use

The Jena Semantic Explorer (JeSeMe) (Hellrich and Hahn, 2017; Hellrich et al., 2018) uses the RSC 2.0 as one of its underlying corpora. With JeSeMe, users can interactively explore similar words, word emotion, typical context, and word frequencies for lemmata.

A study on diachronic word embeddings on the RSC was undertaken by (Fankhauser, 2017) and an interactive visualization of the results is available at *Leibniz Institut für Deutsche Sprache* (IDS) in Mannheim[10].

RSC 2.0 is also included in DiaCollo[11] (Jurish, 2018) maintained at *Berlin-Brandenburgische Akademie der Wissenschaften* (BBAW), which can be used for the extraction of diachronic collocations.

6. Sample Analysis: From KLD to Concordance

KLD provides a useful starting point for the analysis of linguistic similarities and differences between different time periods in the corpus data, helping the analyst to detect distinctive features. The eco-system around the RSC provides the user with various options for exploring candidate features further, e.g. by using CQPweb queries for concordancing, collocations, distributional data, frequency lists

[8] https://www.clarin-d.net/en/accessing/fcs-search-in-resources

[9] https://corpora.clarin-d.uni-saarland.de/cqpweb/

[10] http://corpora.ids-mannheim.de/diaviz/royalsociety.html

[11] https://kaskade.dwds.de/dstar/rsc/diacollo/

and the possibility to investigate larger textual patterns and lexical and grammatical contexts in which distinctive items occur.

In this section we briefly exemplify some salient differences between the first and the most recent 50-year period in the RSC (1665–1699, 1950–1996) illustrated by examples from the word cloud visualization (see Figure 5) highlighting important words which contribute distinctively to the differences between the two time periods under consideration.

The differences between early and contemporary scientific articles are profound and the data confirm what we would expect with regard to the development of English scientific writing (cf. for instance (Atkinson, 1999; Gross et al., 2002; Biber and Gray, 2010)). The scientific paper has evolved from a narrative form with some letter-like features to a strongly standardized article structure and content-focused, compressed structures.

As we can already conclude from the KLD visualizations, the period until 1700 was characterized by few specific content words, but numerous particular function words comprising coordinate and subordinate conjunctions (*and, or, but*), relative pronouns (*which*) as indicators of a distinctive use of clause complexes and lower lexical density and less compressed syntactic structures than in the more recent data. Negation markers and contrastive conjunctions (*not, yet, but*) are typical for the argumentative structure of earlier texts. Additionally, personal pronouns indicate that these earlier texts were characterized by rather explicit interaction between writer and reader (*my, you*), references to other individual (male) scientists (*he*) and the use of long coreference chains and hence a low frequency of new discourse referents within texts (*it, them, their*). The following passage from the 1680s illustrates these typical features of early scientific discourse practices well. Words that occur in the KLD visualization (see Figure 5, right) are highlighted in boldface.

(1) *[. . .]* ***my*** *Reason in short is this: whatever is of sufficient Power to raise the minute Particles of a Heavy Body in a light Fluid, is certainly a sufficient cause to keep* ***them*** *in that state: now* ***my*** *Supposition may give some account of this; what* ***my*** *Brother says, never can; for* ***he*** *must necessarily suppose* ***them*** *first raised;* ***and*** *then* ***he*** *gives the reason of* ***their not*** *sinking: Whereas* ***it*** *is not to be questioned* ***but*** *that that Force* ***which*** *raised* ***them****, is the same that keeps* ***them*** *from falling to the bottom.* (RSC, W. Molyneux, 1686, RSC ID: 101846, DOI: 10.1098/rstl.1686.0015)

The modern texts have a highly standardized article structure with particular sections (e.g. *results*, preferably illustrated with *figures*). Quantitative research methods and means of expressing information symbolically are distinctive for modern articles as indicated, for instance, by nouns referring to general mathematical and scientific expressions such as *measurements, data, values*, and single letters used as abbreviations, e.g. for units, or as variables in formula and mathematical laws. The visualization also reflects the higher thematic specialization of modern scientific journals as we find specific nouns related to the physical and the life sciences (*energy, cell, cells*). Prepositions suggest a distinctive use of phrasal post-modifiers (*in, for, on*) within noun phrases. The verb form *is* could be an indicator for passive use and the importance of linking verbs in the sentence structure. The following passage from an article from the 1960s exemplifies these features that make modern texts quite distinct from historical ones. It is rich in independent clauses with a high frequency of nominal content words and pre- or postmodified noun phrases. Again, words that occur in the visualization (see Figure 5, middle) are highlighted.

(2) *The* ***energy*** *threshold for detection of γ-rays was $\sim$ 30 GeV. Curves of constant* ***energy in*** *the laboratory systems are included.* ***Figure*** *89 shows the* ***results*** *of some of the measurements. Each γ-ray* ***is*** *represented by a point at the appropriate co-ordinate.* (RSC, P. H. Fowler, D. H. Perkins, 1964, RSC ID: rspa_1964_0070, DOI: 10.1098/rspa.1964.0070)

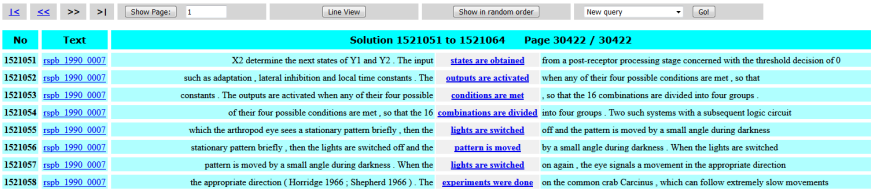
To be able to check in detail the items that are marked as distinctive by KLD, words in the cloud are linked to a CQPweb representation of the corpus, which can be queried by clicking on a word of interest. In CQP(web), more detailed queries can be formulated allowing to further inspect results. Furthermore, query results can be sorted, categorized and downloaded for further analysis as plain-text tables with information on metadata and linguistic annotations.

Figures 6, 7 & 8 show examples from a concordance, a frequency breakdown and a distribution table of nouns followed by a form of *BE* and a passive verb, a grammatical pattern that has become increasingly important over time in the corpus. As illustrated, the RSC on CQPweb allows users to perform various types of sophisticated corpus queries via the web interface and to extract and visualize the results in different ways. Concordances of particular patterns as shown above can be used in the classroom or for linguistic research in order to go beyond information currently available in other types of resources.

7. Summary

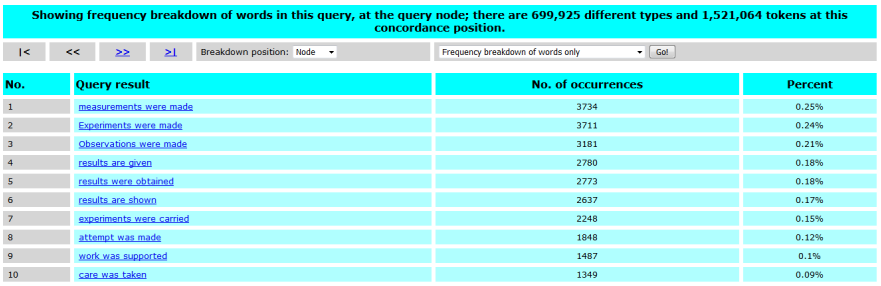
We have presented a new, extended release of the *Royal Society Corpus* (RSC), now covering all publications from the Royal Society of London from 1665 to 1996. We have shown that the RSC fills a gap in the landscape of diachronic, scientific corpora of English (Section 2). Given the role of the Royal Society in scientific publications, the RSC is highly relevant not only for linguistics but also for historical and cultural analysis. The corpus has rich metadata and has been linguistically processed according to best practices. The larger part of the corpus is open and distributed in several formats commonly used by computational and corpus linguists as well as digital humanists. Beyond this, we provide several web services to explore and analyze the corpus that are also freely accessible, such as visualization of differences across time periods on the basis of relative entropy and surprisal.

In our ongoing work, we are enhancing our metadata, e.g. by providing information on disciplines (approximated by topic models) or marking-up the individual authors of texts. This will allow more fine-grained analysis of the linguistic development of disciplines as well as detecting trends due to authors' styles.



No	Text	Solution 1521051 to 1521064 Page 30422 / 30422		
1521051	rspb_1990_0007	X2 determine the next states of Y1 and Y2 . The input	states are obtained	from a post-receptor processing stage concerned with the threshold decision of 0
1521052	rspb_1990_0007	such as adaptation , lateral inhibition and local time constants . The	outputs are activated	when any of their four possible conditions are met , so that
1521053	rspb_1990_0007	constants . The outputs are activated when any of their four possible	conditions are met	, so that the 16 combinations are divided into four groups .
1521054	rspb_1990_0007	of their four possible conditions are met , so that the 16	combinations are divided	into four groups . Two such systems with a subsequent logic circuit
1521055	rspb_1990_0007	which the arthropod eye sees a stationary pattern briefly , then the	lights are switched	off and the pattern is moved by a small angle during darkness
1521056	rspb_1990_0007	stationary pattern briefly , then the lights are switched off and the	pattern is moved	by a small angle during darkness . When the lights are switched
1521057	rspb_1990_0007	pattern is moved by a small angle during darkness . When the	lights are switched	on again , the eye signals a movement in the appropriate direction
1521058	rspb_1990_0007	the appropriate direction (Horridge 1966 ; Shepherd 1966) . The	experiments were done	on the common crab Carcinus , which can follow extremely slow movements

Figure 6: CQPweb: Concordance of nouns followed by a form of BE and a passive verb.

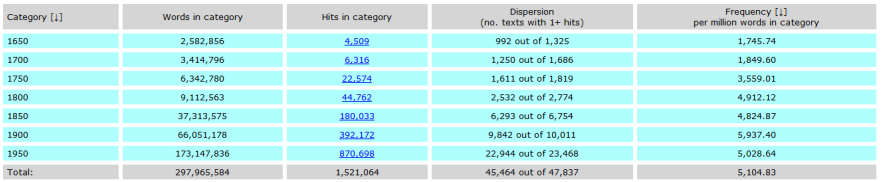


Showing frequency breakdown of words in this query, at the query node; there are 699,925 different types and 1,521,064 tokens at this concordance position.

|< << >> >| Breakdown position: Node — Frequency breakdown of words only — Go!

No.	Query result	No. of occurrences	Percent
1	measurements were made	3734	0.25%
2	Experiments were made	3711	0.24%
3	Observations were made	3181	0.21%
4	results are given	2780	0.18%
5	results were obtained	2773	0.18%
6	results are shown	2637	0.17%
7	experiments were carried	2248	0.15%
8	attempt was made	1848	0.12%
9	work was supported	1487	0.1%
10	care was taken	1349	0.09%

Figure 7: CQPweb: Frequency breakdown of most frequent *noun + BE + passive verb* sequences.



Category [↓]	Words in category	Hits in category	Dispersion (no. texts with 1+ hits)	Frequency [↓] per million words in category
1650	2,582,856	4,509	992 out of 1,325	1,745.74
1700	3,414,796	6,316	1,250 out of 1,686	1,849.60
1750	6,342,780	22,574	1,611 out of 1,819	3,559.01
1800	9,112,563	44,762	2,532 out of 2,774	4,912.12
1850	37,313,575	180,033	6,293 out of 6,754	4,824.87
1900	66,051,178	392,172	9,842 out of 10,011	5,937.40
1950	173,147,836	870,698	22,944 out of 23,468	5,028.64
Total:	297,965,584	1,521,064	45,464 out of 47,837	5,104.83

Figure 8: CQPweb: Distribution of nouns followed by a form of BE and a passive verb across time.

8. Acknowledgements

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 232722074 – SFB 1102 / Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. The creation of the *Royal Society Corpus* (RSC) was also supported by the German Federal Ministry of Education and Research (BMBF) under grant CLARIN-D, the German Common Language Resources and Technology Infrastructure. We are especially grateful to Dr. Louisiane Ferlier, digitisation project manager at Royal Society Publishing, and the Royal Society of London for supporting our corpus building effort with their advice and data.[12]

9. Bibliographical References

Atkinson, D. (1999). *Scientific discourse in sociohistorical context: The philosophical transactions of the Royal Society of London, 1675–1975*. Lawrence Erlbaum, Mahwah, NJ, USA.

Biber, D. and Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9:2–20.

Biber, D., Finegan, E., and Atkinson, D. (1994). ARCHER and its challenges: Compiling and exploring a representative corpus of historical english registers. In Udo Fries, et al., editors, *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993*, pages 1–13, Amsterdam. Rodopi.

Broeder, D., Schonefeld, O., Trippel, T., Van Uytvanck, D., and Witt, A. (2011). A pragmatic approach to XML interoperability—the component metadata infrastructure (CMDI). In *Balisage: The Markup Conference 2011*, volume 7.

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

Crespo-García, B. and Moskowich, I. (2015). A Corpus of History Texts (CHET) as Part of the Coruña Corpus Project. In *Proceedings of Corpus Linguistics*, pages 14–23, St. Petersburg, Russia.

Davies, M. (2010). The Corpus of Historical American English (COHA): 400 million words, 1810–2009. Available online at `https://corpus.byu.edu/coha/`.

De Smet, H., Flach, S., Tyrkkö, J., and Diller, H.-J. (2015). The Corpus of Late Modern English Texts (CLMET), version 3.1: Improved tokenization and linguistic annotation. KU Leuven, FU Berlin, U Tampere, RU Bochum. Available online at `https://hdl.handle.net/21.11119/0000-0002-43F3-0`.

Degaetano-Ortlieb, S. and Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and*

[12] `https://blogs.royalsociety.org/publishing/linguistics-history-of-science/`

Literature at COLING 2018, pages 22–33, Santa Fe.

Degaetano-Ortlieb, S. and Teich, E. (2019). Towards an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory (open access)*, pages 1–33.

Degaetano-Ortlieb, S., Kermes, H., Lapshinova-Koltunski, E., and Teich, E. (2013). SciTex: a diachronic corpus for analyzing the development of scientific registers. In Paul Bennett, et al., editors, *New Methods in Historical Corpora*, volume 3 of *Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP)*, pages 93–104. Narr, Tübingen.

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*, University of Birmingham, Birmingham, UK.

Fankhauser, P., Kermes, H., and Teich, E. (2014a). Combining macro- and microanalysis for exploring the construal of scientific disciplinarity. In *Digital Humanities 2014: Book of Abstracts*, pages 461–463, Lausanne, Switzerland. Alliance of Digital Humanities Organizations (ADHO).

Fankhauser, P., Knappen, J., and Teich, E. (2014b). Exploring and visualizing variation in language resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4125–4128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Fankhauser, P. (2017). Visual correlation for exploring paradigmatic language change. Talk at the workshop *Making effective use of metadata of historical texts and corpora*. Available online at `http://corpora.ids-mannheim.de/diaviz/material/visual_correlation_slides.pdf` (last accessed on 14th November 2019).

Feltgen, Q., Fagard, B., and Nadal, J.-P. (2017). Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. *Royal Society Open Science*, 4(11):170830.

Fischer, S., Fankhauser, P., and Teich, E. (2017). Visualization of corpus frequencies at text level. In *Corpus Linguistics 2017: Conference Abstracts*, University of Birmingham, Birmingham, UK.

Fyfe, A., McDougall-Waters, J., and Moxham, N. (2015). 350 years of scientific periodicals. *Notes and Records: the Royal Society Journal of the History of Science*, 69(3):227–239.

Gross, A., Harmon, J., and Reidy, M. (2002). *Communicating science: The scientific article from the 17th century to the present*. Oxford University Press, Oxford.

Hardie, A. (2012). CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17:380–409.

Hellrich, J. and Hahn, U. (2017). Exploring diachronic lexical semantics with JeSemE. In *Proceedings of ACL 2017, System Demonstrations*, pages 31–36, Vancouver, Canada. Association for Computational Linguistics.

Hellrich, J., Buechel, S., and Hahn, U. (2018). JeSemE: Interleaving semantics and emotions in a web service for the exploration of language change phenomena. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 10–14, Santa Fe, New Mexico. Association for Computational Linguistics.

Hinrichs, E., Hinrichs, M., and Zastrow, T. (2010). WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29, Uppsala, Sweden. Association for Computational Linguistics.

Jurish, B. (2018). Diachronic collocations, genre, and DiaCollo. In Richard J. Whitt, editor, *Diachronic Corpora, Genre, and Language Change*, volume 85 of *Studies in Corpus Linguistics*, pages 41–64. John Benjamins, Amsterdam.

Kermes, H. and Teich, E. (2017). Average surprisal of parts-of-speech. In *Proceedings of the Corpus Linguistics 2017 Conference*, University of Birmingham, Birmingham, UK.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J., and Teich, E. (2016). The Royal Society Corpus: From uncharted data to corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Klaus, C., Klakow, D., and Fankhauser, P. (2019). OCR post-correction of the Royal Society Corpus based on the noisy channel model. In *Proceedings of the 41. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS 2019)*, University of Bremen, Germany.

Knappen, J., Fischer, S., Kermes, H., Teich, E., and Fankhauser, P. (2017). The making of the Royal Society Corpus. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 7–11, Gothenburg, Sweden. Linköping University Electronic Press.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Moxham, N. and Fyfe, A. (2018). The Royal Society and the prehistory of peer review, 1665–1965. *The Historical Journal*, 61(4):863–889.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank project (3rd revision). Technical Report MS-CIS-90-47, University of Pennsylvania, Department of Computer and Information Science.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Taavitsainen, I., Pahta, P., and Mäkinen, M. (2005). Middle English Medical Texts. CD-ROM.

Taavitsainen, I., Pahta, P., Hiltunen, T., Mäkinen, M., Mart-

tila, V., Ratia, M., Suhr, C., and Tyrkkö, J. (2010). Early Modern English Medical Texts. CD-ROM.

TEI Consortium. (2019). TEI P5: Guidelines for electronic text encoding and interchange. Version 3.5.0. Last updated on 29th January 2019. Available online at `https://www.tei-c.org/Guidelines/P5/` (last accessed on 14th November 2019).

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: the virtual language observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Voormann, H. and Gut, U. (2008). Agile corpus building. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251.

Wilkinson, M. D., Dumontier, M., and [...] Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Nature - Scientific Data*, 3:160018.

Yáñez-Bouza, N. (2011). ARCHER past and present (1990–2010). *ICAME Journal*, 35:205–236.