Using Diachronic Corpora of Scientific Journal Articles for Complementing English Corpus-based Dictionaries and Lexicographical Resources for Specialized Languages

Katrin Menzel
Dept. of Language Science and Technology, Saarland University
E-mail: k.menzel@mx.uni-saarland.de

Abstract

As technology and science permeate nearly all areas of life in modern times, there is a certain trend for standard dictionaries to bolster their technical and scientific vocabulary and to identify more components, for instance more combining forms, in technical terms and terminological phrases. In this paper it is argued that recently built diachronic corpora of scientific journal articles with robust linguistic and metadata-based features are important resources for complementing English corpus-based dictionaries and lexicographical resources for specialized languages. The Royal Society Corpus (RSC, ca. 9,800 digitized texts, 32 million tokens) in combination with the Scientific Text Corpus (SciTex, ca. 5,000 documents, 39 million tokens), as two recently created corpus resources, offer the possibility to provide a fuller picture of the development of specialized vocabulary and of the number of meanings that general and technical terms have accumulated during their history. They facilitate the systematic identification of lexemes with specific linguistic characteristics or from selected disciplines and fields, and allow us to gain a better understanding of the development of academic writing in English scientific periodicals across several centuries, from their beginnings to the present day.

Keywords: English diachronic corpora, Graeco-Latin combining forms, scientific journal articles, scientific vocabulary, corpus-based dictionaries, lexicographical resources for specialized languages

1 Introduction

The value of large corpora for the complementation of lexicographical resources has been demonstrated by various scholars, e.g. Podhajecka (2010, 2011) using Google Books. Although this huge book corpus posed some challenges due to partly incorrect metadata (cf. James & Weiss 2012) and poor OCR quality, it proved useful, e.g. for antedating various OED headwords. This paper will argue that recently created, large specialized corpora representing specific text types, technical areas, academic vocabulary and domain-specific words from different stages of the English language have not yet been exploited to their full potential. Such resources will enhance the systematic identification of lexemes and expressions with predefined linguistic features which otherwise may be found only rarely or randomly in general corpora as broad samples of language use that may lack fine-grained annotations. Numerous citations that illustrate the natural usage of words in current dictionaries, such as the *Oxford English Dictionary* – the “definitive record of the English language” – have been drawn from general language texts and from various canonical and culturally important types of sources, such as literary, theological, historical and philosophical works of renowned authors and particular time periods. Nowadays, such lexicographical resources are updated constantly with examples from Present-Day English that are often drawn from the internet, from digital archives and electronic corpora, e.g. from newspapers, magazines or blogs. It is worth investigating whether a more differentiated and complete view of the English language can be achieved by integrating more information retrieved from enriched domain-specific, specialized electronic corpora that are becoming more widely available and searchable for a variety of linguistic patterns. An important source for lexicographical research are scientific periodicals serving as vehicles for the communication of new discoveries and ideas and as repositories of knowledge. In this paper, it will be shown how two large scientific journal corpora can be used as linguistic resources for lexicographers. If we take, for instance, items containing combining forms (CFs), these data offer valuable insights into the general linguistic and terminological developments in scientific writing, particularly with regard to topics related to natural phenomena and the study of the material world.

2 Resources

The analyses presented in this paper are based on two English scientific journal corpora that can be queried via the Saarbrücken CQPweb (Corpus Query Processor) interface: the Royal Society Corpus (RSC; Kermes et al. 2016) and the Scientific Text Corpus (SciTex; Degaetano-Ortlieb et al. 2013).[1] These corpora cover scientific articles from the early stages of the first scholarly journals published in English from the middle of the 17th century onwards to contemporary scientific publications. The latest release of the RSC (V3.7.0) consists of around 9,800 digitized texts (ca. 32 million tokens) published between 1665 and 1869 in the *Philosophical Transactions* (*Philos. Trans*.) and the *Proceedings* (*Proc*.) of the Royal Society of London – publications that figure among the longest-running English scientific journals in the world. These journals used to cover all branches of science of the time. Later, when the number of scientific journals grew considerably, they became increasingly specialized. Each of them split into separate series (series A covering mathematics and the physical sciences, and B focusing on the life sciences). The authors were British fellows of the Royal Society and other outstanding scientists of the time. Early research articles resembled letters addressed to a wider audience in many ways. We see a development in this genre from a more involved to a more informational style, and the development of research community standards with regard to terminological consistency and typical phrases and collocations. Linguistic precision, objective style and the features of plain English became more dominant textual features over time. We are currently extending the RSC corpus in the framework of the project *Information Density and Scientific Literacy in English – Synchronic and Diachronic Perspectives*[2] by integrating all texts from these publications, including their A and B series, from five additional decades into the corpus. (1870-1920). This resource will therefore cover the entire Late Modern English period (LModE, ca. 1700-1900) as well as the transition periods at the beginning and the end of LModE, i.e. the end of the Early Modern English period and the beginning of Present-Day English. The ongoing corpus extension will also increase the corpus size substantially as these journals, particularly *Proc.,* grew in size over time.[3]

The RSC contains annotations at different linguistic levels, for instance, part-of-speech and lemma annotations and information on historical spelling variants. It provides a number of query and graphical

[1] The RSC has been made available for free download and online query from the CLARIN-D center at Saarland University under the persistent identifier http://hdl.handle.net/11858/00-246C-0000-0023-8D1C-0, cf. also http://corpora.clarin-d.uni-saarland.de/cqpweb/. Access to SciTex can be granted to researchers upon request. As it includes texts from the more recent past, there are stricter copyright restrictions than with the RSC and users interested in these data for research purposes cannot easily be given unrestricted access to the full corpus texts.

[2] http://www.sfb1102.uni-saarland.de/?page_id=275

[3] One might argue – given that the *Philosophical Transactions* of the Royal Society of London already are among the top ten sources of the OED with almost 16,000 quotations – that there is no further need to analyse more material from this journal or similar resources, at least not for this dictionary. Indeed, numerous quotations in the OED come from scientific journals, but a considerably much higher number has been drawn from other specific sources representing entirely different text types such as newspapers, poets and literary prose writers (e.g. ca. 42,000 quotations from the Times, ca. 33,000 from Shakespeare's works, more than 5,000 from Milton's *Paradise Lost* etc.).

visualization options.[4] Various types of metadata that were collected or generated are encoded for each text, e.g. information on the author, publication date, text type (e.g. abstract, book review, full article etc.) and text topics (based on probabilistic topic modelling, cf. Fankhauser, Knappen & Teich 2016) so that the texts can be sorted according to subjects such as optics, thermodynamics, botany, metallurgy, medicine, chemistry, etc. To obtain more variables for analysis, we are currently enriching the corpus with more fine-grained syntactic and morphological annotations and more types of high-quality metadata, some of which we recently obtained from the newly indexed and digitized Royal Society Journal Collection. Language models were built to estimate predictability and information density at the level of words and certain word-internal elements. Surprisal, an information-theoretic operationalization, has been calculated for various units (e.g. tokens and various word-internal elements) and has been annotated. Our current results confirm an increasing specialization of scholarly work reflected in the growing number of informationally dense expressions and structures. Additionally, an ongoing conventionalization of various linguistic structures can be observed over time.[5] The RSC is complemented with SciTex, an equally well-annotated large corpus of more contemporary texts from the 1970/80s and the early 2000s. It consists of English scientific journal articles from several disciplines, such as computer science, linguistics, biology, and mechanical engineering. The current version contains approximately 5,000 documents and 39 million tokens. Like the RSC, it is also tokenized, lemmatized and part-of-speech tagged, and contains rich metadata, e.g. information on the author(s), discipline/topic and year of publication.

Working with data from these specialized corpora by taking full advantage of their linguistic and metadata annotations and their sophisticated query options has substantial potential for making a unique contribution to lexicographical knowledge. The data offer possibilities for compiling or enhancing discipline-specific reference sources, e.g. terminology databases, technical dictionaries and encyclopaedias, and for assessing the vocabulary coverage of general dictionaries. Early scientific and technical terms are a particularly important part of English vocabulary. They can give us essential insights into co-developments between language and society. Many early terms have changed with the evolution of science and technology, some have dropped out of use, and others have become common in non-specialized use or in even more specific contexts than in the past. Large historical dictionaries like the OED may benefit from the possibility of incorporating newly retrieved information resulting from the systematic search in the RSC and SciTex for certain types of lexemes and word-formation patterns, as well as for typical collocations and multiword expressions that function as technical terms. If we take, for instance, the adjective *ethereal*, general dictionaries emphasize its chiefly literary or poetic usage or its function as synonym for *beautiful*, *delicate* or *heavenly* in formal contexts. The OED illustrates this with quotes from various literary and other sources. The word is quoted there less frequently in precise technical meanings from scientific contexts where it occurs in specific, term-like adjective-noun-patterns for which we find many historical and contemporary examples in our corpus data. The nouns in this pattern become more and more specific over time in the corpora. Many can be found in the 1740s (*ethereal fire / matter / medium* etc.), in the 1850/60s (*ethereal solution / odour / filtrate / extract* etc.) and in the 20th century (e.g. *ethereal diazomethane*).

The RSC in combination with SciTex as two recently created corpus resources that offer the possibility to provide a fuller picture of the development of specialized vocabulary and of the number of meanings that general and technical terms have accumulated during their history. They facilitate the systematic identification of lexemes with specific linguistic characteristics or from selected disciplines

[4] Cf. also Knappen et al. (2017) for annotation quality, OCR correction methods and the evaluation and improvement of part-of-speech tagging in the RSC.

[5] Surprisal is an information-theoretic notion measuring the probability of a linguistic unit to occur in a given textual context. It has the advantage of accounting for probabilities conditioned on a context, which cannot be achieved by considering mere frequencies (see also Degaetano-Ortlieb and Teich (2016) for a comparison of surprisal and type-token ratio to investigate productivity).

and fields, and allow us to gain a better understanding of the development of academic writing in English scientific periodicals across several centuries, from their beginnings to the present day.

3 Case Study on Combining Forms (CFs)

The following section reports some illustrative findings from a case study on different morpheme types in our corpora which underline the relevance of such resources for lexicographical and terminographical purposes. The main focus of this chapter will be on Graeco-Latinate combining forms[6] in English derived from classical nouns, verbs and adjectives, a morpheme type playing an important role in numerous scientific formations in combination with other elements that are also often of Latinate or Greek origin. One example of these CFs is *-lysis* and its variants, which can be traced back to Greek roots (λύειν 'to loosen' or λύσις 'loosening'). It exemplifies the complexity of this morpheme group as it plays a role in a variety of lexeme-formation processes. This CF co-occurs with various root morphemes and affixes in our data (e.g. *photo+lysis, para+lytic+al, dia+lys+er, hydro+geno+l-ysis*). In earlier texts, it occurs in various neoclassical compounds and derivations. In texts from the more recent past, new lexemes also reflect current trends in word-formation processes, such as clipping and blending involving CFs. There are also backformations with word-class changes and newly coined independent lexemes that have evolved from initially bound CFs into free morphemes (e.g. *lysis* as noun and *lyse* as verb) as well as some hybrid forms with neoclassical and native English elements (e.g. *LysoTracker*, a trademark for a lysosome tracker). Term formation in the early stages was frequently based on adjective-noun patterns with at least one CF involved. The methods and insights from the results on CFs can be generalized to other morpheme types (e.g. pre- and suffixes), specific word-classes, word-formation processes or part-of-speech patterns if they are characterized by formal similarities so that they can be identified via corpus queries.

Various monolingual English dictionaries apply the category of CFs to initial and final bound lexical elements in the description of complex words and their components, but the information available on CFs in dictionaries can only give us a rough estimation of how productive these elements are and how many different words have been coined with them in English as a whole, or in particular time periods or registers. As many words with such forms tend to be rare in average texts and general corpora of modern English, they are sometimes assumed to play only a marginal role in the language as a whole. However, in scientific and technical English their components are very productive. In diachronic English corpora, such as our specialized corpora of English academic writing, they are related to important register-specific word-formation patterns. They are also very interesting from a cross-linguistic perspective, as they have often been incorporated into English as lexical or phraseological borrowings from other European languages or vice versa, and in many cases serve as scientific internationalisms. As argued by ten Hacken and Panocová (2014), it is useful to include entries with such neoclassical formatives in lexicographical resources and to link them in electronic dictionaries with the entire group of words they appear in (for a summary of the treatment of CFs in lexicographic and lexicological resources and studies and the role of combining forms in scientific discourse cf. also Menzel and Degaetano-Ortlieb (2017: 187-209)). McCauley (2006) gave an overview on some aspects of the revision of etymology sections in CFs entries in the 3rd edition of the OED, and explained that dictionary editors hope to take advantage of newly available data to provide more consistent, transparent and informative entries for CFs.

[6] Lexicographers nowadays also include other word-formation elements of native and non-native origins in the list of combining forms, e.g. forms derived from prepositions (e.g. *by-*) or affixes (*-ene*), truncated words or clipped word fragments (*splinters*) such as *-gate* in the sense of *scandal* and pseudo-morphemes that have undergone semantic and structural reanalysis by processes of analogy (e.g., *-aholic, -(a)-thon*) which are productive sources for novel blends in contemporary creative language use, advertisements, media discourse and quasi-technical jargon.

This group of morphemes presents various challenges. As a potentially open-class category due to their lexeme-like semantics, CFs do not share specific semantic features and cannot be identified automatically as a group through a certain length or specific combinations of letters. Around 2,300 elements are currently identified as CFs in the OED, but there seem to be many more that have been used or are still used productively in word-formation processes in English. Other resources, e.g. those edited by Quinion (2005) or Sheehan (2000), focus specifically on lists of word-initial and word-final elements. Such elements are typically presented with no strict dividing line between affixes and CFs, and are labelled with more general terms in their description, e.g. *word parts*, *vocabulary elements* or *word beginnings* and *endings*. In addition to alphabetical lists, word parts are sometimes presented in thematic lists, for instance on anatomy and physiology. In Sheenan (2000: 187) one example for a thematic list are morphemes related to 'breath' such as *afflat- (afflatus); -hale (exhale); halit- (halitosis); ozostom- (ozostomia); -pnea (apnea); pneo- (pneograph); pneumato- (pneumatometer); pneumo- (pneumobacillus); pneumono- (pneumonophorous); pneusio- (pneusiobiognosis); pulmo- (pulmnonary); spiro- (spirograph).* The information on CFs obtained from such existing specialized lexicographic resources and general corpus-based dictionaries can be taken as a starting point for the search for domain-specific or semantically related CFs in order to verify and complement our current knowledge through the analysis of our corpora. This will allow us to identify and record new words, specific meanings and semantic shifts and to check whether first attestations in established dictionaries can be antedated.

If we search, for instance, the morpheme group *pneo-, pneu-, pulmo- spiro-* with a CQP query in the RSC and SciTex, all lexemes and collocational patterns with these components can be retrieved in these annotated corpora of scientific texts from Early Modern English to Present-Day English, currently comprising 71 million tokens. These CFs occur around 1,700 times in about 420 different texts, and are slightly more frequent in the older data from the RSC than in SciTex.[7] Among the earliest words with these components in the data are *pneumatic* and *pneumatical* (e.g. in *pneumatic / pneumatical engine*), and *pneumatiques* (as in *science of pneumatiques*) in the 17th century, indicating some lexical, terminological and spelling variation in scientific writing of this time. Some early terms and collocations were already relatively complex patterns, e.g. *hydraulo-pneumatical fountain* (1660s/70s). The most frequent result found with this query in both the earlier and the more contemporary corpus is the adjective *pulmonary* (typically in adjective-noun patterns such as *pulmonary artery / vein / vessels* etc., again with some similarly structured near-synonyms, e.g. *pneumonique vessels / arteries* etc.). Newer words with one of these CFs (e.g. *electropneumatic* or *pneumoencephalography*) reflect scientific developments and specialization processes. Many lexemes with these elements become established and conventionalized; and increasingly more derivatives and combinations with native forms can be found in the data, e.g. *spirometrically-measured* in a text on bioinformatics from the 1970s. It is possible to focus on selected n-gram types and part-of-speech patterns to find out which are most typical for particular time periods, e.g. trigrams such as adjective-adjective-noun patterns that function as collocations or compounds (as in *pneumatic mercurial apparatus, mercurial pneumatic apparatus* (1800s/1810s), *great pneumogastric nerve, pulmonic capillary vessels, pulmonary semilunar valves* (1830s-1860s)). In earlier texts, there is a noticeably higher lexical variability in our data, later texts are characterized by fewer general-language words and more specific adjectives, more standardized terminological usage and adjective-adjective-noun patterns embedded in more complex nominal structures (e.g. *the anterior aortic and pulmonary semilunar valve-rudiments* (1869)).

Another example is the CF *iso-* and its Latin-based equivalent *equi-* ('equal').[8] They often occur in multimorphemic adjectives. Related lexemes, derivations and frequencies of variants can easily

[7] A few lexemes are filtered out by more precise queries out as spiro- (from Latin *spīrāre* – 'to breathe') can also have a different sense (from Latin *spīra* 'coil, spiral') as in some names of spirally shaped bacteria.

[8] Variants are *is-* (possible before vowels, but does not occur in our dataset) and *aequi-* (in older texts, e.g. *aequivocal*).

be found and extracted (e.g. *isochronous*, *isochronal*, *isochronic*, *isochron*, *isochronously*, *isochronism*). Additionally, the data may be sorted and filtered according to selected categories of metadata to obtain relevant quotations from specific time periods or types of authors if they appear to be underrepresented in the dictionary quotation database. It is possible to select only those quotations, for instance, from early female scientists such as Caroline Herschel (who used this CF in the term *isosceles triangle*), from American scientists from the 18th century such as Benjamin Franklin (using for instance the noun *equinoctial*), authors that may be slightly less well-known today, but published far more texts than others, e.g. Everard Home (using *equivocal* as a predicative adjective) or from specific decades or other time slices (e.g. some words seem to be under-represented in 18th century quotations in dictionaries).[9] Newer texts from SciTex include few coinages with the Latin-based *equi*, but several lexemes that were coined from the end of the 19th century onwards with the CF *iso-* (e.g. *isochore*, *isologous*, *isoleucine*, *isocyanate* and *isooctane*). It is possible to sort the results according to topics and disciplines that are specified in the document metadata. Texts on the solar system in the RSC, for instance, frequently include these CF in words such as *equinoctial* (or sometimes *equinoxial*), *equinox*, *equidistant* / *equi-distant* and in occasionally occurring words such as *equilateral*, *equiangular*, *isosceles* and *isochronal*. Texts from similar time periods in the RSC on chemistry have more types and tokens with these CFs in total. Among the most frequently used lemmata with these forms in chemistry texts are *equivalent*, *isomer*, *isomeric*, *isomerism*, *isomorphous*, *isomorphism*, *isoprene*, *isopropyl*, *equi-diffusive* / *equidiffusive* and some adjectives related to geometry such as *equidistant*, *isosceles* and *equilateral*. Texts with mathematics as the primary topic in the RSC include frequent occurrences of *equivalent*, *equilibrium*, *equidistant*, *equilateral*, *isotropic* and some occurrences of lexemes such as *isodynamic*, *equipotential*, *isoperimetrical*, *equimultiples*, *equiponderant*, *equiradial* and *isobarism*.

CFs like these examples are distributed differently across research articles from different academic disciplines, but there is also variation across different text types. The two morpheme groups *pneo- / pneu- / pulmo- / spiro-* and *iso- / equi-* are most frequently found in abstracts (ca. 63 and 493 occurrences per million words (pmw), respectively) while the lowest frequencies can be found in review papers (ca. 35 and 92 occurrences pmw). Text types with high frequencies of these initial CFs are not only characterized by a large number of low-frequency forms indicating a high productivity rate of these morphemes, but also by a certain use of lexical repetition and high frequency terms. Final CFs, particularly those with less specific or more abstract meanings such as *-(o)logy* and *-(o)graphy*, tend to play the most prominent role in review papers in the data. These two forms taken together occur ca. 214 times pmw in reviews compared to 133 in abstracts and 70 in full research papers. They steadily increase in frequency over time in all corpus registers. Earlier corpus texts contain a relatively low number of hybrid forms. Newer coinages more freely combine neo-classical elements of Greek and Latin origin with each other. In the more contemporary corpus texts, we also find more independent lexemes that have evolved from bound CFs (e.g. *graph*, *photo*, *bio*) and more combinations of CFs with native elements and fully anglicized lexemes of various origins (e.g. *cell-biology*, *mailgram*, *ultrasoundcardiography, polarography*), with proper names and acronyms (*roentgenology*, *galvanoscopic*, *DUV-lithography* etc.). Several low-frequency items, particularly in newer texts, contain more than two CFs within one lexeme indicating increasingly specialized discussions requiring more specific and complex terms (e.g. *psychobiology*, *photolithography*, *electrocardiography*, *cinephotomacrography* and *electrophoresis-homochromatography*).

[9] Texts from female scientists are rare in the time span currently covered by the RSC as the Royal Society did not elect any female fellows until the 1940s. American (colonial) scientists supplied relatively few contributions until the end of the 18th century, but there are a number of articles written by such authors in this period on astronomy and electricity. The person who wrote most articles in the in the journals included our corpora is Everard Home. He published more research papers in *Philos. Trans*. than anyone else but he is not quoted in the OED from this source.

Highly frequent lexemes with CFs that are also used in non-scientific discourse and general language sometimes become semantically bleached over time towards a less literal usage or lose their compositionality (e.g. *apology*, *equivalent*), while many morphologically complex terms with CFs remain relatively transparent with regard to their component morphemes. This does not necessarily mean that the exact meaning of compounds or multiword expressions with CF can easily be derived from the literal meaning of their constituents without knowing a proper definition or the context in which they are used, but they give strong hints as to the meaning of technical and scientific terms. Classically educated scholars and physicists of the 19th century who were familiar, for instance, with the concept of the *luminiferous* (= light-bearing / light-producing) *ether* and with word-formation patterns with combining forms in general probably had no difficulties in decomposing similarly structured complex adjectives in phrases such as *sanguiferous vessels*, *carboniferous limestone*, *mortiferous diseases* or *odoriferous flowers*.

The corpus data may also serve to provide additional information to that contained in dictionaries. *Luminiferous*, for example, is quoted twice in the OED from *Philos. Trans*., but it is not straightforward to see there from the information given with regard to date and author whether these sentences

(1) and (2) are from the same or from different text:
(1) 1801 Young in *Philos. Trans*. (Royal Soc.) 92 22 The actual velocity of the particles of the luminiferous ether.[10]
(2) 1802 T. Young in *Philos. Trans*. (Royal Soc.) 92 14 A luminiferous Ether pervades the Universe, rare and elastic in a high degree.

Queries for these passages in the RSC will enable users to browse detailed lists of document metadata to obtain information on the title of the paper, the author, text topics, and so on, and to find the full text by clicking on a DOI link. Then it can easily be seen that the text from which these two quotes were taken was read before the Royal Society in 1801 as the one of Thomas Young's Bakerian Lectures and published in 1802. Our corpus data indicate in which time periods certain CFs played a particularly important role (e.g. *-scope* was used especially productively in the RSC between 1750 and 1800), and we can find words that are morphologically related and not attested in the OED yet (e.g. *colloidoscope*), or where the OED records quotations only from a specific time period. *Laryngoscope*, for instance, is quoted in the OED only from texts between 1860 and 1880, but has been assigned to Frequency Band 4 indicating a moderately high frequency in current use.[11] In the RSC, it occurs a few times in the 1860s and in Scitex once in the 1970s. Preferred general, domain-specific or time-specific spelling variants can be detected in the data, e.g. in the 19th century the adjective *hypovanadous* was sometimes used in texts on chemistry, most frequently in the collocation *hypovanadous salt*.[12] In the OED, only the differently spelt *hypovanadious* is recorded which does not occur in our dataset.

If new terms are introduced, various types of recurring multiword sequences can be queried in the corpora as indicators for lexical innovation and neologistic forms (e.g. 'propose to call' as in 'I have constructed the instrument which I propose to call the Stereomonoscope' or 'Mr. Faraday proposes to call the acid […] Sulpho-naphthalic Acid.'). Additionally, the data indicate how and by whom scientific terms and collocations were popularized in English in certain time periods, and in which domains they used to play a dominant role. The adjective-noun collocation *galvanoscopic frog(s)* is

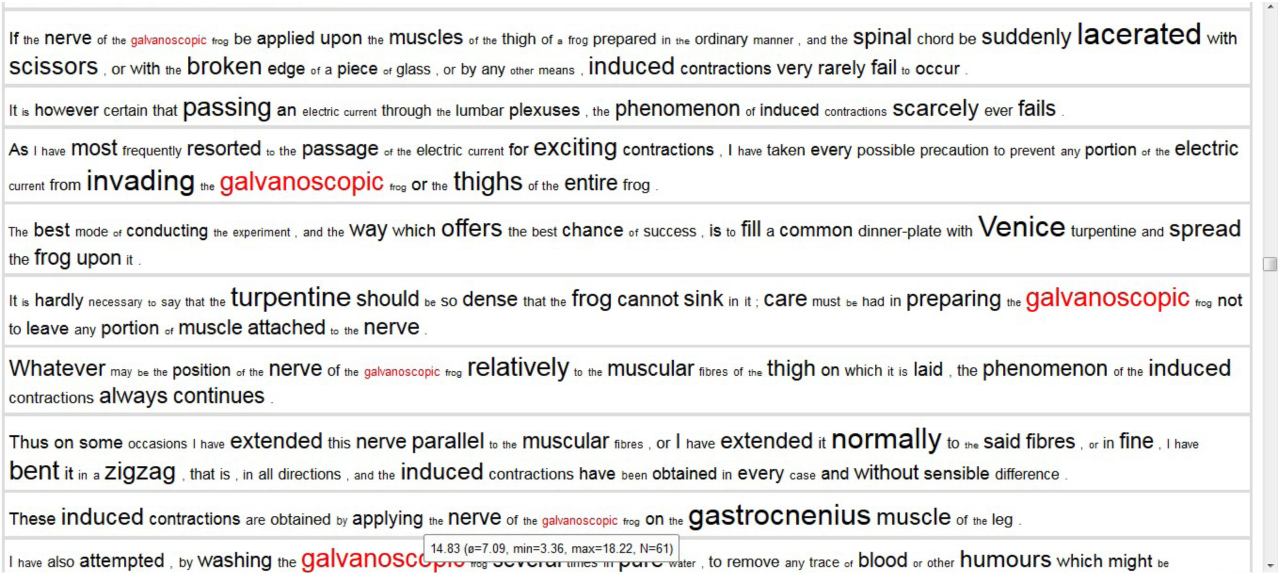
[10] cf. http://www.oed.com/view/Entry/111121 and http://www.oed.com/view/Entry/64728

[11] http://public.oed.com/how-to-use-the-oed/key-to-frequency/

[12] *Hypovanadous salt* in the corpus data refers to another concept than *hypovanadic salt*, which is also indicated by the fact that both terms typically co-occur in the same corpus texts. Moreover, they have cognates in French texts on chemistry from similar time periods (*des sels hypovanadeux / hypovanadiques*). It should be kept in mind that lexemes that may look structurally very similar and could appear to be synonyms or term variants (as in the examples *pneumonic / pulmonary* and *pneumatic / pneumatical* discussed above) might refer to quite different scientific concepts.

an example that was relatively frequent in in our data in texts from the 1840s and 1850s and also occurred a few times in texts on experiments, electromagnetism and physiology from later time periods. It was used repeatedly in various articles in *Philos. Trans*. by Carlo Matteucci, an Italian neurophysiologist. As already mentioned in Section 2, surprisal values as information-theoretic operationalization for measuring information density have been annotated in the data. The information gained from this probability value reveals an additional aspect to observed unconditioned frequencies and type-token ratios of words as indicators of productivity. Surprisal provides us with additional information on the probability of a linguistic unit to occur in a given textual context. It is defined as the negative logarithm of the probability of a unit (e.g. a word) in context (e.g. its preceding words), and it is measured in bits. The value indicates, for instance, whether a given word is more *surprising* after a sequence of preceding words and hence provides more information (see also Degaetano-Ortlieb and Teich (2016) for more details and a comparison of surprisal and type-token ratio). Surprisal values can be downloaded for the corpus query results or visualized with the tool implemented in CQPweb (Fischer, Fankhauser & Teich 2017).

Figure 1 shows a visualization of surprisal values for an extract from a corpus text from 1845. This figure shows, for instance, that in our corpus data, the adjective *galvanoscopic* most typically occurs before the word *frog*, sometimes also before words such as *effects*, *nerve*, *limb* or *leg*. The most typical contextual pattern in which this adjective is used is the noun phrase *the nerve of the galvanoscopic frog*. *Galvanoscopic* in general is not a frequently used adjective in the whole corpus. However, it has different surprisal values in different syntactic and lexical contexts. The probability of this adjective occurring after the word sequence *nerve of the* is relatively high, and therefore the surprisal value of this word in such contexts is low, indicated by the small font size in the visualization of Figure 1. After other preceding contexts it is less expected and thus visualized with a larger font size. However, the word *frog* can be expected with an increasingly high probability after this adjective.



If the nerve of the galvanoscopic frog be applied upon the muscles of the thigh of a frog prepared in the ordinary manner , and the spinal chord be suddenly lacerated with scissors , or with the broken edge of a piece of glass , or by any other means , induced contractions very rarely fail to occur .

It is however certain that passing an electric current through the lumbar plexuses , the phenomenon of induced contractions scarcely ever fails .

As I have most frequently resorted to the passage of the electric current for exciting contractions , I have taken every possible precaution to prevent any portion of the electric current from invading the galvanoscopic frog or the thighs of the entire frog .

The best mode of conducting the experiment , and the way which offers the best chance of success , is to fill a common dinner-plate with Venice turpentine and spread the frog upon it .

It is hardly necessary to say that the turpentine should be so dense that the frog cannot sink in it ; care must be had in preparing the galvanoscopic frog not to leave any portion of muscle attached to the nerve .

Whatever may be the position of the nerve of the galvanoscopic frog relatively to the muscular fibres of the thigh on which it is laid , the phenomenon of the induced contractions always continues .

Thus on some occasions I have extended this nerve parallel to the muscular fibres , or I have extended it normally to the said fibres , or in fine , I have bent it in a zigzag , that is , in all directions , and the induced contractions have been obtained in every case and without sensible difference .

These induced contractions are obtained by applying the nerve of the galvanoscopic frog on the gastrocnenius muscle of the leg .

I have also attempted , by washing the galvanoscopic frog several times in pure water , to remove any trace of blood or other humours which might be

14.83 (ø=7.09, min=3.36, max=18.22, N=61)

Figure 1: Visualization of higher and lower surprisal values by the use of larger and smaller font sizes in a corpus excerpt from the Royal Society Corpus.

The examples presented in this section demonstrate how the RSC and SciTex can be used to effectively to search and identify lexemes, multi-word expressions and collocations in which combining forms play a role. They illustrate the insights which can be gained through a detailed analysis of the corpus resources.

4 Conclusion

The linguistic evidence found in our specialized diachronic corpora can complement English corpus-based dictionaries and lexicographical resources for specialized languages in several ways. The convenient query and visualization options in the Royal Society Corpus and Scientific Text Corpus provide various research routes for lexicographical and lexicological purposes. They allow us to browse an enormous number of specialized texts on a broad spectrum of topics with fine-grained annotations and rich metadata. This enables us to find linguistic structures with certain components, used by specific authors, in particular time spans or with respect to particular topics in a systematic way. This has substantial potential to lead to a more nuanced understanding of the developments and dynamics of language use in specialized registers over time. The examples given in this paper have particularly emphasized the usefulness of diachronic corpora for updating and improving historical dictionaries, but the methods described can also be applied to dictionaries of contemporary language. The information obtained from these corpora can be used to improve various corpus-based general and specialized lexicographical reference works.

References

Degaetano-Ortlieb, S., Kermes, H., Lapshinova-Koltunski, E. and E. Teich (2013). SciTex – A Diachronic Corpus for Analyzing the Development of Scientific Registers. In Bennett, P. et al. (eds.). *New Methods in Historical Corpus Linguistics. Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP, 3.* Tübingen: Narr, pp. 93-104.

Degaetano-Ortlieb, S. & Teich, E. (2016). Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In *Proceedings of Language Technologies for the Socio-economic Sciences and Humanities (LATECH'16), Association for Computational Linguistics (ACL), 7-12 August 2016,* Berlin, Germany, pp. 165-173.

Fankhauser, P., Knappen, J. and Teich, E. (2016). Topical Diversification over Time in the Royal Society Corpus. In *Digital Humanities 2016. Conference Abstracts. 11-16 July 2016.* Jagiellonian University and Pedagogical University, Kraków, Poland., pp. 496-500.

Fischer, S., Fankhauser, P. and E. Teich. 2017. Visualization of Corpus Frequencies at Text Level. In *Proceedings of the Corpus Linguistics Conference, CL2017, Poster Session, 25-28 July 2017,* University of Birmingham, UK.

James, R. and A. Weiss (2012). An Assessment of Google Books' Metadata, In: *Journal of Library Metadata,* 12(1), pp. 15-22.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. and E. Teich (2016). The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC'16, 23-28 May 2016.* Portorož, Slovenia, pp. 1928-1931.

Knappen, J., Fischer S., Kermes, H., Teich, E. & Fankhauser, P. (2017). The Making of the Royal Society Corpus. In: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, 22 May 2017.* Gothenburg, Sweden, pp. 7-11.

McCauley, J. (2006). Technical Combining Forms in the Third Edition of the OED: Word formation in a Historical Dictionary. In *Selected Proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (HEL-LEX), 17-19 March 2005,* Helsinki, Finland, pp. 95-104.

Menzel, K. and S. Degaetano-Ortlieb (2017). The Diachronic Development of Combining Forms in Scientific Writing. In: Lege Artis. Language Yesterday, Today, Tomorrow, The Journal of University of SS Cyril and Methodius in Trnava. Warsaw: De Gruyter Open, vol. 2 (2) 2017, pp. 185-249.

Oxford English Dictionary. (2000- 3rd ed.), Oxford: Clarendon Press. Accessed at: http://www.oed.com/ [28/03/2018]

Podhajecka, M. (2010). Antedating Headwords in the Third Edition of the OED: Findings and Problems. In *Proceedings of the XIV EURALEX International Congress, 6-10 July 2010.* Fryske Akademy, Leeuwarden, Netherlands, pp. 1044-1064.

Podhajecka, M. (2011). Research in Historical Lexicography: Can Google Books Collection Complement Traditional corpora? In S. Góźdź-Roszkowski (ed.), *PALC 2009: Explorations Across Languages and Corpora.* Frankfurt am Main: Peter Lang, pp. 529-546.

Quinion, M. (ed.), (2005). Ologies and isms: A dictionary of word beginnings and endings. Oxford: Oxford University Press.

Sheehan, M.J. (ed.). (2000). Word parts dictionary. Standard and reverse listings of prefixes, suffixes, roots and combining forms. 2nd ed. Jefferson, NC & London: MCFarland.

ten Hacken, P. and R. Panocová (2014). Neoclassical Formatives in Dictionaries. In *Proceedings of the 16th EURALEX International Congress*, *15-19 July 2014.* Bolzano, Italy, pp. 1059-1072.

Acknowledgements

This study contributes to the project Information Density and Scientific Literacy in English – Synchronic and Diachronic Perspectives in the framework of the Collaborative Research Center 1102 with the title Information Density and Linguistic Encoding funded by the German Research Foundation (DFG).