Using relative entropy for detection and analysis of periods of diachronic linguistic change

Stefania Degaetano-Ortlieb
Language Science and Technology
Universität des Saarlandes
Saarbrücken, Germany
s.degaetano@mx.uni-saarland.de

Elke Teich
Language Science and Technology
Universität des Saarlandes
Saarbrücken, Germany
e.teich@mx.uni-saarland.de

Abstract

We present a data-driven approach to detect periods of linguistic change and the lexical and grammatical features contributing to change. We focus on the development of scientific English in the late modern period. Our approach is based on relative entropy (Kullback-Leibler Divergence) comparing temporally adjacent periods and sliding over the time line from past to present. Using a diachronic corpus of scientific publications of the Royal Society of London, we show how periods of change reflect the interplay between lexis and grammar, where periods of lexical expansion are typically followed by periods of grammatical consolidation resulting in a balance between expressivity and communicative efficiency. Our method is generic and can be applied to other data sets, languages and time ranges.

1 Introduction

The awareness of the necessity and possibilities of large scale analysis of the temporal dynamics of cultural phenomena has risen considerably in the last two decades or so in a number of scientific disciplines, including literary studies, musicology, biology and marketing research. One common challenge is to determine the periods of change. For example, to detect periods of stylistic change in popular music Mauch et al. (2015) use data-driven methods from bioinformatics based on a set of predefined audio features; or for periodization of prose texts van Hulle and Kestemont (2016) use stylometric methods with selected function words.

Here, we come from the perspective of language and linguistics. Specifically, we are interested in the formation of discourse types and registers. Focusing on scientific writing in the period of late Modern English (1700-1900) — the period in which scientific writing evolved as a distinctive discourse type (Atkinson, 1999; Bazerman, 1988) — we want to test the hypothesis that scientific writing became increasingly specialized, expert-oriented and geared towards communicative efficiency (Halliday, 1988). Linguistic reflexes are expected in vocabulary expansion, notably in the area of terminology, and consolidation in grammatical usage. Therefore, we consider both the lexical and the grammatical level as well as their interplay.

While there is a long tradition in diachronic, corpus-based analysis (see Nevalainen (2006) for an overview), time periods and linguistic features considered are typically predefined, thus introducing possible biases. To avoid this, we have designed a data-driven approach based on the information-theoretic measure of relative entropy by which we can both detect features involved in diachronic linguistic change and discern the time periods change(s) occur(s). Our approach is generic and can be applied to any diachronic data set with any type of linguistic feature.

Following a more detailed presentation of related work (Section 2), we describe our approach in Section 3. In Section 4, we report our results capturing important aspects of change in language use in the scientific domain. Section 5 provides a brief summary of our main results and an outlook on follow-up studies.

2 Related work

The traditional linguistic method of describing temporal change in language use is to start with a set of preselected linguistic features and inspect their frequency distributions across predefined time spans (cf. Nevalainen (2006) for an overview in the area of corpus linguistics). While clearly providing interesting and relevant descriptive insights on changing language use, this kind of approach is biased in two regards. First, features are selected on the basis of the human analyst's educated guesses about which linguistic features are subject to change with a view to high-frequency features (e.g. Atkinson (1999), Banks (2008), Biber and Finegan (1989), Biber and Gray (2016), Degaetano-Ortlieb et al. (2014), Fanego (1996), Michel et al. (2011), Moskowich and Crespo (2012), Rissanen et al. (1997), Teich et al. (2016)). Other frequency bands, while potentially relevant, are not considered. Second, the time spans are typically predefined, too. For example, for a period of two to three hundred years under consideration, typically 20 or 50-year periods are selected. This approach may obscure trends present in the data and prevent the exact periodization of a given change or set of changes.

To remedy these drawbacks, more exploratory, data-driven approaches have been argued for. In corpus-linguistics, Gries and Hilpert (2008) propose a specific clustering approach which they apply to the historical development of English. In van Hulle and Kestemont (2016) stylometry methods are applied to the periodization of literary works of Samuel Beckett. Popescu and Strapparava (2013) use a statistical approach for the characterization of epochs. Within the field of computational sociolinguistics, various techniques (such as topic modeling, correlations, regression) are tested for applicability to sociological questions and interpretability of the features involved in variation (see e.g., McFarland et al. (2013), O'Connor et al. (2011), Eisenstein (2018)). A recent strand of data-driven approaches for the analysis of diachronic change applies information-theoretic divergence measures. In particular, the use of relative entropy, implemented as Kullback-Leibler Divergence (KLD) or its symmetrical variant Jensen-Shannon Divergence (JSD) (Kullback and Leibler, 1951), as a measure of changes in the probability distribution over linguistic features has proven effective. For example, Hughes et al. (2012) measure stylistic influence in the evolution of literature and Klingenstein et al. (2014) analyze to what extent the ways of talking in criminal trials differed between violent and nonviolent offenses over time. Or Bochkarev et al. (2014) use KLD to compare change in the frequency distribution of words within one language and a symmetric version to compare changes across languages in the Google Books Corpus. Also working on the Google Books data set, Pechenick et al. (2015) use JSD to assess the corpus' validity for analysis of cultural and linguistic evolution. Furthermore, Fankhauser et al. (2014) demonstrate the applicability of KLD for corpus comparison at large, showing KLD on various corpora (including the Brown corpora), and provide an interactive visualization for exploratory inspection of (degrees of) divergence between corpora as well as the items (here: words) contributing to the divergence. In our own previous work, besides other applications such as intra-textual variation across sections of research articles from biology (Degaetano-Ortlieb and Teich, 2017), we have used KLD to analyze the linguistic development of scientific writing over time considering pre-defined time periods (50 years) and comparison to general English to discern change specific to scientific writing (Degaetano-Ortlieb and Teich, 2016; Degaetano-Ortlieb et al., 2018; Degaetano-Ortlieb and Strötgen, 2018). In particular, we found major changes going on around the period of 1750-1800.

We build on this work and extend the existing approaches by capturing periods of change, i.e. determining when a change occurs rather than a priori setting *specific* periods. Most similar to our approach is the recent study of Barron et al. (2018) on debates held in the French Revolution's first parliament using KLD between sequential speeches considering the notions of novelty (abrupt change), transience (novelty of the past), and resonance (novelty minus transience). While they only employ the aggregated KL divergence, we consider the contribution of individual linguistic features to KL divergence. On this basis, we are able to determine which features are involved in change at different linguistic levels and to inspect how different linguistic levels interact to allow for a balanced information density in (scientific) communication. Thus, our approach addresses the above mentioned drawbacks — predefined periods and preselected features — and provides a generic, exploratory method for periodization combining feature detection and period determination on the basis of one mechanism.

3 Method

3.1 Royal Society Corpus

The data set we use is the Royal Society Corpus (RSC) (Kermes et al., 2016), consisting of the journal publications of the Transactions and Proceedings of the Royal Society of London – the first and longest-running English periodical of scientific writing. The RSC has approx. 32 million running tokens and around 10.000 documents, spanning from 1665 (first publication) to 1869. It is encoded for text type (article, abstract), author, title, date of publication, and time periods (decades, fifty years). Linguistic annotation is provided at the levels of tokens (with normalized and original forms), lemmas, and parts of speech using TreeTagger (Schmid, 1995), achieving 95.1% on normalized word forms (normalization is based on VARD; see Baron and Rayson (2008)). The corpus is hosted by a CLARIN-D repository[1], which provides a free download as a vertical text format (vrt). The RSC provides a well-suited test bed for periodization, as it spans approx. two centuries and there are a number of linguistic studies on some parts of this material (e.g. Biber and Finegan (1997), Atkinson (1999), Banks (2008)). As we detected major changes around the period of 1750-1800 in previous studies (Degaetano-Ortlieb and Teich, 2016; Degaetano-Ortlieb et al., 2018), we select texts from 1700 to 1850 for periodization with a cut off of five occurrences per document to exclude (especially for the older documents) OCR errors and other possible biases. Table 1 lists by decade the number of lemmas and part-of-speech (POS) trigrams used to approximate the lexical and grammatical level, respectively, as well as the number of types of lemmas and POS trigrams.

decade	lemmas	POS trigrams
1700	407,801	2,905
1710	261,143	2,605
1720	283,123	2,578
1730	307,049	2,114
1740	538,567	4,494
1750	626,685	8,725
1760	507,820	9,828
1770	855,618	20,751
1780	827,720	18,562
1790	809,482	23,966
1800	971,271	18,550
1810	808,536	15,766
1820	809,682	17,388
1830	1,567,919	43,794
1840	1,237,025	36,633
types	15,611	1,154

Table 1: Number of lemmas and POS trigrams in the RSC per decade and number of types for each.

3.2 Method of periodization at different linguistic levels

We exemplify our proposed method for periodization with Kullback-Leibler Divergence (KLD) considering two linguistic levels – lexical and grammatical – without preselecting single linguistic features. For the lexical level we use all lemmas (unigrams) occurring at least five times in a document. The grammatical level is captured by sequences of three parts of speech (POS trigrams, e.g. noun-preposition-noun) again with a minimum of five occurrences per document. Trigrams were chosen as shorter sequences tend to not reflect grammatical structures, longer sequences lead to quite sparse data. To further avoid possible POS tagging errors, in the extraction procedure nouns were restricted to a size of >2 characters. Furthermore, we exclude POS trigrams consisting of characters constituting sentence markers (e.g. fullstops, colons), brackets, symbols (e.g. equal signs), and words tagged as foreign words. By looking at lemmas, we capture vocabulary changes and by looking at POS trigrams we capture changes in grammatical use. Note however that any kind of linguistic unit could be used (phoneme, morpheme, word, etc).

[1] https://fedora.clarin-d.uni-saarland.de/rsc

Relative entropy or Kullback-Leibler Divergence (KLD; Kullback and Leibler (1951)) is a method of comparing probability distributions measuring the number of additional bits needed to encode a given data set A when a (non-optimal) model based on a data set B is used (cf. Equation (1)).

$$D(A||B) = \sum_i p(feature_i|A) log_2 \frac{p(feature_i|A)}{p(feature_i|B)} \tag{1}$$

Applied to the comparison of language corpora, KLD gives us an indication of the degree of difference between corpora measured in bits as well as the features that are primarily associated with a difference, i.e. those features that need (relatively) high amounts of additional bits for encoding[2]. In the models we employ, difference in vocabulary size is controlled for by representing the data sets by ngram language models smoothed with Jelinek-Mercer smoothing and lambda 0.05 (cf. Zhai and Lafferty (2004) and Fankhauser et al. (2014)).

To detect periods of change using KLD, we slide over the time line of the corpus to find relative peaks or troughs in relative entropy which are taken to indicate a change. For this we select a starting year (e.g. 1720) and a sliding window (e.g. 2 years). We then use KLD to compare preceding (*pre* period) and subsequent years (*post* period) from the sliding window. Defining the size of the sliding window depends on the data set used. In our case, as publications do not appear yearly in the Proceedings of the Royal Society, we use a minimum of a 2-year sliding window. For other text types, such as news texts, for example, the sliding window could be based on months or even days. The bigger the sliding window based on a particular data set the less fine-grained the observed changes will be. A further parameter to be set is the time range of comparison for *pre* and *post* periods, which again has to be set according to the data set used and the aimed periodization. In our case, we use a period range of 20 and 10 years in which we assume changes to occur. Note that KLD is asymmetric and we are only interested in the direction from *post* to *pre* as we aim to determine periodization from past to present in the development of scientific writing. Thus, we measure divergence only between *post* (after sliding window) and *pre* (before sliding window) as shown in Equation (2).

$$D(post||pre) = \sum_i p(unit_i|post) log_2 \frac{p(unit_i|post)}{p(unit_i|pre)} \tag{2}$$

Based on this, we build KLD models for lemmas and POS trigrams to observe divergences at lexical and grammatical levels, respectively. For both linguistic levels, we use all lemmas/POS trigrams for modeling that occur at least five times in a document. For each window, KLD models are created comparing *post* with *pre* periods. For the analysis we use 2-, 5-, and 10-year windows inspecting 10- and 20-year ranges.

Moreover, the individual contribution (discriminative power) of a feature to relative entropy allows us to observe which features are involved in change. The higher the KLD value of a feature (here: lemma or POS trigram), the more discriminative the feature is for the *post* period (see Equation (3)).

$$D_{feature}(post||pre) = p(feature|post) log_2 \frac{p(feature|post)}{p(feature|pre)} \tag{3}$$

We then also test if there is a significant difference between the relative frequencies of a feature in the *pre* and *post* periods by an unpaired Welch's t-test (see equation (3) with var denoting the variance and n the number of documents in a corpus).

$$t = \frac{mean_{pre} - mean_{post}}{\sqrt{(\frac{var_{pre}}{n_{pre}} + \frac{var_{post}}{n_{post}})}} \tag{4}$$

Thus, for each *post* with *pre* period comparison, we obtain a list of those features that contribute most to the distinction of a *post* period (high KLD value) and which pass the significance test (p-value$<$0.05).

[2]Note that KLD is an asymmetric measure, i.e. there may be a significant difference between a data set A and B when B is used as a basis for encoding but not necessarily when A is used as a basis. Also, the features responsible for a difference may be different ones.

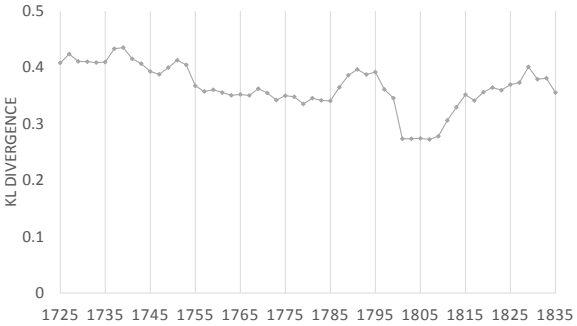


Figure 1: Relative entropy based on KLD for *post* vs. *pre* periods for lemmas (with parameters set at a 2-year window and 20-year period).

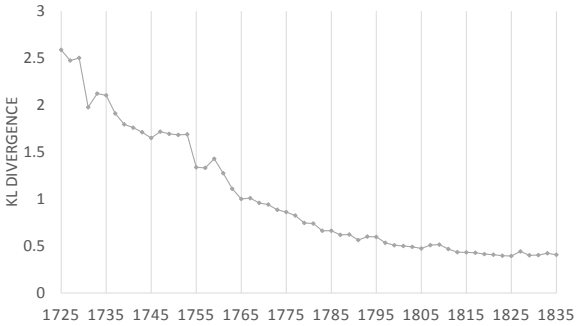


Figure 2: Relative entropy based on KLD for *post* vs. *pre* periods for POS trigrams (with parameters set at a 2-year window and 20-year period).

In addition, at the lexical level to select from these lists those lemmas that show the greatest variation in terms of their contribution over time, we calculate the standard deviation of the KLD value of each lemma across all comparisons. At the grammatical level, as the set of possible options is more confined, we consider all discriminative and significant POS trigrams but still rank them by the standard deviation over time.

Given a data set of the features' occurrences across time (e.g. by years, months, days), the periodization procedure is operationalized with the software environment R (R Development Core Team, 2010) with a script allowing to select the window and period range and run the process automatically. The R code will be released upon publication through Github via a link on `https://stefaniadegaetano.com/`.

4 Change of language use in scientific writing (18th and 19th century)

We present the results of application of the described method on the time period of 1700-1850 of the Royal Society Corpus at the lexical and grammatical levels. The focus is on comparison of *post* periods with *pre* periods, tracing the development of scientific writing over time. For this, we inspect different parameters as described in Section 3.2. For both analyses, we consider (1) overall diachronic trends, (2) when possible changes occur, and (3) which features contribute to changes.

4.1 Overall diachronic trends at the lexical and grammatical level

First, we want to investigate which diachronic trends we can observe over time in scientific writing across different linguistic levels.

At the lexical level, Figure 1 shows KLD values plotted on the time line based on lemmas (with a 2-year window and 20-year slices of *pre* and *post* periods). The general tendency shows peaks and troughs in KLD, i.e. periods of lexical expansion (where a *post* period shows higher KL divergence from a *pre* period) are followed by periods of lexical consolidation (where a *post* period shows lower KL divergence from a *pre* period). We can observe that there are smaller peaks around the 1740s, 1750s, and a major peak around the end of the 18th century followed by a trough in KLD and again an increase in the early 19th century. For example, considering the year 1795 (here: 2-year window (1795-1796)), KL divergence between the 20 years following (*post*) and the 20 years preceding (*pre*) the window is 0.39 bits, while considering 1805 it drops to 0.27 bits. Thus, around 1795 we have a period of lexical expansion, while around 1805 we have a period of consolidation.

Let us now consider the grammatical level approximated by POS trigrams. In Figure 2, we clearly see a declining tendency of KLD at the grammatical level, i.e. over time grammatical usage consolidates in comparison to a more varied usage in the past. Comparing Figure 1 with 2, while at the lexical level we have waves of expansion and consolidation, at the grammatical level there is a strong tendency of consolidation. This tendency becomes even more pronounced from the mid 18th century onwards. The formation of the scientific register and processes of professionalization (Ure, 1982; Biber and Gray, 2011; Halliday, 1988) are presumably reflected here. In fact, from 1751 onwards the Proceedings of the

Royal Society started to have a reviewing process.

In summary, we can deduce that the lexical and grammatical levels play different roles in the development of scientific writing in the RSC. The consolidation of the grammatical level might be a counterbalance to the phases of expansion at the lexical level. In the next sections, we present a detailed account on which lemmas and grammatical structures contribute to periods of change occurring at both levels.

4.2 Lexical contributions to periods of change

In a second step, we investigate which lemmas contribute most to the observed differences by selecting lemmas based on their discriminative power as described in Section 3.2. To consider how different time windows and ranges might impact the results, we will consider windows of 2, 5, and 10 years with 20 and 10 years of *pre* and *post* periods.

Figure 3 shows lemmas contributing to periods of change over time across window sizes (years between a *pre* and *post* period) and ranges (of *pre* and *post* periods compared by KLD)[3]. Across the window sizes, the general trend remains relatively stable, with specific lemmas contributing to periods of change. From around 1725 to 1745 lemmas related to the field of electricity are distinct (light blue: *electricity*, *electrify*, *wire*). From the mid 18th century to the beginning of the 19th century a whole field arises that marks the discovery of oxygen (orange) with *air*, *nitrious*, *dephlogisticated*, *gas* marking the beginning of this research field driven by experiments and *oxide*, *oxygen*, *hydrogen* marking the terminology building process around the new field. In fact, a landmark paper on the discovery of oxygen by Joseph Priestley in 1774 entitled *Observations on different kinds of air* brought about a series of publications in the Royal Society dedicated to this new strand of research. Towards the mid 19th century biology terms arise (*cell*, *corpuscule*).

Tuning the period range allows us to inspect the data further (compare Figure 3 (c) and (d), 20-year range vs. 10-year range, respectively). For example, a period of change related to the solar system (purple: *sun*, *venus*, *limb*, *parallax*) is better captured with a smaller range (10 years, Figure 3 (d)). Thus, while some periods of change are more persistent (e.g. the discovery of oxygen) and can thus be captured by using wider ranges (e.g. 20 years), others (such as new observations on the solar system) are more transient and therefore can be better detected by more narrow comparisons of ranges (e.g. 10 years). Also detectable at the 10-year window (see Figure 3 (d)) are variants of use: *oxygen* vs. *oxygene*, where the latter was distinctively used but only for a small period of time (from 1790-1820), while *oxygen* was increasingly used over time and became the standard variant. Thus, this allows us to observe competing forms or lexemes.

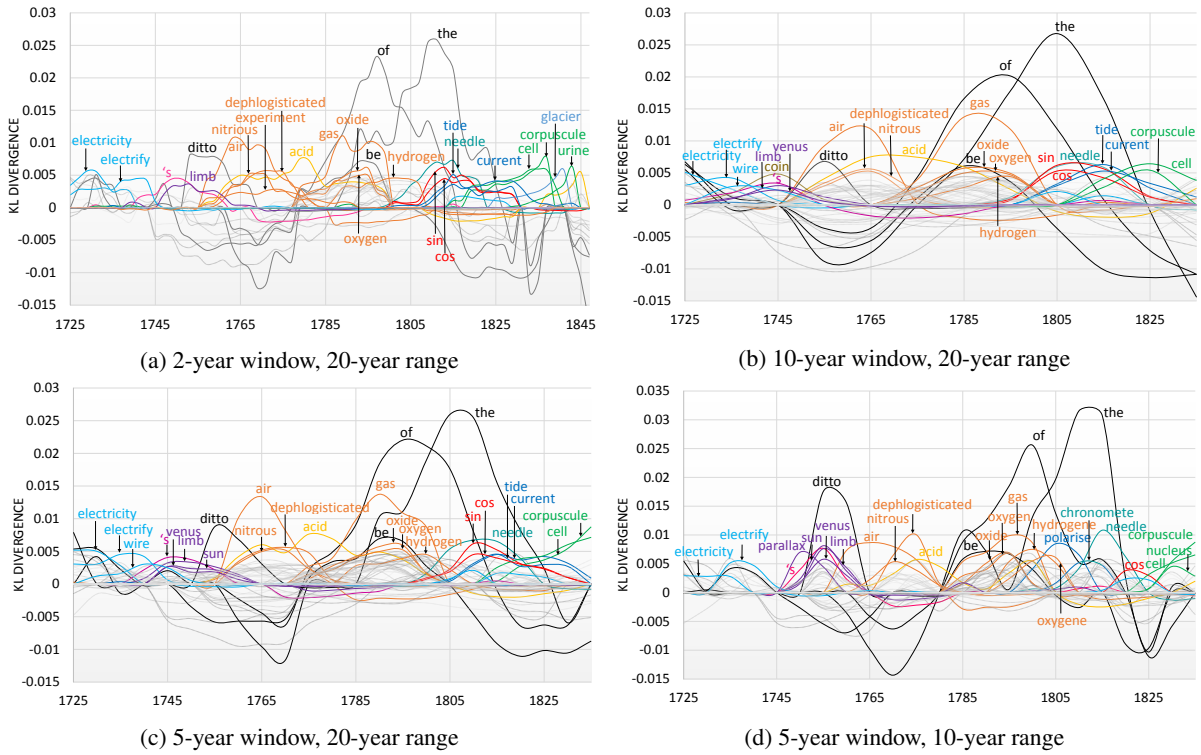
A further major change takes place around the end of the 18th century, where besides words related to terminology as discussed so far, the function words *the* and *of*, reflecting the use of nominal phrases, drastically increase their discriminative power for the *post* period as well as the verb *be*, which might reflect here a relational use (such as e.g. X *is* Y).

These diachronic tendencies across the inspected time frame show how at the lexical level, specific terms become typical of a time period marking strands of terminological evolution which can be attributed to groundbreaking events in the world (such as the discovery of oxygen). The rise and fall in discriminative power of function words seems to indicate changes in the use of grammatical structures. To observe whether this is really the case and which grammatical structures are involved in change, we inspect changes at the grammatical level by approximating grammatical structures with POS trigrams.

4.3 Grammatical structures contributing to periods of change

To inspect which POS trigrams have significantly contributed to changes over time, we again plot the individual KLD value of each discriminative POS trigram on the time line (see Figure 4 showing the 5-year window). A major change in the use of discriminative POS trigrams takes place between the 1740s and 1760s. Here, nominal phrase patterns with prepositions (DT NN IN, NN TO DT), coordinating conjunctions (NN CC NN) and possessives (IN NP POS) are discriminative. Both nominal phrase patterns reflect a conventionalized usage of general nouns combined with the prepositions *of* and *to* (e.g. *the end*

[3]Selection of 100 distinctive lemmas ranked by standard deviation across time are displayed (see Section 3.2).



(a) 2-year window, 20-year range

(b) 10-year window, 20-year range

(c) 5-year window, 20-year range

(d) 5-year window, 10-year range

Figure 3: Lemmas contributing to periods of change for different window sizes and period ranges.
The higher the KLD value the higher the pattern's contribution to the overall KL divergence. The KLD values are based on comparison between a *post* vs. *pre* period ($D(post||pre)$) to inspect change from past to present. Positive values reflect distinctness for *post*, negative values distinctness for *pre*. Selection of 100 distinctive lemmas ranked by standard deviation across time are displayed.

of, a letter from, the time of for the DT NN IN pattern; *with regard/respect to* for the IN NN TO pattern). The possessive pattern reflects the peaks shown at the lexical level where the *'s* and the lemmas related to the solar system mark a period of change.

After this major period of change, from the 1750s onwards individual patterns become typical that can be related to specific grammatical structures. First, a nominal compound pattern followed by a preposition (NN NN IN) becomes typical around the 1750s (e.g. *zenith distance of*, *logarithm sine of*), which appears again as distinct after 1810 (e.g. *knife edge of the pendulum*) with a greater variation of use (around 10 vs. 30 instances of at least 5 occurrences). Around the 1760s a comparative pattern arises (VBZ JJR IN; with realizations such as *is greater than*, *is less than*, *is more than*).

At the same time a particularly interesting pattern reflecting relational or passive clauses (JJ NN VBZ) is discriminative, used to define specific types of materials such as air, acid, fluid etc. (e.g. *inflammable air is pure phlogiston*, *dephlogisticated air is only water deprived of phlogiston*) or to explain what these materials are used for (e.g. *nitrous air is mixed with*, *alkaline air is saturated by*). This pattern is closely related to the beginnings of early modern chemistry marked by the discovery of oxygen as shown at the lexical level. A constant increase of KLD value of this relational pattern might also indicate an increasing need for specification in this new research field (i.e. defining what it is exactly that has been discovered). As the field around oxygen became established, one may think that also the need for this specification pattern may no longer exist. However, in terms of frequency (see Figure 5) it rises steeply from 1760 to 1780, has a period of stagnation between 1780 and 1830, rising again afterwards. Thus, it rises considerably in frequency due to the need to elaborate on the findings of this new research field around chemistry (rising period). This pattern then becomes established in scientific writing (stagnation period). The decline in KLD confirms this tendency: the pattern is no longer discriminative for a *post* vs. *pre* by 1790s, as it is similarly used in both *pre* and *post* slices.

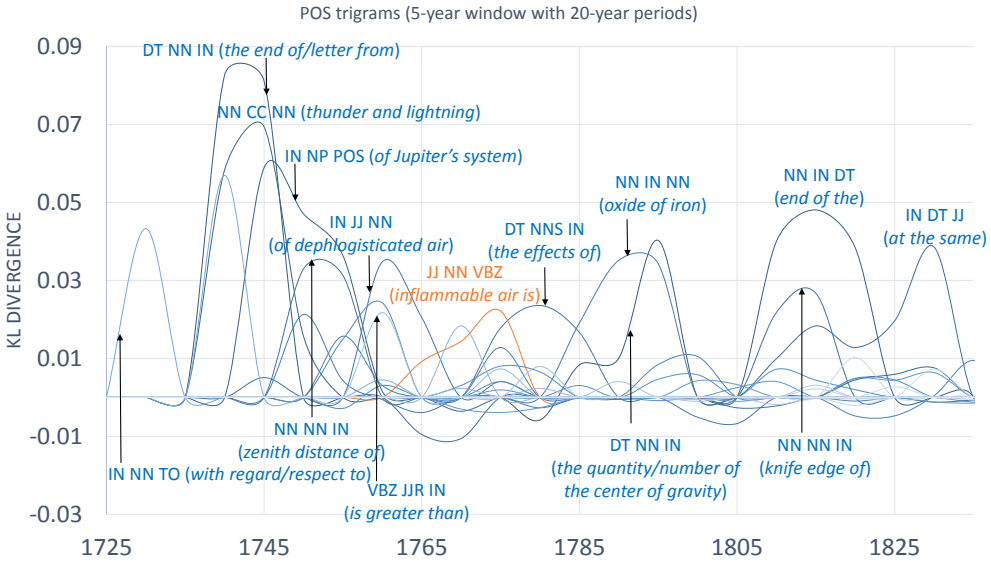


Figure 4: Grammatical structures (approximated by POS trigrams) contributing to periods of change. The higher the KLD value the higher the pattern's contribution to the overall KL divergence. The KLD values are based on comparison between a *post* vs. *pre* period ($D(post||pre)$) to inspect change from past to present. Positive values reflect distinctness for *post*, negative values distinctness for *pre*. Selection of 100 distinctive lemmas ranked by standard deviation across time are displayed.

Starting around the 1780s longer noun phrases with a plural head (DT NNS IN) become discriminative (such as *the effects/observations/results of*) pointing to scientific outcomes. Around the 1790s and 1800s the terminological pattern 'noun of noun' (NN IN NN) is typical (with realizations such as *centre of gravity, carbonate of lime, phosphate of lime, oxide of iron, sulphate of iron*). If we compare this again with the lexical level (see Figure 3), we can see how after specific terms were established, grammatical structures arise around these terms. This terminological pattern is also reflected in the discriminative power of the preposition *of* at the lexical level (see Figure 3). By considering the level of grammar (approximated here with POS trigrams), we have a clearer picture of the changes that have occurred in terms of grammatical structures.

Starting from the late 1780s, the discriminative power of the nominal pattern DT NN IN rises again, this time not only marking conventionalized usage by particular expressions (such as *the quantity of* or *the number of*) but also terminological usage by terms establishing themselves in that period (such as *the center of gravity*, *the bulb of the thermometer*, *the temperature of the air*).

At the beginning of the 19th century, there is again a rise of specific nominal patterns with prepositions (NN IN DT, NN NN IN, IN DT JJ). The first two patterns (NN IN DT and NN NN IN) both reflect longer nominal phrases related to terminology (e.g. *length of the second pendulum*, *part of the nervous system*, *knife edge of the pendulum*). The IN DT JJ pattern instead marks a rise of functional expressions pointing to contrast/comparison (e.g. *at the same time*, *on the other hand*) and elaboration (e.g. *in the same way*, *in the same manner*).

Comparing our findings to previous accounts on the Proceedings and Transactions of the Royal Society (PTRS), we are clearly in line a.o. with Halliday (1988) and Atkinson (1999), who showed a shift from an involved to an informational style of writing between the 17th and 19th century based on a manual and a multi-dimensional analysis, respectively, reflected in higher nominal style in the later productions. In Degaetano-Ortlieb et al. (2018) we confirm this finding using a data-driven approach. In this paper, we were able to show when particular nominal patterns reflecting informational style become distinctive in comparison to earlier periods and how their use is intertwined with changes occurring at the lexical level. In summary, the diachronic tendencies at the grammatical level, first, show a major period of change around the 1750s marked by a strong contribution of patterns related to conventionalized style of writing

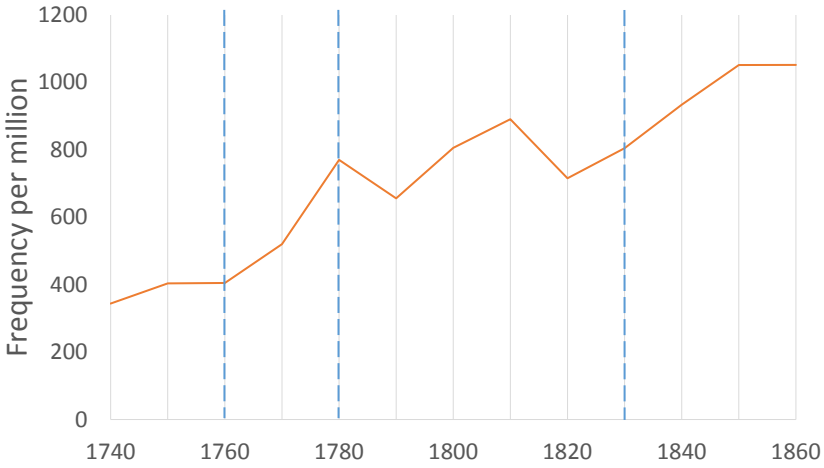


Figure 5: Frequency distribution of the JJ NN VBZ trigram

(*of* and *to* prepositional phrases, conjunctive and possessive phrases). Second, this period is followed by rise of individual lexico-grammatical patterns over time due to needs driven by expansions at the lexical level, on the one hand, and further lexical conventionalization of the patterns, on the other hand. Thus, in comparison to previous work, we are able to detect when and possibly why particular patterns become distinctive showing not only reflections of specialization in the formation of terminology but also of conventionalization in particular lexically confined grammatical patterns.

5 Summary and outlook

We have presented a generic, data-driven approach based on Kullback-Leibler Divergence (KLD) for detecting features involved in diachronic linguistic change and discerning periods of change without pre-selection of features and periods. Our method is illustrated on the Royal Society Corpus, showing which features are involved in change and observing periods of change in scientific writing. The features detected indicate two types of change, *lexical expansion* and *grammatical consolidation*. Note here that while the first type of change relates to low-frequency instances, it is a highly distinctive feature over time — a fact that a traditional frequency-based approach would have missed.

What we can also see from our sample analysis is that changes proceed in waves — a wave of lexical expansion is typically followed or partially paralleled by reduction in grammatical variation, thus indicating the continuous effort to balance expressivity and communicative efficiency. In this way, rational language users make sure that, while language use changes, communication remains successful. Lexis and grammar thus show a nice symbiosis in enhancing expressivity and maintaining communicative efficiency.

In a wider perspective, our research is a contribution to information-theoretic accounts of language use with rational communication as an explanatory framework, adding to it a diachronic perspective (cf. Hume and Mailhot (2013) for related work in phonology). In future work, we plan to look at register-mixed language as a reflection of 'general' language and longer time ranges using the proposed method, going over attested periods of evolution of the English language, starting from Early Modern English to Late Modern English and contemporary English, in order to observe long-term, more persistent grammatical changes. For lexical development, we are currently exploring measures of vocabulary expansion from a paradigmatic perspective on the basis of word embeddings, both for scientific as well as 'general' language (cf. Hamilton et al. (2016), Fankhauser and Kupietz (2017a), and Fankhauser and Kupietz (2017b) for related work).

References

Dwight Atkinson. 1999. *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. Erlbaum, New York.

David Banks. 2008. *The Development of Scientific Writing: Linguistic Features and Historical Context*. Equinox, London/Oakville.

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.

Alexander T. J. Barron, Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. Individuals, Institutions, and Innovation in the Debates of the French Revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612.

Charles Bazerman. 1988. *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. University of Wisconsin Press, Wisconsin.

Douglas Biber and Edward Finegan. 1989. Drift and the Evolution of English Style: A History of Three Genres. *Language*, 65(3):487–517.

Douglas Biber and Edward Finegan. 1997. Diachronic Relations among Speech-based and Written Registers in English. In Terttu Nevalainen and Leena Kahlas-Tarkka, editors, *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, pages 253–276. Société Néophilologique, Helsinki.

Douglas Biber and Bethany Gray. 2011. The Historical Shift of Scientific Academic Prose in English towards Less Explicit Styles of Expression: Writing without Verbs. In Vijay Bathia, Purificación Sánchez, and Pascual Pérez-Paredes, editors, *Researching Specialized Languages*, pages 11–24. John Benjamins, Amsterdam.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge University Press, Cambridge, UK.

Vladimir Bochkarev, Valery D. Solovyev, and Soren Wichmann. 2014. Universals versus Historical Contingencies in Lexical Evolution. *Journal of The Royal Society Interface*, 11(101).

Stefania Degaetano-Ortlieb and Jannik Strötgen. 2018. Diachronic variation of temporal expressions in scientific writing through the lens of relative entropy. In Georg Rehm and Thierry Declerck, editors, *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, volume 10713 of *Lecture Notes in Computer Science*, pages 259–275. Springer International Publishing.

Stefania Degaetano-Ortlieb and Elke Teich. 2016. Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In *Proceedings of the 10th LaTeCH Workshop*, pages 165–173, Berlin. ACL.

Stefania Degaetano-Ortlieb and Elke Teich. 2017. Modeling Intra-textual Variation with Entropy and Surprisal: Topical vs. Stylistic Patterns. In *Proceedings of the Joint LaTeCH and CLfL Workshop*, pages 68–77, Vancouver, Canada. ACL.

Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, Ekaterina Lapshinova-Koltunski, Noam Ordan, and Elke Teich. 2014. Data Mining with Shallow vs. Linguistic Features to Study Diversification of Scientific Registers. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland. ELRA.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2018. An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. In Carla Suhr, Terttu Nevalainen, and Irma Taavitsainen, editors, *From Data to Evidence in English Language Research*, Language and Computers. Brill, Leiden.

Jacob Eisenstein, 2018. *The Handbook of Dialectology*, chapter Identifying Regional Dialects in On-Line Social Media, pages 368–383. Number 21. Wiley-Blackwell.

Teresa Fanego. 1996. The Gerund in Early Modern English: Evidence from the Helsinki Corpus. *Folia Linguistica Historica*, 17:97–152.

Peter Fankhauser and Marc Kupietz. 2017a. Visual Correlation for Detecting Patterns in Language Change. In *Visualisierungsprozesse in den Humanities. Linguistische Perspektiven auf Prägungen, Praktiken, Positionen (VisuHu 2017)*, Zürich.

Peter Fankhauser and Marc Kupietz. 2017b. Visualizing Language Change in a Corpus of Contemporary German. In *Proceedings of the Corpus Linguistics International Conference*, Birmingham, UK.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In *Proceedings of the 9th LREC*, pages 4125–4128, Reykjavik, Iceland. ELRA.

Stefan Th. Gries and Martin Hilpert. 2008. The Identification of Stages in Diachronic Data: Variability-based Neighbor Clustering. *Corpora*, 3(1):59–81.

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.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Models of Semantic Change. In *Proceedings of the EMNLP*, Austin, Texas.

James M. Hughes, Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. Quantitative Patterns of Stylistic Influence in the Evolution of Literature. *Proceedings of the National Academy of Sciences*, 109(20):7682–7686.

Elizabeth Hume and Frédéric Mailhot. 2013. The Role of Entropy and Surprisal in Phonologization and Language Change. In Alan C. L. Yu, editor, *Origins of Sound Change: Approaches to Phonologization*, pages 29–47. Oxford University Press, Oxford.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the 10th LREC*, Portorož, Slovenia. ELRA.

Sara Klingenstein, Tim Hitchcock, and Simon DeDeo. 2014. The Civilizing Process in London's Old Bailey. *Proceedings of the National Academy of Sciences*, 111(26):9419–9424.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Matthias Mauch, Robert M. MacCallum, Mark Levy, and Armand M. Leroi. 2015. The Evolution of Popular Music: USA 1960–2010. *Royal Society Open Science*, 2(5).

Daniel A. McFarland, Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D. Manning, and Daniel Jurafsky. 2013. Differentiating Language Usage through Topic Models. *Poetics - Topic Models and the Cultural Sciences*, 41(6):607–625.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, , Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Isabel Moskowich and Begona Crespo, editors. 2012. *Astronomy Playne and Simple: The Writing of Science between 1700 and 1900*. John Benjamins, Amsterdam/Philadelphia.

Terttu Nevalainen, 2006. *Handbook of the History of English*, chapter Historical Sociolinguistics and Language Change, pages 558–588. Wiley-Blackwell.

Brendan O'Connor, David Bamman, and Noah A. Smith. 2011. Computational Text Analysis for Social Science: Model Assumptions and Complexity. *Proceedings of the Second Workshop on Computational Social Science and Wisdom of the Crowds (NIPS 2011)*.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLOS ONE*, 10(10):1–24, 10.

Octavian Popescu and Carlo Strapparava. 2013. Behind the times: Detecting epoch changes using large corpora. In *International Joint Conference on Natural Language Processing*, pages 347–355, Nagoya, Japan. ACL.

R Development Core Team, 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Matti Rissanen, Merja Kytö, and Kirsi Heikkonen, editors. 1997. *English in Transition: Corpus-based Studies in Linguistic Variation and Genre Analysis*. Mouton de Gruyter, Berlin.

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

Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2016. The Linguistic Construal of Disciplinarity: A Data Mining Approach Using Register Features. *Journal of the Association for Information Science and Technology (JASIST)*, 67(7):1668–1678.

Jean Ure. 1982. Introduction: Approaches to the Study of Register Genre. *International Journal of the Sociology of Language*, (35):5–23.

Dirk van Hulle and Mike Kestemont. 2016. Periodizing Samuel Beckett's Works: A Stylochronometric Approach. *Style*, 50(2):172–202.

Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.