Stefania Degaetano-Ortlieb* and Elke Teich

Toward an optimal code for communication: The case of scientific English

https://doi.org/10.1515/cllt-2018-0088

Abstract: We present a model of the linguistic development of scientific English from the mid-seventeenth to the late-nineteenth century, a period that witnessed significant political and social changes, including the evolution of modern science. There is a wealth of descriptive accounts of scientific English, both from a synchronic and a diachronic perspective, but only few attempts at a unified explanation of its evolution. The explanation we offer here is a communicative one: while external pressures (specialization, diversification) push for an increase in expressivity, communicative concerns pull toward convergence on particular options (conventionalization). What emerges over time is a code which is optimized for written, specialist communication, relying on specific linguistic means to modulate information content. As we show, this is achieved by the systematic interplay between lexis and grammar. The corpora we employ are the Royal Society Corpus (RSC) and for comparative purposes, the Corpus of Late Modern English (CLMET). We build various diachronic, computational n-gram language models of these corpora and then apply formal measures of information content (here: relative entropy and surprisal) to detect the linguistic features significantly contributing to diachronic change, estimate the (changing) level of information of features and capture the time course of change.

Keywords: diachronic change, scientific English, Kullback–Leibler Divergence, surprisal

1 Introduction

Language users expect a message or text to be informative. This assumption has been formulated from different perspectives, e.g. as the maxim of quantity ("be as informative as possible but no more"; Grice 1975) in pragmatics or as one of seven standards of textuality ("informativity"; De Beaugrande and Dressler 1981) in text

***Corresponding author: Stefania Degaetano-Ortlieb,** Language Science and Technology, Saarland University, Saarbrucken, Germany, E-mail: s.degaetano@mx.uni-saarland.de
Elke Teich, Language Science and Technology, Saarland University, Saarbrucken, Germany, E-mail: e.teich@mx.uni-saarland.de

linguistics. Common to such statements are two stipulations: (a) that overall, a message or text should be informative, and (b) that the amount of information conveyed should not exceed a certain limit.

More recently, evidence has accumulated in psycholinguistic and computational linguistic research that informativity is an important condition for successful communication. For instance, Jaeger and Levy (2007) observe that language users distribute information evenly over a message, trying to avoid peaks and troughs, i.e. interlocutors aim at a *uniform information density* of their messages. Aylett and Turk (2004) formulate a similar principle for the phonetic level, the *smooth signal redundancy hypothesis*. Similarly, according to Genzel and Charniak (2002) interlocutors expect a certain (high) level of information and will try to be maximally informative without exceeding some limit on *entropy rate*. Others show that too high and too low information rates may in fact cause processing difficulties (Engelhardt and Ferreira 2011; Kravtchenko and Demberg 2015), thus confirming the crucial role of informativity in linguistic communication.

But how do language users achieve an adequate level of information in their messages? It has been observed that there is a direct link between linguistic choice and informativity and we can assume that interlocutors use the linguistic options available to them to modulate the level of information of their messages striving for some kind of "optimal encoding". We have evidence of this link at all linguistic levels. For example, at the phonetic level we find across languages that word informativity influences acoustic duration (Pellegrino et al. 2011) and vowel space size (Schulz et al. 2016); or at the syntactic level, we encounter omission of syntactic elements (e.g. complementizers or relativizers; Sikos et al. 2017) or condensation (e.g. coercion; Delogu et al. 2017) in lexico-grammatical contexts that are highly predictive. Such studies show that language users seem to be concerned about informativity and try to modulate the amount of information in transmission by their linguistic choices.

The underlying formal basis to this line of research is provided by information theory, according to which information is formalized as *unpredictability in context*, commonly referred to as *surprisal* (cf. eq. (1); see also Section 3). Surprisal measures the information content of a given instance of a *unit* (e.g. a word) in number of bits as the negative log base 2 probability of the unit in *context* (e.g. the preceding word(s)).

$$S(unit) = -log_2 p(unit|context). \quad (1)$$

Consider examples (1) and (2) for illustration. In (1) *book* is highly predictable and not very surprising in the context of *Jane read a* and does not transmit a lot of

information; in the context of *Jane bought a*, however, *book* is less predictable and more surprising and therefore more informative.

(1) *Jane read a book.*

(2) *Jane bought a book.*

Importantly, surprisal is proportional to the cognitive effort required to process a word (or other kind of linguistic unit) (Hale 2001; Levy 2008; Balling and Baayen 2012) – a highly probable word (low surprisal) can be easily processed, a less probable word (high surprisal) incurs a higher processing cost. In fact, surprisal is a stable index of cognitive activity in linguistic processing, showing converging results on both behavioral (e.g. reading time) and neurophysiological (e.g. event-related potential (ERP)) measures (cf. Crocker et al. 2016 for an overview). Furthermore, there is ample evidence that surprisal correlates with linguistic encoding choices, as already pointed out above – in highly predictive (low surprisal) contexts, reduced linguistic forms are the preferred choice, in less predictive (high surprisal) contexts, fully expanded forms are preferred (cf. Mahowald et al. 2013 for evidence across languages).

While surprisal and related notions such as information density and entropy are widely used to model and explain on-line linguistic behavior, they only begin to be picked up in other areas of linguistics. If we accept that the primary function of language is communication, then we should assume that it is in some sense optimized for communication. This has direct implications for the study of language variation and change. First, it allows for the view that variation serves communication by offering ways to adapt to informational expectations by making available specific choices in linguistic encoding (e.g. reduced vs. fully expanded linguistic forms); second, whatever drives variation and change externally (social, political or cultural factors), the concern for informativity in communication acts as an overarching constraint.

In the present article, we undertake some important steps to show that the principle of informativity is at work in language diachrony and variation, focusing on the domain of science. Our overarching hypothesis is that while language users adapt to changing socio-cultural conditions (e.g. in the scientific domain, new discoveries lead to coinage of new words), they seek to modulate information content through specific linguistic choices.[1] Apart from being of interest from a philological perspective, focusing on the scientific domain has a number of advantages. First, it gives us some level of control of relevant factors such

1 Note that efficacy/robustness of communication is not explicitly considered here: if and how language users adapt to characteristics of the (noisy) channel by taking into account error probabilities in transmission warrants a separate study.

as register (field, tenor and mode of discourse) and genre/text type (research essay/article). Second, we can build on a wealth of literature on scientific writing (incl. its diachronic development), against which we can check our own analysis results. And third, we can draw on existing knowledge on the history of science (notably the evolution of modern science, its temporal phasing as well as relevant socio-cultural factors) for extrinsic interpretation of our linguistic findings.

We break down our overarching hypothesis of communicative optimization by diversification and conventionalization into the following two more specific and testable hypotheses. To be functional for specialists,

- *scientific language becomes increasingly distinctive from "general," everyday language* (H1);
- *within scientific language, specific linguistic choices become increasingly conventionalized* (H2).

As we will show, the changes in language use that characterize these developments have particular informational signatures. For H1, we estimate the *relative entropy* between models of general language and scientific language in different time periods between 1700 and 1850 (Section 4.1). As an overall tendency we predict that over time, scientific language is represented decreasingly well by a model of general language, as indexed by relative entropy (increasingly more additional bits are needed for encoding). Importantly, we will see that the linguistic features associated with the increasing distinction are features commonly taken as indicators of oral vs. written mode and involved vs. informational style. For H2, we carry out a two-pronged analysis: First, we detect the relevant linguistic features involved in diachronic change and capture its time course, especially focusing on the interplay of lexis and grammar, again using relative entropy (Section 4.2). Second, to inspect the information content of the features selected through the first analysis, we employ a model of *surprisal*, estimating for a given linguistic item or construction the amount of information it transmits on average (Section 4.3). Here, we observe how lexis and grammar work together in modulating information content: Grammatical patterns (approximated by part-of-speech trigrams) are being conventionalized, showing a fairly stable/slightly decreasing surprisal over time, but their lexical realizations are versatile and exhibit fairly high surprisal at certain points in time. As an example, we will discuss a particular part-of-speech trigram (*noun–preposition–noun*) that becomes a popular host for terminological expressions.

Both developments – greater distinctiveness from general language and increasing conventionalization – are beneficial for communication. The formation of distinctive sublanguages or registers is a departure from equi-probability and has an entropy-reducing effect (cf. Harris 1991: 391), thus facilitating

communication. Regarding conventionalization, the more conventionalized an item or construction is, the more readily accessible it is both in comprehension and production. Crucially, conventionalization/entrenchment has been argued to be an important precondition for innovative use (cf. De Smet 2016), which is in line with communicative explanations of language change and variation.

The remainder of the paper is organized as follows. Section 2 provides a survey of related work on the properties of scientific English and models of linguistic variation and change in language use. Section 3 introduces the data sets/corpora we use – the Royal Society Corpus (RSC) representing scientific writing from 1665 to 1869 (Kermes et al. 2016) and the Corpus of Late Modern English Texts (CLMET), which contains a register mix and spans 1710–1920 (De Smet 2006; Diller et al. 2011). Furthermore, we lay out the modeling methods, proposing to track the effects of (evolving) linguistic change computationally using language models, here: relative entropy (or Kullback–Leibler Divergence (KLD)) and average surprisal. Section 4 is then dedicated to the analyses, testing the two hypotheses formulated above. Section 5 concludes with a summary and discussion.

2 Related work

Overall, our research is placed in the area of *usage-based variation*, notably register theory, according to which language use depends on (the type of) situational context (field/topic, mode/medium and tenor/attitude of discourse) (Quirk et al. 1985; Halliday 1985). From this perspective, there is a wealth of descriptive work on scientific discourse as well as its diachronic development (Halliday 1988; Halliday and Martin 1993), including Halliday's seminal paper on the language of physical science which has inspired our hypotheses H1 and H2 above. The work by Biber and colleagues corroborates previous descriptive insights by corpus-based results in the framework of multi-dimensional analysis (Biber and Gray 2011; Biber and Gray 2016) and goes a step further from pure description to abstracting to "dimensions of variation," such as *involved* vs. *informational production* or *abstract style* (Biber 1988; Biber and Finegan 1989), which are relevant for interpretation of quantitative findings. In addition, there are numerous studies on selected scientific domains, such as medicine or astronomy (e.g. Moskowich and Crespo 2012).

In terms of methods, the prevalent approach in studies of language variation and change in language use is frequency-based with a view to high-frequency features (e.g. Biber and Finegan 1989; Biber and Gray 2016; Degaetano-Ortlieb et al. 2013; Fanego 1996; Moskowich and Crespo 2012; Rissanen et al. 1997; Teich et al. 2016). Other frequency bands, while potentially relevant, e.g. to capture

terminology development, are usually not taken into account. Also, the features considered are typically selected on the basis of the human analyst's educated guesses about which linguistic features are subject to change and therefore have a subjective bias. To remedy these drawbacks, more exploratory, data-driven approaches have been argued for. For example, Gries and Hilpert (2008) propose a specific clustering approach which they apply to the historical development of English. Or, in computational literary studies, statistical or stylometric methods are applied to periodization (e.g. van Hulle and Kestemont 2016; Popescu and Strapparava 2013).

Regarding theory, there are only few attempts at explaining why scientific language settles on specific linguistic options rather than others. Obviously, new discoveries and technical advancement call for new linguistic expressions. But how do we explain that some forms persist but not others? A plausible source of explanation are communicative concerns, as suggested already in Harris' work in the 1980s/90s, which takes an information-theoretic perspective on the development of sublanguages and language change at large (Harris 1991). For instance, Harris maintains that the development of explicit distinctions have positive effects on communication in that transmission becomes more error-free (Harris 1991: 393ff) – this would explain why it is beneficial to have a distinctive code for scientific communication (cf. hypothesis H1). In this view, the persistence of a linguistic change depends on its contribution to the communicative efficacy of the linguistic system (Harris 1991: Chapter 12) (or some subsystem of it). In a related perspective, more recently, there have been attempts to model the effects of accumulated knowledge on language use and the role of linguistic experience in language processing. For example, Baayen et al. (2017) use accumulated frequency and entropy to model effects of learning over individual and historical time within a discriminative learning approach; or Pate and Goldwater (2015) describe the conditions of development of optimal codes considering channel characteristics from an on-line processing perspective. Finally, there have been suggestions to draw on game theory to model and explain communication, including language change and evolution (e.g. Jaeger 2008; Franke and Wagner 2014).

In the field of linguistic variation and change, the awareness of an information-theoretic perspective's strengths has grown steadily in the last few years and information-based measures are endorsed as indicators of variation in language use according to register, style as well as social variables. In particular, relative entropy, implemented as Kullback–Leibler Divergence (KLD) or its symmetrical variant Jensen–Shannon Divergence (JSD), as a measure of difference in the probability distributions over linguistic features, enjoys increasing popularity in fields as diverse as stylistics, literary studies, history and linguistics. For example, Hughes et al. (2012) use relative entropy to measure stylistic influence in the evolution of literature; or Klingenstein et al. (2014) analyze to what extent the

ways of talking in criminal trials differed between violent and nonviolent offenses over time, and Bochkarev et al. (2014) use KLD to compare change in the frequency distribution of words within and across languages in the Google Ngram Corpus (Michel et al. 2011).

More generally, Fankhauser et al. (2014) demonstrate the applicability of KLD for corpus comparison at large, showing its use on various corpora (including the Brown corpora), and provide an interactive visualization for exploratory inspection of degrees of divergence between corpora as well as specific linguistic items (words or syntactic constructions). The approach is especially suited for corpus comparison since it allows treating feature typicality and significance as independent assessments – significance is assessed with Welch's *t*-test and used for feature selection, typicality is assessed by KLD and used for feature ranking. Further, possible effects of differences in vocabulary size (e.g. due to different corpus size) are neutralized by subtracting the entropies of the corpora under comparison. Using KLD as a tool, in our own work, we have analyzed linguistic variation according to register, social variables and time using various corpora, see e.g. Degaetano-Ortlieb (2018) capturing linguistic reflexes of social variables in the Old Bailey corpus or Degaetano-Ortlieb and Teich (2018) and Degaetano-Ortlieb et al. (2019) discerning the specific features of scientific writing from a diachronic perspective on the basis of the Royal Society Corpus (RSC; Kermes et al. 2016), including comparison to "general" (i.e. register-mixed) language as well as detecting periods of change.

Other measures of information, notably surprisal, are widely used in studies of human language processing, notably in comprehension. Since they are geared toward local, on-line processing, their application in language variation and change is not directly obvious. Technically, a surprisal model (cf. again formula (1) above) is a computational language model that estimates the probability of a given unit (e.g. a word) in its context of *n* preceding words (bigrams, trigrams, etc.) logging the probability to a base of 2 or 10. As we are interested here in the role of informativity in variation and change, we need to estimate the amount of information a linguistic unit *typically* conveys and assess whether this changes over time or not. Here, *average* surprisal may serve as an adequate measure. For instance, it is known that lexical words typically transmit more information than function words (Bell et al. 2009; Kermes and Teich 2017) and this is fairly stable over time. On the other hand, in a diachronic study based on the Google Ngram Corpus, Cohen Priva and Gleason (2016) have shown that the average surprisal of unigrams and trigrams does indeed change over time, where change in information on unigrams comes at the expense of trigrams and vice versa, i.e. unigram and trigram surprisal are negatively correlated. The authors explain this on the basis of uniform information rate (Jaeger and Levy 2007). Here, we use surprisal for a similar purpose.

3 Methods

3.1 Data sets/corpora

The corpus of scientific writing we use is the Royal Society Corpus (RSC; version 4.0) (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 version of the RSC we use here has around 32 million tokens and contains around 10,000 documents, spanning from 1665 (first publication) to 1869. The RSC 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 (POS) using TreeTagger (Schmid 1995), achieving 95.1% accuracy on normalized word forms (normalization is based on VARD; see Baron and Rayson 2008). The basic statistics of the corpus is presented in Table 1 based on decades showing number of tokens, lemmas (excluding

Table 1: Corpus statistics of the RSC per decade.

Decade	Tokens	Lemma	Sentences
1660–69	455,259	369,718	10,860
1670–79	831,190	687,285	17,957
1680–89	573,018	466,795	13,230
1690–99	723,389	581,821	17,886
1700–09	780,721	615,770	23,338
1710–19	489,857	383,186	17,510
1720–29	538,145	427,016	12,499
1730–39	599,977	473,164	16,444
1740–49	1,006,093	804,523	26,673
1750–59	1,179,112	919,169	34,162
1760–69	972,672	734,938	27,506
1770–79	1,501,388	1,146,489	41,412
1780–89	1,354,124	1,052,006	37,082
1790–99	1,335,484	1,043,913	36,727
1800–09	1,615,564	1,298,978	45,666
1810–19	1,446,900	1,136,581	42,998
1820–29	1,408,473	1,064,613	43,701
1830–39	2,613,486	2,035,107	81,500
1840–49	2,028,140	1,565,654	70,745
1850–59	4,610,380	3,585,299	146,085
1860–69	5,889,353	4,474,432	202,488
Total	31,952,725	24,866,457	966,469

punctuation, list items, etc.) and sentences. The RSC provides a well-suited test bed for our hypotheses, as it spans more than 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).

The data set representing "general" language is the Corpus of Late Modern English Texts (CLMET) (Diller et al. 2011), a principled collection of public domain texts drawn from on-line archives (Oxford Text Archive and Project Gutenberg). The corpus contains approx. 40 million tokens with approx. 350 texts covering the period of 1710–1920 (see Table 2), including narrative fiction, non-fiction, drama, letters, and treatises. We prepared the corpus with the same tools as the RSC (POS-tagging, lemmatization, normalization) providing a CQP encoding.[2]

Table 2: Corpus statistics of the CLMET per decade.

Decade	Tokens	Lemma	Sentences
1710–19	64,052	52,824	2,869
1720–29	313,013	262,453	8,871
1730–39	698,439	579,495	22,726
1740–49	2,592,931	2,131,155	81,589
1750–59	3,151,732	2,587,202	99,283
1760–69	2,249,929	1,865,845	66,643
1770–79	3,111,968	2,579,883	93,748
1780–89	970,264	773,918	34,402
1790–99	1,342,868	1,091,815	56,147
1800–09	372,959	304,266	15,119
1810–19	1,248,934	1,018,545	44,946
1820–29	2,166,354	1,796,178	70,592
1830–39	2,533,793	2,076,200	94,082
1840–49	4,665,285	3,817,129	164,301
1850–59	2,039,048	1,707,408	72,522
1860–69	1,944,650	1,635,892	67,681
1870–79	1,448,668	1,193,945	62,469
1880–89	2,069,690	1,705,119	80,460
1890–99	3,270,407	2,659,051	166,316
1900–09	2,342,007	1,932,540	110,909
1910–19	1,559,191	1,278,216	81,230
1920–29	184,578	156,579	9,066
Total	40,340,60	33,205,658	1,505,971

2 Both the RSC and the CLMET are hosted by a CLARIN-D repository at https://fedora.clarin-d.uni-saarland.de/rsc and are freely available. CLMET (V3.1): http://hdl.handle.net/21.11119/0000-0002-43F3-0, RSC (V4.0): http://hdl.handle.net/21.11119/0000-0001-7E8B-6.

3.2 Measures of information

We apply two kinds of information-theoretic measures in our studies. First, in order to assess differences across general and scientific language and across time periods and to detect the linguistic features involved in contrast/change, we employ *relative entropy* or Kullback–Leibler Divergence (Kullback and Leibler 1951) (cf. hypothesis H1 and H2). Second, for further analysis of the typical information content of a given linguistic unit or construction over time, we use *average surprisal* (cf. hypothesis H2).

3.2.1 Kullback–Leibler Divergence

Kullback–Leibler Divergence (KLD) is a widely used 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. eq. (2)).

$$D(A||B) = \sum_i p(item_i|A) log_2 \frac{p(item_i|A)}{p(item_i|B)} \quad (2)$$

Here, $p(item|A)$ is the probability of a linguistic item (e.g. a word or a syntactic construction) in A, and $p(item|B)$ is the probability of the given item in B. Relative entropy measures the average amount of additional bits per item needed to encode items distributed according to A by using an encoding optimized for B. KLD is an asymmetric measure and its minimum is at 0 for $A = B$. The individual item weights are calculated by the pointwise Kullback–Leibler Divergence (Tomokiyo and Hurst 2003) (eq. (3)):

$$D(A||B) = p(item_i|A) log_2 \frac{p(item_i|A)}{p(item_i|B)} \quad (3)$$

Let us exemplify how KLD is calculated for a word-based model assuming two data sets A and B that consist of one sentence each:

(*A*) $John_{p(0.09)}$ $is_{p(0.09)}$ $watching_{p(0.09)}$ $the_{p(0.27)}$ $cat_{p(0.09)}$ $hunting_{p(0.09)}$ $the_{p(0.27)}$ $mouse_{p(0.09)}$ $and_{p(0.09)}$ $the_{p(0.27)}$ $dog_{p(0.09)}$.

(*B*) $The_{p(0.23)}$ $cat_{p(0.08)}$ $is_{p(0.15)}$ $hunting_{p(0.08)}$ $the_{p(0.23)}$ $mouse_{p(0.08)}$ $and_{p(0.15)}$ $the_{p(0.23)}$ $dog_{p(0.08)}$ $and_{p(0.15)}$ $John_{p(0.08)}$ $is_{p(0.15)}$ $watching_{p(0.08)}$.

All words are present in each data set (equal vocabulary size) but partly with varying probabilities. Consider the comparison between A and B first, i.e. $D(A||B)$.

The first step is to calculate pointwise KLD of each word. For this, we insert the probability of each word into eq. (3) (exemplified in (4) and (5) for the word *the* and *is*, respectively). For *the*, pointwise KLD is 0.0657, i.e. >0 indicating that *the* is distinctive for *A* (higher occurrence rate than in *B*, i.e. higher probability); pointwise KLD for *is* results in −0.069 bits, i.e. <0 and thus not distinctive for *A*:

$$D(A||B) = p(the|A)log_2\frac{p(the|A)}{p(the|B)} = 0.27log_2\frac{0.27}{0.23} = 0.0657 \tag{4}$$

$$D(A||B) = p(is|A)log_2\frac{p(is|A)}{p(is|B)} = 0.09log_2\frac{0.09}{0.15} = -0.069 \tag{5}$$

To obtain the overall KLD (see eq. (6)), i.e. to calculate how many additional bits are needed when *B* is used to model *A*, the sum of all pointwise KLD values is taken. For simplicity, *sW* in eq. (6) denotes the words occurring only once in *A* and *B* in our example (six words: *John, watching, cat, hunting, mouse, dog*). To this, the pointwise KLD of *is*, *the* and *and* is added resulting in 0.0592 additional bits needed to model *A* with *B*. As mentioned earlier, KLD is asymmetric, i.e. calculating $D(B||A)$ may result in a different score. In fact, $D(B||A)$ results in 0.0667 bits, i.e. more bits are needed to model *B* with *A* than vice versa. Intuitively, this makes sense as *B* is, e.g. a longer sentence than *A*.

$$\begin{aligned} D(A||B) &= (6 * (p(sW|A)\, log_2\frac{p(sW|A)}{p(sW|B)})) + p(is|A)\, log_2\frac{p(is|A)}{p(is|B)} \\ &\quad + p(and|A)\, log_2\frac{p(and|A)}{p(and|B)} + p(the|A)\, log_2\frac{p(the|A)}{p(the|B)} \\ &= (6 * (0.09\, log_2\frac{0.09}{0.08})) + 0.09\, log_2\frac{0.09}{0.15} + 0.09\, log_2\frac{0.09}{0.15} \\ &\quad + 0.27\, log_2\frac{0.27}{0.23} = 0.0592 \end{aligned} \tag{6}$$

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 (e.g. words) that are primarily associated with a difference, i.e. features that need (relatively) high amounts of additional bits for encoding. By estimating a feature's individual contribution to KLD, we can identify those features that are mostly responsible for the divergence. Note again that KLD being an asymmetric measure, there may be a difference between a data set *A* and *B* when *B* is used as a basis for encoding but not necessarily vice versa. Also, the individual items responsible for a difference may be different ones. KLD's inherent asymmetry is especially useful here because it allows to adopt different perspectives.

For example, a speaker of "general" English might less well understand a speaker using the scientific register than vice versa; or, for a contemporary speaker a text from the past might be easier to understand than a text from the present for a historical speaker.

All our KLD models control for differences in vocabulary size by using Jelinek–Mercer smoothing and lambda 0.05 (cf. Zhai and Lafferty 2004; Fankhauser et al. 2014). For our analyses, we build two kinds of models, one is a word-based unigram model (reflecting lexical usage), the other is a model based on part-of-speech (POS) trigrams (approximating grammatical usage). See analyses in Sections 4.1 and 4.2. For comparison of scientific with general language (H1), we use 50-year periods, comparing word-based and POS-trigram language models of RSC and CLMET both ways, i.e. $D(RSC1700||CLMET1700)$ and $D(CLMET1700||RSC1700)$). Following Fankhauser et al. (2014), we use an unpaired Welch's *t*-test on the observed probabilities in the individual documents of a corpus to assess the statistical significance of an observed difference in overall frequencies. This is especially useful when a subcorpus contains only a few documents or an item only occurs in few documents.

For inspecting the diachronic development of scientific language (H2), we slide over the time line of the RSC comparing adjacent time periods and finding peaks or troughs in relative entropy as indicators of change. For this, first, we select a starting year (e.g. 1700) and a time range (e.g. two years) in which we assume linguistic changes will have happened that we use as slider over the time line. KLD is then used to compare preceding (PRE) and subsequent time periods (POST) (e.g. 20-year periods). The concrete size of the slider and time periods depends on the data set used. As there is the odd year without any publication in the RSC, we use a two-year slider. For other text types, such as news texts, a different size (e.g. months or days) may be more appropriate. The bigger the slider, the less fine-grained the observed changes in the data will be. As time periods, we use 20 years, since this turns out to be the most suitable for observing substantial changes in the present corpus (again, for other corpora with other text types, other ranges may be better suited; cf. Degaetano-Ortlieb and Teich 2018). As KLD is asymmetric, we inspect different divergence scores from the perspective of the past looking to the present (i.e. $D(POST||PRE)$) vs. from present to past (i.e. $D(PRE||POST)$) (cf. Section 4.2). To capture lexical as well as grammatical changes, models are built based on lemmas and POS trigrams. A feature's contribution to divergence is again obtained by considering pointwise KLD and the *p*-value of the *t*-test. If a peak or trough is attested in the overall KLD in a particular year, ranking the features based on their pointwise KLD in that year shows which features contribute most to a change.

3.2.2 Average surprisal

The informativity of a given linguistic item can be thought of as the average (i.e. the usual) amount of information that the item carries in a given corpus (its surprisal; cf. see eq. (1) above). Technically, it is the weighted average of the negative log probability of all the occurrences of that item in a given data set (cf. Degaetano-Ortlieb and Teich 2016) (see eq. (7)):

$$AvS(item) = \frac{1}{|item|} \sum_{i=1}^{n} - \log_2 p(item_i|context_i) \tag{7}$$

So rather than estimating how probable an item is in a particular context, as in surprisal in on-line comprehension, we are interested in the predictability of an item *across* occurrences and contexts. Our motivation for using average surprisal is similar to that of Cohen Priva (2015), i.e. to capture the fact that there are items whose general informativity is low, even if their local informativity may be high (cf. Cohen Priva 2015: 248). For the items we select for analysis, we are interested in how they settle in certain ranges of high to low surprisal (cf. analysis in Section 4.3). The items we consider are typically words *w* with a preceding context of three other words, w_{i-1}, w_{i-2}, w_{i-3}, as in a four-gram language model (see eq. (11)):

$$AvS(w) = \frac{1}{|w|} \sum_{i=1}^{n} - \log_2 p(w_i|(w_{i-1}w_{i-2}w_{i-3})) \tag{8}$$

For illustration, let us consider a sentence starting with *Jane read a* (cf. examples (1) and (2) discussed in Section 1 above) and assume that the possible continuations are *book*, *magazine*, *newspaper*, and *article* and that they are equally likely to occur. In this case, the probability of *book* is 0.25 (1 over 4) and its surprisal is:

$$S(book) = -log_2 p(book|Jane\ read\ a) = -log_2 p(0.25) = 2 \tag{9}$$

Usually, the probabilities are not equal but one or some continuations are more likely than others. If, say, *book* is much more likely (e.g. 0.8) than the other options, then *book* is more expected and surprisal will be much lower:

$$S(book) = -log_2 p(book|Jane\ read\ a) = -log_2 p(0.8) = 0.32 \tag{10}$$

In a corpus, a given item will occur in different contexts, as illustrated by our examples (1) and (2) above where *book* will be more likely in (1) than in (2) (a

book is a quite likely item to be read but many things can be bought), resulting in the following probabilities and surprisal values of *book*: 0.8 and 0.322 bits in the first example (*Jane read a book*) and 0.2 and 2.322 bits in the second example *Jane bought a book*. The average surprisal of *book* in our examples will then be:

$$AvS(book) = \frac{1}{|2|}(-\log_2 p(0.8)) + (-\log_2 p(0.2)) = \frac{1}{|2|}(0.322 + 2.322) = 1.32 \quad (11)$$

Furthermore, we estimate surprisal of words as used in grammatical patterns, approximated by POS trigrams, that we observe to be involved in shaping scientific language over time (cf. Section 4.3). Here, we calculate AvS on the words filling a POS trigram (lexical level) as well as on the parts of speech (grammatical level) themselves. To obtain AvS values for POS trigrams at the lexical level, we take the mean of the AvS of the three words filling a POS trigram and averaging over all instances (cf. eq. (12)):

$$AvS(POStrigram) = \frac{1}{|POStrigram|}\sum_{i=1}^{n}(\frac{AvS(w_1) + AvS(w_2) + AvS(w_3)}{3})_i \quad (12)$$

To obtain AvS values for POS trigrams at the grammatical level, we proceed in the same way, i.e. we calculate AvS on each part of speech, take the mean of the AvS of the three POS of a trigram and average over all instances. This allows us to compare surprisal (mean and variance) across POS-trigram instances and time, inspecting surprisal shifts. A decreasing tendency points to a more confined lexical/grammatical usage of a trigram and an increasing tendency to a greater variation in the trigram.

4 Analysis and results

We now address the two hypotheses formulated in Section 1, repeated here for convenience:

- *Scientific language becomes increasingly distinctive from "general," everyday language* (H1);
- *Within scientific language, specific linguistic choices become increasingly conventionalized* (H2).

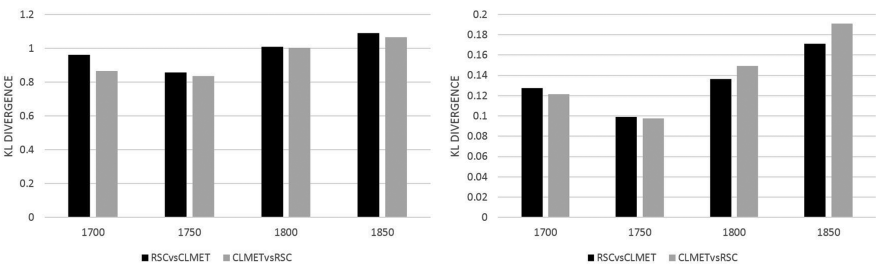
To test H1 we need to compare scientific language with register-mixed, "general" language in the time period considered (see Section 4.1). To test H2 we need to first detect those linguistic features that become typical of scientific language over

time (see Section 4.2) and then estimate their informational contributions (see Section 4.3).

4.1 Scientific language vs. "general" language

We consider KL Divergence between scientific and general language at the lexical and the grammatical level. To this end, we compare smoothed unigram word (lexical level) and part-of-speech (approximating the grammatical level) language models of the RSC (scientific) and the CLMET (general) with KLD as described in Section 3.2. Considering H1 stated above, our main assumption is an increase in divergence between scientific language and general language over time. For diachronic comparison, we consider lexical and grammatical levels at slices of 50 years (i.e. 1700 covering 1700–1749, 1750 covering 1750–1799, etc.) for both scientific vs. general (*RSCvsCLMET*) and general vs. scientific (*CLMETvsRSC*) English, i.e. two KLD scores are calculated, $D(RSC||CLMET)$ and $D(CLMET||RSC)$. Consider Figure 1 showing $D(RSC||CLMET)$ (black) and $D(CLMET||RSC)$ (gray) for words (a) and parts-of-speech (b). Overall, while the scores go down from the first to the second 50-year period, from the 1750 period onward divergence between scientific and general language increases at both levels, confirming our assumption.[3]

The other effect that can be observed, if only a subtle one, is that at the level of parts of speech, in the later periods (1800, 1850) general language is less well modeled by scientific language (gray bars supersede black bars). This may be a



(a) Lexical level based on words (b) Grammatical level based on parts of speech

Figure 1: Relative entropy (KLD) across 50-year time periods based on (a) words and (b) parts of speech (smoothed probability distributions) for RSC and CLMET.

3 Note that the numbers for parts of speech are lower than for words due to a smaller set of data points (part-of-speech instances vs. word tokens).

subtle indication of H2, i.e. of increasing conventionalization in scientific language (here: regarding grammatical usage). In fact, previous work has shown that scientific language converges in grammatical usage on complex NPs and simple clause structure of the kind *X is Y* (relational and passive clauses) cf. Halliday (1988), Atkinson (1999), Banks (2008), as well as more recently Biber and Gray (2016: 17–19). In the following, we inspect more closely these broad diachronic tendencies at both the lexical and the grammatical levels.

4.1.1 Word contributions

Figures 2 and 3 show the words that contribute to overall KLD for *RSCvsCLMET* (left) and *CLMETvsRSC* (right), i.e. those words that are most distinctive for RSC and CLMET, respectively. Color denotes relative frequency and size the individual word's contribution to overall KLD. The words are listed from left to right on each row in descending order of contribution to overall KLD. Here, we focus on the period with the highest overall divergence, 1850. Figure 2 shows the top ranking words, i.e. those with the highest contribution to KLD, Figure 3 the lower ranking ones, i.e. those words contributing less but still significantly by the *t*-test (see Section 3.2). Among the words distinctive of general language (right) are personal pronouns (*you*, *her*, *she*, *I*, *he*, *his*), conjunctions (*and*, *but*), contractions (*'s*, *n't*, *'ll*), and auxiliaries in past tense (*was*, *had*). For scientific language (left) distinctive words are determiners (*the*, *this*, *these*), the postmodification marker *of*, prepositions (*in*, *by*, *from*), auxiliaries in present tense (*is*, *be*, *are*), the relativizer *which* pointing to elaboration, and terms (e.g. *solution*, *fibres*, *acid*). The grammatical markers have been shown previously to be overrepresented in scientific text in a study on the LOB/Brown corpora by Johansson and Hofland (1989). According to Figure 3, which presents the lower ranking items (which are

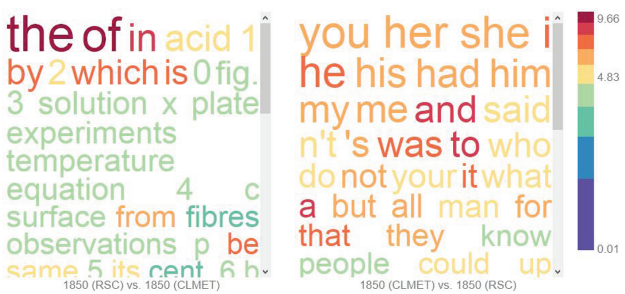


Figure 2: 1850: Top ranking words contributing to overall KLD for scientific language (RSC; left) and general language (CLMET; right). Color denotes relative frequency from high (red) to low (blue). Size denotes an individual word's contribution to overall KLD.

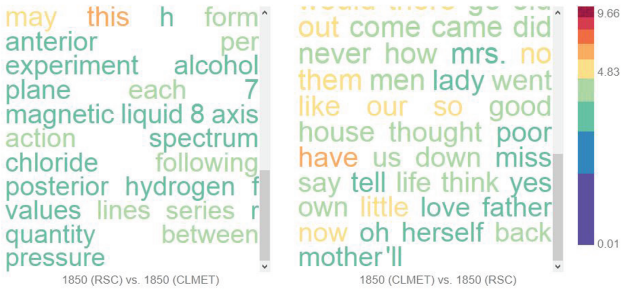


Figure 3: 1850: Lower ranking words contributing to overall KLD for scientific language (RSC; left) and general language (CLMET; right). Color denotes relative frequency from high (red) to low (blue). Size denotes an individual word's contribution to overall KLD.

lower frequency words at the same time) scientific terms are distinctive for RSC (left) and fairly general terms are distinctive for CLMET (right), clearly reflecting specialized vs. general vocabulary. Comparing this to the period of 1700 (see Figure 4), we can see that the words distinctive for RSC are less specialized than in 1850 (compare *observations, glass, experiment(s), air, sun, surface, blood, stone* in Figure 4 with *magnetic, spectrum, chloride, hydrogen* in Figure 3).

Figure 4: 1700: Lower ranking words contributing to overall KLD for scientific language (RSC; left) and general language (CLMET; right). Color denotes relative frequency from high (red) to low (blue). Size denotes an individual word's contribution to overall KLD.

4.1.2 Part of speech contributions

Figure 5 shows the parts of speech (POS) contributing to the overall KLD between scientific and general English, focusing again on the period of 1850 as it shows the highest divergence. Again, color denotes the relative frequency of each POS. The POS are listed from left to right on each row in descending order of their individual contribution to the overall KLD. For general language, personal (PP)

Figure 5: POS contributing to overall KLD for 1850 (RSC: left, CLMET: right). Color denotes relative frequency from high (red) to low (blue). Ordering from left to right denotes the contribution of individual POS to overall KLD.

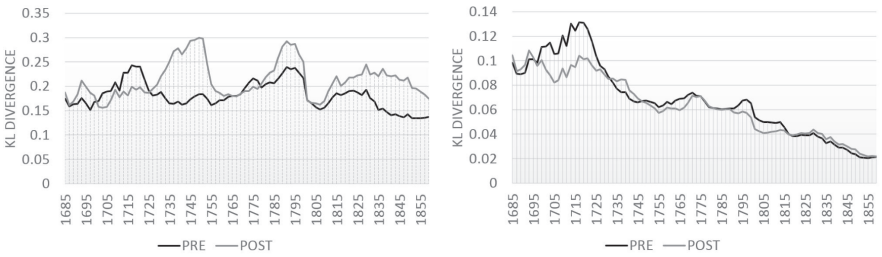
and possessive pronouns (PP$) as well as full verbs (VV), verbs in past tense (VVD, VHD, VBD), and adverbs (RB) are most distinctive. Also quotes (“”), wh-pronouns (WP), conjunctions (CC), modal verbs (MD), *to* (TO) and other particles (RP), interjections (UH), possessives (POS), and existential *there* (EX) are distinctive. For scientific language, proper nouns in singular (NN) and plural (NNS), determiners (DT), adjectives (JJ), prepositions (IN) and wh-determiners (WDT) are distinctive as well as participles (VVN) and verbs in present tense (VVZ and VBZ). Also distinctive are cardinal numbers (CD), parentheses, and list items (LS). Moreover, we can see that the number of POS distinctive for general language is higher than for scientific language. Thus, while general language has a more varied set of distinctively used options, scientific language has a more confined set.

In summary, scientific and general language increasingly diverge from one another over time both at the lexical and the grammatical level. Among the features (lexical items, parts of speech) contributing to the increasing distinction, we observe indicators of involved, verbal style vs. informational, nominal style on the oral-written cline, which is very much in line with previous observations, e.g. by Halliday (1988) or Biber et al. (1999). The communicative implications are that at the lexical level, scientific language becomes harder to understand based on knowledge of “general” language due to specialized vocabulary, while at the grammatical level, conventionalization sets in and eases communication.

4.2 Diachronic development of scientific language

To capture the course of development of scientific language *internally*, we compare adjacent time periods of a specific range (here: 20 years) by relative entropy sliding over the time line with a two-year slider (as described in Section 3.2.1) in the RSC. We consider lemma-based models for the lexical level and POS-trigram models to approximate the grammatical level.

Figure 6a shows the overall results from two perspectives: the black line depicts the past (PRE periods) when compared to the future (POST periods)



(a) Model on lemmas (b) Model on POS trigrams

Figure 6: Relative entropy (KLD) between 20-year periods with a 2-year slider in RSC.

relative to the 2-year slider, the gray line depicts the future when compared to the past. While both distributions show peaks and troughs over time, the peaks are especially prominent for periods following the slider (gray line around 1730–50s, 1790s, and 1810–1850s), indicating periods with a more varied vocabulary for the future (POST) when compared to the past (PRE). This tendency confirms our assumption of periods of lexical innovation/expansion, reflecting times of scientific discovery, where new words are created or existing words are placed in new contexts. In addition, this tendency seems to proceed in waves, a period of lexical expansion with higher KLD being followed by periods of greater lexical similarity when the two lines converge.

To approximate the development at the grammatical level, we use part-of-speech trigrams (see Figure 6b). The major trend is a decrease in KLD over time. The prominent peak around the beginning of the eighteenth century shows a greater divergence of the PRE period to the POST period. Thus, while there was a more varied use of grammatical structures in the past, toward the 1850s there is a rather strong tendency toward grammatical consolidation as the POST and PRE periods diverge less and less from one another.[4]

To analyze which linguistic features contribute to the overall trends observed at the lexical and grammatical levels, we inspect the individual contributions of lemmas and POS trigrams to the overall KLD (i.e. the pointwise KLD).

4.2.1 Lexical level

We start with considering the first major peak at the lexical level in 1739 (see gray line in Figure 6a), marking a change between the lexical distributions in

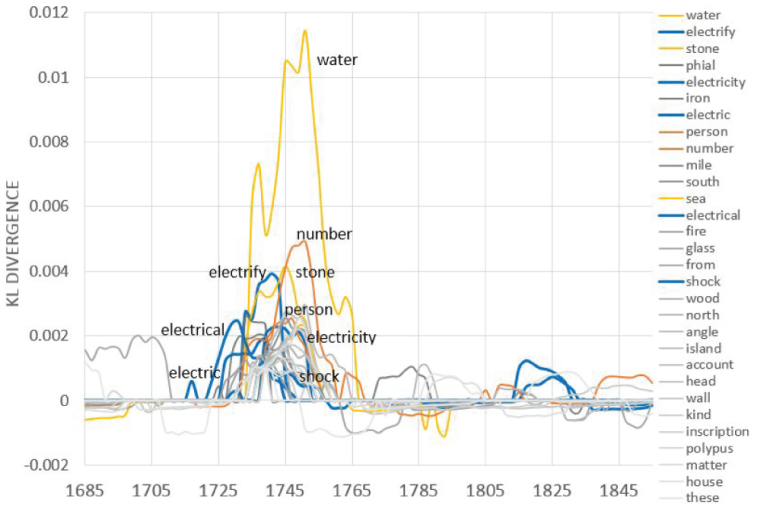
4 Note that we observe a similar, much less pronounced trend in CLMET, but since CLMET is not a balanced corpus, this cannot be further interpreted.

the 20 years preceding and following 1739. In Figure 7a, we plot the top 30 lemmas with the highest KLD contributions. Among these lemmas, ELECTRICITY stands out in particular with *electrify, electrical, electric, electricity*, and *shock*. In fact, in the years following 1739, many experiments reported in the Royal Society were devoted to electricity with one of the world's most famous scientific experiment by Benjamin Franklin in 1751 on the effects of lightning. Other fields of study are related to observations on natural phenomena and resources, such as earthquakes and water springs (*water, sea, stone*, e.g. in *An Extraordinary and Surprising Agitation of the Waters* reporting a.o. on the earthquake of Nov. 1755), and to census statistics of countries and cities (*person, number*, e.g. *A Letter to George Lewis Scot concerning the Number of People in England*).

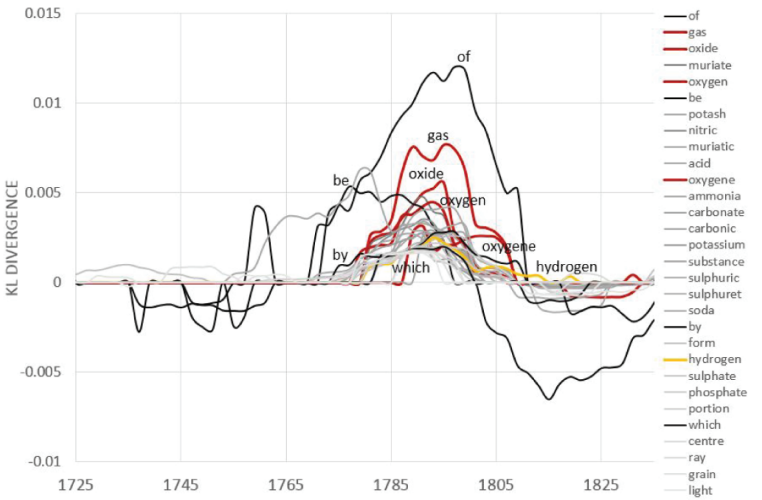
Figure 7b shows the top 30 lemmas contributing to the second major peak in KLD in 1791. This period of lexical innovation/expansion can be attributed to the formation of a new field of chemistry based on the discoveries of oxygen and hydrogen around the mid 1760s and 1770s, the respective terms being only established around the 1790s (cf. Degaetano-Ortlieb and Teich 2018). Besides other lemmas related to substances in this new scientific field (e.g. *muriatic*, *acid*, *potash*, *ammonia*, *carbonate*, etc.), particular function words are distinctively used around this period in comparison to the past. This is particularly interesting because parallel to a period of lexical innovation (related to new scientific discoveries), after inspecting the distinctive function words (e.g. *be, of, which*), specific expressions for definition and elaboration seem to become distinctive: relational *be* as in *X is Y*, passive voice, longer nominal phrases with the preposition *of* and relative clauses with the relative pronoun *which*.

4.2.2 Grammatical level

As function words are clearly shown to be involved in shaping scientific language over time, we more closely inspect changes at the grammatical level using POS trigrams and comparing again 20-year periods using a 2-year slider. Here, our focus is on the period around the end of the eighteenth century, in which the new subfield of chemistry established itself (as shown in Section 4.2.1). Figure 8 shows the top 30 POS trigrams contributing to KL divergence in 1791. The blue dashed lines denote trigrams with a preposition (IN), the most distinctive POS trigram being the noun–preposition–noun trigram (NN.IN.NN, such as *degree of fire, quantity of water*). This is in line with the rise in contribution to KLD of the preposition of that we observed with a lemmabased model (cf. Figure 7b). The black lines mark nominal patterns, the most prominent being the determiner–adjective–noun trigram (DT.JJ.NN, e.g. *the inflammable/dephlogisticated air*), its



(a) Peak in 1739: blue lines marking lemmas related to ELECTRICITY, yellow lines marking other subjects.



(b) Peak in 1791: red lines marking lemmas related to OXYGEN, yellow to HYDROGEN; black lines show function words.

Figure 7: Lemma contribution to KLD peaks in 1739 and 1791 (comparison between 20-year PRE and POST periods with 2-year slider).

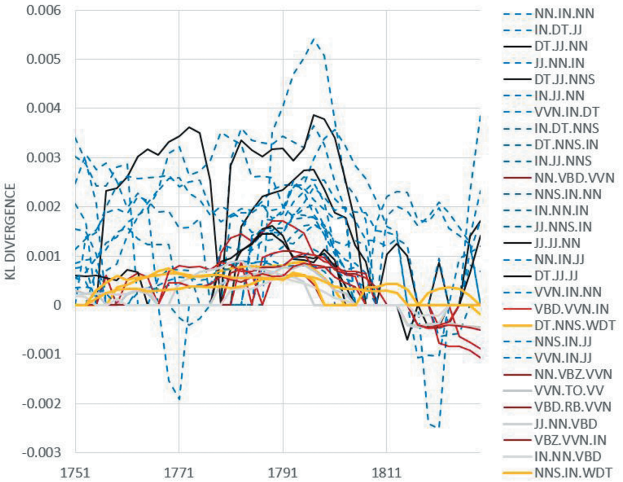


Figure 8: POS trigram contribution to KLD peak in 1791 (comparison of 20-year PRE and POST periods with 2-year slider); blue shades for use of prepositions, black for nominal phrases, red for passive, yellow for relative clauses.

lexical realizations reflecting terminological usage (especially around chemistry). The red lines mark passive clauses, the most distinctive pattern being the noun-BE(past tense)-participle trigram (NN.VBD.VVN, e.g. *rod was electrified*, *sediment was deposited*). This reflects the rise in the verb be observed at the lexical level. The yellow lines mark the use of relative clauses, with the determiner–noun (plural)–wh-trigram (DT.NNS.WDT; e.g. *the substances/vapours which*) ranking highest in terms of contribution to KLD. This reflects the rise in the relativizer *which* seen with the lemma-based model.

While these patterns reflect the single function words' contribution to KLD at the lexical level (see again Figure 7b), using POS trigrams gives a much better insight into which grammatical patterns exactly contribute to KLD in this period.

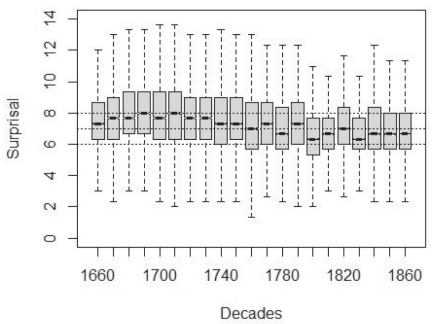
4.3 Toward an optimal code: linguistic patterns for modulating information content

We have shown which lexico-grammatical features figure prominently in the development of scientific language, observing two linguistic motifs, lexical innovation/expansion and grammatical consolidation. Lexis reacts directly to external pressures, such as new scientific insights or discoveries by introducing new words or using known words in new contexts (indexed by peaks of

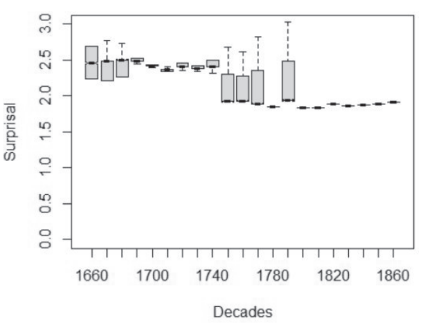
KLD) while grammatical usage is being conventionalized (indexed by decreasing KLD over time) (cf. Section 4.2). To show that this interplay has the effect of modulating information content – a necessary property of an optimal code – we inspect the average surprisal of one relevant lexico-grammatical pattern, indexed by the *noun–preposition–noun* trigram (NN-IN-NN), on an exemplary basis. For simplicity we henceforth use "surprisal" for average surprisal (cf. Section 3.2).

At the lexical level, considering the lexical instantiations of the pattern by decade (Figure 9a), we can see ups and downs in surprisal with a very slight decreasing tendency over time (from around 8 to 7 bits; a decrease of around 13% from the 1660s to 1860s). At the grammatical level, considering surprisal of parts of speech (Figure 9b),[5] we can see that surprisal of the pattern steadily decreases over time (by around 30%) with a major drop between the 1740s and 1750s, the pattern becoming highly predictable. Thus, while the *noun–preposition–noun* pattern as such becomes more predictable grammatically – its frequency of occurrence rises significantly from 2406.25 per million words in the beginning to 9409.76 per million words in the final decades – surprisal at the lexical level is fairly stable.

Inspecting the lexical level by considering the surprisal mean of lexical items at each position (NN1, IN, NN2), we observe a high level of surprisal on the nouns (around 10 bits) and low surprisal on the prepositions (around 3 bits), reflecting the natural difference in information content between lexical words and function words (see Figure 10 left). The general tendency over time is a marginal



(a) Surprisal of NN.IN.NN (lexical)



(b) Surprisal of NN.IN.NN (parts of speech)

Figure 9: Surprisal of *noun–preposition–noun*.

5 Note that surprisal scores are naturally much lower at the level of parts of speech compared to words due to sparsity.

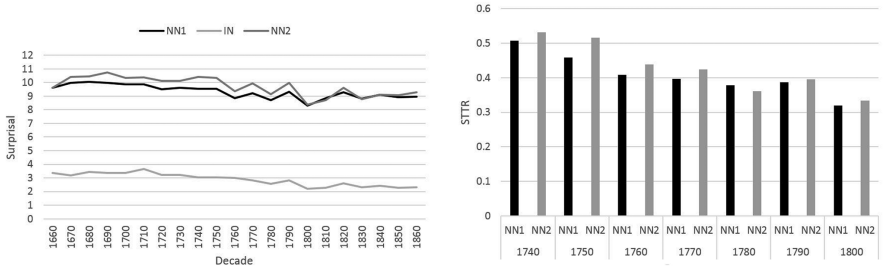


Figure 10: Surprisal (left) and standardized type token ratio (right) of NN.IN.NN.

decrease in surprisal, starting in the 1740s with a major drop between the 1790s and 1800s from 9.33 to 8.29 bits (around 10%) for the first noun (black line) and 9.97 to 8.38 bits (around 16%) for the second noun (dark gray line). The development of standardized type token ratio (STTR) of both nouns (see Figure 10 right) confirms this tendency: variability in the pattern in that period declines steadily (decrease in STTR for both noun positions). However, in the period of lexical innovation/expansion in the 1790s (cf. Section 4.2), STTR and surprisal rise, i.e. there is more variation and less predictability at the lexical level. After that period, STTR as well as surprisal drop considerably (1800s).[6]

Thus, besides the observed grammatical conventionalization over time, this leads us to the following assumptions: (a) in periods of lexical expansion innovative lexical usages of the pattern arise, (b) after periods of lexical expansion further lexical conventionalization sets in, possibly indicating a process of terminology formation. To investigate (a) and (b) further, we select the noun *oxide*, a relatively specific noun arising in the period of lexical expansion shown in Figure 7b. Figure 11 (left) shows surprisal from *oxide*'s first realization in the pattern up to 1869. For the preposition, surprisal is relatively low and stable over time (around 2 bits). Considering the nouns, in the 1790s, surprisal of *oxide* and the second noun is quite high (around 13 and 11 bits, respectively; see Example (3)). This indicates a lexically innovative use of the pattern in the lexical expansion period. In the following decade (1800s), surprisal for *oxide* drops to around 7 bits, rising in predictability. At specific points in time *oxide* is followed by more and less pre-

6 Note also that till the 1800s, the second noun is less predictable than the first one, indicating a difference in specificity: the first noun is a more general noun, the second a more specific one (e.g. *proportion of oxygen*, *quantity of sulphur*). Similarly, STTR shows more variability of the second noun. Interestingly, from the 1800s both nouns converge in surprisal, being equally predictable. This development may indicate a process of term formation.

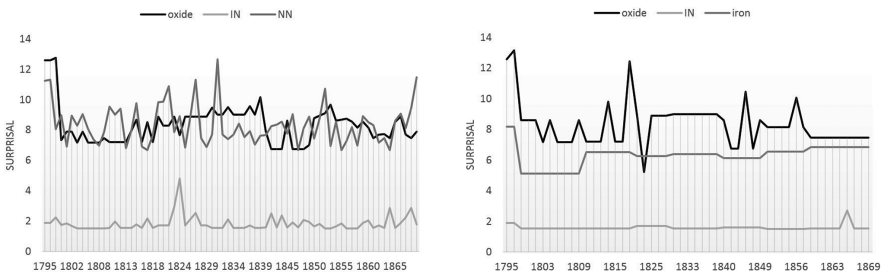


Figure 11: Surprisal of the *noun-preposition-noun* pattern with *oxide*-IN-NN (left) and *oxide*-IN-*iron* (right).

dictable nouns (periods of peaks and troughs in surprisal for the second noun; see Examples (4) and (5) for *oxide* with mid surprisal followed by *nickel* with high surprisal and *iron* with low surprisal.

(3) *A_5.306 little_6.066 green_11.011 or_6.038 blackish_11.741* ***oxide_12.587 of_1.894 copper_10.546*** *adhered_13.363 to_3.004 their_6.937 surfaces_10.713.*
(lexical realization_surprisal; 1796, George Pearson, *Observations on Some Ancient Metallic Arms and Utensils.*)

(4) *The_2.895* ***oxide_7.169 of_1.536 nickel_10.995*** *was_4.793 precipitated_8.235 by_3.985 hydrogenized_14.240 sulphuret_9.966 of_1.536 ammonia_6.799.*
(lexical realization_surprisal; 1802, Edward Howard, *Experiments and Observations on Certain Stony and Metalline Substances [...]*)

(5) *The_2.895* ***oxide_7.169 of_1.536 iron_5.131,*** *_2.224 precipitated_8.235 by_3.985 ammonia_6.799,_2.224 weighed_9.726 8_9.408 grains_6.263;*
(lexical realization_surprisal; 1802, Edward Howard, *Experiments and Observations on Certain Stony and Metalline Substances [...])*

To further observe possible lexical conventionalization (b), we select the most frequent lexical realization, *oxide*-preposition-*iron* (see Figure 11 right). This realization shows higher surprisal in the 1790s, followed by a period of lower surprisal from 1800 to 1810, especially for *iron* (around 5 bits). Considering the mutual information (MI) between *oxide* and *iron* over time, MI is highest in 1800 (0.0297) and a bit lower in 1810 (0.0172), and much lower in the other decades (<0.009). This is an indication of its development toward a term in that period. But it is not until the 1860s that the term is also contextually conventionalized, indicated by a stable mid-surprisal range of both nouns.

5 Summary and conclusions

We have presented an account of change in language use focusing on the linguistic development of scientific English in the late modern period. Specifically, we have suggested to conceptualize change in language use as information-theoretic optimization, opening up the opportunity of a communicative explanation. At the same time, our approach does not deny the influence of other factors at work in change in language use, notably social and cultural factors. While information-theoretic approaches are wide-spread in psycholinguistic and computational linguistic research and have brought many relevant insights on on-line language processing, we here showcase its benefits for the study of change in language use and language variation. An information-theoretic perspective maintains that language users are rational and strive for an "optimal encoding" of their messages by using specific linguistic options to modulate the level of information transmitted. Using scientific language as an example, we have demonstrated that this mechanism is also at work in language diachrony.

In the continuous dynamics of language use, there is a steady external pressure to adapt to new experiences, pushing for innovation and increase in expressivity; at the same time, communicative concerns pull toward convergence on particular options and exclusion of others. We have shown that such forces are at work in scientific language from two perspectives. First, in terms of the overall diachronic development of scientific English, there is convergence on selected grammatical options, i.e. they become conventional and thus more expected (less surprising). In contrast, while showing a slight trend toward conventionalization too, what is characteristic of the lexical level is its versatility, which is indexed by informational highs in phases of lexical innovation/expansion and mid to low levels of information in phases of stability/consolidation. Second, phases of lexical innovation/expansion are typically accompanied or followed by grammatical conventionalization. Over time, particular grammatical patterns and their realizations, such as *noun–preposition–noun*, become habitual hosts for lexical innovation and terminology formation. This interplay has specific informational signatures, conventionalized grammatical patterns settling on lower surprisal levels over time, while their lexical instantiations showing varying levels of information over time (as explained above). In sum, what emerges is a distinctive code that converges on specific linguistic means to modulate information content – in other words: an optimal code. We have shown this for scientific English but strongly assume that the same mechanism applies to other registers and to language overall; see e.g. Hundt and Mair (2012) on the versatility of registers/genres or Leech et al. (2009) on the recent change in English being

characterized by a mix of more economic verbal expressions (e.g. *wanna*, *gonna*) and greater complexity in nominal expressions.

For our analyses, we have employed two information-theoretic measures, *relative entropy* (Kullback–Leibler Divergence), a common measure to compare probability distributions, and *surprisal*, a widely-used measure of information correlated with cognitive effort on on-line language processing. We applied KLD to test hypothesis H1 – scientific language and general language become increasingly distinct over time – estimating the divergence over time between the RSC and the register-mixed CLMET on the basis of words and parts of speech (Section 4.1). The results show that both with a word-based model and a POS-based model KLD increases over time, i.e. scientific language and "general" language become more distinct from one another over time. For hypothesis H2 – scientific language becomes increasingly conventionalized over time – we first employed KLD to detect the features driving diachronic change (Section 4.2). Second, to inspect more closely the informational contributions of typical grammatical patterns (POS trigrams) and their lexical instantiations, we estimated their (average) surprisal over time (Section 4.3). The results confirm conventionalization both at the lexical and the grammatical level.

At the methodological level, we have demonstrated the relevance and effectiveness of information-theoretic measures for modeling linguistic variation and change, adding two more measures to the established corpus-linguistic and computational-linguistic repertoires (loglikelihood, (pointwise) mutual information, information gain, perplexity, etc). In variational linguistics, we need procedures for detecting patterns of variation according to a given variable and methods for assessing the contribution of a given linguistic feature to a distinction. The most common approach is a frequency-based one with predefined features and a bias toward high-frequency occurrences. Effect size is often not considered or estimated by complex, separate procedures. The information-based measures we have employed here support feature detection as well as feature evaluation: Estimating the (relative) amount of information on linguistic units (e.g. words, parts of speech) in context (e.g. preceding word trigrams), as commonly implemented in computational language models, we can detect discriminatory features as well as assess the discriminatory strength of features. Apart from being effective, information-theoretic measures have been shown to be very reliable. For example, Goodkind and Bicknell (2018) demonstrate that a model's predictive power improves as a linear function of language model quality, so even if model quality could be better, the kinds of effects we can observe are the same. This means that surprisal estimates are fairly robust and even lower-quality language models provide usable results.

While we focused on the scientific domain, the methods we have employed can be applied to any kind of diachronic or otherwise contrastive investigation and thus open up the possibility of communicative explanations for language variation and change at large. For instance, regarding the formation of distinctive varieties, registers are beneficial to communication because they reduce entropy by settling on a subset of linguistic options and skewing overall probabilities, thus easing communication. In our ongoing work, we analyze e.g. terminology formation from the perspective of development of an optimal code, using information-theoretic measures (entropy, mutual information) to capture the life cycle of terms. Regarding (long-term) change of the language system, whether a change persists or not will depend on its contribution to the communicative efficiency and efficacy of the system as a whole (or some subsystem of it) (cf. again Harris 1991). Here, information theory may well provide a suitable basis for modeling and explaining grammaticalization processes in terms of code optimization.

Necessarily, there is room for improvement and we leave open a few questions. Regarding analysis, we are aware that POS-trigrams are a mere approximation of grammatical structure. If more complex structures (syntactic embedding, long-distance dependencies) are to be investigated, then an n-gram approach clearly starts to falter. In the long term, we need syntactically richly annotated diachronic data sets, such as e.g. the Penn-Helsinki Parsed Corpus of Early Modern English or the Parsed Corpus of Early English Correspondence[7] and interlink them with information-theoretic approaches such as the one shown in the present paper. Here, we are currently exploring several approaches for syntactic parsing, including models trained on contemporary scientific language (Nguyen and Verspoor 2019). In terms of computational modeling, what we have not addressed in this paper are effects of the noisy channel, i.e. to what extent linguistic choices are adapted to efficacy and robustness of communication (i.e. reduction of error probability in transmission). This would warrant a study on its own in which we can tease apart source and channel coding (for an example see again Pate and Goldwater 2015). Also, effects of convergence would need further investigation regarding the language-external conditions under which interlocutors converge or not (e.g. prestige, time period and intensity of interaction; see e.g. the study by Danescu-Niculescu-Mizil et al. (2013) on the linguistic reflexes of membership phases in on-line communities). Given that what we observe here are all predictability effects, information theory offers a unifying framework to formally model such effects as well as a theoretical basis for an explanation of language use, variation and change as information-theoretic optimization.

7 cf. www.helsinki.fi/varieng/CoRD/corpora/

Funding: This work was supported by the Deutsche Forschungsgemeinschaft, Funder Id: http://dx.doi.org/10.13039/501100001659, Grant Number: SFB1102: Information Density and Linguistic Encoding.

References

Atkinson, Dwight. 1999. *Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. New York: Erlbaum.

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56.

Baayen, R. Harald, Fabian Tomaschek, Susanne Gahl & Michael Ramscar. 2017. The Ecclesiastes Principle in language change. In Marianne Hundt, Sandra Mollin & Simone E. Pfenninger (eds.), *The Changing English language: Psycholinguistic perspectives*, 21–48. Cambridge, UK: CUP.

Balling, Laura Winther & R. Harald Baayen. 2012. Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition* 125(1). 80–106.

Banks, David. 2008. *The development of scientific writing: Linguistic features and historical context*. London/Oakville: Equinox.

Baron, Alistair & 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.

Bell, Alan, Jason M. Brenier, Michelle Gregory, Cynthia Girand & Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60. 92–111.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65(3). 487–517.

Biber, Douglas & Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In Terttu Nevalainen & Leena Kahlas-Tarkka (eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, 253–276. Helsinki: Société Néophilologique.

Biber, Douglas & Bethany Gray. 2011. The historical shift of scientific academic prose in English toward less explicit styles of expression: Writing without verbs. In Vijay Bathia, Purificación Sánchez & Pascual Pérez-Paredes (eds.), *Researching specialized languages*, 11–24. Amsterdam: John Benjamins.

Biber, Douglas & Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Studies in English Language. Cambridge, UK: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow, UK: Longman.

Bochkarev, Vladimir, Valery D. Solovyev & Soren Wichmann. 2014. Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface* 11(101). 20140841.

Cohen Priva, Uriel. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6(2). 243–278.

Cohen Priva, Uriel & Emily Gleason. 2016. Simpler structure for more informative words: A longitudinal study. In *8th annual conference of the cognitive science society*, 1895–1900.

Crocker, Matthew W., Vera Demberg & Elke Teich. 2016 Feb. Information density and linguistic encoding (IDeaL). *KI – Künstliche Intelligenz* 30(1). 77–81.

Danescu-Niculescu-Mizil, Cristian, Robert West, Dan Jurafsky, Jure Leskovec & Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in on-line communities. In *Proceedings of the 22nd international World Wide Web conference (WWW)*, Brazil: Rio de Janeiro.

De Beaugrande, Robert-Alain & Wolfgang Dressler. 1981. *Introduction to text linguistics*. London, New York: Longman.

De Smet, Hendrik. 2006. A corpus of late modern English texts. *ICAME Journal* 29. 69–82.

De Smet, Hendrik. 2016. How gradual change progresses: The interaction between convention and innovation. *Language Variation and Change* 28. 83–102.

Degaetano-Ortlieb, Stefania. 2018. Stylistic variation over 200 years of court proceedings according to gender and social class. In *Proceedings of the 2nd workshop on stylistic variation at NAACL*, New Orleans, USA.

Degaetano-Ortlieb, Stefania & Elke Teich. 2016. Information-based modeling of diachronic linguistic change: From typicality to productivity. In *Proceedings of the 10th LaTeCH workshop at ACL*, 165–173.

Degaetano-Ortlieb, Stefania & Elke Teich. 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 literature at COLING2018*, 22–33, NM, USA: Santa Fe.

Degaetano-Ortlieb, Stefania, Hannah Kermes, Ekaterina Lapshinova-Koltunski & Elke Teich. 2013. SciTex – A diachronic corpus for analyzing the development of scientific registers. In Paul Bennett, Martin Durrell, Silke Scheible & Richard J. Whitt (eds.), *New methods in historical corpus linguistics*, *volume 3 of Corpus linguistics and interdisciplinary perspectives on language*, 93–104. Tübingen: Narr.

Degaetano-Ortlieb, Stefania, Hannah Kermes, Ashraf Khamis & Elke Teich. 2019. An information-theoretic approach to modeling diachronic change in scientific English. In Carla Suhr, Terttu Nevalainen & Irma Taavitsainen (eds.), *From data to evidence in English language research*, Language and Computers, 258–281. Leiden: Brill.

Delogu, Francesca, Matthew Crocker & Heiner Drenhaus. 2017. Teasing apart coercion and surprisal: Evidence from ERPs and eye-movements. *Cognition* 116. 49–59.

Diller, Hans-Jürgen, Hendrik De Smet & Jukka Tyrkkö. 2011. A European database of descriptors of English electronic texts. *The European English Messenger* 19. 21–35.

Engelhardt, Paul E. Ş. Barış Demiral & Fernanda Ferreira. 2011. Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition* 77(2). 304–314.

Fanego, Teresa. 1996. The gerund in early modern English: Evidence from the Helsinki Corpus. *Folia Linguistica Historica* 17. 97–152.

Fankhauser, Peter, Jörg Knappen & Elke Teich. 2014. Exploring and visualizing variation in language resources. In *Proceedings of the 9th language resources and evaluation conference (LREC)*, 4125–4128, Reykjavik.

Franke, Michael & Elliott O. Wagner. 2014. Game theory and the evolution of meaning. *Language and Linguistics Compass* 8(9). 359–372.

Genzel, Dmitriy & Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th ACL*, 199–206. Philadelphia, PA, USA.

Goodkind, Adam & Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, 10–18. Salt Lake City, UT, USA.

Grice, H. Paul. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan, (eds.), *Syntax and Semantics*, vol. 3. New York: Academic Press.

Gries, Stefan Th. & Martin Hilpert. 2008. The identification of stages in diachronic data: Variability-based Neighbor Clustering. *Corpora* 3(1). 59–81.

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd meeting of the North American chapter of the association for computational linguistics on language technologies*, 1–8.

Halliday, M. A. K. 1985. *Written and spoken language*. Melbourne: Deakin University Press.

Halliday, M. A. K. 1988. On the language of physical science. In Mohsen Ghadessy (ed.), *Registers of written English: Situational factors and linguistic features*, 162–177. London: Pinter.

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

Harris, Zellig. 1991. *A theory of language and information. A mathematical approach*. Oxford: Clarendon Press.

Hughes, James M., Nicholas J. Foti, David C. Krakauer & 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.

Hundt, Marianne & Christian Mair. 2012. "Agile" and "uptight" genres: The corpus-based approach to language change in progress. In Douglas Biber & Randi Reppen (eds.), *Corpus Linguistics. Varieties*, vol. 3, 199–218. London: Sage.

Jäger, Gerhard. 2008. Applications of game theory in linguistics. *Language and Linguistics Compass* 2(3). 406–421.

Jaeger, T. Florian & Roger P. Levy. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schölkopf, John C. Platt & T. Hoffman (eds.), *Advances in Neural Information Processing Systems 19*, 849–856. Cambridge, MA: MIT Press.

Johansson, Stig & Knut Hofland. 1989. *Frequency analyses of English vocabulary and grammar*. Oxford, UK: Clarendon Press.

Kermes, Hannah & Elke Teich. 2017. Average surprisal of parts of speech. In *Proceedings of Corpus Linguistics*. Birmingham, UK.

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

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

Kravtchenko, Ekaterina & Vera Demberg. 2015. Semantically underinformative utterances trigger pragmatic inferences. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Austin, TX, USA.

Kullback, Solomon & Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1). 79–86.

Leech, Geoffrey, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in contemporary English: A grammatical study*. Cambridge, UK: Cambridge University Press.

Levy, Roger P. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177.

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

Michel, Jean-Baptiste, 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 & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014). 176–182.

Moskowich, Isabel & Begona Crespo (eds.). 2012. *Astronomy Playne and simple: The writing of science between 1700 and 1900*. Amsterdam/Philadelphia: John Benjamins.

Nguyen, Dat Quoc & Karin Verspoor. 2019. From pos tagging to dependency parsing for biomedical event extraction. *BMC Bioinformatics* 20(1). 72.

Pate, John & Sharon Goldwater. 2015. Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language* (78). 1–17.

Pellegrino, François, Christophe Coupe & Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language* 87(3). 539–558.

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

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rissanen, Matti, Merja Kytö & Kirsi Heikkonen (eds.). 1997. *English in transition: Corpus-based studies in linguistic variation and genre analysis*. Berlin: Mouton de Gruyter.

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Kyoto, Japan.

Schulz, Erika, Yoon Mi Oh, Zofia Malisz, Bistra Andreeva & Bernd Möbius. 2016. Impact of prosodic structure and information density on vowel space size. In *Proceedings of Speech Prosody*, 350–354, Boston, MA, USA.

Sikos, Les, Clayton Greenberg, Heiner Drenhaus & Matthew Crocker. 2017. Information density of encodings: The role of syntactic variation in comprehension. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci 2017)*, 3168–3173.

Teich, Elke, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes & 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.

Tomokiyo, Takashi & Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment – Volume 18*, MWE '03, 33–40, Sapporo, Japan.

van Hulle, Dirk & Mike Kestemont. 2016. Periodizing Samuel Beckett's works: A stylochronometric approach. *Style* 50(2). 172–202.

Zhai, Chengxiang & John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems* 22(2). 179–214.

Bionotes

Stefania Degaetano-Ortlieb is a postdoctoral researcher and lecturer (tenured) at Saarland University in the department of Language Science and Technology. Since 2014, she is a member of the Collaborative Research Center *Information Density and Linguistic Encoding* (CRC1102), investigating, in particular, trends of linguistic densification and communicative efficiency in scientific writing. Degaetano-Ortlieb has a background in Linguistic and Literary Computing and Translatology. In previous projects, she worked on register variation in scientific writing and German-English contrasts in cohesion. She uses text mining and data analytics for research questions from sociolinguistics, register as well as change in language use and language variation. Since 2017, she is a co-organizer of the ACL SIGHUM workshop LaTech-CLfL aimed at a cross-disciplinary exchange between the humanities and computational linguistics community.

Elke Teich is a full professor of English Linguistics and Translation at the Department of Language Science and Technology at Universität des Saarlandes, Saarbrücken, Germany. Since 2014 she has been the head of the Saarbrücken Collaborative Research Center (SFB 1102) *Information Density and Linguistic Encoding* funded by the German Research Foundation (DFG). She is a principal investigator in SFB 1102 as well as in the Cluster of Excellence *Multimodal Computing and Interaction* (MMCI and the German CLARIN (*Common Language Resources and Technology Infrastructure*).

Teich's expertise ranges from descriptive grammar of English and German over (multi-lingual) register analysis (with a special focus on scientific language) to translatology. She has worked in machine translation, automatic text generation, corpus linguistics and digital humanities at a number of academic institutions, including *Fraunhofer-Gesellschaft*, Information Sciences Institute/USC Los Angeles, University of Sydney, Macquarie University and Technical University Darmstadt. Recently, her research focuses on computational approaches to modelling language variation and change. She has published two monographs and over 80 peer-reviewed papers.