CHAPTER 11

Measuring informativity

The rise of compounds as informationally dense structures in 20th-century Scientific English

Stefania Degaetano-Ortlieb
Saarland University

By applying data-driven methods based on information theory, this study adds to previous work on the development of the scientific register by measuring the informativity of alternative phrasal structures shown to be involved in change in language use in 20th-century Scientific English. The analysis based on data-driven periodization shows compounds to be distinctive grammatical structures from the 1920s onwards in *Proceedings A of the Royal Society of London*. Compounds not only increase in frequency, but also show higher informativity than their less dense prepositional counterparts. Results also show that the lower the informativity of particular items, the more alternative, more informationally dense options might be favoured (e.g., *of*-phrases vs. compounds) – striving for communicative efficiency thus being one force shaping the scientific register.

Keywords: diachronic change in scientific English, compounds, information density, informationally dense structures, detection of linguistic change, *Royal Society Corpus*, standardization, specialization

1. Introduction

Endeavours to better understand the temporal dynamics of language use are steadily increasing in various research fields such as computational social sciences and digital humanities, offering insights, for example, into changing social norms or the history of ideas and cultures (Michel et al. 2011; Muralidharan & Hearst 2013; Hamilton et al. 2016; Garg et al. 2018; Degaetano-Ortlieb & Piper 2019). So far, however, most work is focussed on the analysis of changing semantic and lexical properties. The interest of diachronic corpus linguistics goes beyond lexico-semantics, as it is also devoted to analysing grammatical changes (e.g., Kawaguchi et al. 2011; Biber & Gray 2013; Hilpert & Gries 2016; Teich et al. 2016; Gray & Biber 2018).

https://doi.org/10.1075/scl.103.11deg

Taking up a corpus-linguistic perspective, we investigate change in grammatical use in scientific writing – a register shown to be a locus of grammatical change according to Gray and Biber (2018: 117) when compared to other spoken and written registers. In fact, Gray and Biber (2018: 118) have shown that "academic writing has […] been the leader of specific innovations in the use of" particular grammatical features (see also their discussions in Biber & Gray 2011: 223–225, and Biber & Gray 2016). The type of change we are considering here is not language change in terms of changes to the language system but change in language use (cf. Hilpert & Mair 2015). Our work is rooted in the area of usage-based variation, in particular register theory (Quirk et al. 1985; Halliday 1985). Considering previous work on register formation processes, descriptive accounts on the development of a scientific register have been provided by Halliday (1988) and by Halliday and Martin (1993). Biber and colleagues support these findings by corpus-based quantitative results within the framework of multi-dimensional analysis (Biber & Finegan 1989; Biber & Gray 2011). Studies on written scientific language have pointed to a diachronic tendency towards structural compression (Biber & Gray, 2013) accompanied by informationally dense phrasal constructions (e.g., Halliday & Martin 1993; Mair 2006; Rubino et al. 2016; Degaetano-Ortlieb et al. 2019; Degaetano-Ortlieb & Teich 2018, 2019). In our work, we also take up an information-theoretic perspective. As already suggested by Harris (1991), having a distinctive code, e.g., for scientific communication, is beneficial as transmission of information becomes more error-free (Harris 1991: 393ff). Over time, the scientific register is continuously adapted to efficiently communicate scientific knowledge. Mechanisms of specialization and standardization seem to act as balancing forces to modulate the amount of information transmitted (Degaetano-Ortlieb & Teich 2018 and 2019; Bizzoni et al. 2020). Specialization processes match the constant need for increased expressivity, originating from the extra-linguistic context, e.g., in the case of new discoveries. Standardization processes act as a balancing force to specialization, i.e., conventions arise such as terminology formation or formulaic expressions (Degaetano-Ortlieb & Teich 2018 and 2019). A code is being continuously created that is sufficiently conventionalized, while leaving room for innovation – an ongoing process to form an optimal code at any particular point in time.

In this paper, we analyse change in grammatical use in *Proceedings A of the Royal Society of London*, a leading academic periodical covering approximately 100 years (1905–1996). We pursue the following research questions:

1. When do changes in use of grammatical structures occur in the 20th century and which types of changes do we observe? (Section 3)
2. Do these trends reflect structural compression strategies in terms of the use of more informationally dense structures (e.g., when comparing compounds to prepositional post-modification alternatives)? (Section 4)

In our first analysis, we investigate when changes in the use of grammatical structures occur and which structures are involved in change. As an approximation of grammatical structures, we use part of speech (POS) sequences. Results show that POS sequences encompassing compounds predominantly impact change at the beginning of the 20th century. In our second analysis, we further investigate compound sequences as one possible way of packing information in shorter encodings (e.g., in comparison to prepositional alternatives), measuring the amount of information they carry, i.e., their informativity, and comparing it to prepositional alternatives.

Methodologically, for our first research question, we take a text-linguistic exploratory perspective and employ a data-driven periodization technique based on Kullback-Leibler divergence to detect when changes occur (Degaetano-Ortlieb and Teich 2018 and 2019). In line with recent developments, we here address a common challenge in diachronic analysis: determining periods of change rather than using pre-defined periods (see Nevalainen and Traugott 2012: 3). For our second research question, we consider informativity by the measure of surprisal of those POS sequences (here: compounds) contributing to change (selected through data-driven periodization) and indicating possibly informationally dense structures.

The paper is structured as follows. After introducing our corpus and methodology in Section 2, we present two analyses in line with our research questions: We trace diachronic changes at the grammatical level in 20th-century Scientific English (Section 3) and investigate the use of compounds as informationally dense structures shown to be involved in change (Section 4). Section 5 concludes the paper with a summary and outlook.

2. Methods

2.1 Data

For our investigation, we use a subpart of the *Royal Society Corpus* (RSC) (Kermes et al. 2016; Fischer et al. 2020). The RSC is built from the Proceedings and Transactions of the Royal Society of London – the first and longest-running periodical of English scientific articles. The corpus is continuously enhanced/updated and different versions are freely available in accordance with copyrights (cf. Kermes et al. 2016). We use the RSC Version V6.0 (Fischer et al. 2020), which includes *Proceedings A* ranging from 1905 to 1996 and devoted to the mathematical, physical and engineering sciences. Annotations are based on extra-linguistic (authors, titles, dates of publication, journal series, etc.) and linguistic (words as normalized and original forms, lemmas, and parts of speech)[1] information. Based on part-of-speech tagging,

1. Evaluation of the pos tagging was performed on a manually annotated subcorpus (56,000 tokens) achieving 95.1% accuracy (Kermes et al. 2016).

sentence boundaries are annotated via a Perl script. Table 1 gives an overview of the corpus size of *Proceedings A*.

Table 1. Corpus size of *Proceedings A of the Royal Society of London*

Decade	Tokens
1905–09	4,843,426
1910–19	4,563,889
1920–29	7,873,713
1930–39	11,774,985
1940–49	6,282,360
1950–59	14,478,279
1960–69	15,528,318
1970–79	18,988,777
1980–89	19,643,957
1990–96	14,318,263
Total	**188,295,967**

2.2 Data-driven periodization with Kullback-Leibler divergence

Over the past few years, research applying data-driven modelling of language variation and change has increasingly relied on information-theoretic measures. The measure of Kullback-Leibler divergence (KLD) has been effectively applied for measuring differences in probability distributions over linguistic features (Fankhauser et al. 2014). For example, Klingenstein et al. (2014) investigate variation in language use in criminal trials, Bochkarev et al. (2014) apply KLD to compare word distributions across languages, Fankhauser et al. (2014) use KLD for corpus comparison at large, and Barron et al. (2018), whose work is most similar to ours, use KLD to model temporal dynamics of sequential speeches of parliamentary debates on the French Revolution, considering how much speeches diverge over time. In our own work, we have applied data-driven periodization based on KLD to trace the linguistic development of Late Modern Scientific English (Degaetano-Ortlieb & Teich 2018 and 2019). In this paper, we apply KLD to model temporal dynamics in Scientific English of the 20th century.

Data-driven periodization based on KLD allows us to detect when periods of change occur and which linguistic features are contributing to change. The approach has the following elements: (1) comparison of adjacent years by Kullback-Leibler divergence (KLD), (2) relatively unconstrained feature selection, and (3) inspection of relevant features involved in change, i.e., showing high contribution to the overall divergence.

Basically, with KLD we compare language use of the future with the past considering probability distributions of linguistic features. As we are interested in change

in grammatical use, we select as linguistic features part-of-speech sequences, in particular trigrams (POStrigrams), a choice made after experimenting with several ngram sizes.[2] Comparison of adjacent years by KLD is illustrated in Figure 1. We slide over the timeline, comparing a range of years preceding and following a selected year, with KLD. Peaks in KLD indicate periods of change, where the future diverges from past language use; troughs point to periods of consolidation, where future and past are more similar to each other. This allows us to inspect at which particular point in time changes occur. The procedure is operationalized as follows (also illustrated in Figure 1):

1. Select a year *i* (or range of years, if the publication is not yearly) as a gap and a window size *n* of preceding (PAST: *i-1,…,i-n)* and following (FUTURE: *i+1,…,i+n)* years (e.g., 10 years);
2. Calculate KLD for the future given the past *D(future||past)*;
3. Slide to the next year and repeat (2).

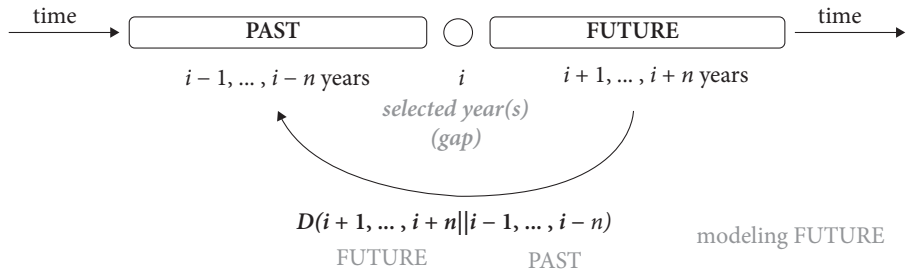


Figure 1. Data-driven periodization with KLD

In Degaetano-Ortlieb & Teich (2018), we experiment with different window sizes: more fine-grained selections help detect more subtle changes (e.g., a window size of 5 years), while coarser selections (e.g., a window size of 20 years) lead one to inspect more general trends. After experimenting with different window sizes for the data at hand, we observe general trends across window size selection. Thus, we opt to show results for a window size of 10 years to depict the general trends in the data.

2. Note that bigrams proved to be too short to depict phrase/clause structure, while four- and fivegrams lead to sparse data. See also work of others considering trigrams for diachronic analyses (Culpeper & Kytö 2010; Kopaczyk 2013). Also note that while parsing would be a much better choice, parsing the large amount of data in the RSC corpus is not a trivial task. Our ongoing parsing experiments have shown low accuracies (around 91%) on the RSC data. Evaluations indicate several problem sources, e.g. correct sentence detection in scientific texts (due to formulas, lists, tables, etc.). In fact, parsing works slightly better for sentences of a particular length (around 50 tokens with 92.6%). Here, we opt to rely on pos sequences as an approximation, but aim to work on parses in the future, when reasonable accuracies are achieved.

Formally, KLD measures the number of additional bits needed to encode a probability distribution based on a POS feature set of the future with a probability distribution of the past (Equation 1). This allows us to investigate whether there is a particular use of grammatical structures (approximated by POStrigrams) in the future that cannot be captured by a model of the past, indicating change in use of grammatical structures.

$$(1) \quad D\,(future \,||\, past) = \sum_i p\,(POStrigram_i|future) \log_2 \frac{_p(\text{POS}trigram_i|future)}{_p(\text{POS}trigram_i|past)}$$

The probability of the *i*th POStrigram in the future dataset, $p(POStrigram_i \mid future)$, and the *i*th feature's probability in the past dataset, $p(POStrigram_i \mid past)$, are used to measure the amount of additional bits needed. The sum over all POStrigrams gives an overall divergence measure, namely KLD $D(future||past)$. The higher the KLD for the future given the past, the more the two datasets diverge (cf. Degaetano-Ortlieb & Teich 2019). As feature size often differs across datasets, i.e., one has to deal with sparse data leading to imprecise probability estimations, Jelinek-Mercer smoothing is used to obtain more accurate models (cf. Zhai & Lafferty 2004; Fankhauser et al. 2014).

Beyond these overall trends, we investigate individual contributions of POStrigrams, obtaining more profound insights into the kinds of change in grammatical use. Whenever we observe a peak in the overall KLD at a particular point in time, we inspect the feature ranking of that particular year. As we are dealing here with multiple bi-class comparisons, i.e., at each year one comparison of the past and future 10 years in a corpus covering approximately 100 years, one has to carefully choose how to inspect all feature rankings in a meaningful way: e.g., (a) rank KLD contribution of individual features by year or (b) consider which features show high variation in their contribution (e.g., words with high contribution to KLD only at particular points in time) using standard deviation calculated across the feature rankings of each year.

2.3 Determining informativity: Surprisal

To address our second research question (informationally dense structures) and assuming a rational communication intent, we apply the notion of surprisal, measuring informativity by considering probabilities of words given their preceding context (see Equation 2).

$$(2) \quad S(word) = -log_2 p(word|context)$$

Surprisal is mainly applied in psycholinguistic studies for on-line comprehension tasks. Levy and Jaeger (2007) have shown how surprisal is linked to variation by analysing the ways in which speakers use variation to optimize the amount of

information of an utterance (e.g., comparing full vs. reduced relative clauses). Also, there is evidence from comprehension studies on how particular linguistic choices are associated with specific levels of surprisal (Levy 2008; Schulz et al. 2016; Delogu et al. 2017; Sikos et al. 2017). We apply surprisal to measure the amount of information a word carries given a particular context (Equation 2) or across all contexts in a corpus, i.e., the average surprisal (Equation 3) (cf. Degaetano-Ortlieb and Teich 2019). Similar to comprehension studies, we are interested in how particular words settle in certain ranges of high to low surprisal.

$$(3) \quad AvS(word) = \frac{1}{|word|} \sum_{i=1}^{n} - \log_2 p(word_i | context_i)$$

To consider less or more contextual information, we can narrow down or widen the context. We opt for a fourgram model, which allows us to consider local lexical and some grammatical information (cf. Degaetano-Ortlieb & Teich 2019), i.e., surprisal of a *word* with a preceding context of three words $word_{(i-1)}$, $word_{(i-2)}$, $word_{(i-3)}$ (Equation 4).

$$(4) \quad AvS(word) = \frac{1}{|word|} \sum_{i=1}^{n} - \log_2 p(word_i | word_{i-1} word_{i-2} word_{i-3})$$

Considering academic writing and its diachronic development, we apply surprisal to investigate how particular compounds, shown to be relevant features of variation based on data-driven periodization, change in terms of their informativity over time. In addition, we also investigate how their informativity differs between compounds and alternative encodings. High surprisal indicates less predictable words given their previous context and higher informativity, while low surprisal indicates more predictable words given their previous context and lower informativity. Within scientific writing, we assume a development towards the use of particular structures with higher surprisal.

3. Tracing change in grammatical use in 20th-century Scientific English

In our first analysis, we investigate whether there are changes in use of grammatical structures, applying data-driven periodization. We ask the following questions:

a. Do we observe changes in use of grammatical structures over time for Scientific English in the 20th century? (Section 3.1)
b. If changes occur, what are the kinds of change in grammatical use at particular points in time? (Section 3.2)
c. Do these changes point to the increased use of more informationally dense structures? (Section 3.3)

3.1 The temporal dynamics of grammatical use in 20th-century Scientific English

With data-driven periodization, we model the course of change in grammatical use for *Proceedings A of the Royal Society of London* from 1905 to 1996 using part-of-speech trigrams (POStrigrams) as an approximation of grammatical structures. We consider how well a grammatical model of the future is captured by a grammatical model of the past (see Section 2.2). Peaks indicate change, as the models diverge from each other, while troughs indicate periods of consolidation, where grammatical use of the future and past is relatively similar. Figure 2 shows overall KLD. A major peak occurs around the 1920s. The curve starts rising in 1916 with a steady increase until 1922. Thus, 1922 marks a period of change, where the 10 years after 1922 (1923–1932) diverge the most from the 10 years preceding 1922 (1912–1921). Considering KLD to measure additional bits of information needed to encode the future with the past, instead of around 0.04 bits in 1915, around 0.08 bits are needed by the 1920s, i.e., an increase of 50%, a relatively high increase. The decrease in divergence afterwards points to a converging tendency, where grammatical use becomes increasingly more similar between the future and past. By the 1930s, there is relatively strong convergence.

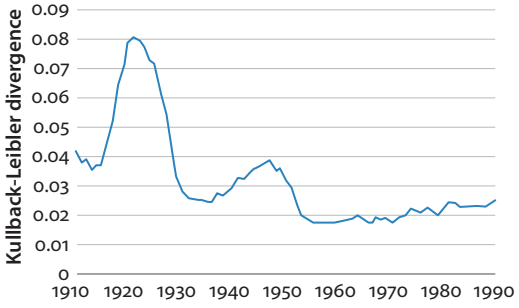


Figure 2. Overall KLD from data-driven periodization

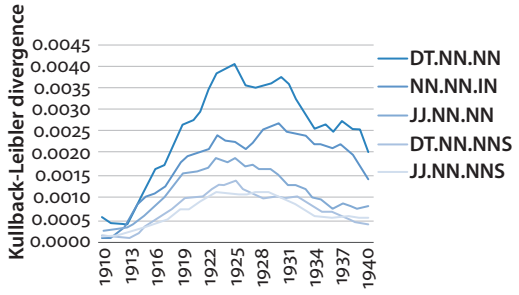


Figure 3. KLD for the five top-ranking POStrigrams

3.2 Kinds of change in grammatical use: Inspecting distinctive patterns

We consider the highest peak in overall KLD as a point in time worthy of inspection, looking at POStrigrams with the highest contribution to overall KLD in the 1920s. This allows us to detect which POStrigrams are distinctive for the future, as they are less well captured by a model of the past. Figure 3 shows the five top-ranking features. Considering the types of grammatical structures indicated by POStrigrams, all include compounds (sequences of at least two nouns). Looking at the most frequent lexical realizations of the top three patterns, they are all terminological patterns (see in Table 2 e.g. *the discharge tube* with 27.18 FpM for the DT.NN.NN pattern, *temperature coefficient of* with 15.49 FpM for the NN.NN.IN pattern), Considering their frequency distributions, we see a rise from the 1920s onwards (Figure 4).

Table 2. Lexical realizations of the top three distinctive patterns (showing three lexical realizations each in bold), 1920s

Trigram	Freq.	FpM	Example
DT.NN.NN	214	27.18	*A Gaede rotary mercury pump was attached to the exit of* ***the discharge tube***
	204	25.91	*if the arrangement in* ***the unit cell*** *is like that for naphthalene and anthracene*
	162	20.57	*the* ***surface tension of*** *water is greater than the sum of the tensions*
NN.NN.IN	122	15.49	*The* ***temperature coefficient of*** *nickel sulphide is thus about 1.*
	122	15.49	*We may write as a* ***power series of*** *the arguments*
	122	15.49	*It was intended for examining* ***absorption spectra of*** *small objects,*
JJ.NN.NN	42	5.33	*the angle between the* ***perfect cleavage plane*** *and the axis of the rod varied*
	40	5.08	*the added electron occupies an orbit of higher* ***total quantum number***
	40	5.08	*we shall use to denote the* ***principal quantum number*** *of the shell*

DT: determiner, NN: sg. noun, JJ: adjective, IN: preposition

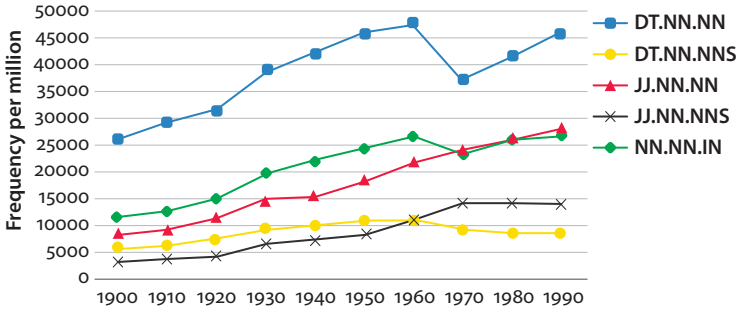


Figure 4. Frequency distribution for the five top-ranking POStrigrams

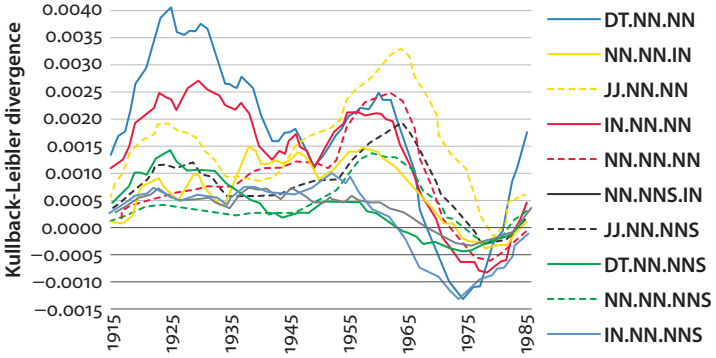


Figure 5. Top 10 POStrigrams by KLD with NN.NN sequences

As compounds are shown to shape change in grammatical use around the 1920s and beyond, we select all compound patterns among the POStrigrams contributing to KLD by searching for patterns of noun sequences with at least two nouns. We then rank them by standard deviation to observe the temporal dynamics of those trigrams showing the most prominent changes. Figure 5 presents the top 10 ranking compound POStrigrams. The types of compounds covered by these top 10 POStrigrams are definite/indefinite compounds (DT.NN.NN(S)), e.g., *the discharge tube*), post-modification with prepositional phrases (NN.NN(S).IN, IN.NN.NN(S), e.g., *power series in*), adjectival pre-modification (JJ.NN.NN(S), e.g., *reciprocal lattice vector*), and compounds with three nouns (NN.NN.NN(S), e.g., *electron spin resonance*). Looking at their contribution to KLD over time, we observe a trend delineating changes related to structural compression strategies: most distinctive around the mid-1920s are definite/indefinite compounds and prepositional post-modification, while towards the 1960s adjectival pre-modification and three-noun compounds are shown to make the highest contribution (JJ.NN.NN and NN.NN.NN).

3.3 Tracing changes towards the use of informationally dense structures

To obtain a better insight of changes towards the use of more informationally dense structures, we confine our feature set to noun phrases with modification variants from least to most dense structures (cf. Biber & Gray 2016: 207) and re-run data-driven periodization. The following noun phrase modification variants are selected: noun phrases modified by finite and non-finite relative clauses, noun phrases post-modified by prepositional phrases, and noun phrases pre-modified by adjectives and nouns (cf. also Biber & Gray 2016: 207). As post-modification features, we extract a noun followed by either a relativizer (considering restrictive and non-restrictive relative clauses as well as object and subject relative clauses) or a preposition. As pre-modification features, we extract noun sequences with zero to five pre-nominal modifiers. These sequences are not overlapping in terms of nouns (i.e., noun-noun is neither preceded nor followed by a noun), but might encompass adjectival pre-modification. Using these features, we re-run data-driven periodization. From Figure 6, we again see a high peak in the 1920s (compare to Figure 2) followed by converging and diverging trends afterwards and an overall decreasing tendency in divergence.

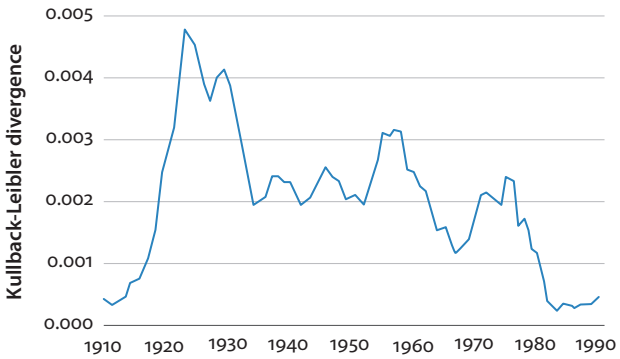


Figure 6. Overall KLD from data-driven periodization based on selected noun phrases with least to most compact modification variants (cf. Biber & Gray 2016: 207)

Considering each feature's contribution to divergence, Figure 7 shows a high impact of two-noun compounds (NN.NN) starting in the 1920s, while single noun phrases (NN) are not distinctive. At the same time, the contribution of three-noun compounds steadily rises with a peak from the mid-1950s to the 1980s. Longer compound sequences have a much lower contribution (only four-noun compounds are shown in Figure 7). Considering post-modification, we clearly see how prepositional post-modification (NN.IN) becomes less distinctive for the future over time, being more distinctive for the past. Considering finite and non-finite relative

clauses, they have very low contribution in comparison to noun compound sequences. Comparing both types of relative clauses, the more compact non-finite variant (N.nfrel) shows a slightly higher contribution from the 1950s to the 1980s in comparison to the less compact finite variant (NN.rel).

In summary, nominal pre-modification with at least two nouns is distinctive of future grammatical use, while prepositional post-modification is distinctive for past grammatical use. This seems to indicate a replacement by more dense vs. less dense structures.

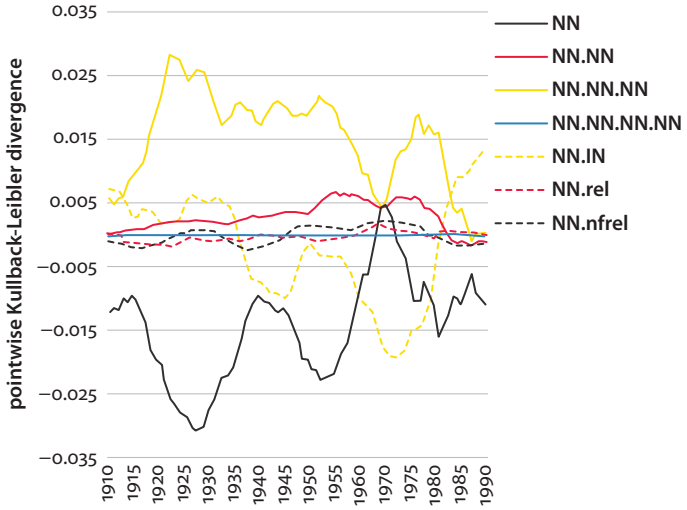
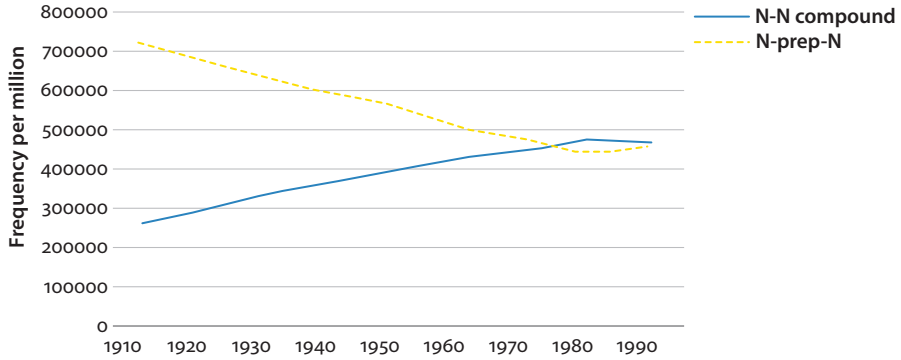


Figure 7. Pointwise KLD for noun phrases with least to most compact modification variants

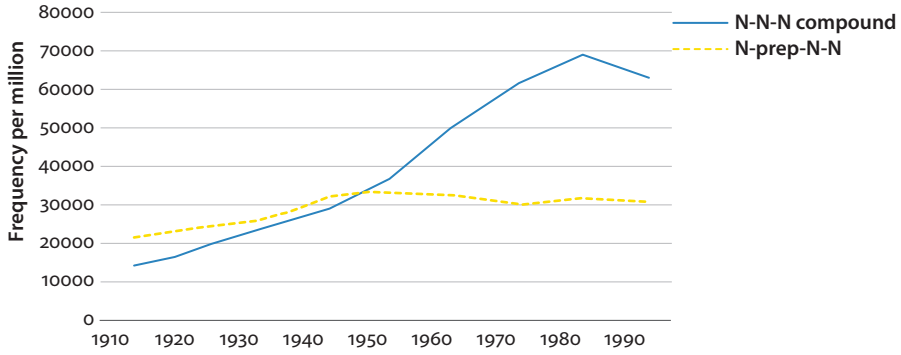
4. Tracing the development of informationally dense structures

In this second analysis, we focus on compounds as pre-modification variants and prepositional post-modification variants shown to be involved in change in the RSC, measuring their informativity with surprisal to observe whether there is a tendency towards the use of more informationally dense grammatical structures over time. We select compounds of bi- to trigram sequences of nouns (NN.NN and NN.NN.NN),[3] which have been shown to make a strong contribution to an increase in divergence (see Section 3.3), and their prepositional counterparts (NN.IN.NN and NN.IN.NN.NN).

3. These noun sequences are not overlapping, e.g., a NN.NN sequence is neither followed nor preceded by a noun. Note also that NN here denotes singular as well as plural nouns.



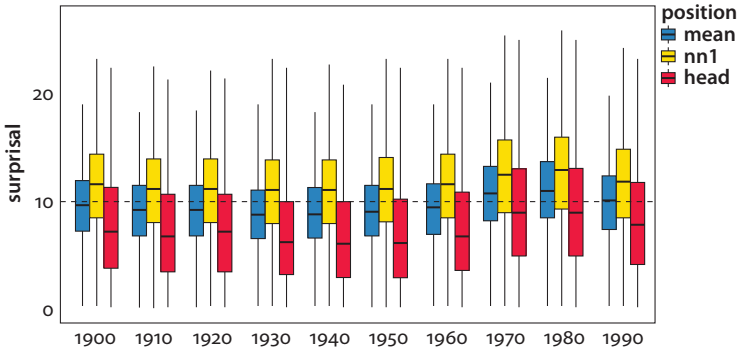
a. N-N compound vs. N-prep-N



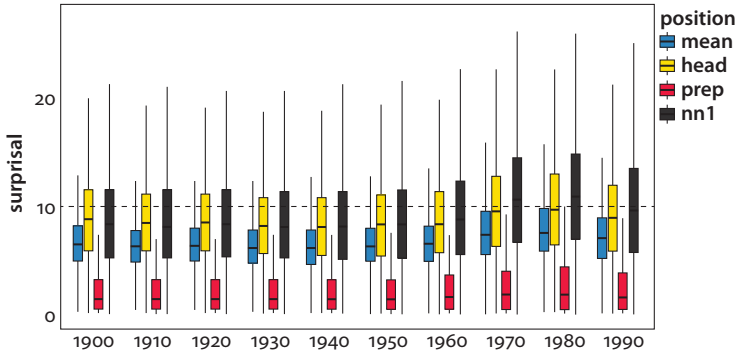
b. N-N-N compound vs. N-prep-N-N

Figure 8. Frequency distribution of compounds and prepositional pre-modified alternatives across decades

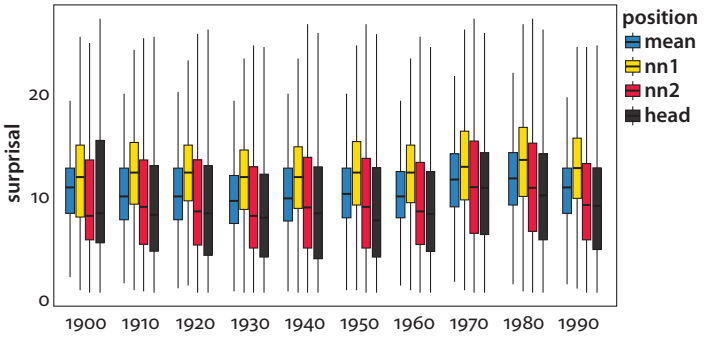
First, we compare their frequency distributions. Diachronically, we can observe from Figure 8 that the prepositional post-modification variants either decrease or remain relatively stable, while the compounds clearly increase. In a second step, we use surprisal to calculate the informativity of both the compound and prepositional variants. Surprisal is calculated on words (cf. Section 2.3, i.e., the predictability of each word given its previous three words) and averaged for each variant across decades. High surprisal equals high informativity, while low surprisal equals low informativity. Figure 9 shows surprisal for the compound and prepositional variants. Considering mean surprisal of each variant (white box plots), while surprisal for the compound variants is around 10 bits, the prepositional variants clearly show a lower median of around 7/8 bits – the compound variants having higher informativity. Thus, the compounds rise in frequency over prepositional variants and carry higher informativity, which is in line with our assumption of an increase in use of more informationally dense variants over time.



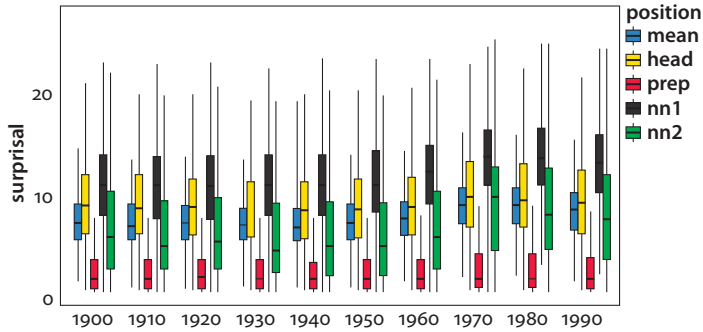
a. N-N compound



b. N-prep-N



c. N-N-N compound



d. N-prep-N-N

Figure 9. Surprisal across decades

Comparing each element of the variants, the syntactic head, especially for the N-N compound variant, has relatively low surprisal compared to the pre-modifying noun. Similarly, for the N-N-N compound variant, the last two nouns have lower surprisal than the first noun. Thus, the most informative part is the first noun (above 10 bits), the following nouns being more predictable and less informative (below 10 bits). Table 3 shows the most frequent three lexical realizations of each pattern for 1910 and 1990 with surprisal (averaged across the occurrences of each variant). The syntactic heads seem to be relatively general words (e.g., *ray, tube, condition*), while the modifying nouns seem to be more specific (e.g., *cathode, ionisation, mercury*). This tendency is maintained over time (compare 1910 to 1990). Surprisal, in fact, is lower for heads and higher for the more specific modifying nouns. This is a tendency also reflected by mid- and low-frequency items. Exceptions include very specific syntactic heads (e.g., as in *cathode*_8.95 *phosphorescence*_14.96 occurring twice in 1910).

For both prepositional variants, the preposition has relatively low surprisal (see Figure 9b and 9d). The noun preceding the preposition has moderate surprisal (below 10 bits). For the N-prep-N pattern, the noun following the preposition is almost equal in surprisal to the preceding noun, slightly increasing in surprisal over time. For the N- prep-N-N pattern, where a compound follows the preposition, we clearly see the compound behaviour, the syntactic head of the compound being lowest in surprisal, while the pre-modifying noun is highest in surprisal. The top-frequency items (see again Table 3) of prepositional variants seem to show a different trend than compounds, as the syntactic head has more informativity than the following nouns. However, the informativity of the head is, for almost all items, relatively low (< 10 bits).[4] Also, considering mid- and low-frequency items, the syntactic head is on average lower in surprisal than the noun(s) following the preposition (8.9 vs.

4. Note that *order of coincidence* has relatively high informativity; inspection has shown that this is a term used in one paper only, authored by J. C. Fields and A. R. Forsyth.

9.8 bits for occurrences < = 50, 9.1 vs. 10.1 bits for occurrences = 1; e.g., *weight*_5.5 *of*_1.6 *iodine*_10.5). Thus, frequent occurrences become established over time (e.g., *point of view*), where the sequence of lexical items is relatively predictable (the last noun having lower surprisal than the head), while less frequent occurrences follow the trend of having a more predictable head (lower surprisal, less informative) and less predictable modifier (higher surprisal, more informative).

Table 3. Most frequent sequences for compound and prepositional variants for 1910 and 1990 with surprisal

	Dec	Freq	Word	Srp	Word	Srp	Word	Srp	Word	Srp
N-N	1910	200	cathode	11.8	ray	3.3				
		150	discharge	9.4	tube	2.6				
		148	ionisation	11.1	chamber	2.0				
	1990	1620	boundary	8.4	condition	2.8				
		778	boundary	7.6	layer	1.7				
		427	time	8.0	series	3.9				
N-N-N	1910	14	zinc	11.1	sulphide	3.1	screen	0.9		
		12	carbon	8.2	disulphide	4.9	vapour	4.5		
		11	mercury	9.8	arc	4.9	spectrum	4.0		
	1990	108	stress	8.6	intensity	4.0	factor	0.7		
		103	boundary	8.3	value	4.5	problem	0.5		
		65	probability	9.6	density	3.6	function	1.3		
N-prep-N	1910	154	point	5.2	of	1.1	view	1.2		
		141	series	6.8	of	0.8	experiments	2.7		
		127	order	13.2	of	1.5	coincidence	9.4		
	1990	314	equation	7.7	of	2.7	motion	2.5		
		268	degree	9.8	of	0.6	freedom	2.5		
		252	order	6.4	of	0.9	magnitude	1.1		
N-prep-N-N	1910	16	form	4.0	of	0.4	carbon	0.7	monosulphide	3.7
		12	solution	2.7	of	0.9	didymium	15.2	nitrate	5.8
		16	hydrolysis	11.4	of	0.7	cane	5.2	sugar	0.2
	1990	26	bifurcations	5.2	of	0.05	vector	0.7	fields	0.1
		22	force	7.3	per	5.3	unit	0.2	length	1.8
		19	range	5.1	of	0.6	length	10.2	scales	1.3

Thinking of formulaic expressions, such as *point of view*, these are captured well by surprisal, showing low informativity as the tokens of the expression are quite predictable (*view* having only around 1 bit of information; see Table 3). Similarly, very well-established terminology will show the same trend (see in 1990 instances of N-N-N, e.g., where the syntactic head also has surprisal < 1 bit, such as *factor* and *problem*, or for the N-prep-N-N variant *bifurcations of vector fields* with very low surprisal of *fields*).

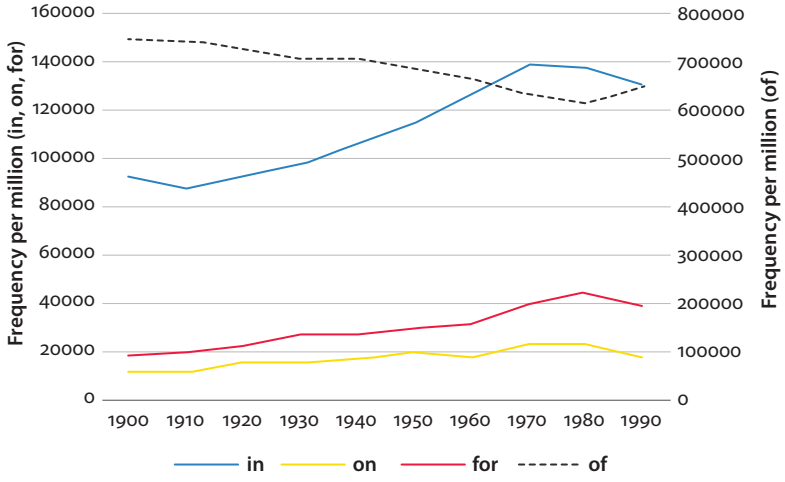


Figure 10. Frequency distribution for *of*(right scale) vs. *in, on* and *for* (left scale) within N-prep-N and N-prep-N-N

Considering the meaning relations expressed by both variants, for compounds the relations are not explicit, while for prepositional post-modification variants the prepositions mark the relation more explicitly. In fact, the overall informativity of the prepositional variants is extremely reduced, in particular due to the low surprisal of the preposition. Comparing different prepositions, Biber and Gray (2016: 151) have shown that while the preposition *of* as a noun modifier decreases over time, the prepositions *in, on* and *for* have been steadily increasing. This tendency is also observable in our data (see Figure 10): while the preposition *of* within N-prep-N and N-prep-N-N decreases, the others increase in use within these patterns. Looking at the prepositions' surprisal (see Figure 11), *of* has the lowest informativity with around 1 bit, while the other prepositions are higher in surprisal with around 4 to 5 bits.

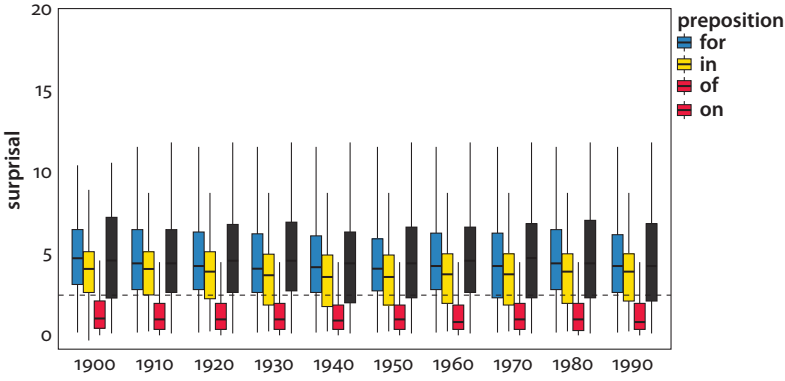


Figure 11. Surprisal of selected prepositions across decades within N-prep-N and N-prep-N-N

Considering informativity, the highly frequent preposition *of*, having very low surprisal, carries very little information, which makes it a good candidate for omission and replacement by nominal pre-modification. In fact, Biber and Gray (2016: 212) have shown that *of*-genitives, as one *of*-phrase type, modifying a particular set of nouns decrease, while nouns modifying the same set of nouns increase. The other prepositions are higher in informativity and in fact rise in frequency within our data as well (see Biber & Gray's 2016: 191) discussion on the expansion of function and meaning of prepositions other than *of*, which increase in frequency).

In summary, compounds are more informationally dense than their longer prepositional counterparts, and also rise in frequency. This confirms a tendency towards an increased use of informationally dense structures when considering these variants. Thus, when information becomes very predictable (very low surprisal), more compact alternatives seem to be the favoured choice, as is the case for prepositional variants, especially with the very predictable *of*, being replaced by compounds.

5. Conclusion

Our aim in this paper was to analyse the temporal dynamics in use of grammatical structures within 20th-century Scientific English. Taking up an information-theoretic perspective, we employ measures from information theory to observe whether changes in grammatical structures shape language use in Scientific English towards more informationally dense productions. We started by investigating common diachronic tendencies within *Proceedings A* of the RSC corpus from 1905 to 1996. A common development is the distinctive use of compounds around the 1920s. Methodologically, we observed these changes by data-driven periodization with Kullback-Leibler divergence, allowing us to detect when changes occur, as well as which grammatical structures are involved. We inspect the observed grammatical changes further, considering compounds vs. longer, less dense variants (relative clauses, prepositional phrases). Results have shown that compounds consisting of two to three nouns become quite distinctive of future grammatical use, while prepositional phrases strongly decrease in distinctiveness for the future, thus becoming distinctive for past grammatical use. Measuring the informativity of prepositional and compound variants with surprisal, we have shown a trend towards the increased use of compounds as more informationally dense variants. With our work, we hope to have shown how an information-theoretic perspective, which allows us to measure the informativity of linguistic devices, adds to findings in previous work showing change in grammatical use towards structural compression in scientific

writing. Compounds are not only increasingly used in comparison to prepositional variants, but also carry higher informativity.

Considering the balancing mechanisms of specialization and standardization processes in scientific writing, we have shown that higher informativity within nominal phrases indicates specialization, e.g., with the use of specialized terms which are high in surprisal. Lower informativity of nominal phrases, instead, indicates standardization reflected in formulaic expressions and well-established terminology. From a communicative perspective, items with low informativity are easy to process, while for items with higher informativity, more processing effort is needed. The lower the informativity of items, the more alternative, more compact options might be favoured (as seen for *of*-phrases vs. compounds), while items with higher informativity are kept (e.g., prepositional phrases with *in, of*, and *for*).

These findings, however, also elicit many interesting open questions, which might be pursued in future work. For example, do experts process discipline-specific language with less effort than non-experts (see work on inexperienced students exposed to highly informational productions in McNamara et al. 1996; McNamara 2001)? How do experts converge on more informationally dense alternatives (see, e.g., work on converging mechanisms in Hawkins et al. 2020)? Combining corpus-based and experimental evidence is one way which will allow us to obtain a more comprehensive picture of the underlying mechanisms of change in language use considering a communicative perspective.

References

Barron, Alexander T. J., Huang, Jenny, Spang, Rebecca L. & DeDeo, Simon. 2018. Individuals, institutions, and innovation in the debates of the French Revolution. *Proceedings of the National Academy of Sciences* 115(18): 4607–4612. https://doi.org/10.1073/pnas.1717729115

Biber, Douglas & Finegan, Edward. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65(3): 487–517. https://doi.org/10.2307/415220

Biber, Douglas & Gray, Bethany. 2011. The historical shift of scientific academic prose in English towards less explicit styles of expression: Writing without verbs. In *Researching Specialized Languages* [Studies in Corpus Linguistics 47], Vijay Bhatia, Purificación Sa´nchez & Pascual Pe´rez-Paredes (eds), 11–24. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.47.04bib

Biber, Douglas & Gray, Bethany. 2013. Nominalizing the verb phrase in academic science writing. In *The Verb Phrase in English: Investigating Recent Language Change with Corpora*, Bas Aarts, Joanne Close, Geoffrey Leech & Sean Wallis (eds), 99–132. Cambridge: CUP. https://doi.org/10.1017/CBO9781139060998.006

Biber, Douglas & Gray, Bethany. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511920776

Bizzoni, Yuri, Degaetano-Ortlieb, Stefania, Fankhauser, Peter & Teich, Elke. 2020. Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence, section Language and Computation*. https://doi.org/10.3389/frai.2020.00073

Bochkarev, Vladimir, Solovyev, Valery D. & Wichmann, Soren. 2014. Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface* 11(101): 1–8. https://doi.org/10.1098/rsif.2014.0841

Culpeper, Jonathan & Kytö, Merja. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: CUP.

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

Degaetano-Ortlieb, Stefania & Piper, Andrew. 2019. The scientization of literary study. In *Proceedings of the 3nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at NAACL*, 18–28, Minneapolis, MN, June. East Stroudsburg PA: ACL. https://doi.org/10.18653/v1/W19-2503

Degaetano-Ortlieb, Stefania & Teich, Elke. 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 COLING*, 22–33, Santa Fe, NM, September. East Stroudsburg PA: ACL.

Degaetano-Ortlieb, Stefania & Teich, Elke. 2019. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic*, 1–33. https://doi.org/10.1515/cllt-2018-0088

Delogu, Francesca, Crocker, Matthew & Drenhaus, Heiner. 2017. Teasing apart coercion and surprisal: Evidence from ERPs and eye-movements. *Cognition* 116: 49–59. https://doi.org/10.1016/j.cognition.2016.12.017

Fankhauser, Peter, Knappen, Jörg & Teich, Elke. 2014. Exploring and visualizing variation in language resources. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, 4125–4128, Reykjavik, Iceland, May.

Fischer, Stefan, Knappen, Jörg, Menzel, Katrin & Teich, Elke. 2020. The Royal Society Corpus 6.0. Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC)*, 794–802, Marseille, France, May.

Garg, Nikhil, Schiebinger, Londa, Jurafsky, Dan & Zou, James. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16): 3635–3644. https://doi.org/10.1073/pnas.1720347115

Gray, Bethany & Biber, Douglas. 2018. Academic writing as a locus of grammatical change: The development of phrasal complexity features. In *Diachronic Corpora, Genre, and Language Change* [Studies in Corpus Linguistics 85], Richard J. Whitt (ed.), 117–146. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.85.06gra

Halliday, Michael A. K. 1985. *Written and Spoken Language*. Melbourne: Deakin University Press.

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

Halliday, Michael A. K. & Martin, James R. 1993. *Writing Science: Literacy and Discursive Power*. London: Falmer Press.

Hamilton, William L., Leskovec, Jure & Jurafsky, Dan. 2016. Cultural shift or linguistic drift? Comparing two computational models of semantic change. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, 2116–2121, Austin, Texas, November.

Harris, Zellig. 1991. *A Theory of Language and Information. A Mathematical Approach*. Oxford: Clarendon Press.

Hawkins, Robert D., Goodman, Noah D., Goldberg, Adele E. & Griffiths, Thomas L. 2020. Generalizing meanings from partners to populations: Hierarchical inference supports convention formation on networks. In *Proceedings of the 42nd Virtual Annual Conference of the Cognitive Science Society*.

Hilpert, Martin & Gries, Stefan T. 2016. Quantitative approaches to diachronic corpus linguistics. In *The Cambridge Handbook of English Historical Linguistics*, Merja Kytö & Päivi Pahta (eds), 36–53. Cambridge: CUP. https://doi.org/10.1017/CBO9781139600231.003

Hilpert, Martin & Mair, Christian. 2015. Grammatical change. In *The Cambridge Handbook of Corpus Linguistics*, Douglas Biber & Randi Reppen (eds), 180–200. Cambridge: CUP. https://doi.org/10.1017/CBO9781139764377.011

Kawaguchi, Yuji, Minegishi, Makoto & Viereck, Wolfgang. 2011. *Corpus-based Analysis and Diachronic Linguistics* [Tokyo University of Foreign Studies 3]. Amsterdam: John Benjamins. https://doi.org/10.1075/tufs.3

Kermes, Hannah, Degaetano-Ortlieb, Stefania, Khamis, Ashraf, Knappen, Jörg & Teich, Elke. 2016. The Royal Society Corpus: From uncharted data to corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, 1928–1931, Portorož, Slovenia, May.

Klingenstein, Sara, Hitchcock, Tim & DeDeo, Simon. 2014. The civilizing process in London's Old Bailey. *Proceedings of the National Academy of Sciences* 111(26): 9419–9424. https://doi.org/10.1073/pnas.1405984111

Kopaczyk, Joanna. 2013. *The Legal Language of Scottish Burghs: Standardization and Lexical Bundles*. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780199945153.001.0001

Levy, Roger P. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3): 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

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

Mair, Christian. 2006. *Twentieth-century English: History, Variation and Standardization*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511486951

McNamara, Danielle S. 2001. Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology* 55(1): 51–62. https://doi.org/10.1037/h0087352

McNamara, Danielle S., Kintsch, Eileen, Butler Songer, Nancy & Kintsch, Walter. 1996. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction* 14(1): 1–43. https://doi.org/10.1207/s1532690xci1401_1

Michel, Jean-Baptiste, Shen, Yuan Kui, Presser Aiden, Aviva, Veres, Adrian, Gray, Matthew K., Pickett, Joseph P., Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A. & Lieberman Aiden, Erez. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014): 176–182. https://doi.org/10.1126/science.1199644

Muralidharan, Aditi & Hearst, Marti A. 2013. Supporting exploratory text analysis in literature study. *Literary and Linguistic Computing* 28(2): 283–295. https://doi.org/10.1093/llc/fqs044

Nevalainen, Terttu & Closs Traugott, Elizabeth. 2012. *The Oxford Handbook of the History of English*. Oxford: OUP. https://doi.org/10.1093/oxfordhb/9780199922765.001.0001

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

Rubino, Raphael, Degaetano-Ortlieb, Stefania, Teich, Elke & van Genabith, Josef. 2016. Modeling diachronic change in scientific writing with information density. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, 750–761, Osaka, Japan, December.

Schulz, Erika, Oh, Yoon Mi, Malisz, Zofia, Andreeva, Bistra & Möbius, Bernd. 2016. Impact of prosodic structure and information density on vowel space size. In *Proceedings of Speech Prosody*, 350–354, Boston, MA, USA, May. https://doi.org/10.21437/SpeechProsody.2016-72

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

Teich, Elke, Degaetano-Ortlieb, Stefania, Fankhauser, Peter, Kermes, Hannah & Lapshinova-Koltunski, Ekaterina. 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. https://doi.org/10.1002/asi.23457

Zhai, Chengxiang & Lafferty, John. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems* 22(2): 179–214. https://doi.org/10.1145/984321.984322