Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal

Julius Steuer, Marie-Pauline Krielke, Stefan Fischer
Stefania Degaetano-Ortlieb, Marius Mosbach and Dietrich Klakow
Department of Language Science and Technlogy, Saarland University, Saarbrücken, Germany
{jsteuer,mmosback,dietrich.klakow}@lsv.uni-saarland.de
s.degaetano@mx.uni-saarland.de
{mariepauline.krielke,stefan.fischer}@uni-saarland.de

Abstract

This study presents an analysis of diachronic linguistic changes in English scientific writing, utilizing surprisal from transformer-based language models. Unlike traditional n-gram models, transformer-based models are potentially better at capturing nuanced linguistic changes such as long-range dependencies by considering variable context sizes. However, to create diachronically comparable language models there are several challenges with historical data, notably an exponential increase in no. of texts, tokens per text and vocabulary size over time. We address these by using a shared vocabulary and employing a robust training strategy that includes initial uniform sampling from the corpus and continuing pre-training on specific temporal segments. Our empirical analysis highlights the predictive power of surprisal from transformer-based models, particularly in analyzing complex linguistic structures like relative clauses. The models' broader contextual awareness and the inclusion of dependency length annotations contribute to a more intricate understanding of communicative efficiency. While our focus is on scientific English, our approach can be applied to other low-resource scenarios.

Keywords: Scientific English, Digital Humanities, Language Change, Evaluation, Language Modeling, Transformer

1. Introduction

Language models, particularly those rooted in machine learning and neural networks, have revolutionized the way we analyze and understand the intricacies of linguistic change (Kim et al., 2014; Hamilton et al., 2016; Dubossarsky et al., 2019). Different models, such as n-gram, LSTM or transformer models, offer diverse possibilities for analyzing language variation and change due to their underlying architectures. N-gram models are usually based on rather small and fixed-size context windows, excelling in capturing local patterns of variation. Transformer models, instead, employ attention mechanisms and deep neural networks capturing long-range dependencies and global context in language data. While they are less efficient in terms of training compared to n-gram models, they excel at capturing complex syntactic and semantic relationships, making them well-suited for analyzing possibly broader and more complex linguistic trends.

Various studies, especially concerned with lexical semantic change, already employ transformer-based models successfully (Giulianelli et al., 2020). However, comparability of the models over time is not a trivial task as the data sets often vary greatly in terms of corpus and vocabulary size, especially for historical material where the data cannot be extended.

In this paper, we apply transformer-based models to explore diachronic linguistic change in 300 years of English scientific writing. In particular, we create models of surprisal (the predictability of a word given its previous context, Shannon, 1948), which are comparable over time. Surprisal models allow us to investigate how change in language use is possibly driven by optimization effects, given that surprisal is proportional to cognitive effort (Hale, 2001; Levy, 2008). A major assumption for the evolution of scientific writing is that it becomes more informationally dense over time (Biber and Gray, 2016) due to the increasing specialization and diversification of scientific disciplines. On the other hand, conventionalization effects are at play which modulate the informational load. This balance between highly informative content and conventionalized ways of expression allows for an optimal code for expert-to-expert communication (Degaetano-Ortlieb and Teich, 2019). Our overarching aim is to create robust diachronic language models capable of capturing and quantifying optimization effects in language use. We begin by outlining previous research on changes in English scientific writing, emphasizing the use of information-theoretic notions (specifically surprisal) to capture changes related to efficiency in communication. We continue by elaborating on the challenges in using models with restricted window sizes (n-gram models) and the motivation to apply transformer-based models as well as the challenges associated with

BUCC 2024: The 17th Workshop on Building and Using Comparable Corpora, pages 12–23

the implementation of these models to diachronic data. Next, we describe the dataset used in our study and discuss the methods we adopt for ng-modeling changes over time using transformer-based models. Working with historical linguistic data presents unique challenges, and we address some of these in detail (vocabulary shifts and train set bias). In our analysis section, we assess our transformer-based surprisal models, comparing surprisal trends with those found in earlier studies. We supplement this with a focused study on relative clauses, which require understanding long-range dependencies. These dependencies can be effectively captured by transformer-based models as they consider larger context windows.

The contributions of this study lie in addressing key modeling challenges inherent in historical data analysis, however, our approach has broader applications, extending to other areas where resources are limited.

2. Previous Work and Rationale

Diachronic change in the English scientific register has received ample attention in previous work. Earlier, descriptive (Halliday, 1988; Halliday and Martin, 1993) as well as corpus-based studies (e.g. Biber et al., 1999; Biber and Gray, 2011, 2016) report on a central mechanism in scientific language which shifts grammatical complexity from the clausal level (subordination and coordination) to the phrasal level (see also Hundt et al.) leading to an increasingly nominal instead of verbal style. Another central development in the scientific register is the conventionalization of lexico-grammatical features, which has been detected to be a necessary condition for innovation on the one hand and grammaticalization on the other (**?**Schmid, 1994). Innovation is probably the most obvious mechanism, as a natural reaction to the need to create new vocabulary for newly arising concepts. Furthermore, diversification of certain features in increasingly distinct contexts has been observed to be at play in the course of the creation of new scientific disciplines and the formation of their respective sublanguages (Halliday, 1988; Harris Sabbettai, 1991).

While the mentioned studies are either qualitative in nature or at most frequency-based, more recent studies have employed information-theoretic measures such as n-gram-based surprisal to detect diachronic changes in the register (Degaetano-Ortlieb and Teich, 2016, 2018; Teich et al., 2021). Surprisal is formalized as the negative log probability of a unit in context $Surprisal(unit_i) = -log_2 P(unit_i|Context)$, which results in bits of information (Shannon, 1948). The motivation to abandon a mere frequency-based approach in favour of n-gram-based surprisal is the assumption that linguistic change underlies the rational strive for communicative efficiency. Since surprisal is a widely-used measure of information, which has been shown to be correlated with cognitive effort in online language processing (e.g. Levy, 2008; Demberg and Keller, 2008) it is well suited to giving a communicative explanation for changes in the lexical as well as grammatical level. Degaetano-Ortlieb and Teich (2019), for instance, show that in scientific writing certain grammatical patterns become less surprising over time, i.e. increasingly conventionalized in their contexts, while specific lexical items show a trend toward "innovation and increase in expressivity" (Degaetano-Ortlieb and Teich, 2019, p. 26) indicated by phases of high surprisal when new concepts enter the language and phases of stability/consolidation when items of lexical usage become conventionalized in their contexts. An example of the interplay between lexical innovation and grammatical consolidation is the noun–preposition–noun pattern (e.g. *oxide of iron*) becoming extremely predictable as a grammatical pattern while serving as a "habitual host for lexical innovation and terminology formation" Degaetano-Ortlieb and Teich (2019, p. 26). The mentioned studies give a plausible explanation for the underlying motives of language change, however, their underlying 4-gram language models (i.e. surprisal based on a word's predictability given its previous three words as context) are fairly restricted in terms of context size. While the narrow context of only three preceding words is well suited for detecting optimization of shorter linguistic units such as the above-mentioned noun-preposition-noun pattern, it is less well suited for drawing conclusions about the diachronic development of linguistic structures exceeding this window size. A possible solution to this is to replace the n-gram with a model covering a larger context window such as an LSTM (Hochreiter and Schmidhuber, 1997) or Transformer (Radford et al., 2018), which is the approach we present in the present paper. However, applying such models, especially transformer-based ones, poses several challenges (e.g., varying corpus and vocabulary sizes or selection of data for modeling). In this paper, we work towards addressing some of these challenges (cf. Section 3.2). While there is a growing body of research on which language model architecture to choose if restricted to a limited compute budget (e.g. Scaboro et al., 2021) or a specific dataset size (e.g. Hoffmann et al., 2022), it is less clear what approach one should take if either one's compute budget or dataset size is very small. This becomes especially pressing in the case of historical corpora as there are only limited

ways of extending the data. While it is possible to train large language models on historical English data (e.g. Hosseini et al., 2021), doing so might be undesirable for a number of reasons. When the reason for training a language model is to create a computational model which serves as an approximation – ideally a cognitively plausible one – of a speaker of a specific time period, the option of training the language model on large-scale text data is not available, since the training data should, for example, be restricted to a time period *preceding* any text from that time period, with the basic assumption that a speaker did not have access to future text productions. This is of course a simplified assumption; in the case of written text, it is plausible to assume that *every* reader has access to some amount of data that lie in the future from the perspective of any given text. However, there are domains in which this amount can plausibly be assumed to be small, such as scientific English writing.

3. Data and Methods

3.1. The Royal Society Corpus

We use the Royal Society Corpus (RSC) (Fischer et al., 2020; Kermes et al., 2016) as a data set. The RSC is based on the *Philosophical Transactions* and the *Proceedings* of the Royal Society of London. In total, it comprises 295 895 749 tokens and 47 837 texts, which were published between 1665 and 1996. The RSC incorporates a comprehensive set of metadata such as text categories (e.g., articles, abstracts), authorship, title, publication date, and historical periods (ranging from decades to half-centuries), along with linguistic annotations at multiple layers including tokens (featuring to some extent both normalized and original forms), lemmas, and parts of speech, utilizing TreeTagger (Schmid, 1994), and Universal Dependency parsing achieved by Stanza (Qi et al., 2020) with combined models. Given that the texts underwent OCR, several preliminary procedures were employed to counteract OCR inaccuracies to the greatest extent feasible (for an in-depth explanation, refer to Kermes et al., 2016; Menzel et al., 2021).

Even though the RSC is large enough for language modeling, the distribution of texts and tokens poses a challenge. Figures 1 and 2 show that texts and tokens are not equally balanced across time. This can be attributed to an increase in publication activity as well as significantly longer texts in recent time periods (see Figure 3).

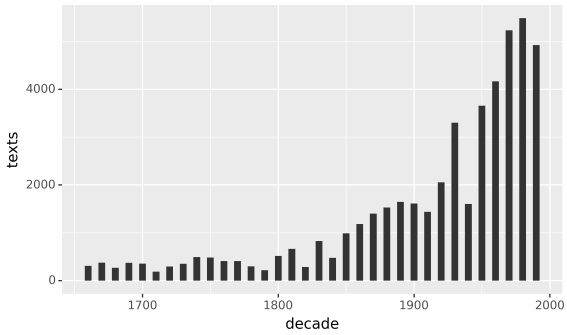

Figure 1: Distribution of texts in the RSC.

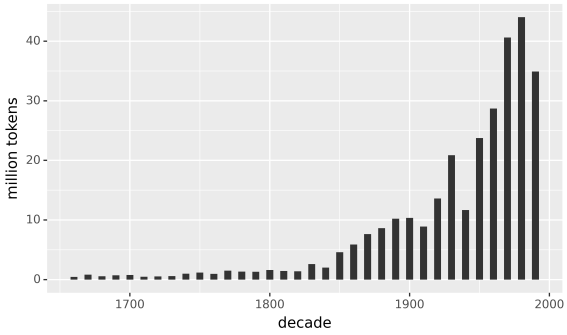

Figure 2: Distribution of tokens in the RSC.

3.2. Modeling Diachronic Change with Transformer-Based Models

Problem Statement Previous research on the diachronic linguistic development of English scientific writing (Degaetano-Ortlieb et al., 2018; Degaetano-Ortlieb and Teich, 2018, 2019) has employed surprisal derived from n-gram language models as a proxy of a model's linguistic knowledge. The hypothesis is that as syntactic structures are conventionalized (i.e. become more predictable) over time, surprisal from n-gram language models fitted to texts from successive time slices of the RSC decreases. While previous work indeed found such an effect (Krielke, 2021), n-gram language models are only a good approximation for local effects and ideally, a more cognitively plausible model should have access to the full sentence-level context.

A possible solution is to replace the n-gram by a large language model with a larger context window (LSTM or Transformer). However, for historical data such as the RSC, this is far from trivial as the number of texts and the number of tokens per year decrease exponentially as we go back in time (see Figure 2). In such a setting, it becomes increasingly difficult to train and compare language models for two reasons:

1. **Vocabulary Shift.** When sampling from peri-

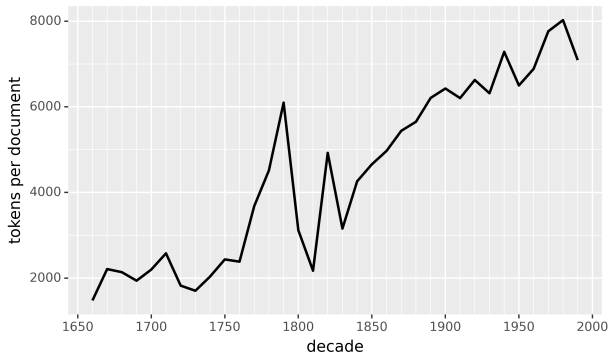

Figure 3: Average number of tokens per document in the RSC.

ods t and t^{+1}, the sets of word types or vocabularies V^t, V^{t+1} will be partially disjoint, i.e. $V_t \cap V^{t+1} \neq \emptyset$. When training language models M^t, M^{t+1} on t and $t+1$, the probability distributions P_{M^t}, $P_{M^{t+1}}$ cannot be directly compared since they are defined over different sets of events.

2. **Train Set Bias.** Let C^t be the set of texts from period t. Since $|C^t| << |C^{t+1}|$, M^{t+1} will see much more training data than M^t when naively sampling from the corpus. The probability estimates derived from M^{t+1} will be tighter than those derived from M^t as a function of the train set size, if the vocabulary stays constnat i.e. for an identical prefix $w_{0...i-1}$ we expect that $P_{M^{t+1}} w_i | w_{0...i-1} \geq P_{M^t} w_i | w_{0...i-1}$.

Approach

Continuous Pre-training While vocabulary shifts can be addressed by sharing a unified vocabulary over all models, train set bias requires sampling the train set such that M^t and M^{t+1} are trained on a similar number of tokens, which is problematic because for earlier periods we may have only very little data. In order to alleviate the effects of trains set bias, we make use of the default NLP pipeline of pretraining a transformer model on a more general dataset D^0_{PT} sampled uniformly from the each time period C^t, and then continue pre-training on the documents of a specific year C^t. In our experiments we use the the smallest version decoder-only OPT architecture (Zhang et al., 2022), with randomly initialized weights.

Pretraining Dataset We sample a pretraining dataset D^0_{PT} from all documents in the corpus such that an equal number of tokens is sampled from the documents C^t. Sampling 10^5 tokens yields $|D^0_{PT}| \approx 2 \times 10^6$. We derive a unified vocabulary by training a BPE tokenizer with $|V| = 5 \times 10^4$ on D_{PT}. We then pre-train on D^0_{PT}, obtaining a set of pre-trained weights θ_{PT}. Surprisal for words that are split into subwords by the tokenizer is calculated by summing their respective log probabilities.

Pre-training on Individual Years We sample datasets D^t_{PT} for each year t in the corpus. Each D^t_{PT} consists of the documents from a period of k years prior to t such that starting from $t_0 = t - k$:

$$D^t_{PT} = \bigcup_{t'=t_0}^{t-1} C^{t'} \quad (1)$$

We then initialize the model with θ_{PT} and fine-tune on D^t_{FT} until validation loss converges. Hyperparameters for continuous pre-training can be found in Table 1. This results in a similar number of training steps (300-400) on each D^t_{PT}, independent of $|D^t_{PT}|$.

Hyperparam	Value
Batch Size	128
Learning Rate	1^{-3}
Warmup	Linear, 10%
Optimizer	AdamW

Table 1: Pre-training hyperparameters

Implementation We used HuggingFace Transformers (Wolf et al., 2020) to train the BPE tokenizer and to pretrain and fine-tune the OPT model. The source code will be made available on GitHub alongside instructions to replicate the result upon publication.

4. Analyzing Linguistic Change

One main assumption regarding the development of scientific English is a balance between informationally dense scientific content and conventionalized scientific style (cf. Degaetano-Ortlieb and Teich, 2019). In fact, it has been shown that scientific English changes from a verbal involved style with embedded clausal structure (see Example (1)) toward a heavy nominal style with long nominal phrases (see Example (2)) and predictable grammatical structures (Biber and Gray, 2016; Degaetano-Ortlieb and Teich, 2019).

(1) *And if the greatest part of these Vessels are Arteries, or other Vessels,* ***that*** *immediately* ***receive*** *liquors from them;* ***I may prove***, ***I think***, *from another Experiment, made by Injection into a part of the Arteria praeparans, before* ***I began***

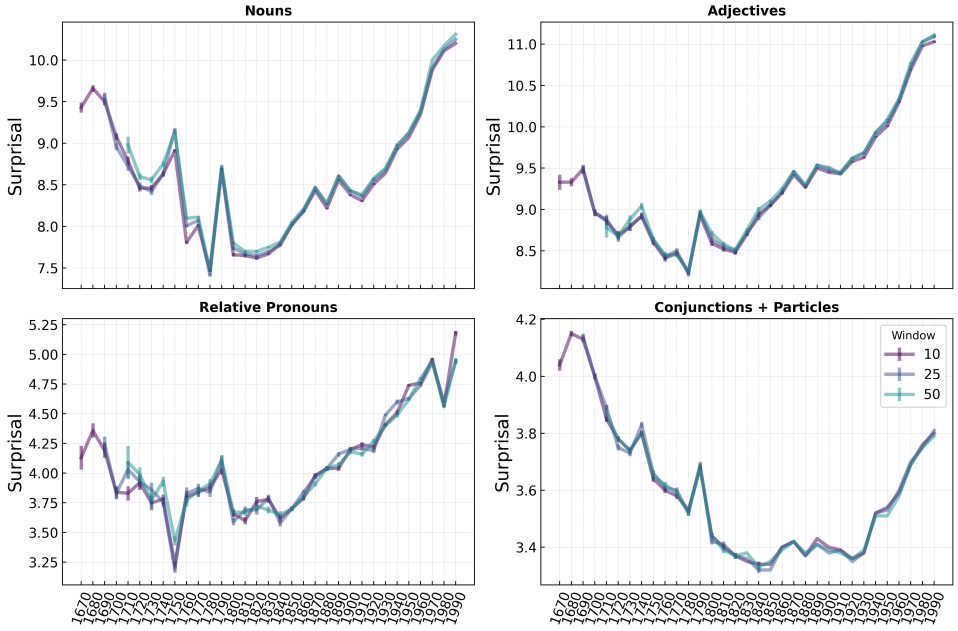

Figure 4: Surprisal on nouns, adjectives, relative pronouns ('which'+'that' with XPOS tag 'WDT') and conjunctions/particles as defined by the UPOS annotations of the RSC. Error bars show standard error.

to expand the Body of the Testis; ***whereupon*** *opening the part,* ***which I saw*** *discoloured,* ***I found****, that many of these Tubes* ***had received*** *some of the fine particles of that matter,* ***which I tinged*** *my injected Spirit with.* (King and de Grieff, 1669)

(2) *On the other hand,* ***a clear red-green stripe pattern of predominantly positive or negative response*** *emerges in* ***the vertical motion signal****.* (Zanker, 1996)

We start by testing whether the results obtained by transformer-based surprisal models are in line with previous findings. We employ the Mann-Kendall trend test (in the Python implementation by Hussain and Mahmud (2019)) to confirm visually salient increases or decreases of surprisal (either on specific words or averaged over POS tags). In particular, we report the direction of the trend (increasing, decreasing, or no trend), its slope s and the p-value p associated with it. Figure 4 shows surprisal of the lexical word classes nouns and adjectives as well as the grammatical function words relative pronouns (e.g., *which*, *that*) and conjunctions/ particles (e.g., *and*, *to*). The surprisal values are calculated as the mean surprisal of all words belonging to a class per decade, determined by the word's POS tag in the CoNLLU. From the 1800s onward we can see an increase in surprisal for word classes associated with nominal style, nouns ($s = 0.0181, p < 0.001$) and adjectives ($s = 0.174, p < 0.001$). Function words, instead, show a steady decrease during the 17th up to the 1840s ($s = -0.01, p < 0.001$) followed by a slight but not significant increase from the 1930s onward ($s = 0.0093, p = 0.2097$). Thus, while nouns and adjectives in general carry more information (higher surprisal, between 7.5 - 12 bits) on average, function words carry much less information (between 3-5 bits). While these findings are in line with the general trend observed in previous work (cf. Kermes and Teich, 2017), we should state that considering average lexical surprisal of words belonging to specific word classes only shows a very aggregated picture as a result of the confluence of different factors, e.g. word frequency, vocabulary diversity, collocational behaviours of words per decade.

4.1. Modeling Convention and Innovation with Surprisal

A more thorough inspection at the lexical level (nouns and adjectives in Figure 4) seems to indicate a wave-like tendency with periods of alternating peaks (e.g., 1680, 1750, 1790, 1890, 1990) and troughs (e.g., 1730, 1780, 1820, 1910) in surprisal. Considering that peaks in surprisal indicate an increase in the use of unpredictable words given their previous context and troughs stand for more predictable usages, the observed changes could be related to discoveries triggering new vocabulary in the corpus at hand, especially at points with abrupt changes such as in the 1790s for nouns. In fact, this peak coincides with the chemical revolution where the 100-year-old phlogiston theory was replaced by the ev-

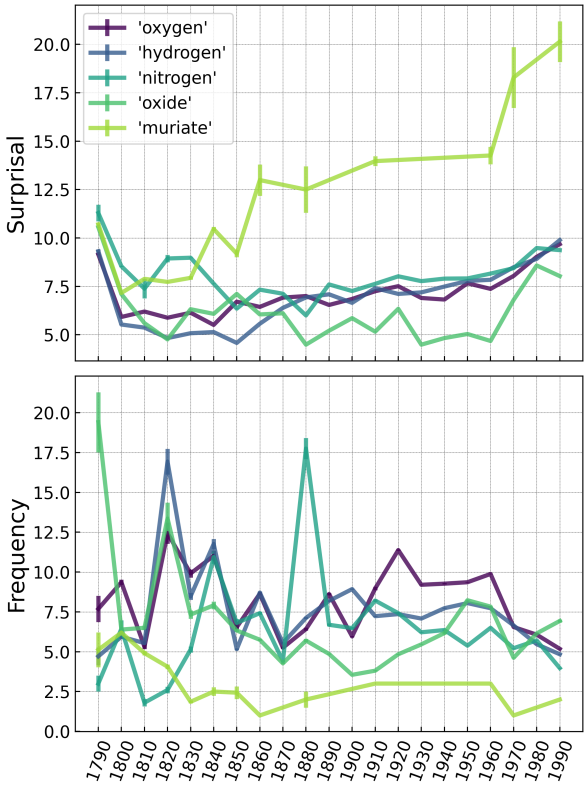

Figure 5: Average surprisal and frequency of nouns contributing to the surprisal increase and related to the chemical revolution. Error bars show standard error over documents.

idence around the discovery of oxygen and hydrogen. Troughs, on the other hand, reflect periods of consolidation, where new vocabulary is integrated into language use possibly becoming established terminology (cf. Degaetano-Ortlieb and Teich, 2019). To test this assumption, we further inspect the nouns contributing most to the surprisal increase in the 1790s (see Figure 5 showing surprisal and frequency). At their first mentioning in the 1790s, these nouns are relatively high in surprisal, but strongly decrease in the decade 1800, when their frequency increases, stabilizing at a mid-surprisal range in the coming decades (1800-1840). This is clearly an indication of a point in time (1790s) of innovation in terms of the use of new lexemes followed by a period of conventionalization, where new terminology was established in the new chemistry field. Thereafter, the chemical elements oxygen ($s = 0.031$, $p < 0.001$), hydrogen ($s = 0.0424$, $p < 0.001$), and nitrogen ($s = 0.0188$, $p < 0.05$) show a continuous increase in surprisal, with a clear peak from the 1970s to the 1990s. This tendency seems to indicate two distinct mechanisms that might have an impact on the nouns' surprisal: (1) given that the frequency is not decreasing until the 1960s, the nouns might be used in more diverse contexts which would explain their increase in surprisal (e.g., *thought that/permeable to/the ketonic oxygen* with high surprisal of *oxygen* >10 bits), (2) in the period of the 1970s to the 1990s, their frequency decreases, which might explain their even stronger increase in surprisal in that later period.

4.2. Modeling Surprisal for Long-Range Dependencies

In this second analysis, we focus on relative clauses (RCs), which inherently involve long dependencies (see Example (3)), necessitating models that can appropriately handle such complexities. In this regard, transformer-based models are arguably more effective than n-gram or LSTM models as they can make use of very long contexts, offering deeper, context-sensitive analysis that is crucial for accurately capturing the nuanced aspects of these linguistic structures.

(3) *The **protein**, for which the detailed biochemical pathway analysis conducted by the researchers identified several novel interaction partners, **exhibits** properties consistent with increased metabolic resistance.*

4.2.1. Surprisal of Relativizer

We start by considering the surprisal of the relative pronouns in Figure 4, which show a clear turn in 1920, while conjunctions and particles seem to stabilize in surprisal for more than 100 years between 1800 and 1920. Interestingly for relativizers, the development until the 19th century is not exactly in line with previous research on the diachronic development of relativizer informativity (cf. Krielke, 2021), which reports on an overall slightly increasing surprisal of relativizers (*which* and *that*). Krielke (2021) explains the upward trend with the frequency decrease of relativizers in scientific writing. However, they report on an increasing conventionalization of *which* as the preferred relativizer in scientific writing occurring in increasingly uniform contexts accounting for some very low surprisal values. An explanation for the differences in our results might be the different modeling of surprisal since Krielke (2021) uses a 4-gram surprisal model based on probability estimates of relativizers within 50-year periods. As our estimates are based on a dynamic context size and models are more precise in terms of surprisal prediction in different time slices due to the balanced vocabulary size of each slice, we can assume that our results reflect changes in relativizer informativity more reliably. Moreover, the increase in surprisal in the 20th century in our data may have two mutually non-exclusive explanations: (a) rela-

tivizers become less frequent in the 20th century, and (b) they occur in increasingly diverse contexts as a reaction to specialization, e.g., they might follow a higher diversity of nouns, which are also less predictable (cf. Figure 4).

4.2.2. Relativizers in Context

We furthermore inspect more thoroughly the reasons for the increasing surprisal values of relativizers in the 20th century by examining a) the frequency distributions of relativizers over time (Figure 6) as well as the immediate lexical contexts of relativizers (*that* and *which*, Table 2).

The frequencies show that the overall strongest decrease in relativizers over time happens roughly in the 19th century. This shows that the sudden increase in surprisal of relativizers in the 20th century does not seem to be motivated by a major frequency drop but rather by a specialization of contexts that relativizers tend to occur in. The most frequent part-of-speech 3-grams preceding the relativizers *that* and *which* reflect this: *that* in the decade 1990 mostly occurs after complex noun phrases (Table 2), which can be assumed to decrease the predictability of the relativizer. The preceding contexts of *which* are more predictive since they introduce either prepositional RCs (e.g., *the way in which*) or restrictive RCs separated from the matrix clause with a comma (e.g. *the experiment, which*). What is interesting here is the fact that the more predictable *which* steadily drops in frequency while *that* steadily increases from 1900 onward. This could explain the overall increase in surprisal since the strongly conventionalized *which* becomes less influential in the average surprisal while *that* becomes less predictable in context and more frequent.

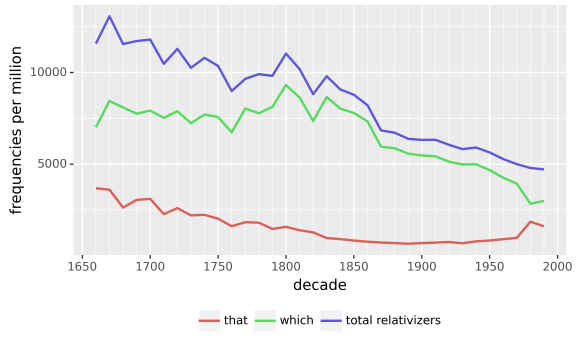

Figure 6: Distribution of relativizers in the RSC.

While this kind of pattern-based context analysis would also be possible using 4-grams, our surprisal model should also account for larger contexts and better surprisal estimates compared to the 50-year-based 4-gram surprisal model used in previous studies allowing more reliable interpretations of diachronic trends.

freq.	trigram	example
3034	IN DT NN	of the acid *that*
2893	DT JJ NN	the muriatic acid *that*
1627	NN IN NNS	number of experiments *that*
1473	IN JJ NNS	in various instances *that*
1415	DT NN NN	the iron particles *that*
6392	DT NN IN	the way in *which*
3123	JJ NN ,	unique solution, *which*
2866	JJ NN IN	special case in *which*
2818	NN NN ,	length scale *which*
2722	DT JJ NN	the complex plane *which*
1767	(CD)	(3.1) *which*

Table 2: POS trigrams preceding *that* and *which*

4.2.3. Surprisal in Long-Range Dependencies

Here we ask whether syntactic context has an influence on the predictability of syntactic elements relying on longer distances to their syntactic heads. As a plausible measure to define syntactic context, we make use of the well-established metric *dependency length* (DL) describing the distance from an element X to its syntactic head (Heringer et al., 1980; Hudson, 1995). We apply this metric to measure the distance between the head noun of an RC and its embedded verb to find out whether the distance correlates with the predictability of the embedded verb. One assumption could be that the closer the relevant information (head noun) is located to the upcoming dependent (embedded verb) the lower the surprisal of the upcoming word (see Example (4)).

(4) a. *The **woman** who **ate** the sandwich was hungry.* $(DL = 2)$
b. *The **woman** whom the manager of operations wanted to talk to **was** upset.* $(DL = 10)$

We start by inspecting the diachronic development of DL and the surprisal of embedded verbs in RCs by decades (see Figure 7). DL and surprisal are negatively correlated, i.e. the opposite of our assumption is the case: in decades where the embedded verb on average is further away from its syntactic head the verb's average surprisal is lower, and in decades where the verb is closer to its syntactic head the verb is more surprising (compare Examples (5-a) and (5-b)).

(5) a. *The **questions** that we might **ask** are not easy to answer.* $(DL = 4)$
b. *The **questions** that **lie** on the table are not easy to answer.* $(DL = 2)$

One explanation is that at the point of encountering

the relativizer the entropy (i.e. uncertainty about the rest of the sentence, Hale, 2006) is higher than at a later point in the relative clause.

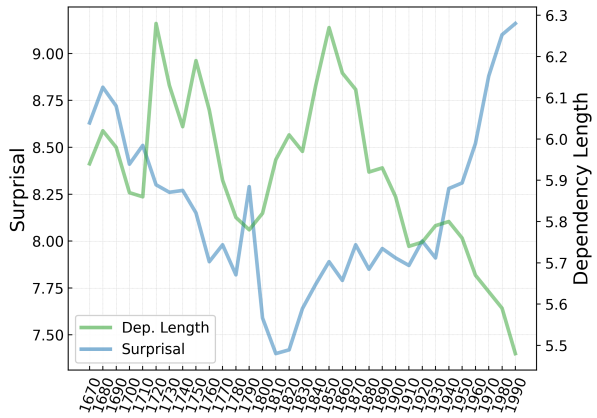

Figure 7: Average surprisal on the verb in the RC and dependency length to its head noun in the RSC. Negative correlation for surprisal and dependency length (Spearman's $\rho = -0.35$, $p < 0.05$).

To get a better intuition about the reasons for the lower predictability in shorter syntactic contexts and the higher predictability in longer syntactic contexts, we extract the most surprising contexts of embedded verbs in RCs in 1990 plus the least surprising contexts in 1820. Example (6) reflects two aspects we have mentioned so far. First, the conventionalized pattern of prepositional RCs (i.e. *determiner noun preposition*) contexts not only seem to increase the predictability of the upcoming relativizer but also that of the embedded verb. Second, surprisal at the main verb (participle) in the RC might be reduced due to being in a highly predictable passive construction (*has been + participle*).

(6) *the first **case** in which a quantitative attempt has been **made*** ($Surprisal = 0.0142$) Conversely, the high-surprisal contexts are extremely short where the head noun is directly followed by the relativizer and the embedded verb (Example (7)).

(7) *recordings in the region of the pontine **nuclei** that **VERB*** ($Surprisal = 9.9998$) The latter, high-surprisal cases syntactically belong to the subject RC type, while the first, low surprisal cases belong to the oblique relative clause type. The negative correlation between surprisal and DL can thus also be explained with Hale's Entropy Reduction Hypothesis (ERH, Hale, 2006) – uncertainty about the rest of the sentence tends to decrease as new words are introduced, and the degree of this reduction aligns with the information that the word conveys within the context of the current sentence (cf. Frank, 2013, p. 476). Thus, a high surprisal value at the verb of a subject RC directly following the relativizer is equivalent to a strong reduction of uncertainty about the rest of the sentence since at this point the relativized grammatical relation can be resolved. The low surprisal at the embedded verbs of RCs extracted from other positions (e.g., oblique) implies that entropy reduction here is much lower since a lot of information for disambiguation about the rest of the sentence has been given before.

For comparison, we consider the 1820s where DL is comparatively long, while surprisal is fairly low. Example (8-a) shows a particularly long dependency relation between the head noun (*bundles*) and the embedded verb (*composed*) with extremely low surprisal.

(8) a. *denote the **bundles** of fibres of which the brain is **composed*** ($Surprisal = 0.1132$)

Since the immediate left context (i.e. *the brain is*) of *composed* is not particularly predictive, while the head noun *bundles* is much more so, our surprisal estimates seem to rely on a context beyond the immediately preceding 3-gram. Thus, our modeling approach allows us to capture long-range syntactic relations. Together with the fact that our results align with observations from psycho-linguistic studies (e.g. Frank, 2013), we conclude that our methodology for generating diachronically comparable surprisal estimates provides plausible metrics to investigate the interplay between information content and syntactic context in diachronic language change.

5. Conclusion

We demonstrated the efficacy of transformer-based surprisal models in analyzing diachronic linguistic change, highlighting their capacity to accommodate long-range dependencies and global context. The application of these models to historical data is not without challenges given the exponential increase in texts and tokens over time, leading to partially disjoint vocabularies and non-comparable probability distributions across different periods. Significant disparities in training set sizes between time periods further complicate the modeling process. To address this, we implement two key strategies: (1) sharing a unified vocabulary over all models, (2) pre-training on a more general dataset sampled uniformly from the whole corpus and then continue pre-training on documents of a specific year. Our models also uniquely incorpo-

rate a temporal aspect, restricting training data to texts published before the target period (which can be adapted for other research questions).

Our empirical analysis, compared against prior studies using n-gram models on the Royal Society Corpus, revealed both corroborative and novel insights. Specifically, the examination of linguistic phenomena, such as relative clauses, underscored the superiority of transformer-based models in predicting changes not solely based on changes related to frequency distributions. These models, with their broader contextual awareness, facilitated a more nuanced exploration of communicative efficiency, aligning with theoretical frameworks like Hale's Entropy Reduction Hypothesis. The inclusion of DL annotations further enriched our analysis, allowing for a more granular examination of syntactic structures. This provided deeper insights into the adaptive mechanisms of language, reflecting shifts in complexity and efficiency within scientific discourse.

While our research focused on diachronic scientific English, the methodologies employed are universally applicable, especially in low-resource environments facing similar challenges with vocabulary consistency and corpus size disparities. This universality significantly broadens the potential impact of our findings, suggesting that our approaches could be instrumental in diverse linguistic and historical analyses.

6. Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID 232722074 – SFB 1102.

7. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Douglas Biber and Bethany Gray. 2011. Grammatical change in the noun phrase: the influence of written language use. 15(2):223–250.

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

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

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Stefania Degaetano-Ortlieb. 2021. The rise of compounds as informationally dense structures in 20th-century scientific English: Chapter 11. measuring informativity. In Elena Seoane and Douglas Biber, editors, *Corpus-based Approaches to Register Variation*, Studies in Corpus Linguistics, pages 291–312. John Benjamins Publishing Company.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2018. *An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English*. Brill. Pages: 258-281 Section: From Data to Evidence in English Language Research.

Stefania Degaetano-Ortlieb and Elke Teich. 2016. Information-based modeling of diachronic linguistic change: from typicality to productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 165–173.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb and Elke Teich. 2019. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. 109(2):193–210.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 794–802, Marseille, France. European Language Resources Association.

Stefan L. Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5(3):475–494.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. 112(33):10336–10341. Publisher: National Academy of Sciences.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 446–457. Association for Computational Linguistics.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

M.A.K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of written English: Situational factors and linguistic features*, pages 162–177. Pinter. Tex.date-added: 2010-03-09 11:19:11 +0100 tex.date-modified: 2010-03-30 15:16:26 +0200.

M.A.K. Halliday and James R. Martin. 1993. *Writing science: Literacy and discursive power*. Falmer Press.

William L. Hamilton, J. Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zellig Harris Sabbettai. 1991. *A Theory of Language and Information: A Mathematical Approach*. Oxford University Press, Oxford, New York.

Hans Jürgen Heringer, Bruno Strecker, and Rainer Wimmer. 1980. *Syntax: Fragen, Lösungen, Alternativen*. Fink.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural Language Models for Nineteenth-Century English. ArXiv:2105.11321 [cs].

Richard Hudson. 1995. Measuring syntactic difficulty. *Manuscript, University College, London*.

Marianne Hundt, David Denison, and Gerold Schneider. Relative complexity in scientific discourse. 16(2):209–240.

Md. Hussain and Ishtiak Mahmud. 2019. pymannkendall: a python package for non-parametric Mann Kendall family of trend tests. *Journal of Open Source Software*, 4(39):1556.

Tom S Juzek, Marie-Pauline Krielke, and Elke Teich. 2020. Exploring diachronic syntactic shifts with dependency length: the case of scientific English. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 109–119, Barcelona, Spain (Online). Association for Computational Linguistics.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich.

2016. The Royal Society Corpus: From uncharted data to corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1928–1931, Portorož, Slovenia. European Language Resources Association (ELRA).

Hannah Kermes and Elke Teich. 2017. Average Surprisal of Parts-of-(s)peech. Birmingham, UK. Corpus Linguistics 2017.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, Detroit, MI, USA. IEEE.

Marie-Pauline Krielke. 2021. Relativizers as markers of grammatical complexity: A diachronic, cross-register study of English and German. *Bergen Language and Linguistics Studies*, 11(1):91–120.

Marie-Pauline Krielke, Luigi Talamo, Mahmoud Fawzi, and Jörg Knappen. 2022. Tracing syntactic change in the scientific genre: Two Universal Dependency-parsed diachronic corpora of scientific English and German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4808–4816, Marseille, France. European Language Resources Association.

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

Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. Generating Linguistically Relevant Metadata for the Royal Society Corpus.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. An Analysis of Neural Language Modeling at Multiple Scales. Publisher: arXiv Version Number: 1.

Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukas Burget, and Jan Honza Cernocky. 2011. RNNLM - Recurrent Neural Network Language Modeling Toolkit. In *IEEE Automatic Speech Recognition and Understanding Workshop*. Edition: IEEE Automatic Speech Recognition and Understanding Workshop.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Simone Scaboro, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2021. NADE: A benchmark for robust adverse drug events extraction in face of negations. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 230–237, Online. Association for Computational Linguistics.

Hans-Jörg Schmid. 2015. A blueprint of the Entrenchment-and-Conventionalization model. 3(1):3–26. Publisher: De Gruyter Mouton.

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

Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904. ISCA.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. Less is more/more diverse: On the communicative utility of linguistic conventionalization. 5:142.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick

von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

8. Language Resource References

Fischer, Stefan and Knappen, Jörg and Menzel, Katrin and Teich, Elke. 2020. *The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study*. European Language Resources Association. [link].

Kermes, Hannah and Degaetano-Ortlieb, Stefania and Khamis, Ashraf and Knappen, Jörg and Teich, Elke. 2016. *The Royal Society Corpus: From Uncharted Data to Corpus*. European Language Resources Association (ELRA). [link].