Modeling the Memory-Surprisal Trade-Off over Time: Communicative Efficiency Decreases with Lexico-Grammatical Change in Scientific English

Julius Steuer[†,*], **Marie-Pauline Krielke**[*], **Stefania Degaetano-Ortlieb**[*], **Elke Teich**[*] **and Dietrich Klakow**[*]

[†]Heidelberg Institute for Theoretical Studies
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg
[*]Universität des Saarlandes
Department of Language Science and Technology
Campus A2.2, 66123 Saarbrücken, Germany

Abstract

The memory-surprisal trade-off (MST) has been shown to hold cross-linguistically as a general principle of communicative efficiency: languages that exhibit information locality tend to have word orders that allow for efficient memory use, i.e., lower surprisal at a fixed memory budget. In this paper, we explore the influence of diachronic variation on the MST. We compare scientific English in the Royal Society Corpus (RSC, 18thc. – 20thc.) to "general language" in the Corpus of Historical American English (COHA) to assess the impact of intra-linguistic variation (register). We find that both time and register influence the shape of the tradeoff: Over time, vocabulary expansion raises minimal surprisal, while the shape of the MST curves changes. Decreasing distances between syntactic dependencies due to more local nominal encodings change how predictive information is distributed across memory scales. The effects are stronger for RSC than for COHA.

Keywords: Memory-Surprisal Trade-Off, Scientific English, Diachronic Variation

1. Introduction

The development of scientific English over the last 300 years was characterized by a shift from more intricate sentence structure with a high degree of clausal embedding towards increasingly informationally packed noun phrases, shorter sentences, and decreasing dependency length (DL). These changes have led to the conclusion that scientific English has become syntactically less complex at sentence level and more complex at noun phrase level over time (Juzek et al., 2020; Krielke et al., 2022; Krielke, 2024). At the same time, scientific English has expanded its vocabulary drastically from ca. 1900 onward, as seen, e.g., in the exponential increase of noun types in the Royal Society Corpus (RSC; Fischer et al. 2020) (see Figure 3). In this paper, we set out to model the impact of lexical expansion and syntactic change on communicative efficiency in terms of the memory-surprisal trade-off (MST, Hahn et al., 2021). The MST unifies two perspectives relevant to communicative efficiency in language processing: surprisal theory (Levy, 2008), which models the cost of processing words as a function of their predictability from prior context, and dependency locality theory (Gibson, 2000), which models the memory costs associated with syntactic distance. It does so by quantifying how much average surprisal can be reduced when more memory of the preceding context is available. The resulting tradeoff curve shows the minimal surprisal achievable for a given memory budget, or conversely, how much memory is required to reach a certain level of predictability.

In their seminal paper, Hahn et al. (2021) showed that the word order of a typologically diverse set of languages lies close to the efficient frontier of the memory-surprisal tradeoff. Building on this framework, we ask how the tradeoff of scientific English has changed over time as a result of lexical and syntactic developments. In particular, we investigate which aspects of linguistic change shift the shape of the tradeoff curve, for example, whether additional memory yields continuous reductions in surprisal, as in earlier scientific English characterized by information spread across long-distance dependencies (compare example in Figure 2 (a)), or whether surprisal is minimized locally, as in later periods with a preference for compact noun phrases packaging more information more locally as in Figure 2 (b). This perspective allows us to separate the predictable effect of vocabulary growth on minimal surprisal from the structural effects of syntactic change on the distribution of predictive information. We also ask whether any changes in MST are register-specific, i.e., whether scientific English is affected more than general English, and whether there are differences between subdisciplines. The paper is structured as follows. In Section 2, we review extant literature on change in scientific English and previous work on memory-surprisal models. We describe our data – the scientific English cor-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11309–11319
11–16 May 2026.

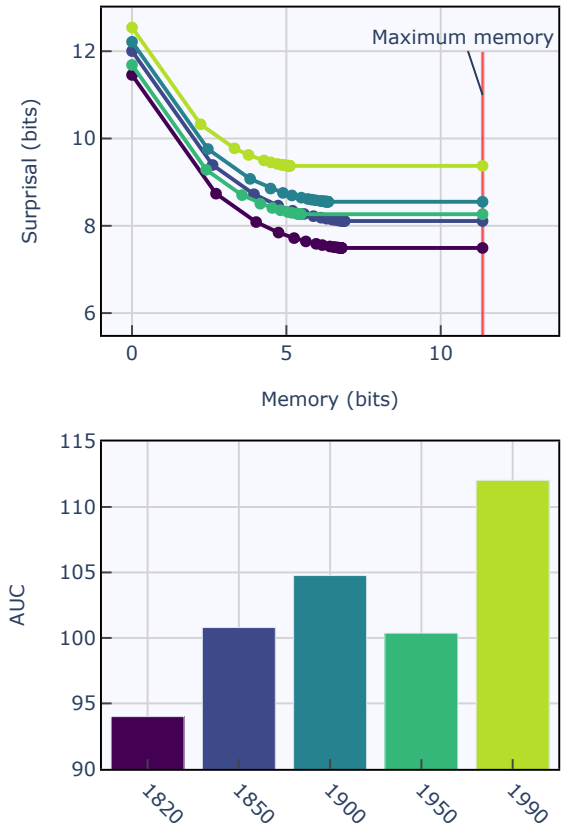


Figure 1: MST curves (1820-1990) and their respective areas under the curves (AUC). MST curves where extended to the maximum per-document memory in RSC.

pus(RSC) and a mixed language corpus (COHA, Davies 2021) (Section 4) – and our methods including language model training for surprisal estimation and MST calculation (Section 5). In Section 6, we present the results of our MST analyses regarding the influence of time, register, and word class as predictors of variation in MST with a stronger effect in RSC, reflecting the increasingly dense nominal style (Example 2). Section 8 closes with a discussion of the diverging development of the scientific metaregister with its specialization into subdisciplines and the language in the COHA corpus.

2. Related Work

2.1. Diachronic development of scientific English

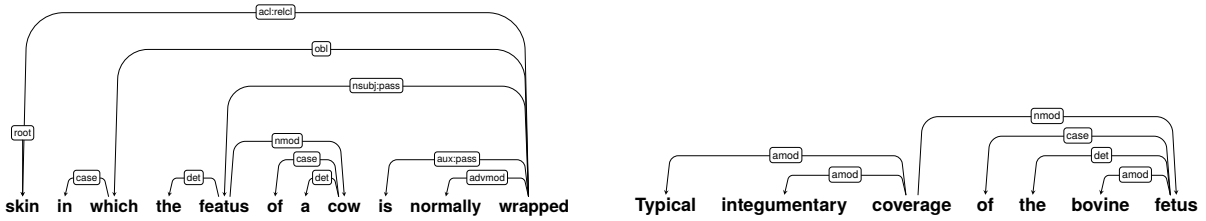
In the past 300 years, scientific English has undergone substantial changes on the lexical and grammatical levels (e.g., Banks, 2003; Halliday, 1988). Lexis is continuously expanded with new technical terms (Halliday and Martin, 1993; Wang et al., 2023), and due to the increasing shared background knowledge within individual scientific disciplines, grammatically explicit constructions such as clausal subordination (Example in Figure 2a) become less frequent (Hundt et al., 2012; Krielke, 2024) in favor of a dense, implicit nominal style with heavy noun phrase constructions (Biber and Gray, 2011; Biber and Clark, 2002) (cf. Example in Figure 2b).

According to rational communication, diachronic change is a continuous process of adapting the linguistic system to emerging communicative needs while holding processing effort stable. Information-theoretic approaches have shown that periods of *lexical expansion* are associated with increased surprisal (e.g. Steuer et al., 2024), while *grammatical conventionalization* leads to increasingly predictable grammatical constructions (Degaetano-Ortlieb et al., 2019; Teich et al., 2021).

To cognitively assess syntactic phenomena, dependency locality (the distance between syntactically related words) has been used to approximate the processing difficulty of working memory (Gibson, 2000; Lewis and Vasishth, 2005). While overall, languages tend to minimize the length of their syntactic dependencies (Futrell et al., 2015; Liu, 2008) compared to random baseline word orders, this also applies diachronically (Gulordava and Merlo, 2015; Lei and Wen, 2020) and specifically in the register of English (Juzek et al., 2020) and German scientific writing (Krielke, 2024). In the present paper, we set out to measure communicative efficiency by applying the MST over time as well as by register (scientific vs. non-scientific language).

2.2. Memory-surprisal models

Hahn and Futrell (2020) extend expectation-based processing models (Levy, 2008) and lossy compression theory (Cover and Thomas, 2006) to propose an information-theoretic framework for memory efficiency in language. They define memory efficiency as a trade-off between surprisal and memory usage where reducing average surprisal per word requires storing more information about past context. Applying this to 54 languages, they find that word order optimizes processing efficiency under memory constraints, supporting the idea that syntax facilitates efficient online processing. Hahn et al. (2021) extend the notion of the MST proposing the Efficient Tradeoff Hypothesis, which suggests that word order in natural language is shaped by pressures to optimize this tradeoff. They further derive that languages achieve more efficient tradeoffs when they exhibit information locality, i.e. predictive information about a word is concentrated in its immediate preceding linguistic context. While these approaches have proven a cross-linguistic tendency to order words and mor-



(a) (b)

Figure 2: Dependency structures of (a) noun phrase with relative clause postmodification (15 words) and (b) noun phrase with multiple premodication (9 words).

phemes to achieve a maximally efficient tradeoff between memory and surprisal, to date, the approach has not been applied to intralinguistic or diachronic studies.

2.3. MST Intuition

Figure 1 illustrates the relationship between MST curves and area under the curve (AUC) for five years from RSC: the curve for the last year (1990) starts at the highest unigram surprisal and then converges to the highest surprisal level after the maximum amount of memory bits, resulting in the highest AUC value. Conversely, the first year (1820) starts at the lowest unigram surprisal and reaches the lowest surprisal level overall after the maximum number of memory bits, resulting in the lowest AUC and thus a better MST. Between those years, 1950 still follows the (expected) trend of increasing unigram surprisal, but intersects the MST curve of 1850, leading to a *lower* AUC.

3. Rationale and Hypotheses

Over time, scientific English has shifted toward a nominal style with high lexical density and syntactic conventionalization, promoting local but implicit dependencies. Simultaneously, vocabulary growth increases lexical variability, suggesting an interaction between lexis and grammar. Nominal constructions represent implicit but highly local dependency relations, depending on low-memory prediction, whereas verbal constructions represent explicit, less local dependencies and thus benefit from more information stored in memory. As vocabulary expands, the average lexical surprisal rises, implying that the minimum achievable surprisal in later periods exceeds that of earlier ones.

Specifically, we expect that changing preferences for specific syntactic constructions will lead to different shapes of the MST. For instance, a language variety (e.g., register and/or period) with a high usage of subordinate constructions leads to longer dependencies generating longer predictive contexts (e.g., Figure 2a). Such constructions benefit from higher memory usage to predict the next word, since more memory helps to reduce surprisal. In contrast, varieties using highly dense constructions (e.g., Figure 2b), less memory should be enough on average to predict the next word, while more memory should not necessarily improve the prediction.

To quantify the quality of the MST over time, we calculate the area under the curve (AUC) of the memory–surprisal graph per decade in scientific and general English. We compare the AUCs calculated for both corpora per 50-year periods to find diverging trends between the registers. To detect the trends for verbal vs. nominal constructions, we calculate the MST nouns, verbs and "other" and compare the AUCs respectively.

Since the AUC can only give us a reduced picture of the actual shape of the MST, we also consider the actual MST graphs and interpret their slopes. If a graph flattens at a low memory budget, this means, more memory does not contribute to improving the prediction of the next word. If a graph decreases steadily, this means that every further token held in memory improves the prediction of the next word further.

Based on the attested developments in scientific English, we form the following **hypotheses**:

- **H1: Impact of time** We expect a general temporal increase of the AUCs in both registers due to vocabulary expansion.
- **H1.1: Impact of register** We expect to find a higher and steeper increase in the AUCs for RSC than for COHA due to a stronger vocabulary increase in scientific English.
- **H1.2: Difference between POS** The vocabulary expansion affects the MST of nouns (i.e. increasing AUC) more than other POS, especially in scientific English.
- **H2: Effects of nominal style on shape of the MST curves** Over time, we expect to find a weaker surprisal reduction at bigger memory budgets, especially in RSC.

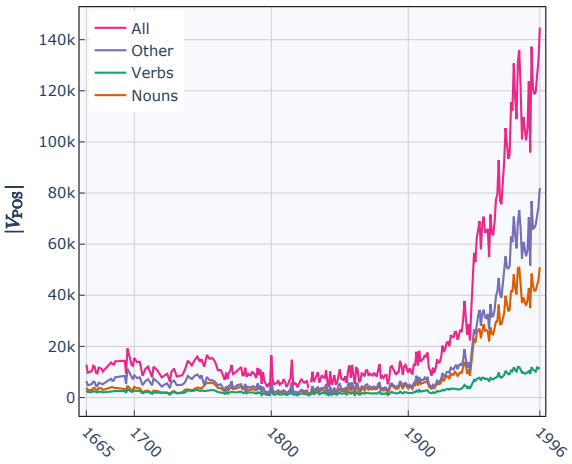


Figure 3: Increase of types per part-of-speech (POS) in RSC over time; "Other" contains all POS except nouns and verbs.

4. Data

4.1. Royal Society Corpus

We use two English diachronic corpora covering the time between 1750 and 2000. For scientific English we use the Royal Society Corpus, (RSC; Fischer et al., 2020)[1], consisting of the publications of the Royal Society of London with 47K documents and 300M tokens. We evaluate the evolution of the MST in 3 sub-samples, given that RSC was split into subjournals around the 1900: (1) **RSC** encompasses all documents from 1665 to 1900, and from 1900 onward (2) **RSC-A** includes the Proceedings and Transactions of the Mathematical, Physical and Engineering Sciences, and (3) **RSC-B** containing publications of the Proceedings and Transactions of the Biological Sciences. Documents from a forth category containing, e.g., obituaries were excluded from the analysis.

4.2. Corpus of Historical American English

For general English, we use a reduced version (masked words) of the multi-genre, diachronic COHA corpus (Davies, 2021). The full COHA comprises over 475 million words spanning the 1820s to 2010s. To make the linguistic annotation comparable in both corpora, we parse and POS-tag the corpus with the Stanza software package (Qi et al., 2020), using the default English parser.

[1]The RSC is available for download at https://fedora.clarin-d.uni-saarland.de/rsc_v6/

4.3. Corpus subsampling

We follow the diachronic language modeling approach introduced by Steuer et al. (2024) by subsampling train sets of approximately identical size for each year in a corpus (see Table 1 in Appendix A). For the tokenization methods not based on subwords, we apply a post-processing step that reduces the number of vocabulary items to obtain approximately similar vocabulary sizes of ≈ 80.000. For each tokenization method, we choose a separate threshold frequency t_{REPL} that any token in the train set must exceed to be included in the tokenizer's vocabulary. We split the train set by white spaces and replace all words that occur only t_{REPL} times in the train set by an "unknown" token that corresponds to its POS tag as given by the UPOS column in the conllu file. Then, we replace all OOV items in the validation and test sets in the same way.

5. Methods

5.1. Tokenization

We tested several tokenization methods for both corpora. These methods are described in detail in Appendix A. For the results in the main paper, we used a *lempos-based* tokenization: We first split the train corpus by whitespaces, and then replace each word with a concatenation of its lemma form as given by the UPOS tag as given by the respective columns of the conllu file. In case the absolute frequency of a word did not exceed the threshold value t_{REPL} it is replaced by its UPOS tag. We then replace all out-of-vocabulary (OOV) items in the validation and test set in the same way. The final tokenizer (used for all models trained on that corpus) is trained on the concatenated train sets of each corpus. This dampens the effect of the exponential increase of noun types in RSC, and allows a closed, word-based vocabulary sampled equally from all years of the corpus.

5.2. Language models

For each tokenization method, we use Hugging Face transformers (Wolf et al., 2020) to train the base version of the OPT architecture (Zhang et al., 2022) on each subset of the training corpus (i.e., the train set of pertaining to a single year in either RSC or COHA) for 10 epochs with a batch size of 256, a learning rate of 5×10^{-5} and a linear learning rate warmup over 50% of training steps. Word-level models were trained with a context window of 32, and the BPE model with a context window of 64. Training was done on a cluster of 8 Nvidia A100 GPUs with 40GB of memory and took about 2 hours per model. We then used the language models to

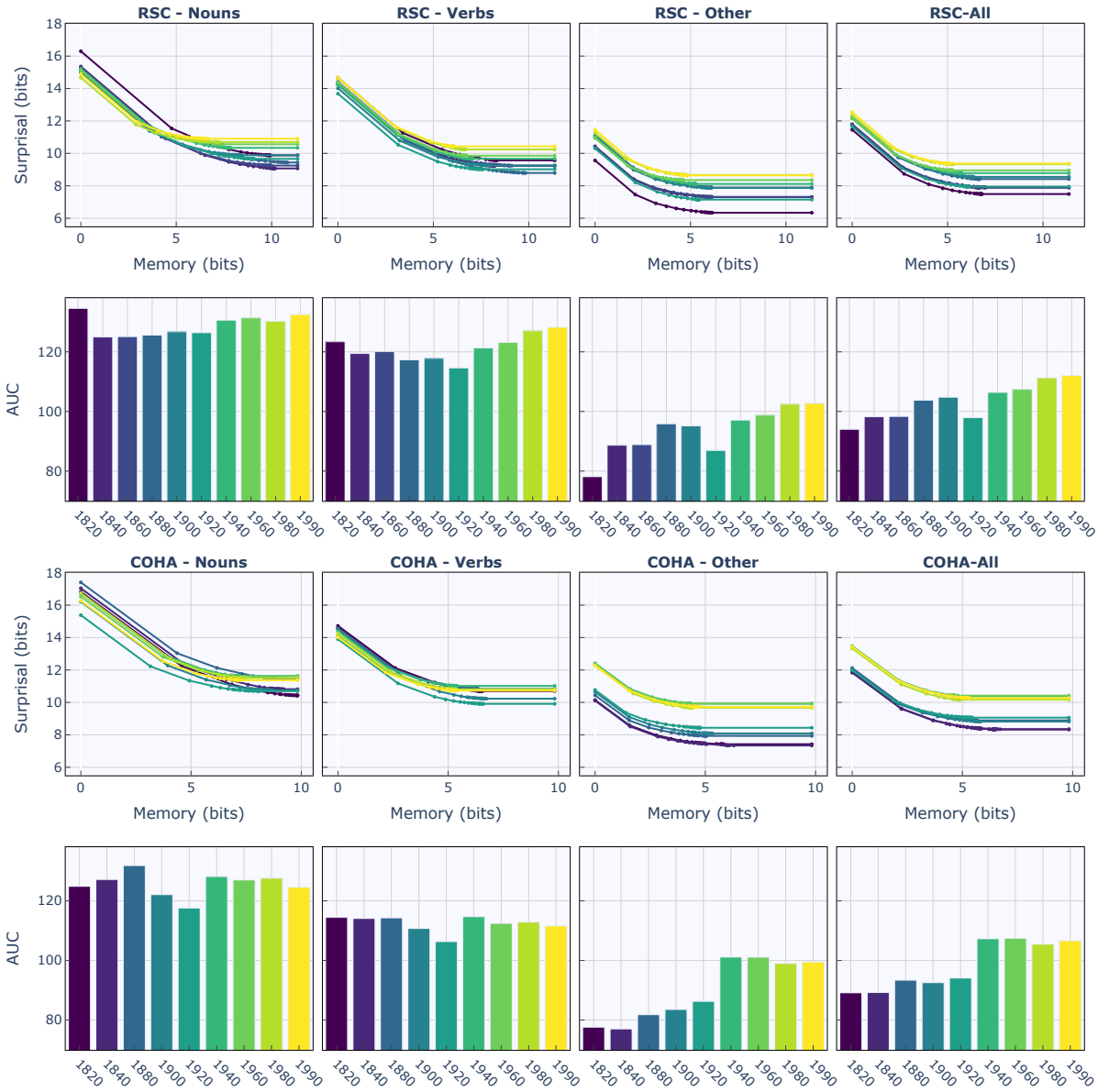


Figure 4: Memory-surprisal trade-off curves for 10 years from RSC. Surprisal was averaged over documents and cross-validation folds. Each dot on a curve corresponds to a surprisal - memory pair, starting with unigram surprisal and no memory. All curves where extended to the maximal amount of memory available.

estimate surprisal values on all documents from each test year of the two corpora. The code for language model training and MST estimation are made available on GitHub[2].

5.3. Surprisal estimation

For each context size T ranging from $T = 0$ (unigram surprisal) to $T_{Max} = 20$, we estimate average surprisal $\hat{S}_T$ on a document D of $|D|$ words following Hahn et al. (2021):

$$\hat{S}_T = \frac{1}{|D| - T} \sum_{t=T}^{|D|} -\log_2 p(w_t | w_{t-T}, ..., w_{t-1}) \quad (1)$$

We estimate $p(w_t | w_{t-T}, ..., w_{t-1})$ directly from a transformer model averaging $\hat{S}_T$ on the documents from a single year over five models trained on different cross-validation splits as described in Section 4. Since the model may overfit for larger values of T due to data sparsity, we stop estimating $\hat{S}_T$ if $\hat{S}_T > \hat{S}_{T-1}$ and substitute $\hat{S}_{T-1}$ for $\hat{S}_T$. Since we want to compare the MST of different POS tags, we calculate $\hat{S}_T$ for a a given set of POS tags

[2] https://github.com/uds-lsv/rsc-mst.git

$P = \{p_1, ..., p_{|P|}\}$ and a subset of words $D_P \subseteq D$ as:

$$\hat{S}_T^P = \frac{1}{|D_P| - T} \sum_{t=T}^{|D_P|} -\log_2 p(w_t | w_{t-T}, ..., w_{t-1}) \quad (2)$$

5.4. AUC calculation

We then use surprisal estimates $\hat{S}_T^P$ to calculate mutual information I_T^P for each context size T as $I_T^P = \hat{S}_{T-1}^P - \hat{S}_T^P$, and memories M_T^P as $\sum_{t=0}^{T} t I_T^P$. We chose the following POS tag sets: Nouns (UPOS = "NOUN"), verbs (UPOS = "VERB") and other (all other POS). After estimating $\hat{S}_T^P$s and I_T^P, we calculate the area under the memory-surprisal trade-off curve (AUC) by applying the trapezoidal rule using the corresponding function of the scikit-learn Python package (Pedregosa et al., 2011).

5.5. Statistical modeling

To assess the temporal development of the MST in the two corpora, we fit linear mixed-effects models (LMEs) via the lmerTest R package with AUC as response variable and average per-document 7-gram surprisal (surprisal estimated from the transformer model with 6 words in the context), journal, period, and POS as dependent variables. As we calculate the AUC for each document in the corpus, we include the document ID as a random effect nested in the corpus variable. We fit a separate LME for each tokenization method. We used the following formula to fit all regression models:

```
lmer(auc ~ year * pos * journal
         + surprisal-7 * year
         + (1|corpus/doc_id),
         data = .)
```

We normalized auc, year and surprisal-7 to the interval $[0, 1]$. We chose "coha" (that is, the whole COHA corpus) as the base level of the journal variable, and "other" (neither noun nor verb) as the base level of the POS variable.

6. Analysis

6.1. General effects

Figure 5 shows all effects of interest. The strongest effect is plausibly found for 7-gram surprisal (Estimate: 0.8525; CI: 0.8492, 0.8557; $t = 522.77$), since AUC is correlated with the surprisal values at different memory budgets. We also see main effects of POS, with both nouns (Estimate: 0.44; CI: 0.43, 0.45; $t = 96, 17$) and verbs (Estimate: 0.27, CI: 0.26,0.28; $t = 61.47$) having on average higher AUCs than other POS, which is in line with their generally higher surprisal.

6.2. Effect of time

AUC increases gradually over time (main effect of the "Year" variable; Estimate: 0.0798, CI: 0.0731, 0.0866; $t = 23.26$). We find significant interactions of time and POS, with both nouns (Estimate: -0.15; CI: -0.17, -0.14; $t = -32.94$) and verbs (Estimate: -0.11; CI: -0.12, -0.1; $t = -24.96$) showing a markedly slower increase in AUC than other POS. This effect is stronger in RSC than in COHA, with triple interactions between time, POS and journal indicating a slower increase for nouns compared to other POS in RSC (Estimate: -0.7; CI: -0.10, -0.04, $t = -4, 41$), RSC-A (Estimate: -0.04; CI: -0.07, -0.03; $t = -4.37$), and RSC-B (Estimate: -0.05; CI: -0.07, -0.03; $t = -3.82$).

6.3. Effect of register

Comparing the language of the three subjournals of RSC to COHA, we find a mixed picture. AUC is higher for the early RSC (up until 1900), though this effect is small (Estimate: 0.03; CI: 0.002,0.06; $t = 2.11$), while we find no significant effect for RSC-B. For RSC-A we find a large negative effect (Estimate: -0.1, CI: -0.11, -0.09; $t = -15.41$) indicating stronger divergence from general language.

6.4. Effect of POS

Apart from the main effect of POS, we also find significant interactions of POS and journal: Verbs generally have a higher AUC than other POS in RSC (Estimate: 0.16; CI: 0.12, 0.2; $t = 7.95$), RSC-A (Estimate: 0.19, CI: 0.17, 0.21; $t = 21.33$), and RSC-B (Estimate: 0.06; CI: 0.04, 0.08; $t = 5.19$) compared to COHA, while nouns are overall associated with lower AUC. This is in line with our findings for the interaction of time, POS and journal: Not only do nouns in RSC generally have a lower AUC, but the increase in AUC over time is not as large as may be expected based on the overall increase. Thus, while the number of nominal vocabulary items in RSC increases drastically, they seem to increase in local predictability, resulting in lower surprisal values at small memory budgets compared to earlier years, while over time benefiting less from increased memory budgets (compare Figure 4).

7. Effect of nominal style on the shape of the MST curves

In the previous section, we have analyzed the overall development of the MST in the two corpora as measured by the AUC. However, the AUC is an aggregate measure of slope and average level of surprisal values per memory. Even when two curves

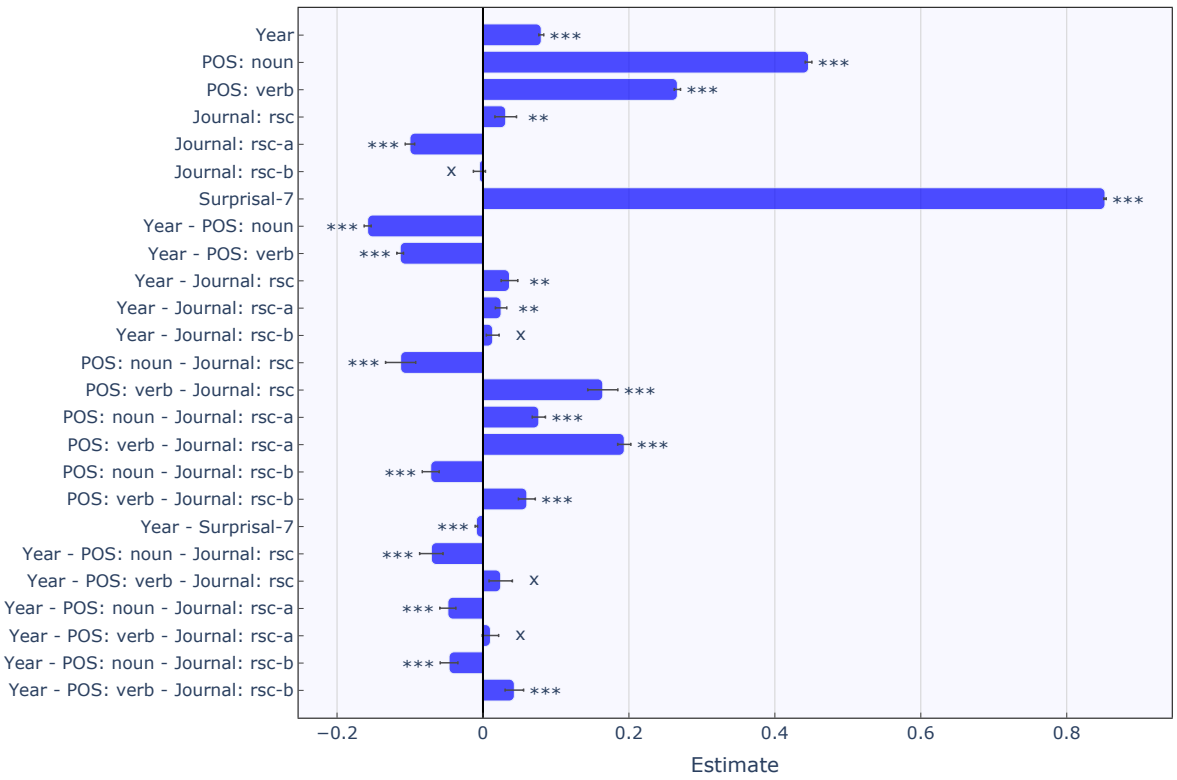


Figure 5: Effects of part of speech, journal and time on AUC for the period from 1820 to 1996. Reference levels for factor variables are "other" (POS) and "coha" (journal). Significance levels: '***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$, 'x' $p >= 0.05$. Error bars show standard error of the coefficient estimate.

do not cross, the degree to which more bits of memory reduce surprisal is not covered by the AUC. We therefore analyze the individual shapes of the MST curves as well as the surprisal reduction rate per every additional bit of memory.

Figure 4 shows that for nouns in RSC surprisal drops continuously in the first 100 years (1820 - 1920) per additional bit of memory, while in the last 60 years, surprisal shows minimal reduction with memory budgets above 5 bits. A similar trend can be observed for verbs in RSC and other POS, however, not as pronounced as for nouns. At the same time, nouns show decreasing unigram surprisal, which is surprising given that the number of nominal vocabulary items increases over time. It shows, however, that in the case of nouns, the temporal increase in AUC is not owed to rising unigram surprisal but instead to the decreasing surprisal reduction per bit of information held in memory. A meta-interpretation of this would be that increasingly dense structures typical for nominal style lead to a decreasing information gain through additional memory, or in other words: If information is packed in dense constructions, only locally placed information helps reduce surprisal of the next word, while with less dense constructions, longer context windows are beneficial for prediction of the next word.

To show the effect of the MST slope on the overall information gain for each word class, we calculate the average slope of the memory-surprisal curves at equidistant memory intervals of one bit (see Figure 6). Comparing RSC and COHA, the slope of the MST curves levels out faster for COHA than for RSC, i.e., the language models trained on RSC data can make use of more bits of memory. This difference may be a result of generally longer sentence lengths in scientific English than in general English. Over time, information gain per bit of memory decreases more strongly in RSC possibly due to a stronger syntactic restructuring across periods in the register. Comparing POS in RSC overall, less and less information is gained (or surprisal reduced) per additional bit of memory for each step of 20 years. In both corpora, the temporal effect is strongest for nouns and especially pronounced in RSC. For verbs, the slope is fairly similar across time in RSC, indicating that there has been less change in predictive contexts of verbs than for nouns. This is plausible given that most changes in scientific English are known to have affected the structure of noun phrases, which have become increasingly dense over time.

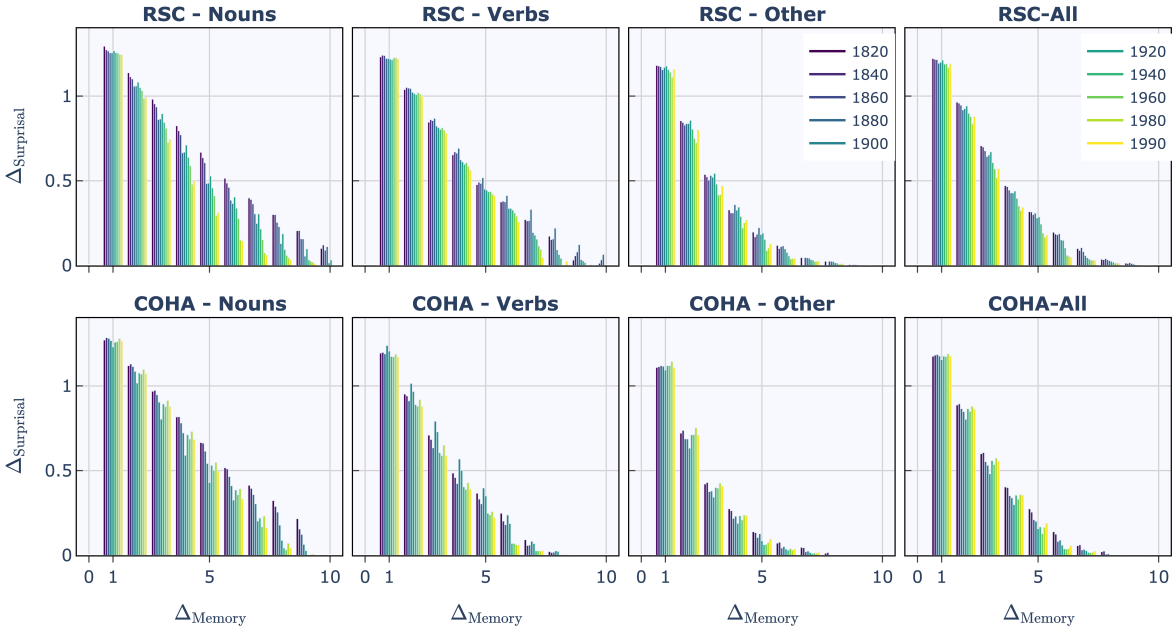


Figure 6: Average slope of the MST curve at equidistant memory intervals of 1 bit.

8. Conclusion

We calculated the Memory-Surprisal Tradeoff (MST) of scientific vs. general English over time. Our central question was whether the MST has changed diachronically and, if so, whether such changes vary across registers. This inquiry was motivated by the well-documented shift in English toward nominal rather than verbal style, manifested in complex, informationally dense noun phrases. While the Efficient Tradeoff Hypothesis predicts that more optimal word order w.r.t. locality should yield more efficient MSTs, our findings for diachronic data show that minimally achievable surprisal increases, resulting in increasing AUCs over time. While this may seem like a deterioration in efficiency, these results rather point to an interplay of syntax and lexis (vocabulary expansion and more local, nominal encodings) affecting the shape of the MST. The first key factor, vocabulary size, leads to higher average surprisal. The second factor, local nominal encodings, decreases the predictive power of larger memory budgets over time. The main difference between the two registers is the impact of POS on the shape of the MST curve: Both nouns and verbs in RSC can make better use of larger memory budgets (reduce surprisal more per bit of memory) than nouns and verbs in COHA. Our statistical analysis in Section 6 suggests that in RSC, the more local encodings of noun phrases (but also verb phrases) contribute to the less pronounced deterioration of the MST compared to other POS. We take this as an indication that scientific English reacts in a specific way to increased processing effort due to expanding vocabulary: Firstly, through more local encodings of noun and verbal phrases, and secondly, through conventionalization, enabling readers to make efficient use of memory in semantically dense constructions. We also analyzed the average slope of the MST curves, capturing the information gained per bit of memory independently of absolute surprisal levels. This analysis revealed that for short memory budgets (1–3 bits), the MST remains relatively stable over time, while the MST is primarily affected at longer context windows, consistent with a broader shift toward more local encodings.

9. Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 Information Density and Linguistic Encoding.

Limitations

There are several limitations to our study. First, our analysis of scientific English distinguishes between three journals within RSC (RSC, Journal A, and Journal B). These journals reflect both different disciplines (biology and mathematics) and represent different time periods (Journals A and B are only published from 1900 onward, RSC contains all earlier publications). A more detailed analysis of the three journals could reveal variation among scientific disciplines.

Furthermore, more fine-grained distinctions by topic, author, or subfield could give additional insights into how the MST varies along these lines.

Second, our comparison contrasts these journals with the entirety of COHA, rather than with more carefully matched subsets of general English. A more nuanced comparison might better isolate register-specific effects.

Third, we did not investigate which specific documents or genres drive the observed increase in surprisal over time, nor did we examine which texts could be considered to have a particularly strong effect on the shape of the MST curve. Addressing these points would provide a more detailed understanding of the interaction between register, vocabulary growth, and communicative efficiency. Finally, although Scientific English may appear decreasingly efficient w.r.t. the MST, in our surprisal models, we have not accounted for the factors of specialization and background knowledge. This is because our modeling is based on the entire corpus, which may mask discipline-specific effects. These effects could become apparent if we were to model each discipline separately. Additionally, psycholinguistic studies on expert text processing would be necessary to draw more definitive conclusions.

10. Bibliographical References

David Banks. 2003. The evolution of grammatical metaphor in scientific writing. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pages 127–148.

Douglas Biber and Victoria Clark. 2002. Historical shifts in modification patterns with complex noun phrase structures. *English Historical Morphology. Selected Papers from 11 ICEHL, Santiago de Compostela, 7–11 September 2000*, 11:43–66.

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

TM Cover and Joy A Thomas. 2006. *Elements of information theory*. Hoboken, NJ: Wiley-Interscience.

Mark Davies. 2021. Corpus of Historical American English (COHA).

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2019. An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. In *From Data to Evidence in English Language Research*, Language and Computers, pages 258–281. Brill.

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.

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

Edward Gibson. 2000. The dependency locality theory: a distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first Mind Articulation Project Symposium*, pages 95–126. Cambridge, MA: MIT Press.

Kristina Gulordava and Paola Merlo. 2015. Diachronic Trends in Word Order Freedom and Dependency Length in Dependency-Annotated Corpora of Latin and Ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 121–130. Uppsala University, Uppsala, Sweden.

Michael Hahn, Judith Degen, and Richard Futrell. 2021. Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychological Review*, 128(4):726–756.

Michael Hahn and Richard Futrell. 2020. Crosslinguistic Word Orders Enable an Efficient Tradeoff of Memory and Surprisal. *Society for Computation in Linguistics*, 3(1).

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

Michael 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.

Marianne Hundt, David Denison, and Gerold Schneider. 2012. Relative complexity in scientific discourse. *English Language and Linguistics*, 16(2):209–240.

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. Association for Computational Linguistics.

Marie-Pauline Krielke. 2024. Cross-linguistic Dependency Length Minimization in scientific language: Syntactic complexity reduction in English

and German in the Late Modern period. *Languages in Contrast*, 24(1):133–163.

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.

Lei Lei and Ju Wen. 2020. Is dependency distance experiencing a process of minimization? A diachronic study based on the State of the Union addresses. *Lingua*, 239:102762.

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

Richard L. Lewis and Shravan Vasishth. 2005. An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval. *Cognitive Science*, 29(3):375–419.

Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2):159–191.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.

Julius Steuer, Marie-Pauline Krielke, Stefan Fischer, Stefania Degaetano-Ortlieb, Marius Mosbach, and Dietrich Klakow. 2024. Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, pages 12–23.

Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. Less is More/-More Diverse: On the Communicative Utility of Linguistic Conventionalization. *Frontiers in Communication*, 5.

Gui Wang, Hui Wang, Xinyi Sun, Nan Wang, and Li Wang. 2023. Linguistic complexity in scientific writing: A large-scale diachronic study from 1821 to 1920. *Scientometrics*, 128(1):441–460.

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. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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. Publisher: arXiv Version Number: 4.

11. Language Resource References

Davies, Mark. 2021. *Corpus of Historical American English (COHA)*. UCLA Dataverse.

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.

A. Tokenization

A.1. Tokenization Methods

In order to mitigate the problem of vocabulary expansion, we employ and independently evaluate three tokenization strategies, which all drastically reduce the number of tokens in the vocabulary and do not require a model whose parameters are mostly in the embedding layer (which would happen in

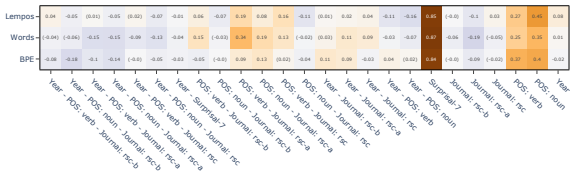


Figure 7: LME coefficients for AUC, surprisal from language models trained on lempos, word-level and BPE tokenizations. Non-significant effects in parentheses.

case of a vocabulary of about 500K tokens, as in COHA).

Word-level tokenization: This is the simplest tokenization approach and requires a few tweaks to work. We use word-level tokenization with replacement of OOV items instead of a subword tokenization method because words that are split into many subtokens due to high tokenizer fertility would be assigned higher surprisal values by default. The surprisal of these de-facto OOV items would artificially inflate our AUC measure and obscure the impact of word order on AUC.

Lempos tokenization: This tokenization approach is derived from word-level tokenization, but reduces the size of the unigram vocabulary even further by replacing word forms with a combination of the corresponding lemma and UPOS tag.

BPE tokenization: We use the default implementation of the BPE algorithm in the Hugging Face tokenizers Python package to train a tokenizer with a vocabulary size of 100K on the subsampled version of each corpus. We did not replace OOV words, as those are handled by the tokenization algorithm. An overview of the tokenization methods, thresholds and examples of a tokenized sentence can be found in Table 1.

Corpus	Tokenizer	Tokens	t_{REPL}	$\lvert V \rvert$
RSC	Lempos	2.5M	1	79K
	Word		1	74K
	BPE		0	100K
COHA	Lempos	3.5M	1	83K
	Word		3	98K
	BPE		0	100K

Table 1: Corpus sizes and data preprocessing parameters. 10% of the sampled tokens were used as a development set.

A.2. Consistency across tokenization methods

We re-fitted all LMEs with surprisals and AUCs from language models trained on word-level and BPE-tokenized versions of the subsampled corpora. We found that, while effect sizes vary greatly between tokenizations, the direction of the effects is consistent. See Figure 7 for a detailed overview of the LME coefficients for all three tokenization methods.