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

Anonymous ACL submission

Abstract

The memory-surprisal trade-off (MST) has been shown to hold cross-linguistically as a general principle of communicative efficiency that provides a processing explanation to some basic properties of language. In this paper, we explore the influence of diachronic variation on the MST. We investigate scientific English in the Royal Society Corpus (RSC) spanning from the 18$^{\text{th}}$ century to modern time; to assess the impact of intra-linguistic variation (register), we compare scientific English with "general language" using parts of the Corpus of Historical American English (COHA). We observe a clear diachronic effect for scientific English towards decreased efficiency as scientific texts shift from verbal to nominal style and the lexicon in the scientific domain expands, while in general language the effect is less pronounced.

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).

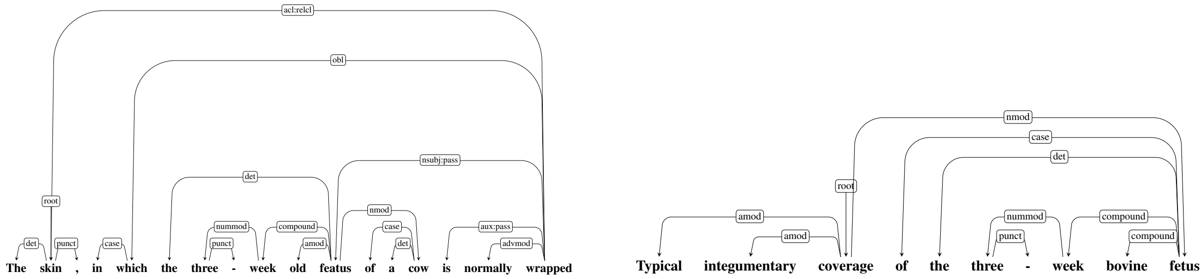
In this paper, we set out to model the impact of lexical expansion and syntactic change on communicative efficiency in terms of the memory-surprisal trade-off (MST, Hahn et al., 2021). The MST unifies two competing approaches to communicative efficiency: Surprisal theory (Levy, 2008), focusing on expectation-based efficiency, assumes that a word w_t becomes easier to predict the more context information (e.g., preceding words $w_1, \ldots, w_{t-1}$) is available. Dependency Locality Theory (Gibson, 2000) assumes that memory-based efficiency is optimized if words that are close to each other in a dependency tree (see Figure 1) are also close to each other in the surface form of a sentence, i.e., if dependency lengths between words on average are small. The MST combines these approaches by positing that for a given language, the actual word order in a sufficiently large corpus balances the requirements of predictive processing (surprisal theory) and communicative efficiency (dependency locality) by optimizing the amount of information that needs to be stored in memory to reach an average surprisal level. In their seminal paper, Hahn et al. (2021) showed that the syntax of a typologically diverse set of languages is optimized with respect to this trade-off.

We ask whether the communicative efficiency of scientific English as measured by the MST changes over time due to linguistic change, and if so, whether it becomes better or worse and which factors are most influential. We also ask whether any changes in MST are register-specific, i.e., whether scientific English is affected more than general English, and whether the language of different scientific disciplines reacts differently (due to domain-specific lexical expansion).

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 back-



(a) (b)

Figure 1: Dependency structures of (a) noun phrase with relative clause postmodification (15 words) and (b) noun phrase with multiple premodication (9 words).

ground knowledge within individual scientific disciplines, grammatically explicit constructions such as clausal subordination (Example in Figure 1a) 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 1b).

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 (Degaetano-Ortlieb and Teich, 2019) have shown that periods of *lexical expansion* are associated with increased surprisal (and thus increased processing load) (e.g. Steuer et al., 2024), while *grammatical conventionalization* leads to optimization of expectation-based processing for increasingly predictable grammatical constructions (Degaetano-Ortlieb and Teich, 2019; Degaetano-Ortlieb et al., 2019; Teich et al., 2021; Bizzoni et al., 2020).

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, 1998, 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 in specific registers (Juzek et al., 2020; 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 morphemes 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 2 illustrates the relationship between MST curves and area under the curve (AUC) for five years from the RSC: the curve for the last year (1900) 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 more optimal MST. Be-

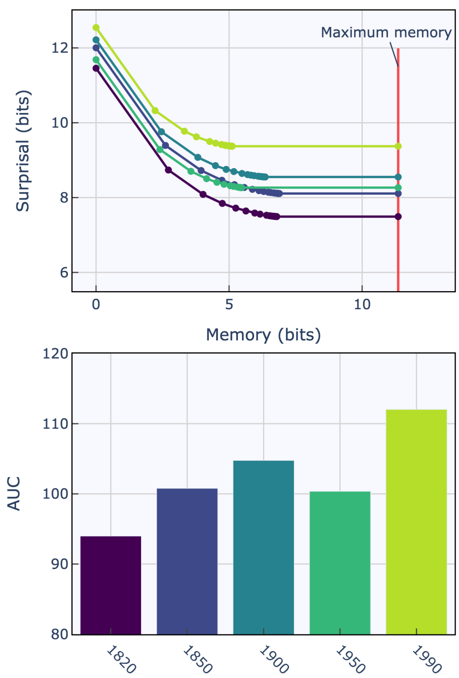


Figure 2: MST curves (1820-1990) and their respective areas under the curves (AUC). MST curves where extended to the maximum per-document memory in the RSC.

tween those years, 1950 still follows the (expected) trend of increasing unigram surprisal, but intersects the MST curve of 1850, leading to a *lower* AUC and a temporary increase in optimality w.r.t. the MST.

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 reduce the efficiency of memory-based prediction due to implicit dependencies, whereas verbal constructions support explicit, less local dependencies and benefit more from 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 1a). 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 1b), 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 out if optimization on memory efficiency develops differently in scientific vs. general English. For a more fine-grained analysis, we calculate the MST for word classes (nouns, verbs, 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**:

H1.1: Impact of average surprisal We expect the MST to deteriorate (i.e. increasing AUC) in both RSC and COHA (scientific vs. "general" English) due to the general increase in surprisal through vocabulary expansion.

H1.2: Impact of register We expect the MST to deteriorate (i.e. increasing AUC) more strongly in the RSC than for COHA due to a stronger vocabulary increase in scientific English.

H2: Difference between POS The vocabulary expansion affects the MST of nouns (i.e. increasing AUC) more than other POS, especially in scientific English.

H3: Effects of nominal style on shape of the MST curves Over time, we expect to find a weaker surprisal reduction with more bits of memory, especially in the RSC.

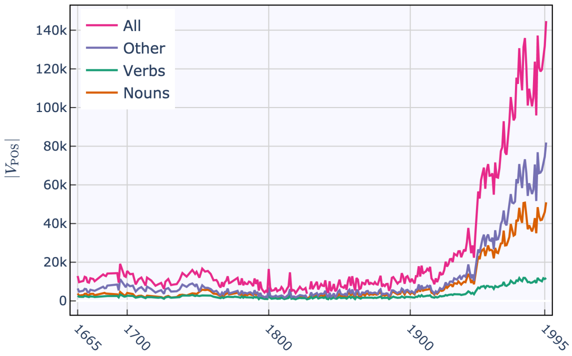


Figure 3: Increase of types per part-of-speech (POS) in the RSC over time; "Other" contains all POS except nouns and verbs.

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), 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 the 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.

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 the 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 pertainging 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 estimate surprisal values on all documents from each test year of the two corpora.

Figure 4: Memory-surprisal trade-off curves for 10 years from the 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.

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 5 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 $P = \{p_1, ..., p_{|P|}\}$ 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

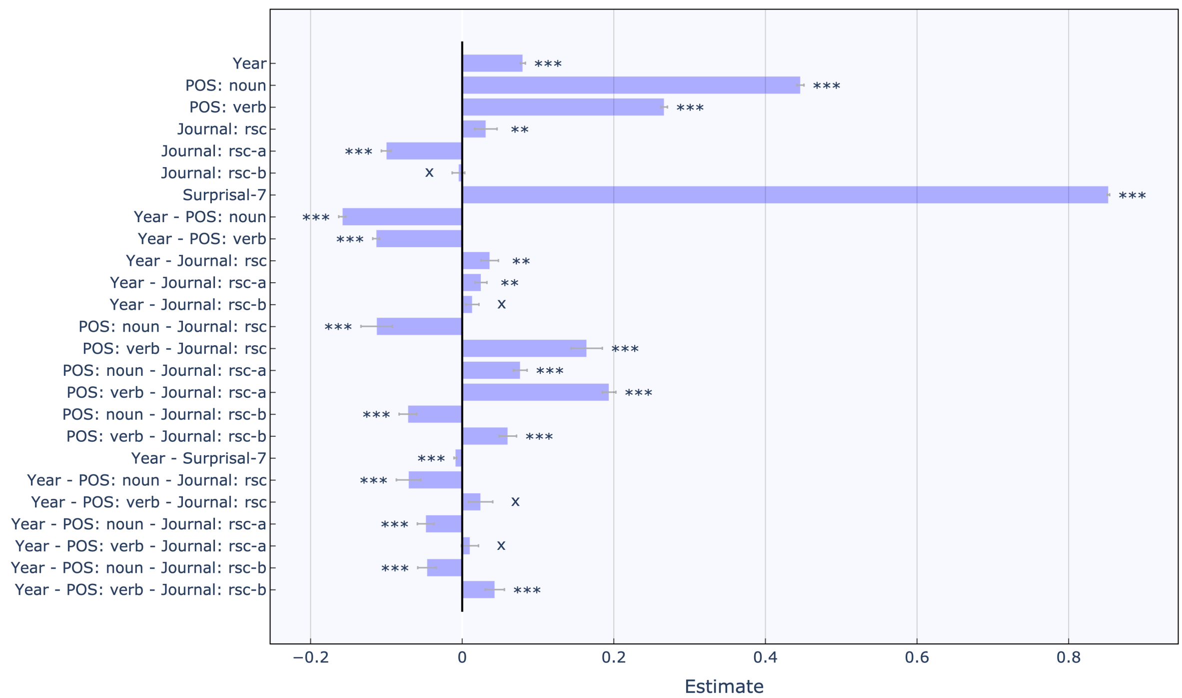


Figure 5: Effects of part of speech, journal and time on AUC for the period from 1820 to 1996. Reference levels for factor variables are "other" (POS) and "coha" (journal). Significance levels: '***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$, 'x' $p >= 0.05$. Error bars show standard error of the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

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" (not noun or verb) as the base level of the POS variable.

6 Analysis

6.1 Effect of surprisal

Overall, the observed effects are in line with our expectations. We find the strongest effect for 7-gram surprisal (Estimate: 0.8525; CI: 0.8492, 0.8557; $t = 522.77$). A positive estimate corresponds to an increase in AUC, i.e., a worse MST, while a negative estimate corresponds to a decrease in AUC and a better MST compared to the base level of the variable. Given the fact that AUC is correlated with the surprisal values at different memory budgets, the strong association between the predictor and response is plausible. Figure 5 is a detailed overview of all effects of interest, besides surprisal. 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 RSC subjournals

Comparing the language of the three subjournals of the 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

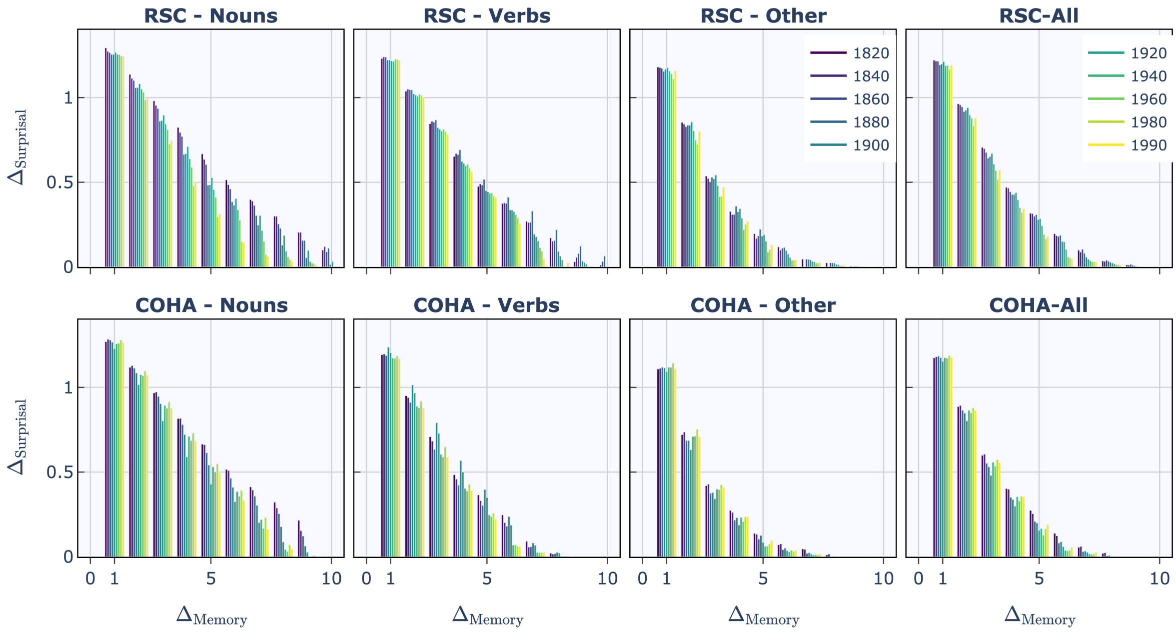


Figure 6: Average slope of the MST curve at equidistant memory intervals of 1 bit.

RSC-B. For RSC-A we find a large negative effect (Estimate: -0.1, CI: -0.11, -0.09; $t = -15.41$), showing that the language of this subjournal is optimized w.r.t. the MST compared to general English.

6.3 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 the 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; $t = -3.82$).

6.4 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 the RSC generally have a lower AUC (see Section 6.1), 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 increases drastically over time in the RSC, the syntax of scientific English is still in some sense optimized w.r.t. the MST for nouns and verbs.

7 From AUC to Shape of the MST curves

7.1 Effect of nominal style

In the previous section, we have analyzed the overall development of optimality in the two corpora as measured by the AUC. However, the AUC is only an approximation of optimality, given that MST curves whose AUC is compared are parallel in time. Furthermore, even when two curves do not cross, the degree to which more bits of memory reduce surprisal is not covered by the AUC. We will therefore analyze in more detail the individual shapes of the MST curves as well as the surprisal reduction rate per every additional bit of memory.

Looking at Figure 4, we see that the MST curves show different shapes in different years. Especially for nouns in the RSC, MST curves show an interesting picture: While surprisal in the first 100 years (1820 - 1920) continuously drops per additional bit of memory, in the last 60 years, surprisal shows very little reduction with less than 5 bits of memory. A similar trend can be observed for verbs in the

RSC and other POS, however, not as pronounced as for nouns. At the same time, nouns show a decreasing unigram surprisal, which is surprising given the fact that the number of nominal vocabulary items increases over time. It shows, however, that in the case of nouns, the increase in AUC over time is not owed to increasing 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 as 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.

7.2 Effect of NP density

This interpretation is backed by the calculation of the average slope of the memory-surprisal curves at equidistant memory intervals of one bit (see Figure 6). For nouns, less and less information is gained (or surprisal reduced) per additional bit of memory for each step of 20 years. Compared to verbs and other POS, this is especially pronounced. Comparing RSC and COHA, the slope of the MST curves levels out faster for COHA than for the RSC, i.e., the language models trained on the 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. Comparing POS, in both corpora, the temporal effect is strongest for nouns and especially pronounced in the RSC. For verbs, the slope is fairly similar across time in the 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.

8 Conclusion

We examined the communicative efficiency of scientific vs. general English over time, as measured by the Memory-Surprisal Tradeoff (MST). Our central question was whether MST optimality 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 orderings with respect to locality should yield more efficient MSTs, our findings indicate the opposite: denser encodings and vocabulary expansion over time appear to reduce optimality. Specifically, we identified two key factors: growing vocabulary size leads to higher average lexical surprisal, and less predictive contexts result in less efficient memory usage. The observed trends in AUC values suggest a general decline in optimality over time. However, this interpretation must be qualified, as vocabulary growth is an inherent feature of language evolution. Although our subsampling strategy was designed to mitigate the influence of vocabulary size on surprisal estimates, the overall trend persists. This raises important questions regarding the comparability of surprisal values across historical stages. To address this, 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 contexts (1–3 bits), the tradeoff remains relatively stable over time, suggesting that efficiency has declined primarily for longer contexts. We interpret this as evidence that English has become less optimal in terms of long-range predictability, consistent with a broader shift toward shorter, denser encodings.

Limitations

There are several limitations to our study. First, our analysis of scientific English distinguishes between three journals within the RSC (RSC, Journal A and Journal B). It is important to note that 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 efficiency 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 are driving the observed increase in surprisal over time, nor did we examine which texts could be considered particularly (non-)optimal w.r.t. the MST. Addressing these points in future work would provide a more detailed understanding of the interaction between register, vocabulary growth, and communicative efficiency. Finally, although Scientific English may appear less optimal, 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.

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.

Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. 2020. Linguistic Variation and Change in 250 years of English Scientific Writing: A Data-driven Approach. *Frontiers in Artificial Intelligence, section Language and Computation*.

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.

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

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. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

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. *Preprint*, arXiv:2003.07082.

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.

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 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 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.

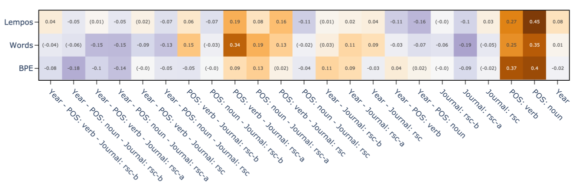


Figure 7: LME coefficients for AUC, surprisal from language models trained on lempos, word-level and BPE tokenizations. Non-significant effects in parentheses.

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.

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.