Predicting Comprehensibility in Scientific Text Based on Word Facilitation

Moritz Hartstang *Osnabrück University* mhartstang@uos.de
Martyna Plomecka *University of Zürich* martyna.plomecka@uzh.ch
Nicole Gotzner *Osnabrück University* nicole.gotzner@uos.de
Sebastian Musslick *Osnabrück University, Brown University*
sebastian.musslick@uos.de

Abstract

Comprehensible scientific writing is the basis for interdisciplinary work and science literacy. Common metrics for measuring text comprehensibility are based on word length, frequency, and predictability. Since scientific writing often contains long, rare, and low predictability words, such linguistic metrics negatively assess them without sufficiently taking the context into account. Here, we introduce an attention-based metric derived from transformer models to capture how much a word facilitates the processing of others. We find that such facilitation complements traditional linguistic metrics by explaining late reading times and neural correlates of text understanding, capturing unique variance for words that the existing linguistic metrics rate as difficult. These results hold promise for the development of an interpretable comprehensibility metric, usable for scientific writing, and potentially adaptable to individual comprehension based on vocabulary knowledge.

Keywords: text understanding, computational linguistics, transformer attention

1 Introduction

Improving the comprehensibility of scientific writing is increasingly vital. Today's scientific challenges demand interdisciplinary collaboration and closer interaction between science and society. Yet, despite the growth of open and accessible science [Klebel et al., 2025] levels of scientific literacy are declining [OECD, 2023]. Addressing this gap requires writing that communicates complex ideas in ways that remain accessible across disciplines and to broader audiences. A key step toward this goal is the ability to quantify and predict text comprehension, providing objective measures that can guide efforts to make scientific writing more accessible.

Traditional approaches to measuring comprehensibility have clear limitations for scientific texts. Readability formulas, which rely on word length and frequency [Flesch, 1948, Dale and Chall, 1948], often misclassify texts as difficult simply because they contain long or rare words. However, long and rare words are common in scientific contexts, even when the text is understandable. Predictability metrics, based on how likely humans or language models (LLMs)

are to anticipate a word [de Varda et al., 2024, Goodkind and Bicknell, 2018], also fall short: in scientific writing, a word may be hard to predict without domain expertise but still easy to comprehend once it was explained.

In this work, we derive a novel method for predicting text comprehension for scientific texts based on the attention mechanism. Attention weights quantify how strongly each word contributes to processing other words in context [Vaswani et al., 2017]. While primarily used to predict the next word, attention captures richer structural information than word probabilities alone [Clark et al., 2019]. Although the interpretability of attention weights remains debated [Jain and Wallace, 2019, Wiegreffe and Pinter, 2019], they offer a promising signal for estimating how much a word facilitates the processing of surrounding words. Here, we evaluate attention-based word facilitation as a metric of text comprehension. By testing its ability to explain variance in behavioral and neural correlates of text understanding beyond traditional metrics, we find that it provides additional information, opening a new AI-driven lens through which to study text comprehension.

2 Methods

We quantify word facilitation as the total amount of attention it allocates to other words in a transformer model, with L layers and H attention heads per layer. For a word at position i in a sentence of N words (denoted w_i), i indexes the target word, while j indexes all other words in the sentence. The score thus reflects how much attention the target word w_i allocates to each context word w_j. Its facilitation score is

$$\text{Facilitation}(w_i) = z_s\left(\sum_{j=1}^{N} D_{ij} \sum_{\ell=1}^{L}\sum_{h=1}^{H} \alpha_{\ell h}\, \widehat{A}_{ij}^{(\ell,h)}\right),$$

where $z_s(x) = \frac{x-\mu_s}{\sigma_s}$ is the sentence-level z-score with μ_s, σ_s computed across words in the same sentence; D_{ij} encodes directional weights and excludes self-attention; $\widehat{A}^{(\ell,h)}$ is token-level attention aggregated to words (row-normalized); and $\alpha_{\ell h}$ are entropy-based weights defined as

$$\alpha_{\ell h} = \frac{(\bar{H}_\ell + \varepsilon)^{-\gamma}}{\sum_r (\bar{H}_r + \varepsilon)^{-\gamma}} \cdot \frac{(\bar{H}_{\ell h} + \varepsilon)^{-\gamma}}{\sum_k (\bar{H}_{\ell k} + \varepsilon)^{-\gamma}}.$$

We apply such inverse-entropy weighting via $\alpha_{\ell h}$ to give more weight to attention heads with focused (low-entropy) distributions, thus reducing the influence of diffuse attention patterns. Critically, this weighting is directional, emphasizing facilitation flowing from earlier to later words, reflecting the natural progression of reading:

$$D_{ij} = \begin{cases} w_r, & j > i \ \ \text{(right/future)} \\ w_\ell, & j < i \ \ \text{(left/past)} \\ 0, & j = i \ \ \text{(self)} \end{cases} \qquad \text{with } w_r \geq w_\ell \geq 0.$$

Punctuation and self-attention contributions are excluded. For words split into multiple subword tokens, attention values are averaged across tokens to ensure comparability. Finally, scores are normalized within each sentence to control for sentence-level variability.

We compared the explanatory power of the facilitation metric (parameterized with $\gamma = 1.5$, $w_r = 1.0$, $w_\ell = 0.5$) against that of traditional linguistic metrics, such as length, frequency (SUBTLEX-US corpus [Brysbaert et al., 2012]), surprisal (negative log probability derived from GPT2), and predictability (human rating on a 1–5 scale) using the data set from [de Varda et al., 2024]. The data set contains behavioral measures of text comprehension (including eye-tracking measures) for 205 English sentences, as well as neural correlates of text comprehension recorded for each word. Because later stages of reading and corresponding neural signals are thought to reflect the integration of a word into its broader sentence context [Radach and Kennedy, 2013, Kutas and Federmeier, 2011, Thornhill and Van Petten, 2012, Aurnhammer et al., 2021], we focused our comparisons on the following dependent measures of late reading time and neural activity:

- **Go-past reading time**: total fixation time on a word, including regressions (i.e., time from first entering a word until moving past it, including rereading; reflects effort to integrate the word into the sentence) [Radach and Kennedy, 2013].
- **N400 amplitude**: a centro-parietal negativity occuring around 400 ms after word onset, increases when a word is unexpected or hard to fit semantically with prior context [Kutas and Federmeier, 2011].
- **EPNP**: an early frontal positivity after the N400 (500 ms), reflecting violations of specific lexical expectations about which word would occur [Thornhill and Van Petten, 2012].
- **P600 amplitude**: a late positivity (600 ms) after word onset, associated with syntactic reanalysis and integrating a word into sentence structure [Aurnhammer et al., 2021].
- **PNP**: a later positivity (600–900 ms) after word onset, occurs when an unexpected but plausible word forces re-evaluation of the sentence meaning [Thornhill and Van Petten, 2012].

As traditional linguistic metrics negatively assess scientific writing with long, less frequent, and less predictable words without taking the context into account sufficiently, we examined whether the facilitation metric captures additional variance of dependent measures. To address this, we computed the incremental R^2 (ΔR^2) of a metric after controlling for the others. Critically, we performed this analysis for different words, grouping them based on their length, frequency, surprisal, or predictability into quantiles. This results in unique variance explained for each metric, as a function of these word features.

3 Results

Figure 1 depicts that the facilitation metric explains unique variance for long, less frequent, and less predictable words. Facilitation even explained the most unique variance for first fixation reading times (RTfirstfix) in bins with the longest words. A measure that reflects initial lexical access [Demberg and Keller, 2008] and is normally dominated by length and frequency.

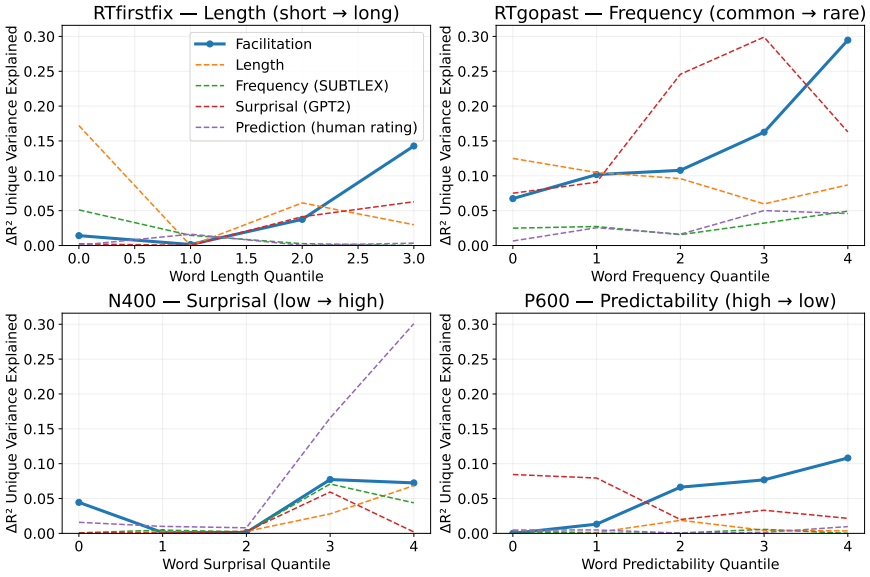

Figure 1: The unique variance (ΔR^2) explained by a predictor for different dependent measures of text comprehensibility (first fixation duration (RTfirstfix), go-past reading time (RTgopast), N400 and P600 amplitude). Unique variance is shown across word property quantiles.

4 Conclusions

We introduced an attention-based predictor of text comprehension, capturing how strongly each word contributes to processing other words in context within a transformer model. Facilitation complements the traditional linguistic metrics used to predict text comprehensibility by explaining late reading times and neural measures reflecting integrative processes. Facilitation especially explained unique variance for words that the existing linguistic metrics rate as difficult, suggesting that it may be well suited predicting comprehensibility of scientific texts. Taken together, these findings highlight word facilitation as a promising and interpretable metric that not only advances the measurement of scientific text comprehensibility but also may offer a pathway for tailoring texts to readers' vocabulary by limiting reliance on words beyond their knowledge.

5 Limitations

Our initial analyses rely on a psycholinguistic dataset of short English sentences. While this setup enables precise mapping between model-derived metrics and behavioral or neural correlates, it limits the ecological validity of our findings. Future work should extend the approach to longer and more complex scientific and medical texts to test its robustness in naturalistic contexts.

Other limitations pertain to the robustness of the introduced metric. The current implementation relies on a single transformer architecture and static attention representations. Because attention patterns can differ substantially across layers, model architectures, and training objectives, cross-model validation will be essential to establish whether facilitation reflects a general property of contextual processing or merely a model-specific artifact. Moreover, the introduced metric treats comprehensibility as uniform across readers, neglecting individual differences. In practice, comprehension depends on prior knowledge and linguistic expertise. Incorporating personalized comprehension profiles would allow the metric to better capture real-world variability and predict individual text adaptation.

Finally, the investigated relationship between the attention-based metric and comprehension remains correlational. Although this metric reveals links between the attention mechanism and traditional measures of human text comprehension, its explanatory power remains to be studied.

6 Outlook

This study represents an initial step toward linking the attention mechanism implemented in transformers to psycholinguistic measures of human text comprehension. In future work, we seek to extend these findings in two directions: (1) by translating the facilitation metric into applied tools for scientific and health communication, and (2) by strengthening its theoretical foundation as an interpretable measure of comprehension in both humans and LLMs.

On the applied side, integrating the metric into writing and reading support systems could facilitate understanding of (and compliance with) complex information. For example, scientific writing assistants might highlight terms that impede comprehension, while adaptive health information systems could automatically tailor explanations to readers' vocabulary and domain knowledge. Such applications move beyond the broad goal of increasing science literacy toward concrete, user-centered interventions.

References

C. Aurnhammer, F. Delogu, M. Schulz, H. Brouwer, and M. W. Crocker. Retrieval (n400) and integration (p600) in expectation-based comprehension. *Plos one*, 16(9):e0257430, 2021.

M. Brysbaert, B. New, and E. Keuleers. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44(4):991–997, 2012.

K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

E. Dale and J. S. Chall. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54, 1948.

A. G. de Varda, M. Marelli, and S. Amenta. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, 56(5):5190–5213, 2024.

V. Demberg and F. Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210, 2008.

R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3): 221, 1948.

A. Goodkind and K. Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18, 2018.

S. Jain and B. C. Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

T. Klebel, V. Traag, I. Grypari, L. Stoy, and T. Ross-Hellauer. The academic impact of open science: a scoping review. *Royal Society Open Science*, 12(3): 241248, 2025.

M. Kutas and K. D. Federmeier. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62(1):621–647, 2011.

OECD. *PISA 2022 Results (Volume I) The State of Learning and Equity in Education: The State of Learning and Equity in Education*. OECD Publishing, 2023. ISBN 9789264351288.

R. Radach and A. Kennedy. Eye movements in reading: Some theoretical context. *Quarterly Journal of Experimental Psychology*, 66(3):429–452, 2013.

D. E. Thornhill and C. Van Petten. Lexical versus conceptual anticipation during sentence processing: Frontal positivity and n400 erp components. *International journal of psychophysiology*, 83(3):382–392, 2012.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

S. Wiegreffe and Y. Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.