Quantifying the Development of Communicative Efficiency in Scientific English

Julius Steuer, Marie-Pauline Krielke, Stefania Degaetano-Ortlieb, Elke Teich, Dietrich Klakow

jsteuer@lsv.uni-saarland.de

Scientific English is characterized by high informational density, technicality and abstractness, making it efficient for expert-to-expert communication (Banks, 2003; Biber & Gray, 2011, 2016). Over time, scientific English has evolved to balance lexical innovation (e.g., new technical terms) with grammatical conventionalization to ensure communicative efficiency, e.g., favoring nominal over verbal structures (Degaetano-Ortlieb & Teich, 2019; Teich et al., 2021). In this work, we explore the diachronic mechanism(s) of communicative efficiency focusing on sentence processing.

Incremental sentence processing is assumed to depend on two factors: working memory (Gibson, 1998; Lewis & Vasishth, 2005) and expectation (Hale, 2001; Levy, 2008). Both are involved in linguistic change in scientific English (Degaetano-Ortlieb & Teich, 2019; Juzek et al., 2020),but how they interact diachronically is still an open question. To address this, we use the Memory-Surprisal Tradeoff (MST; Hahn et al., 2021), which specifically models the interaction between these two factors. The MST indicates how much information a reader from a specific period needs to store in memory to reduce surprisal maximally compared to a reader from another period. We assume the MST to change over time as the linguistic code adapts to periods of innovation and conventionalization, that is, we expect the MST of some time periods to be less optimal than the MST in others, depending on the rate of innovation.

As a data set, we use the Royal Society Corpus (RSC; Fischer et al., 2020), covering scientific publications from the Royal Society from 1665 to 1996. We split each decade into a train and test section, and then estimated token-level surprisal on the test set from a language model trained on the train set using the base version of the OPT architecture (Zhang et al., 2022).

Figure 1 shows MST curves for four decades (each 100 years apart). The 17thc. shows the best MST, achieving with one bit of memory the lowest average surprisal ($<$7). In 1785-1795, the decade of the chemical revolution (Degaetano-Ortlieb & Teich, 2019), the MST deteriorates drastically: with the same amount of memory (1 bit), a much higher surprisal is needed on average (around 8 bits), possibly due to a vocabulary expansion resulting from the new discoveries at the time. In 1885-1895, the MST improves, which might be related to a period of conventionalization in the 19thc. (cf. Degaetano-Ortlieb & Teich, 2019). In 1985-1995, the MST deteriorates again, reflecting the immense increase in scientific activities in the 20thc. leading to the further expansion of a specialized vocabulary (Steuer et al., 2024) indicating specialization and diversification trends.

Overall, our findings suggest that during periods of innovation and specialization lexical expansion is rather disadvantageous to the MST. To obtain a more comprehensive picture, we want to compare (a) rather conventionalized patterns with a high degree of formulaicity (e.g., it is ADJECTIVE to/that, passive constructions), which should show an improvement of the MST, and (b) lexically productive nominal constructions (e.g. nominal compound, noun-of-noun pattern), which should show a comparatively less favourable MST. This comparison will allow us to further inspect the diachronic mechanisms of communicative efficiency at work over time.

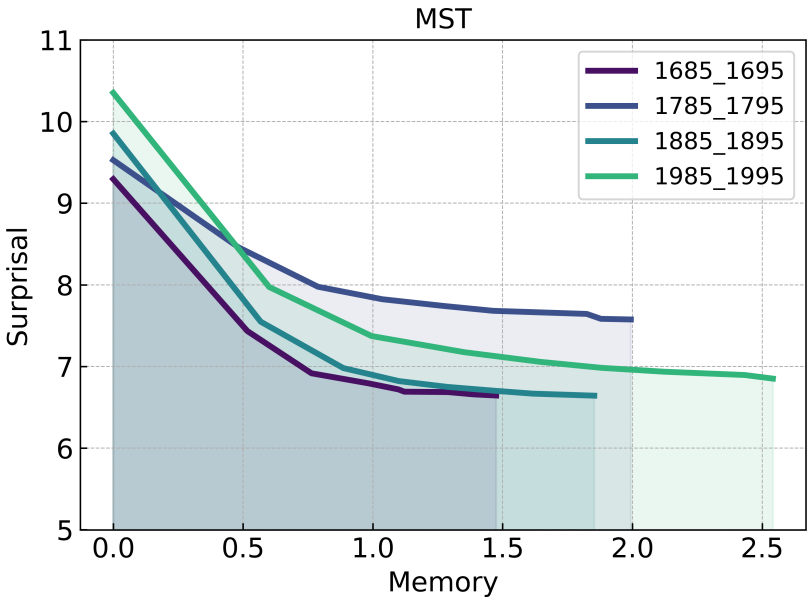

Figure 1: Memory-Surprisal Trade-Off (in bits) for four selected decades in the RSC, including the decade marking the end of the chemical revolution (1785-1795).

References

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

Biber, D., & Gray, B. (2011). The historical shift of scientific academic prose in English towards less explicit styles of expression. *Researching specialized languages*, *47*, 11.

Biber, D., & Gray, B. (2016). *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge University Press.

Degaetano-Ortlieb, S., & Teich, E. (2019). Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*.

Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020). The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 794–802.

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

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

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Juzek, T. S., Krielke, M.-P., & Teich, E. (2020). Exploring diachronic syntactic shifts with dependency length: The case of scientific English. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, 109–119.

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

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Steuer, J., Krielke, M.-P., Fischer, S., Degaetano-Ortlieb, S., Mosbach, M., & Klakow, D. (2024, May). Modeling diachronic change in English scientific writing over 300+ years with transformer-based language model surprisal. In P. Zweigenbaum, R. Rapp, & S. Sharoff (Eds.), *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024* (pp. 12–23). ELRA; ICCL.

Teich, E., Fankhauser, P., Degaetano-Ortlieb, S., & Bizzoni, Y. (2021). Less is more/more diverse: On the communicative utility of linguistic conventionalization (A. Benîtez-Burraco, Ed.). *Frontiers in Communication, Section Language Sciences*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models [Publisher: arXiv Version Number: 4].