Communicative efficiency in scientific English through the lens of memory-surprisal trade-off

Julius Steuer, Marie-Pauline Krielke, Stefania Degaetano-Ortlieb, Elke Teich & Dietrich Klakow
Universität des Saarlandes
jsteuer@lsv.uni-saarland.de, mariepauline.krielke@uni-saarland.de,
s.degaetano-ortlieb@uni-saarland.de, e.teich@uni-saarland.de,
dietrich.klakow@lsv.uni-saarland.de

Through its evolution towards highly dense, technical, and abstract structures, Scientific English has become optimized for expert communication. Its improved communicative efficiency is a result of balancing lexical innovation (e.g., new technical terms) with grammatical conventionalization favoring nominal over verbal structures (Degaetano-Ortlieb & Teich 2019).

This study investigates diachronic mechanisms of communicative efficiency, focusing on sentence processing, shown to reliy on working memory (Gibson 1998) and expectation (Hale 2001). We employ the Memory-Surprisal Tradeoff (MST; Hahn et al. 2021) to model their interaction over time.

Using the Royal Society Corpus (RSC; Fischer et al. 2020) covering 1665–1996, we trained language models by decade to estimate token-level surprisal. MST curves reveal diachronic trends: optimal in the 17th century (<7 bits with 1-bit memory), deteriorating during the chemical revolution (1785–1795) due to vocabulary expansion (Degaetano-Ortlieb & Teich 2019), improving in the 19th century during conventionalization, and deteriorating again in 1985–1995 with increased scientific activity and vocabulary growth.

Our findings suggest that lexical expansion during innovation and specialization negatively impacts the MST. Future comparisons of conventionalized patterns (e.g., passive constructions) and lexically productive nominal structures will refine understanding of communicative efficiency mechanisms over time.

References. Degaetano-Ortlieb, S. & E. Teich (2019). Toward an optimal code for communication: the case of scientific English. *Corpus Linguistics and Linguistic Theory* 18(1): 175-207. • Fischer, S., J. Knappen, K. Menzel & E. Teich (2020). The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 794-802. • Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68(1), 1-76. • Hahn, M., J. Degen & R. 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. • Hale, J. (2001). A probabilistic Earley Parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics.*