Information-Theoretical Insights into the Diachronic Behavior of Multi-Word Expressions in Scientific English

Diego Alves*[1] and Bagdasarov Sergei*[1]

[1]University of Saarland / Universität des Saarlandes [Homburg, Germany] – Allemagne

Résumé

Multi-word expressions (MWEs) are sequences of words perceived either as wholes or with highly predictable transitions from one word to the next. Their use in scientific writing is particularly interesting because MWEs help smooth the information load more evenly across sentences (Conklin & Schmitt, 2012). Previous studies (e.g., Teich et al., 2021; Degaetano-Ortlieb & Teich, 2016) have used various information-theoretical principles and measures to describe phenomena such as conventionalization and standardization, showing that scientific English has evolved into an optimized code for expert-to-expert communication. However, these works do not consider multi-word or formulaic expressions in the context of diachronic change.

Our aim, therefore, is to provide an overview of diachronic shifts in scientific English, incorporating different information-theoretical measures while treating multi-word expressions as single linguistic units. To this end, we use the Royal Society Corpus (RSC; Fischer et al., 2020), a diachronic corpus of scientific English spanning the period from 1665 to 1996. The corpus is based on the *Philosophical Transactions and Proceedings of the Royal Society of London*, and comprises 47,837 texts (295,895,749 tokens).

First, we show that, for different categories of MWEs, the surprisal delta (defined as the surprisal of the last MWE token minus that of the first, following Shannon, 1948) is typically negative, confirming our hypothesis regarding the predictability of transitions within MWEs. We then demonstrate that the paradigmatic variability of MWEs decreases over time. This refers to the range of linguistic alternatives available in similar syntagmatic contexts and is measured using entropy over a probability distribution based on the likelihood of selecting one word over others within a given lexical neighborhood in the embedding space. In contrast, the syntagmatic productivity of MWEs increases over time. This reflects the ability of MWEs to combine with other words in various syntactic contexts and is measured using entropy in left and right contexts within a three-word window on each side (Alves et al., 2025).

These findings suggest that MWEs tend to become more semantically specific, which reflects a process of standardization. At the same time, the increased syntagmatic productivity points to a process of conventionalization, understood here as the widespread use of MWEs by the scientific community.

References:

*Intervenant

Alves, D., Fischer, S., & Teich, E. (2025, May). Syntagmatic Productivity of MWEs in Scientific English. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)* (pp. 1-6).

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual review of applied linguistics*, *32*, 45-61.

Degaetano-Ortlieb, S., & Teich, E. (2016, August). Information-based modeling of diachronic linguistic change: from typicality to productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 165-173).

Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020, May). 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* (pp. 794-802).

Teich, E., Fankhauser, P., Degaetano-Ortlieb, S., & Bizzoni, Y. (2021). Less is more/more diverse: on the communicative utility of linguistic conventionalization. *Frontiers in Communication*, *5*, 620275.
Shanon, C. (1948). A mathematical theory of communication. The Bell systems technical journal, 27. *Mathematical Reviews (MathSciNet).*