Paradigmatic Variability of Multi-word Expressions in Scientific English

Diego Alves[1] Stefan Fischer[1] Elke Teich[1]

[1]Saarland University, Germany

In this study, we analyze the paradigmatic variability (i.e., the sets of linguistic options available in a given or similar syntagmatic contexts) of different categories of multi-word expressions (MWEs) in the domain of scientific writing, inspecting diachronic changes from the mid-17th century to today. 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 contribute to smoothing the information load over a message (Conklin & Schmitt, 2012). Teich et al. (2021), using embedding spaces and entropy measures to estimate paradigmatic variability, observed a reduction in this dimension for different parts-of-speech, indicating a continuous, diachronic process of conventionalization that serves to manage linguistic variability in the interest of cognitive resource efficiency. Our hypothesis is that different categories of MWEs present lower paradigmatic variability due to their semantic characteristics compared to analogous expressions, thus, contributing even more to conventionalization.

To test this hypothesis, we first extracted and classified the MWEs from an extensive diachronic dataset of English scientific texts, the Royal Society Corpus (RSC) into six categories following the work proposed by Alves et al. (2024): (1) compounds, composed of sequence of nouns (e.g., *orange juice, sea salt*); (2) flat, sequences of proper nouns and names of places and institutions (e.g., *Isaac Newton, New York*); (3) phrasal verbs (e.g., *carry out, shut down*); (4) fixed, used for certain fixed grammaticalized expressions which tend to behave like function words (e.g., *due to, in spite of*); (5) academic formulaic expressions, list of MWEs from the Academic Formulas List (Simpson-Vlach & Ellis, 2010) (e.g., *on the other hand, a kind of*); and (6) miscellaneous MWEs extracted from the RSC using the Partitioner tool (Williams, 2016) (e.g., at first sight, give rise). Then, we processed the RSC texts by connecting the tokens belonging to MWEs and proceeded with the calculation of the embedding space using structured skip-grams. The paradigmatic variability of a word over time was calculated following the method introduced by Teich et al., (2021), which defines it as the entropy over a probability distribution, based on the probability of a word from a specific neighbourhood being chosen instead of the other words in the same area.

Figure 1a shows that up to 1940, compounds have lower paradigmatic variability than nouns, with the same decreasing tendency, and flat MWEs present lower values than proper nouns, however, with peaks in 1810 and 1820. In Figure 1b, it is possible to notice that although phrasal verbs start with a higher paradigmatic variability when compared to other verbs, from 1750 on, the inverse is observed, with phrasal verbs presenting a considerable decreasing tendency regarding paradigmatic variability in the twentieth century. As shown in Figure 2a, academic formulaic expressions and fixed MWEs present a quite stable paradigmatic variability in time, with lower values when compared to adverbs and function words. Finally, Figure 2b shows that the other MWEs category presents similar behavior to function words and adverbs. Thus, overall, we can conclude that the conventionalization process throughout time regarding the lexicon in the scientific domain is even more evident when MWEs are considered as whole units.

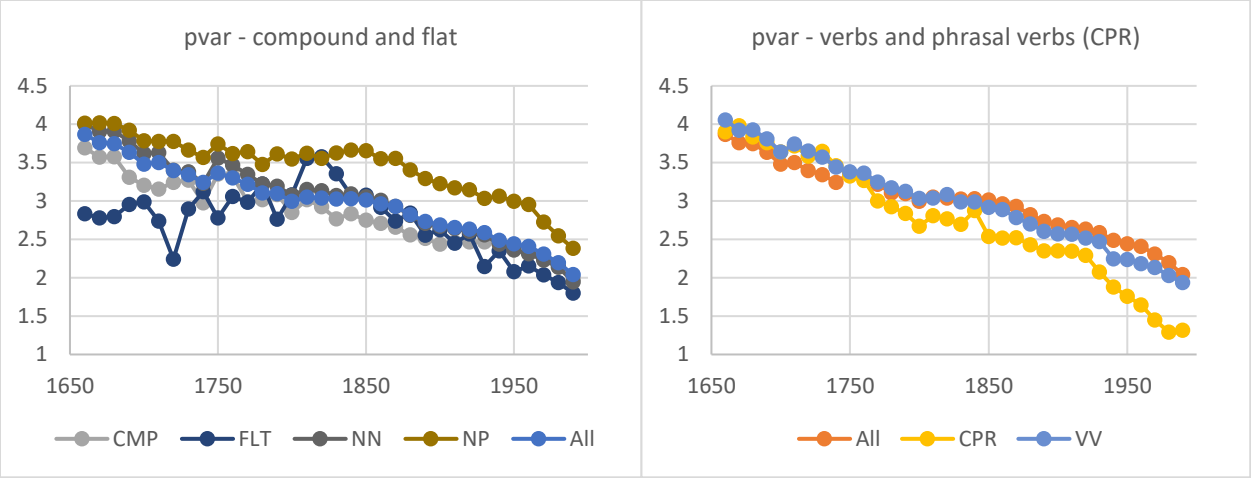

Figure 1. Paradigmatic variation per decade of: a) Compounds (CMP), Flat MWEs (FLT), Nouns (NN), Proper Nouns (NP), and All words in the embeddings space; and b) Phrasal Verbs (CPR), other verbs (VV), and All words in the embeddings space.

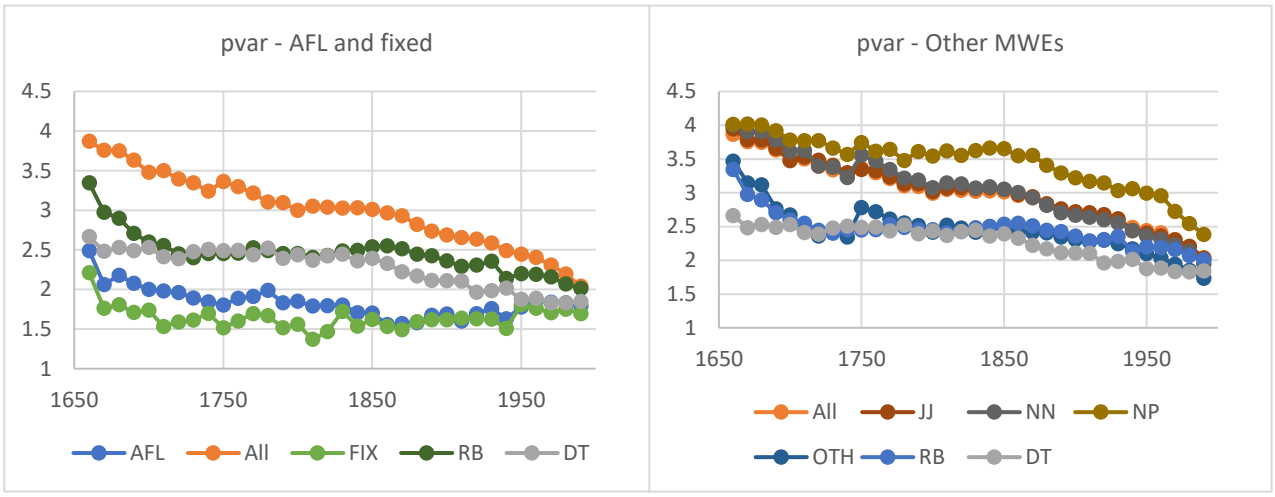

Figure 2. Paradigmatic variation per decade of: a) Academic formulaic expressions (AFL), fixed MWEs (FIX), Adverbs (RB), function words (DT), and All words in the embeddings space; and b) other MWEs (OTH), Adjectives (JJ), Nouns (NN), Proper Nouns (NP), Adverbs (RB), function words (DT), and All words in the embeddings space.

References:

Alves, D., Fischer, S., Degaetano-Ortlieb, S., & Teich, E. (2024, March). Multi-word expressions in English scientific writing. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)* (pp. 67-76).

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

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.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics*, *31*(4), 487-512.

Williams, J. R. (2016). Boundary-based MWE segmentation with text partitioning. *arXiv preprint arXiv:1608.02025*.