Syntagmatic Productivity of MWEs in Scientific English

Diego Alves[1], Stefan Fischer[2], Elke Teich[3]

Saarland University, Saarbrücken - Germany

diego.alves@uni-saarland.de, stefan.fischer@uni-saarland.de, e.teich@mx.uni-saarland.de

Abstract

This paper presents an analysis of the syntagmatic productivity (SynProd) of different classes of multiword expressions (MWEs) in English scientific writing over time (mid 17th to 20th c.). SynProd refers to the variability of the syntagmatic context in which a word or other kind of linguistic unit is used. To measure SynProd, we use entropy. The study reveals that, similar to single-token units of various parts of speech, MWEs exhibit an increasing trend in syntagmatic productivity over time, particularly after the mid-19th century. Furthermore, when compared to similar parts of speech (PoS), MWEs show a more pronounced increase in SynProd over time.

1 Introduction

In this paper, we examine the syntagmatic productivity of multiword expressions (MWEs) in English scientific writing, focusing on diachronic changes from the mid-17th century to the present. The syntagmatic productivity of a word refers to its ability to combine with other words in various syntactic contexts to form meaningful and coherent expressions. We use entropy to measure how often and in which ways a word can appear in different syntagmatic (sequential) relationships within larger constructions.

From a communicative perspective, multiword expressions play an important role in language efficiency because they are usually highly conventionalized. MWEs consist of combinations of words that are mutually highly predictable and often processed as single chunks, providing a significant processing advantage for language users. Their use in scientific writing is particularly noteworthy, given the high informational load typical of the scientific domain, where MWEs function as tools to smooth the information density over a message (Conklin and Schmitt, 2012).

It has been shown that scientific writing becomes increasingly conventionalized over time (see e.g., Degaetano-Ortlieb and Teich (2019) and Teich et al. (2021)), and that different classes of MWEs exhibit distinct diachronic tendencies in terms of association measures (Alves et al., 2024b) and discourse functions (Alves et al., 2024a).

In this study, our aim is to use entropy as a measure to analyze changes in the syntagmatic productivity of different classes of MWEs over time, comparing them to changes in individual tokens within similar parts of speech (e.g., compounds compared to nouns, and phrasal verbs compared to single-token verbs). Our hypothesis is that, due to a conventionalization process regarding the usage of MWEs, the syntagmatic productivity of these constructions presents a more pronounced increase over time when compared to their single-token counterparts.

The remainder of the paper is organized as follows. In Section 2, we discuss related work on the characterization of MWEs in English scientific writing. Sections 3 and 4 present our methods and results, respectively. We conclude with a summary of our findings and perspectives for future work in Section 5.

2 Related Work

From a linguistic standpoint, numerous corpus-based studies have explored MWEs across various registers, including the scientific domain (e.g., Biber and Barbieri (2007); Hyland (2008); Liu (2012)). Some of these studies provide lists of MWEs used in academic texts, which are extracted using corpus-based methods such as frequency and mutual information (e.g., Simpson-Vlach and Ellis (2010)). However, these studies primarily focus on synchronic analysis and provide valuable data for manuals aimed at improving writing skills.

Regarding NLP research, most studies focus on

Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), pages 1–6

the correct identification and extraction of MWEs (Ramisch et al., 2023). The PARSEME initiative (Savary et al., 2015) provides valuable corpora and guidelines for annotating MWEs, however, their approach is restricted to verbal MWEs and the available corpora concern only recent texts.

A characterization of different classes of MWEs in scientific English, based on dimensions of information (i.e., dispersion and association), was proposed by Alves et al. (2024a). The authors demonstrated that specific formulaic expressions commonly used in scientific writing exhibit a stronger diachronic tendency to increase the association between the units forming the MWEs.

Moreover, Alves et al. (2024b) demonstrated that different types of MWEs, used for specific discourse functions (e.g., referential expressions and discourse organizers), exhibit distinct diachronic changes that are linked to the linguistic needs of different time periods.

As shown by Ramisch et al. (2023), the identification and evaluation of MWEs can be highly problematic, especially when dealing with specific registers, as is the case in our study. The studies in the last two paragraphs demonstrate that the methods used in our analysis are quite robust for identifying MWEs in a diachronic scientific corpus of English.

Regarding the syntagmatic productivity of different parts of speech in scientific writing, it has been shown that from 1660 to 1920, all parts of speech exhibit an increasing tendency, with a more pronounced slope starting around 1840 (Fankhauser, 2025). However, in this study, MWEs were not considered.

3 Methods

3.1 Data

In our analysis, we use the Royal Society Corpus (RSC) 6.0[1], a diachronic corpus of scientific English spanning the period from 1665 to 1996. This resource consists of 47,837 texts (295,895,749 tokens), primarily scientific articles from various fields, including mathematics, physical sciences, and biology. It is based on the Philosophical Transactions and Proceedings of the Royal Society of London (Fischer et al., 2020). The distribution of texts per discipline over time was not controlled in this analysis; this issue will be addressed in future work.

The corpus was parsed with the Stanza tool (Qi et al., 2020) using the combined model for the English language trained on different UD corpora (i.e., EWT, GUM, GUMReddit, PUD, and Pronouns). To identify the different classes of MWEs in the RSC, we followed the methodology proposed by Alves et al. (2024a). Once identified, the MWEs were combined into a single token (with spaces between tokens replaced by a character not seen in the corpus: ||) and labelled according to the classes described below.

- compound - combinations of tokens that morphosyntactically behave as single words (e.g., *water content, sea waves*)
- flat - this relation combines elements of an expression where none of the immediate components can be identified as the sole head using standard substitution tests. For example: *Hillary Clinton and San Francisco*
- phrasal verb (e.g., *shut down* and *find out*)
- fixed - used for certain fixed grammaticalized expressions which tend to behave like function words (e.g., *because of, in spite of, as well as*).
- Academic Formulas List (AFL) - list of formulaic expressions proposed by Simpson-Vlach and Ellis (2010) automatically extracted from academic texts (e.g., *in terms of, at the end of, whether or not*)

In total, 3,147,703 types of MWEs were identified in our corpus. The distribution of these types across the different classes is presented in Table 1.

Class	Number of Types
`compound`	2,523,696
`flat`	604,057
`phrasal verb`	16,337
`fixed`	3,107
`AFL`	506

Table 1: Distribution of the MWEs types in the RSC according to their MWE class.

3.2 Syntagmatic Productivity

As previously mentioned, the syntagmatic productivity of a word refers to its ability to combine

[1] https://fedora.clarin-d.uni-saarland.de/rsc_v6/

with other words in various syntactic contexts and form meaningful and coherent expressions. This can be measured using entropy as described by Fankhauser (2025): The syntagmatic productivity of a term is the entropy over all syntagmatic neighbours of a word x within a contextual window C_x of +/- 3 (see Formula 1).

$$SynProd(x) = -\sum_{c_i \in C_x} p(c_i|x) log(p(c_i|x)) \quad (1)$$

Entropy is a measure of uncertainty or variability in a system, and in the context of syntagmatic productivity, it quantifies the diversity of words that co-occur with a given term. A higher entropy value indicates that a word appears in a wide range of syntactic contexts with many different neighbors, suggesting greater syntagmatic flexibility. On the other hand, lower entropy implies that the word tends to co-occur with a more limited set of words, reflecting restricted combinatory potential. By capturing the distributional diversity of a word's syntagmatic associations, entropy provides a numerical representation of how productively a word participates in different constructions within a given corpus.

For each class of MWEs, we calculated the average syntagmatic productivity per decade of the RSC. Using a contextual window of 3, we define L3 as the syntagmatic productivity in the left context of each textual unit (single tokens and MWEs), and R3 as the syntagmatic productivity in the right context of each textual unit.

4 Results

4.1 Overall Syntagmatic Productivity

Figure 1 shows the average syntagmatic productivity of different classes of MWEs identified in the RSC, analyzed per decade. These results are compared to the average overall syntagmatic productivity of all other tokens in the text (i.e., tokens that are not part of MWEs, labelled as *All*).

As expected, all classes of MWEs exhibit an increasing tendency regarding both R3 and L3, with a more pronounced rise beginning in the mid-19th century. This pattern aligns with the observations reported by Fankhauser (2025) in their analysis of different parts of speech up to 1920. The graphs show that this rapid increase continues throughout the entire 20th century, not only for the different classes of MWEs but also for all other parts of speech. This suggests an expansion in the range of contexts where MWEs are employed.

Moreover, we observe that, although not identical, the R3 and L3 curves exhibit similar patterns across all analyzed cases. When compared to the curve representing the syntagmatic productivity of other parts of speech, it becomes evident that fixed and AFL MWEs display higher SynProd values, indicating more diverse usage. This can be attributed to the domain-independent, functional nature of these expressions, as they do not refer to a specific entity or action and can therefore be used in a wider variety of contexts. Additionally, from the mid-18th century onward, the average SynProd values of these two classes diverge even further from the *All* values, suggesting a growing conventionalization in the usage of these expressions in the scientific register.

Compounds and flat expressions, due to their more restricted meanings, exhibit lower SynProd values compared to other parts of speech. However, the difference between these two classes of MWEs and the *All* curve becomes less pronounced, especially in the 20th century.

It is interesting to note that phrasal verbs, often described as less common in academic prose (see, e.g., Biber et al. (2021) and Brown et al. (2015)), exhibit lower average values of L3 and R3 compared to other parts of speech (*All*) until the mid-19th century. However, they show an increasing trend in the more recent decades, surpassing the *All* curve in the final decades of the 20th century.

4.2 Syntagmatic Productivity per Class

To better understand the diachronic changes in syntagmatic productivity across different classes of MWEs, we compared them to the average SynProd of single-token units with comparable parts of speech. Figures 2 and 3 present the L3 and R3 graphs, comparing: a) phrasal verbs to other verbs; b) compounds and flat expressions to nouns and proper nouns; c) fixed and AFL MWEs to function words.

In all cases, changes are observed around the mid-19th century. Regarding phrasal verbs, their syntagmatic productivity is generally lower than that of other verbs. However, there is a reduction in the SynProd difference in more recent texts.

Compounds exhibit the lowest SynProd values up to 1730, being used in more specific contexts. However, after this decade, their average SynProd value increases, bringing it much closer to the pro-

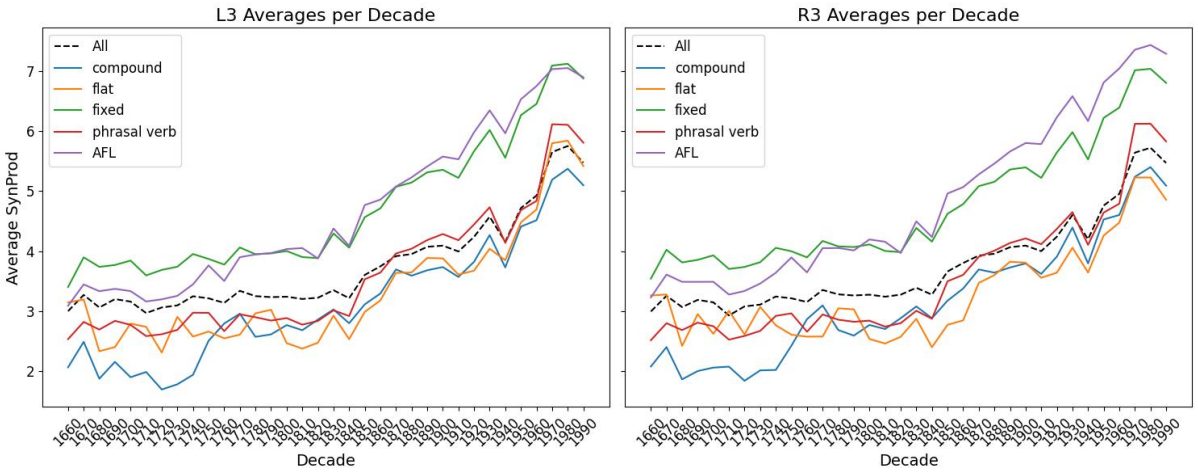

Figure 1: Average syntagmatic productivity of the different classes of MWEs per decade of the RSC.

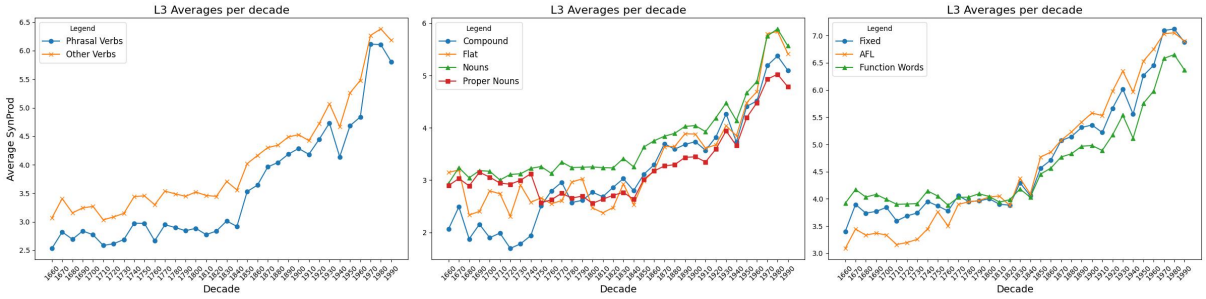

Figure 2: Average syntagmatic productivity considering the left context (L3) comparing MWEs with similar PoS.

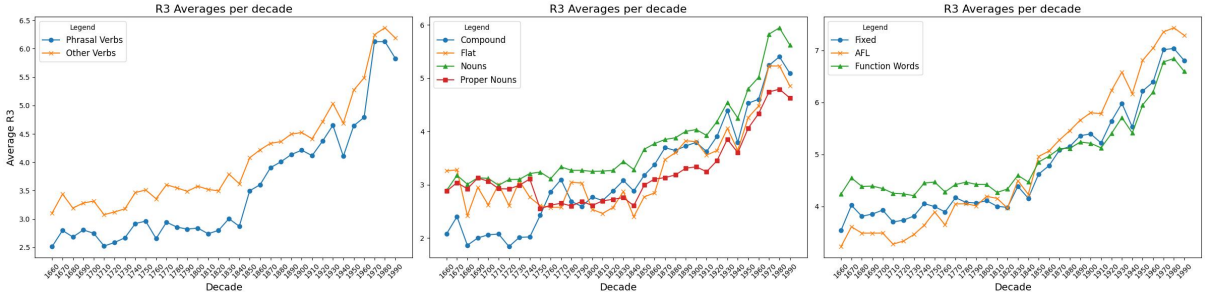

Figure 3: Average syntagmatic productivity considering the right context (R3) comparing MWEs with similar PoS.

ductivity of nouns in the RSC. In contrast, flat expressions start with higher SynProd averages but do not show a significant increase until the mid-19th century. When compared to proper nouns, flat expressions have higher SynProd values from the mid-19th century onward, even approaching the L3 SynProd of nouns in the final decades of the 20th century.

Finally, fixed and AFL expressions are the classes that surpass similar parts of speech in terms of SynProd after the mid-19th century, confirming the widespread conventionalized usage of these constructions in this register. In the later periods, we observe that, with regard to L3 values, fixed and AFL expressions exhibit quite similar averages. However, this is not the case for R3, where AFL expressions show higher syntagmatic productivity.

These results demonstrate that the use of MWEs in scientific English broadens in terms of context over time, exhibiting stronger increasing tendencies compared to similar parts of speech. Furthermore, they confirm the conventionalized and recurrent usage of fixed and formulaic expressions in this register.

It is important to mention that the size of the sub-corpora representing each time period was not controlled, which may affect the entropy values. As future work, we intend to conduct the same analysis using equal-sized samples for each period.

5 Conclusion and Future Work

In this paper, we have presented an analysis of the syntagmatic productivity of different classes of MWEs in scientific writing. Our investigation reveals that, like other single-token units with comparable parts of speech, MWEs exhibit an increasing tendency in SynProd, especially after the mid-19th century, considering both left and right contexts. We have also shown that, when comparing each class of MWE with corresponding parts of speech, MWEs tend to exhibit a more considerable increase in syntagmatic productivity over time. In most cases, the average SynProd values for MWEs are lower; however, over time, the delta decreases, or even reverses, as is the case for AFL and fixed expressions. In future work, we intend to compare the tendencies regarding syntagmatic productivity of MWEs to other information-theoretical measures such as paradigmatic variability (i.e., the sets of linguistic options available in a given or similar syntagmatic contexts) and typicality.

Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Diego Alves, Stefania Degaetano-Ortlieb, Elena Schmidt, and Elke Teich. 2024a. Diachronic analysis of multi-word expression functional categories in scientific English. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 81–87.

Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024b. 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)*, pages 67–76.

Douglas Biber and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for specific purposes*, 26(3):263–286.

Douglas Biber, Stig Johansson, Geoffrey N Leech, Susan Conrad, and Edward Finegan. 2021. *Grammar of spoken and written English*. John Benjamins.

David West Brown, Chris C Palmer, Michael Adams, Laurel J Brinton, and Roger D Fulk. 2015. The phrasal verb in American English: Using corpora to track down historical trends in particle distribution, register variation, and noun collocations. *Studies in the history of the English language VI: Evidence and method in histories of English*, 85:71–97.

Kathy Conklin and Norbert Schmitt. 2012. The Processing of Formulaic Language. *Annual Review of Applied Linguistics*, 32:45–61.

Stefania Degaetano-Ortlieb and Elke Teich. 2019. Toward an Optimal Code for Communication: The Case of Scientific English. *Corpus Linguistics and Linguistic Theory*, 0(0):1–33.

Peter Fankhauser. 2025. Measuring and visualizing diachronic word use. Accessed: 2025-01-26.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. 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*, pages 794–802.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27(1):4–21.

Dilin Liu. 2012. The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, 31(1):25–35.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Carlos Ramisch, Abigail Walsh, Thomas Blanchard, and Shiva Taslimipoor. 2023. A survey of MWE identification experiments: the devil is in the details. In *Proceedings of the 19th workshop on multiword expressions (MWE 2023)*, pages 106–120.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, et al. 2015. Parseme–parsing and multiword expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Rita Simpson-Vlach and Nick C Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied linguistics*, 31(4):487–512.

Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. Less is More/More Diverse: On The Communicative Utility of Linguistic Conventionalization. *Frontiers in Communication*, 5.