

A corpus-based study of paradigmatic variability of multiword expressions in scientific English from the 17th to the 20th century

Research Article

Diego Alves, Elke Teich

Saarland University, Saarbrücken

Received 1 December 2025; accepted 24 February 2026

Abstract: This study extends a paradigm-based framework for lexical conventionalization to multiword expressions (MWEs) in diachronic scientific English using the Royal Society Corpus (1665–1996). Building on prior work that quantified paradigmatic variability, the set of contextually available lexical alternatives, via distributional entropy, we model MWEs as single lexical units in distributional vector spaces to examine their role in long-term language standardization. Our aim is to verify whether MWEs undergo the same conventionalization process as the one previously observed for most word classes. Our results confirm that the overall decline in paradigmatic variability, indicating increasing conventionalization, continues through the 20th century. Different MWE classes follow distinct trajectories: compounds and phrasal verbs show substantial conventionalization in the 18th–19th and 19th–20th centuries, converging with or dropping below the variability of single nouns and verbs. In contrast, fixed expressions and academic formulas become conventionalized early and remain stably low thereafter. Flat expressions remain more variable, patterning with general nouns. These findings demonstrate that MWEs play an active role in the register-specific development of scientific English, with different types contributing differently to the reduction of paradigmatic choice.

Keywords: *paradigmatic variability • multiword expressions • scientific English • diachronic linguistics*

© Author(s)

1 Introduction

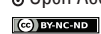
Language use varies according to pragmatic, cognitive, and social factors. From a diachronic perspective, recurrent patterns of linguistic choice may accumulate into conventionalization over time, that is, the gradual stabilization and social diffusion of form–meaning pairings through repeated use. This process widely considered a prerequisite for innovation (De Smet 2016) and a component of persistent change, including grammaticalization (Bybee 2010; Schmid 2015). Complementary to the stabilizing effects of conventionalization is the diversification phenomenon, whereby a word or word form shifts away from its original usage context and becomes specialized semantically (Halliday & Martin 1993; Harris 2002) or grammatically. For example, this happens in cases where lexical verbs become increasingly associated with particular syntactic environments, potentially as a step toward grammaticalization.

Conventionalization and diversification reflect a single underlying mechanism of change, namely the reduction of paradigmatic variability in specific usage contexts. While diversification involves semantic or grammatical specialization of forms across different contexts, it simultaneously restricts the range of competing alternatives within each context. Conventionalization further stabilizes these specialized form–meaning pairings through social diffusion and entrenchment. Together, these processes progressively narrow the set of paradigmatic options available in a given context.

While Teich et al. (2021) demonstrate that scientific English undergoes systematic conventionalization and diversification over time, captured through diachronic changes in paradigmatic variability, they do not consider multiword expressions in their analysis. Yet MWEs often encode specific entities or domain-

* Corresponding author: Diego Alves, E-mail: diego.alves@uni-saarland.de

Open Access. © 2026 Author(s).

 This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

relevant concepts (e.g., *sodium chloride*, *carry out*) or exhibit a high degree of mutual information (e.g., *due to*, *such as*), making them functionally comparable to single lexical units. Given their semantic and structural cohesion, as well as their centrality in scientific communication, we therefore extend the paradigmatic variability framework by treating MWEs as single linguistic units. Our aim is to verify whether MWEs are subject to the conventionalization mechanisms previously attested for single-word units, but with potentially distinct diachronic dynamics (e.g., stronger effects or different temporal phases). On this basis, we hypothesize that different MWE classes exhibit specific patterns of paradigmatic variability and that these patterns actively contribute to the processes of conventionalization and diversification identified by Teich et al. (2021). As our data-base, we use the Royal Society Corpus (Fischer et al. 2020) and automatic methods to identify MWEs. Paradigmatic variability over time is then measured using word embeddings and entropy, following Teich et al. (2021).

The remainder of the paper is structured as follows. Section 2 presents the theoretical background. Section 3 describes the dataset used for the analysis, our methods for MWE identification and extraction, and the calculation of paradigmatic variability. Section 4 reports the results of the analysis, with a specific focus on each MWE class, and Section 5 discusses overall diachronic trends. Section 6 concludes the paper with a summary and perspectives for future work.

2 Theoretical background

Regularity in language is not limited to structural domains such as syntax or morphology; it also concerns the recurrent ways in which words combine into recognizable patterns carrying specific meanings or connotations. MWEs range from formally fixed, often figurative idioms (e.g., *night owl*) to more compositional and productive combinations, such as compounds (e.g., *sea water*) and phrasal verbs (e.g., *pull up*) (Avgustinova & Iomdin 2019). MWEs are ubiquitous because they contribute to communicative efficiency: their internal transitions tend to be highly predictable, and highly conventionalized MWEs can be retrieved holistically rather than processed incrementally, which affords them a processing advantage (Conklin & Schmitt 2012). Beyond processing, MWEs play a crucial role in register formation, providing conventional encodings of context-specific meanings. Legal language, for instance, employs characteristic expressions such as *in no event* or *said therein*, while scientific terminology relies heavily on domain-specific MWEs, including productive noun–noun or adjective–noun combinations such as *iron oxide* or *sulphuric acid*. Recent work by Alves et al. (2024), using the Royal Society Corpus, confirms that the use of MWEs in scientific English has increased steadily over time and that different classes of MWEs exhibit distinct structural and distributional behaviours.

Definitions of multiword expressions vary across authors. Here, we adopt the syntactic approach proposed by De Marneffe et al. (2021) in the dependency relations annotation guidelines of the Universal Dependencies (UD) framework. In this framework, four types of dependency relations describe multiword expressions:

- 1) compound (CMP) – combinations of tokens that behave morphosyntactically as a single word. Not all lexicalized or idiomatic expressions are classified as compounds. In English, compounds specifically involve combinations of nouns (e.g., *hard drive*, *iron oxide*). Consequently, adjective–noun combinations, prepositional phrases, and light-verb constructions are not considered compounds and are instead annotated using other dependency relations in the Universal Dependencies (UD) framework (Universal Dependencies 2024a);
- 2) compound:prt (CPR) – phrasal verbs (e.g., *shut down*, *find out*). Phrasal verbs are constructions in which a verb combines with a particle that can modify or change the verb's meaning (e.g., *take up*, *fly away*), and the particle is often movable (*take the book up* vs. *take up the book*), whereas prepositional verbs consist of a verb followed by a preposition that takes a noun phrase complement (e.g., *wait for*, *delight in*), with the preposition being fixed and inseparable from the verb (Claridge 2000: 46–64). In the UD framework (Universal Dependencies 2024b), the compound:prt relation is intended for phrasal verbs only;

- 3) flat (FLT) – The UD framework defines that this relation applies to sequences in which none of the components can be identified as a head using standard substitution tests like sequences of proper nouns and items separated for readability (e.g., *Hillary Clinton, San Francisco*) (Universal Dependencies 2024c). However, the parser used to annotate the corpus also labels sequences composed of a noun followed by a specifier with this category (e.g., Table 1). These sequences are formulaic and serve specific discourse functions, a characterization supported by previous research on formulaic language (Simpson, Vlach & Ellis 2010). Moreover, they constitute the most frequent type of flat expressions identified by the parser in our dataset. In line with this theoretical background and their prevalence, we retained them in our analysis, treating them as flat expressions;
- 4) fixed (FIX) – certain fixed grammaticalized expressions behaving like function words (e.g., *because of, in spite of, as well as*). Fixed expressions generally do not allow intervening words. The formation of fixed multiword expressions represents the final stage of grammaticalization, and the UD framework recommends restricting annotation to the most fully grammaticalized cases, treating fixed expressions as a closed class documented with language-specific lists (Universal Dependencies 2024d).

Although this approach does not cover all types of MWEs (e.g., idiomatic expressions), it enables the identification of multiword units that are particularly relevant to the scientific register, such as compounds widely used as technical terminology.

In addition to the MWEs identified using the UD framework, we supplemented our list with the Academic Formulas List (AFL), a set of 607 frequently used formulaic expressions extracted by Simpson-Vlach & Ellis (2010) using quantitative measures such as mutual information and frequency (e.g., *in terms of, in order to, whether or not*). To our knowledge, no comparable resource using this methodology exists specifically for scientific English. The relevance of these academic formulaic expressions to scientific texts is supported by the presence of most of them in our dataset. Conventionalization is the process by which a linguistic form or construction becomes regularly used and socially entrenched, making its use and interpretation predictable within a speech community. The more conventionalized a linguistic item or construction is, the more readily accessible it becomes in both comprehension and production (De Smet 2016). Crucially, conventionalization, has been argued to constitute an important precondition for innovative use (cf. De Smet 2016). This position aligns with communicative accounts of language change and variation. From this perspective, innovation does not arise in opposition to convention but is grounded in highly entrenched patterns that speakers can exploit and extend.

Within the scientific register, particular linguistic choices become increasingly conventionalized over time, reflecting shared communicative goals and disciplinary norms (Degaetano-Ortlieb & Teich 2019). Such conventionalization supports efficiency, precision, and intersubjective alignment, while simultaneously providing a stable foundation for controlled variation and innovation. At the same time, scientific language is shaped by diversification, as external pressures such as increasing specialization and disciplinary differentiation promote the development of more expressive and distinctive linguistic resources. Thus, diversification drives language toward greater expressivity and differentiation, whereas communicative pressures encourage convergence on a smaller set of commonly preferred forms through conventionalization. This tension can be conceptualized as a process of communicative optimization, driven jointly by diversification and conventionalization.

Conventionalization and diversification in scientific language can be captured by paradigmatic variability, a distributional property reflecting the range of alternative lexical choices available to express a given meaning. Recent work conceptualises word embeddings as weakly neural, probabilistic models of paradigmatic relations (Hamilton et al. 2016; Dubossarsky et al. 2017). In these, paradigmatic variability can be quantified by calculating the entropy of a word's local neighbourhood in vector space and tracking its change over time. Here, entropy is defined following Shannon (1948) as a measure of uncertainty or unpredictability in a probability distribution: the higher the entropy, the more evenly distributed the alternatives and the greater the variability; conversely, lower entropy indicates stronger convergence on particular options and greater conventionalization.

Teich et al. (2021) applied this entropy-based measure to scientific English using the Royal Society Corpus (1665–1996; Fischer et al. 2020) and showed that paradigmatic variability decreases diachronically in this register. While entropy reduction is normally a cognitively costly process in online production (Venhuizen et al. 2019; Tourtouri et al. 2019), in scientific English it occurs as a diachronic process through which language use is dynamically optimized, constraining otherwise excessive lexical variation to maintain communicative function (Teich et al. 2021; Degaetano-Ortlieb & Teich 2022).

3 Methodology

In this section, we describe the data set of our study, outline our methodological framework, detail the procedure for extracting MWEs, and explain the approach used to compute paradigmatic variability.

3.1 Data

As our data source, we use the Royal Society Corpus (RSC) 6.0,¹ a large diachronic corpus of scientific English spanning the period from 1665 to 1996. The corpus contains 47,837 texts (295,895,749 tokens), primarily scientific articles covering a broad range of disciplines, from mathematics to the physical and biological sciences, from the Philosophical Transactions and Proceedings of the Royal Society of London (Fischer et al. 2020). Table 1 presents the number of texts and tokens in the corpus across successive 50-year periods.

Table 1. Distribution of texts and tokens in the Royal Society Corpus (RSC) 6.0 by period (Fischer et al. 2020).

Period	Texts	Tokens
1665–1699	1,325	2,582,856
1700–1749	1,686	3,414,795
1750–1799	1,819	6,342,489
1800–1849	2,774	9,112,274
1850–1899	6,754	36,993,412
1900–1949	10,011	65,431,384
1950–1996	23,468	172,018,539

The RSC, as presented by Fischer et al. (2020), includes many texts, particularly from earlier periods (17th–19th centuries), whose base text was generated via Optical Character Recognition (OCR) by the corpus compilers and therefore contains recognition noise typical of historical OCR processing, which can affect the identification of MWEs. This potential bias is discussed further in the Results section.

The RSC has proven highly relevant for diachronic studies of scientific English as it provides a unique window into the historical evolution of this specialized register. Previous research has shown that the scientific register developed from an involved, orally oriented style, reflecting early practices in which papers were read aloud by fellows at the Royal Society, into a highly informational, written expert-to-expert mode characteristic of modern scientific communication (Degaetano-Ortlieb et al. 2018). This trajectory appears specific to scientific writing and is not mirrored in register-mixed corpora (Degaetano-Ortlieb & Teich 2019). Studies have also documented increasing disciplinary diversification, such as the emergence of modern chemistry in the 18th century (Bizzoni et al. 2020), alongside a broader trend toward conventionalization in scientific discourse (Teich et al. 2021). These findings demonstrate that the RSC is an ideal resource for investigating long-term linguistic changes within a single, highly conventionalized register, allowing researchers to trace how language adapts to the communicative demands of modern science over centuries.

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

3.2 MWE extraction

To identify and extract MWEs from the RSC, we combined two complementary strategies. First, we leveraged specific dependency relations defined in the Universal Dependencies (UD) framework (De Marneffe et al. 2021). Second, we incorporated a list of formulaic expressions identified by Simpson-Vlach & Ellis (2010), derived from a collection of academic corpora containing a substantial amount of text related to the disciplines represented in the RSC. The details of each strategy are described in the following subsections.

3.2.1 Universal dependencies method

As previously mentioned in Section 2, the Universal Dependencies (UD) framework (De Marneffe et al. 2021) provides guidelines for annotating syntactic dependency relations, including four labels relevant to MWEs: i) compound (CMP), ii) compound:prt (CPR), iii) flat (FLT), and iv) fixed (FIX). Texts annotated according to UD guidelines use the CoNLL-U format, which is supported by state-of-the-art dependency parsers such as Stanza (Qi et al. 2020). This makes it possible to automatically identify word sequences corresponding to the four MWE types.

For our study, the RSC 6.0 was parsed using Stanza's combined English model. We then developed a Python script using the `pyconll` library² to extract and count MWEs per year, producing a unified list of UD-based MWEs across the corpus.

Parser accuracy is a critical component of the reliability of our results. A manual evaluation conducted by Alves et al. (2024) considering 140 sentences randomly selected from the parsed corpus (i.e., 20 from each 50-year period) indicated $\geq 85\%$ accuracy regarding head identification and dependency label for compound, compound:prt, and fixed MWEs, and 75% accuracy for flat MWEs, with performance consistent across time periods.

Regarding compounds, the remaining errors mostly concern adjective–noun combinations (e.g., *uniaxial crystal* and *olfactory bulb*). For compound:prt, the parser distinguishes phrasal from prepositional verbs (only phrasal verbs are annotated as compound:prt), and the few errors primarily involve prepositional verbs being mislabeled as phrasal (e.g., *flow up* and *set over*). For flat expressions, the parser labels structures such as noun followed by a specifier sequences; as previously discussed, we retain these as flat expressions due to their formulaic nature and discourse-functional role.

3.2.2 Academic Formulas List (AFL)

As previously explained, the Academic Formulas List (AFL) (Simpson-Vlach & Ellis 2010) was used as an external, lexicographically defined inventory of multiword expressions. It provides a curated set of highly frequent formulaic sequences in academic English and consists of three components: (i) a core list of 207 expressions attested across written and spoken academic registers, (ii) a list of 200 expressions drawn from written academic corpora, and (iii) a list of 200 expressions extracted from spoken academic English.

Using a Python script, we identified and counted AFL MWEs in the RSC. Of the 607 AFL expressions, 506 were found in the corpus. Most of the missing 101 expressions come from the spoken-list subset (e.g., *I'll talk about*, *gonna talk about*, *let's look at*), indicating that not all items from the spoken-list subset are represented in the RSC.

3.3 Paradigmatic variability

To quantify diachronic changes in the paradigmatic variability of scientific English, we follow the methodology of Teich et al. (2021), measuring the degree of choice available in a given context or set of similar contexts. A reduction in paradigmatic variability is interpreted as an index of increasing conventionalization and, potentially, grammaticalization in some cases. The method is based on word embeddings (Ling et al. 2015), where a simple measure of paradigmatic variability for a word is the number of its close neighbours within a given radius. As proposed by Teich et al. (2021), we apply a

² <https://pypi.org/project/pyconll/>

more refined measure that weights each neighbour x_i in the neighbourhood C_x of a target word x by both its frequency ($\text{freq}(x_i)$) and its cosine similarity ($\cos(x_i, x)$) to the target. This allows us to estimate the probability ($p(x_i|C_x)$) that a neighbour could be chosen instead of x , giving higher weight to more frequent and more similar neighbours. The paradigmatic variability ($pvar$) of x is then defined as the entropy over this probability distribution:

$$(1) \ pvar(x) = H(P(.|C_x)) = - \sum_{\cos(x_i, x) > \theta} p(x_i|C_x) \log(p(x_i|C_x))$$

$$\text{with } p(x_i|C_x) = \frac{\cos(x_i, x) \text{freq}(x_i)}{\sum_{x_j} \cos(x_j, x) \text{freq}(x_j)} \quad (2)$$

Intuitively, a word exhibits high paradigmatic variability if it has many close neighbours that are evenly distributed in terms of similarity and frequency, whereas words with fewer or more skewed neighbours show lower variability. For example, in the word embedding space derived from the full RSC, the adverb *considerably* is surrounded by close neighbours such as *enormously*, *notably*, and *extraordinarily*. In contrast, technical terms such as *capillary* and *electrostatics* are more isolated, with fewer and more distant neighbours. Following Teich et al. (2021), we set the similarity threshold θ to 0.6 and limited the neighbourhood size to a maximum of 30 neighbours per word. These parameter values were established in Teich et al. (2021) on the basis of qualitative inspection, which showed that they yield coherent and interpretable neighbourhoods. We adopt the same configuration here to ensure methodological consistency and comparability with their results.

For all analyses, we compute word embeddings using structured skip-grams (Ling et al. 2015), which model each context position separately and are particularly effective for characterizing function words. Period-specific (i.e., 10-year intervals) embeddings are aligned by initializing the first period with embeddings trained on the complete corpus, and initializing each subsequent period with the embeddings of the immediately preceding period. This sequential approach stabilizes the representations and helps mitigate the impact of sparsely represented periods, addressing potential reliability issues due to the uneven distribution of texts across time. All models use a hidden-layer dimension of 100, a context window of ± 5 , 10 negative samples, and a minimum complete-corpus frequency threshold of 5. Period-specific embeddings are trained for 25 iterations, and the complete-corpus embeddings for 5.

To build the embeddings, we use tokens annotated with their part of speech (e.g., *water_NN*). Part-of-speech information is provided in the RSC (Fischer et al., 2020), which was annotated using TreeTagger (Schmid 2013) and the Penn Treebank Tagset (Santorini 1990) with minor modifications. As noted earlier, multiword expressions (MWEs) are treated as single tokens, and instead of a part-of-speech (PoS) tag, they receive their MWE class label (CMP, CPR, FIX, FLT).

Thus, using Equations 1 and 2 together with the vectors from the embedding spaces, we computed paradigmatic variability for each decade for all linguistic units (i.e., tokens and MWEs) that occur 50 times or more in the complete corpus. This frequency threshold ensures that each unit appears in sufficient quantity to contribute meaningfully to diachronic comparisons across the periods in the RSC. Also, it mitigates OCR-related noise by filtering out low-frequency artifacts introduced during the digitization process done by the corpus compilers (Fischer et al. 2020). Additionally, to analyse statistically the differences in paradigmatic variability, we employed a multi-faceted statistical approach. For each part-of-speech category, we tested changes across decades using Kruskal-Wallis tests, followed by pairwise Wilcoxon tests with Bonferroni correction when significant differences were detected. Conversely, for each decade, we examined variability across PoS categories using the same non-parametric framework.

4 Results

In this section, we first present the main statistics regarding MWE extraction, then we report the overall paradigmatic variability across all PoS categories and MWE classes, and finally, we analyse in detail the diachronic changes in paradigmatic variability for each MWE class.

4.1 Extracted MWEs

Table 2 presents an overview of different classes of multiword expressions (MWEs), including their frequency and representative examples. The table distinguishes between compounds (CMP), phrasal verbs (CPR), fixed expressions (FIX), flat expressions (FLT), and formulaic expressions from the Academic Formulas List (AFL), showing both the total number of types identified in the RSC and the number of high-frequency types occurring more than 50 times used for the paradigmatic variability analysis. Representative examples are provided for each class to illustrate the variety and usage patterns of MWEs in the corpus.

Table 2. Overview of MWE categories, type counts, and examples.

MWE class	Number of types	Number of types (> 50 occurrences)	Examples
compound (CMP)	2 523 696	6 823	<i>absorption band; meson field; sucrose density</i>
phrasal verb (CPR)	604 057	417	<i>added up; building up; cancels out; mixed up</i>
fixed (FIX)	16 337	112	<i>about to; above all; as far as; such as</i>
flat (FLT)	3 107	87	<i>figure i; table iii; table 12</i>
AFL	506	307	<i>allow us to; all sorts of; function of the</i>

Examination of the representative examples for each class shows that the flat (FLT) class mainly consists of nouns followed by a numerical specifier, which are typical in the scientific register. By considering these collocations as MWEs, we can analyse them as single linguistic units, differentiating these numbers from other numerical expressions in the texts. Although sequences of proper nouns are present in the corpus, they generally occur at low frequencies. Moreover, although the parser is trained to classify phrasal verbs as CPR, in a few instances, certain prepositional verbs are also erroneously labelled as such due to parser limitations.

It is evident that the number of MWE types occurring more than 50 times in the corpus is much lower than the total number of extracted MWEs, suggesting that many expressions appear only a few times, likely in specific terminological contexts. Applying this frequency filter allows us to focus on linguistic units with sufficient data for diachronic analysis, ensuring a better overall distribution in the corpus. The filter also helps reduce bias from expressions affected by Optical Character Recognition (OCR) errors or parser mistakes.

4.2 Overall paradigmatic variability trends

Figure 1 shows the paradigmatic variability across decades for all linguistic units in the corpus, including both single tokens and multiword expressions (MWEs). By considering the full set of tokens, the graph provides an overview of general trends in lexical variability over time.

The results in Figure 1 show that the general decreasing trend in pvar over time, observed by Teich et al. (2021), is also apparent when MWEs are treated as single linguistic units. While the previous study analysed RSC data only up to 1920, our analysis indicates that the conventionalization process reflected by the decrease in pvar continued throughout the 20th century. Notably, there are significant

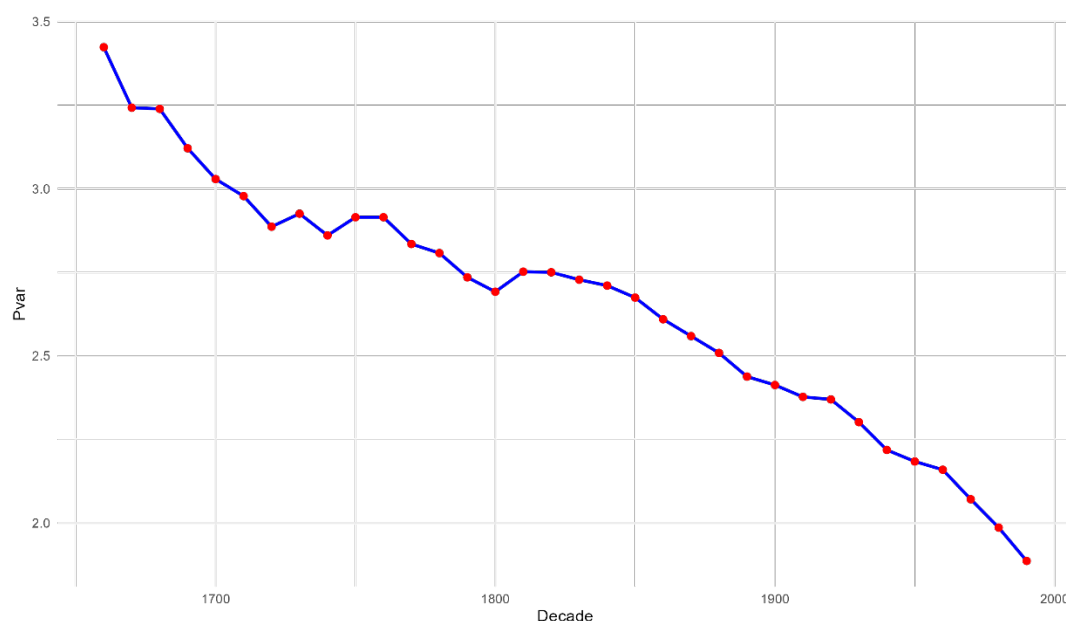


Figure 1. Overall paradigmatic variability (pvar) of linguistic units over time.

differences ($p < 0.05$) in pvar when comparing the early decades of the 20th century with those towards the end, suggesting ongoing changes in the usage and variability of linguistic units over time.

The diachronic changes of each part of speech (PoS) and MWE class are illustrated in the heatmap shown in Figure 2, where darker shades correspond to lower paradigmatic variability values.

A general overview of the results presented in Figure 2 shows that, in later periods, the heatmap tends to display darker shades, indicating an overall decrease in pvar across the different PoS and MWE classes. However, content words generally exhibit higher pvar values in the earlier decades (yellow shades), reflecting a stronger tendency toward conventionalization over time. This pattern is particularly evident for verbs, including base form (VV), past tense (VVD), *ing*-forms (VVG), past participle (VVP), and 3rd person singular present (VVZ), as well as common and proper nouns (singular NN/NP, plural NNS/NPS), adjectives (JJ, comparative JJR, superlative JJS), and adverbs (RB). Singular nouns (NN) and plural nouns (NNS) were considered separately because the label “compound” applies to both forms. Since paradigmatic variability over time was calculated for NN and NNS individually, we plotted them separately to allow comparison of the broader compound category with its singular and plural constituents.

Among the different classes of MWEs, compounds, phrasal verbs, and, to a certain extent, flat expressions show trends comparable to those of nouns and verbs, whereas fixed expressions and AFL units display a distinct behavior, having low pvar values even in the earlier decades. A detailed analysis of these tendencies for each MWE class will be presented in the following subsections. These patterns indicate that the conventionalization process affects PoS and MWE classes differently, in line with the findings reported by Teich et al. (2021).

The box plot in Figure 3 provides an overview of the mean values of paradigmatic variability across specific PoS and MWE classes.

The results shown in Figure 3 complement those observed in Figure 2. Verbs and nouns exhibit higher mean paradigmatic variability values along with relatively high variability, while function words (except TO) display lower pvar with less variation. Notably, AFL and fixed expressions are the classes with both the lowest pvar and the lowest variability, indicating a high degree of conventionalization and relative stability over time.

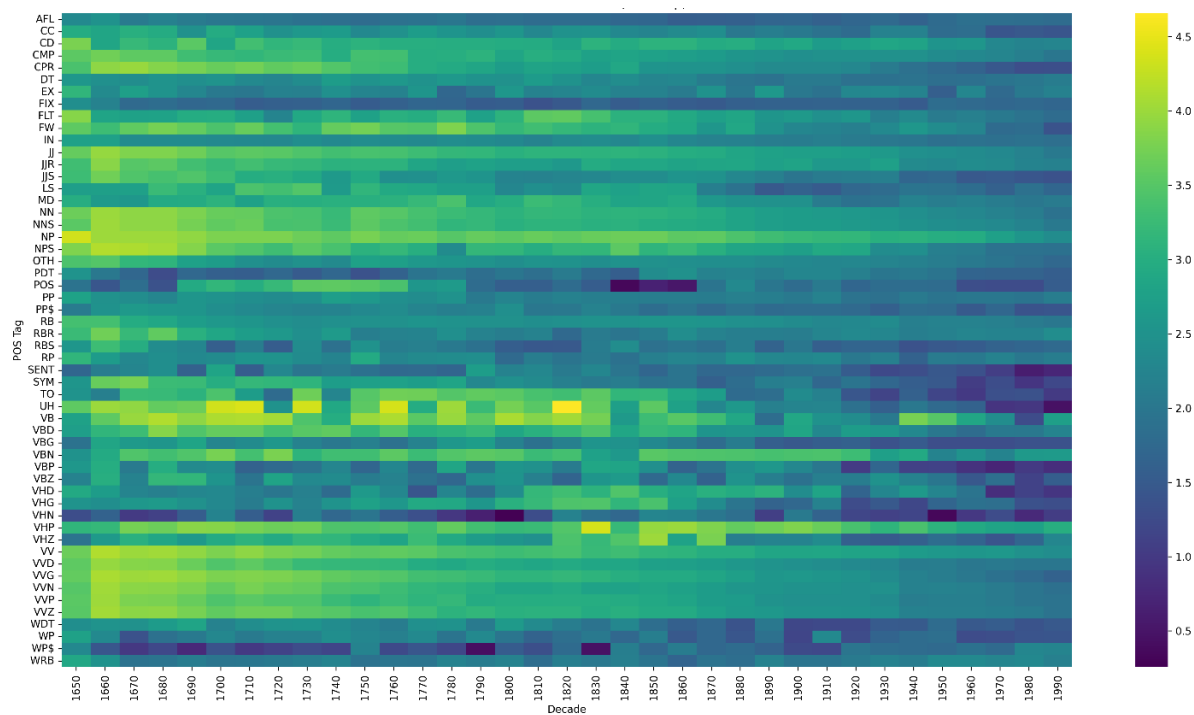


Figure 2. Diachronic paradigmatic variability across PoS and MWE classes.

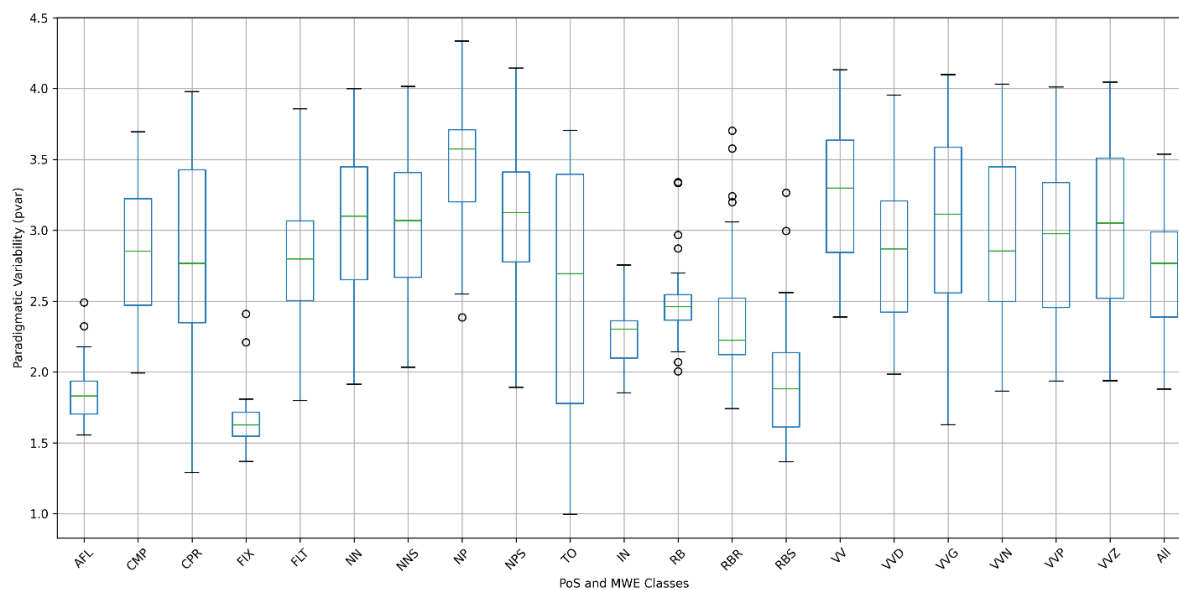


Figure 3. Mean and spread of paradigmatic variability (pvar) for PoS and MWE categories.



Figure 4: Mean paradigmatic variability (pvar) for compounds (CMP), singular nouns (NN), and plural nouns (NNS) across decades.

4.3 Paradigmatic variability of compounds

As previously described, within the Universal Dependencies framework, English compounds are typically analysed as combinations of nouns. English compounding is generally right-headed: the rightmost element functions as the head, determining the syntactic category, whereas the left element acts as a modifier or specifier (Bauer et al. 2015; Plag 2018). Thus, in order to analyse shifts in the paradigmatic variability of compounds (CMP), we compare them to singular and plural common nouns (NN and NNS, respectively). Figure 4 presents the decade-by-decade mean paradigmatic variability values for the CMP, NN, and NNS categories.

It is possible to observe a similar decreasing tendency in pvar for all three classes. This decrease is statistically significant ($p < 0.05$) when comparing earlier periods with more recent ones. When comparing single-token nouns (NN and NNS), they exhibit relatively similar pvar values across the decades. Statistical tests confirm significant differences only in 1650, 1850, and the last three decades of the RSC (1970–1990), with singular nouns showing a relatively higher degree of conventionalization. Regarding compounds, we observe that highly specific terms such as *vector potential* (example (1)), *radius vector*, and *potential gradient* show a strong decrease in paradigmatic variability, whereas some compounds, such as *heat capacity* (example (2)) with synonyms like *thermal capacity* (example (3)) and *water molecules*, also referred to as *aqueous molecules*, exhibit an increase in paradigmatic variability.

- (1) We shall first examine the nature of the **vector potential** inside a space due to magnets or currents outside that space. (RSC, 1881)
- (2) The specific **heat capacity** of sodium is approximately one third of that of water. (RSC, 1990)
- (3) The errors in the measured **thermal capacity** are about for the sample with its low self-heating. (RSC, 1976)

Comparing compounds (CMP) to the two noun categories, there are few statistically significant differences in the earlier periods (1650, 1680, 1770). However, from the nineteenth century up to 1930, the differences between CMP and single-token nouns are consistently significant. From 1930 onward, CMP and NN show similar pvar values, differing only from NNS.

Thus, compounds appear to have undergone a more accelerated process of conventionalization during the late eighteenth and nineteenth centuries, likely because they are more semantically specific

due to the presence of modifiers. Moreover, their use has increased over time in general English (Schneider et al. 2017) and particularly in scientific English (Alves et al., 2024b), accompanied by a decrease in paradigmatic variability. However, when considering the overall processes of diversification and conventionalization in scientific English, the data show that after 1930, nouns, especially singular ones, reach the same level of conventionalization as compounds, indicating that this tendency has been consolidated for both simple and more complex nominals.

4.4 Paradigmatic variability of flat expressions

As described in Section 2, in the RSC, the parser identified flat (FLT) expressions primarily as noun phrases composed of a noun and a numerical specifier (e.g., *Table 1*) or sequences of indexes (e.g., *i, ii*). Although in the UD framework “flat” is usually used for combinations of proper nouns, these were excluded from our analysis because paradigmatic variability, as stated above, was calculated only for multiword expressions (MWEs) occurring more than 50 times in the corpus, ensuring that the included MWEs contribute meaningfully to the diachronic change. Most proper noun combinations did not meet this frequency threshold. Consequently, for comparison with PoS, FLT expressions were compared with nouns (NN and NNS), as shown in Figure 5.

The diachronic shifts of pvar for FLT expressions show greater variability compared to NN and NNS, which exhibit a clear decreasing trend. For FLT expressions, from 1650 to 1900, there are very few statistically significant differences. However, during the twentieth century, pvar decreases consistently, particularly for terms such as *Table II* (example (4)), *Figure 2* (example (5)), and *Section 3* (example (6)), which are used as part of discourse-organizing constructions, showing significant differences compared to earlier periods.

- (4) At the foot of **Table II** we have entered the mean for January in the 24 years, 1848 to 1871. (RSC, 1883)
- (5) As also another Crystal Particle represented by **Figure 2**. (RSC, 1706)
- (6) This later condition of the eruption is represented in **section 3**, where the level of I in the strata p falls below the sea-level. (RSC, 1886)

When comparing FLT pvar values with NN and NNS, no consistent statistical differences are observed over time. Although the graph suggests that FLT expressions exhibit lower paradigmatic variability in the early periods of the RSC, these differences are not statistically significant ($p > 0.05$ for all periods), likely due to the high variability of FLT expressions reflecting the diverse items within this class. In the later periods, FLT expressions also considered in this study show a more stable trend, suggesting gradual conventionalization.

4.5 Paradigmatic variability of phrasal verbs

Figure 6 presents the diachronic changes in paradigmatic variability of phrasal verbs (CPR) compared to different verb PoS categories, excluding auxiliaries and copulas, including base forms (VV), past tense (VVD), *ing*-forms (VVG), past participles (VVN), singular present non-3rd person (VVP), and 3rd person singular present (VVZ).

For all verb PoS categories and phrasal verbs, there is a statistically significant ($p < 0.05$) decreasing tendency in paradigmatic variability over time. Comparing the verbal PoS, the base form (VV) generally exhibits higher pvar than other verbal forms. Statistically significant differences between VV and some other verbal forms are already observed in earlier periods; however, consistent significance across all comparisons ($p < 0.05$) emerges from the mid-19th century onwards. Among the inflected verb forms, pvar values are usually similar and tend to decrease over time.

Phrasal verbs (CPR) show a distinct trend: although their pvar is similar to that of base forms in the earliest periods of the RSC, they undergo a strong process of conventionalization during the eighteenth and nineteenth centuries, which becomes even more pronounced in the twentieth century. By this period, phrasal verbs exhibit the lowest pvar of all classes in the graph, with statistically significant differences. These results corroborate those reported by Alves et al. (2025), who, using information-

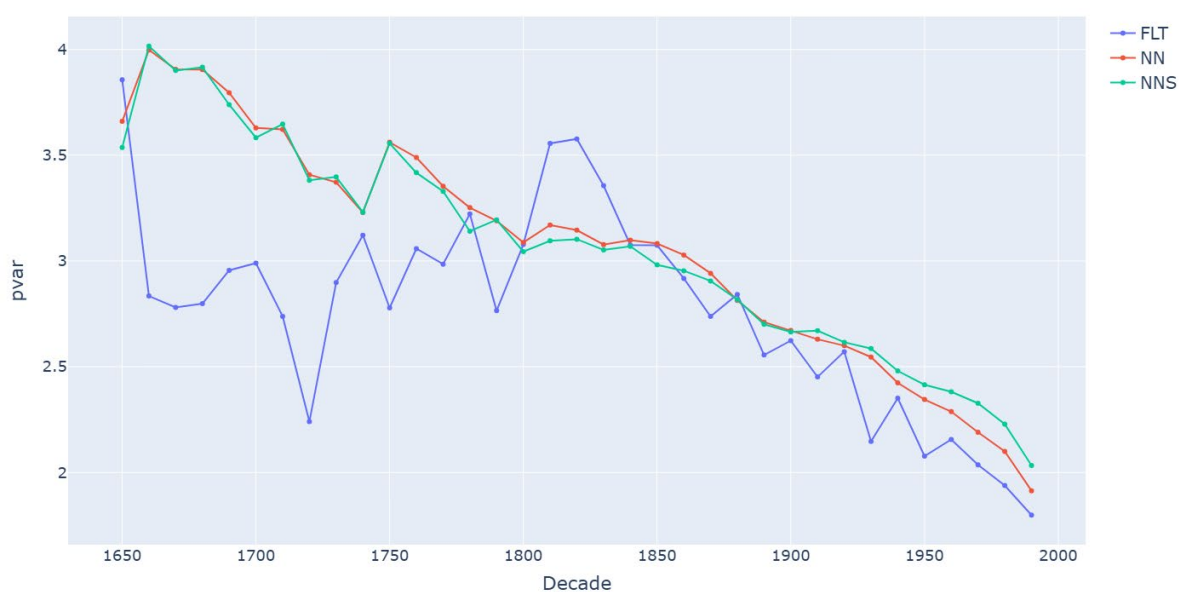


Figure 5: Mean paradigmatic variability (pvar) for flat expressions (FLT), singular nouns (NN), and plural nouns (NNS) across decades.

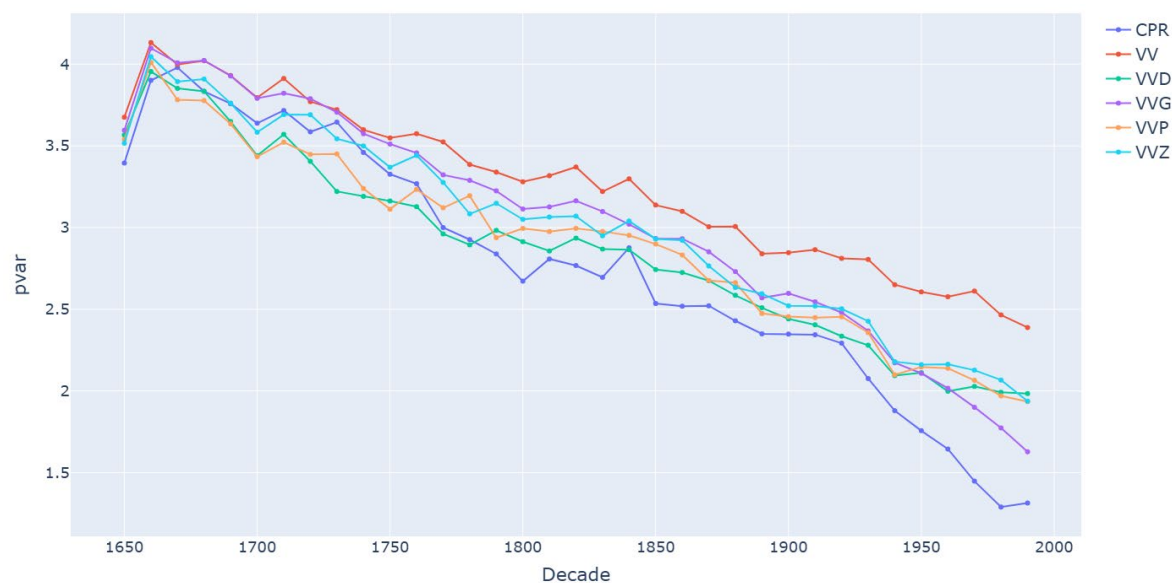


Figure 6. Mean paradigmatic variability (pvar) for phrasal verbs (CPR) and verb categories (VV, VVD, VVG, VVN, VVP, VVZ) across decades.

theoretic measures, showed that over the centuries phrasal verbs undergo cycles of expansion and contraction: periods of increasing type diversity are followed by periods of decreasing diversity accompanied by the conventionalization of specific forms driven by communicative needs (e.g., phrasal verbs used in experimental descriptions such as *carry out* and *filter out*, as in examples (7) and (8), or in the presentation of results, such as *find out*, as in example (9)).

- (7) To **carry out** the experiments I prepared a series of small tubes made of very thick and well-annealed glass, each tube about four centimetres in length, and having a bore of five millimetres. (RSC, 1870)
- (8) But it is tempting to suppose that the geometry of the flight path has been shaped to aid the capture of relevant information and to **filter out** irrelevancies. (RSC, 1992)
- (9) And I know no surer or better way to **find out** the nature and properties of those steams, than by observing the effects, and all the alterations of the air, as they did. (RSC, 1675).

Thus, although phrasal verbs are generally considered less frequent in scientific and academic English (Biber et al., 2000) and show a slight decreasing trend in relative frequency within scientific English (Alves et al. 2024b), those that do occur in this register appear to be highly conventionalized and restricted to specific contexts where no alternative expression is possible, as reflected by the low pvar values observed for this class of MWEs.

4.6 Paradigmatic variability of fixed and AFL expressions

As previously described, fixed expressions are grammaticalized sequences of tokens, usually functioning as function words (e.g., *as if in example (10)* and *such as in example (11)*), while AFL expressions are formulaic expressions. Both AFL and fixed expressions have different discourse functions, as shown by Alves et al. (2024a), with distinctive patterns of usage over time in scientific English. We compare these expressions to prepositions (IN), adverbs (RB), and nouns (NN) because these word classes provide meaningful points of reference: prepositions and adverbs exhibit lower paradigmatic variability, smaller standard deviations, and less variation over time (Figures 2 and 3), and most fixed expressions behave as prepositions. Many AFL expressions, on the other hand, function as adverbials (e.g., *in this paper* and *in such a way*, as in examples (12) and (13)) or include at least one noun (e.g., *the relationship between and the presence of*, as in examples (14) and (15)), making IN, RB, and NN relevant for comparison (Figure 7).

- (10) He was awaked this night by pain, **as if** burning irons had been clapped to the insides of his legs, with anxiety and a sense of tremor over the whole body. (RSC, 1764)
- (11) Other substances, **such as** platinum and carbon filament, were also tried, but the glow was not obtained. (RSC, 1909)
- (12) I have only referred to a few of the memoirs that contain the figures and descriptions of the parts referred to **in this paper**. (RSC, 1887)
- (13) That the contrivance be such that the moving power may be applied **in such a way** as to act to the greatest advantage. (RSC, 1781)
- (14) **The relationship between** the two is not therefore one of symbiosis. (RSC, 1899)
- (15) So by the second Experiment we find **the presence of** the air to be no ways necessary to assist in the ascent of the water, in small tubes, or between the planes. (RSC, 1708)

Analysing the curves in Figure 7, it is evident that AFL, FIX, IN, and RB exhibit distinct patterns of paradigmatic variability change compared to nouns. For these four categories, a reduction in pvar is observed during the seventeenth century, followed by greater variability (ups and downs) in subsequent centuries. In contrast, nouns show a more stable decreasing trend, as noted previously. Additionally, AFL, FIX, IN, and RB generally present lower paradigmatic variability values than nouns across most decades, with the difference becoming less pronounced in the second half of the twentieth century.

These results suggest that the conventionalization of FIX and AFL expressions occurs early, reaching an optimal level relatively quickly. Nouns, on the other hand, start with higher pvar values and undergo a more gradual conventionalization process over the centuries. When comparing the different classes, AFL and FIX do not show significant differences over time, whereas both exhibit lower pvar values than IN and RB up to 1950, with all four categories showing similar values in the later periods of the RSC.

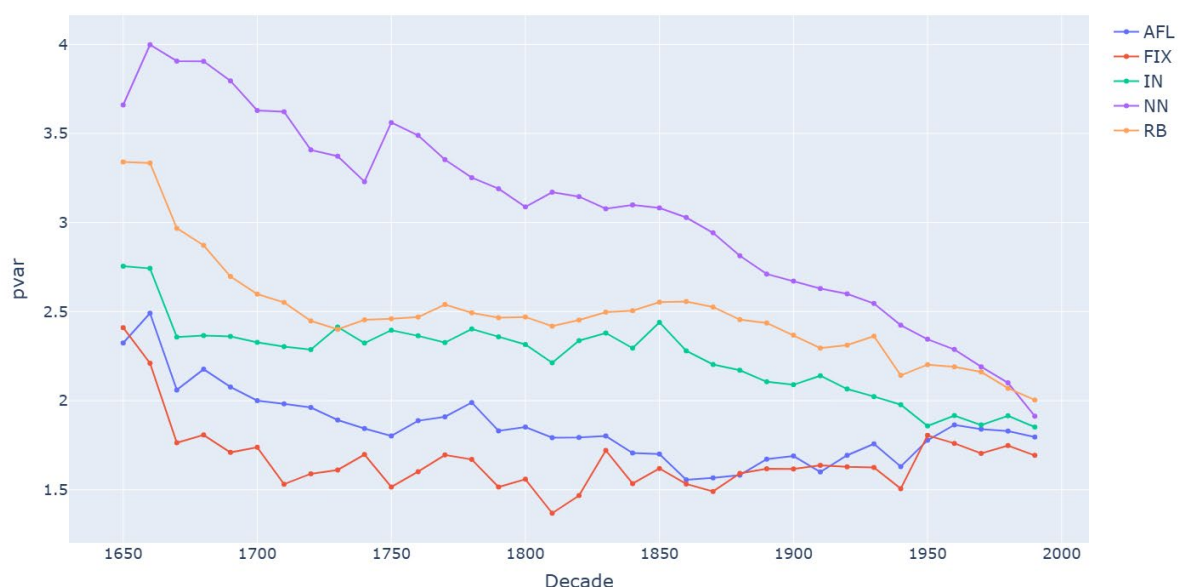


Figure 7. Mean paradigmatic variability (pvar) for AFL and fixed (FIX) expressions, prepositions (IN), adverbs (RB), and Nouns (NN) across decades.

5 Discussion

The processes of lexical diversification and conventionalization in scientific English, as measured using paradigmatic variability, were previously documented by Teich et al. (2021) up to 1920. Our study confirms these findings, extends them to later decades of the twentieth century, and adds statistical validation, providing a more robust and comprehensive view of diachronic trends. Importantly, by considering multiword expressions (MWEs) as single tokens, we were able to examine how different MWE classes behave over time with respect to paradigmatic variability.

The various classes of MWEs exhibit distinct patterns and periods of particularly strong conventionalization. Compounds (CMP), for example, show a rapid decrease in pvar, reaching lower values than nouns until the twentieth century. In the most recent decades, however, the paradigmatic variability of compounds converges with that of nouns, indicating that their conventionalization stabilizes over time. In contrast, phrasal verbs (CPR) display a pronounced conventionalization process in the last century of the RSC, achieving highly specific and conventionalized usage compared to other verb forms. Fixed expressions (FIX) and AFL constructions, by contrast, undergo a very early process of conventionalization in the seventeenth century and exhibit relatively stable pvar values thereafter, without further drastic changes.

The FLT expressions considered here are somewhat more complex, as they appear to consist of a large diversity of types, resulting in variable pvar values. However, these differences are not statistically significant when compared to nouns, indicating that FLT expressions maintain high variability across the decades.

Overall, these results show that MWEs have contributed substantially to the general decrease of pvar in scientific English over time, with statistically validated trends. However, each class contributes differently: compounds and phrasal verbs reflect dynamic changes across centuries, while fixed and AFL expressions are conventionalized from early periods.

6 Conclusion

In this paper, we analysed paradigmatic variability in scientific English using entropy-based measures derived from diachronic word embeddings, with a specific focus on multiword expressions (MWEs), aiming to determine whether these MWEs behave similarly to single-word units across different parts of speech. Our study confirms and extends the findings of Teich et al. (2021), demonstrating a significant long-term decrease in paradigmatic variability in the Royal Society Corpus, a trend that continued throughout the 20th century. Additionally, by treating MWEs as single linguistic units, we found that they also participate in the process of lexical conventionalization, although different MWE classes display distinct diachronic profiles. We found that compounds and phrasal verbs underwent dynamic conventionalization, with phrasal verbs, in particular, becoming highly fixed in modern scientific writing. In contrast, fixed and formulaic expressions were conventionalized early and remained stable, while flat expressions displayed higher variability.

As future work, a deeper qualitative analysis of the specific lexical neighbours in the embedding spaces for different MWE classes could provide more fine-grained, functional explanations for their variability trends. Moreover, the extraction of MWEs can be optimized also considering more recent frameworks defining MWEs (e.g., PARSEME (Savary et al. 2017)). Additionally, it would be valuable to investigate whether the variability trends observed in scientific English also hold in other registers, such as legal or journalistic English, as this could help determine the extent to which our findings generalize across different domains.

Acknowledgments

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

References

- Alves, Diego. 2025. Diachronic analysis of phrasal verbs in English scientific writing. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, 8–16, 2025, Tallin, Estonia.
- Alves, Diego, Stefania Degaetano-Ortlieb, Elena Schmidt & Elke Teich. 2024a. Diachronic analysis of multiword expression functional categories in scientific English. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, 81–87. Torino, Italy.
- Alves, Diego, Stefan Fischer, Stefania Degaetano-Ortlieb & Elke Teich. 2024b. Multiword 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)*, 67–76. St Julian's, Malta.
- Avgustinova, Tania & Leonid Iomdin. 2019. Towards a typology of microsyntactic constructions. In *International Conference on Computational and Corpus-Based Phraseology*, 15–30. Malaga, Spain.
- Bauer, Laurie, Rochelle Lieber & Ingo Plag. 2015. *The Oxford reference guide to English morphology*. Oxford: Oxford University Press.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 2000. *Longman grammar of spoken and written English*. Harlow: Longman.
- Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Peter Fankhauser & Elke Teich. 2020. Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence* 3. 73.
- Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.
- Claridge, Claudia. 2000. *Multi-word verbs in Early Modern English: A corpus-based study*. Amsterdam: Rodopi.
- Conklin, Kathy & Norbert Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics* 32. 45–61.

- Degaetano-Ortlieb, Stefania, Hannah Kermes, Ashraf Khamis & Elke Teich. 2018. An information-theoretic approach to modeling diachronic change in scientific English. In Carla Suhr, Terttu Nevalainen & Irma Taavitsainen (eds.), *From data to evidence in English language research*, 258–281. Leiden: Brill.
- Degaetano-Ortlieb, Stefania & Elke Teich. 2022. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory* 18 (1), 175–207.
- De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47 (2). 255–308.
- De Smet, Hendrik. 2016. How gradual change progresses: The interaction between convention and innovation. *Language Variation and Change* 28 (1). 83–102.
- Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1136–1145. Copenhagen, Denmark.
- Fischer, Stefan, Jörg Knappen, Katrin Menzel & 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 (LREC)*, 794–802. Marseille, France.
- Halliday, Michael Alexander Kirkwood & James R. Martin. 2003. *Writing science: Literacy and discursive power*. London: Routledge.
- Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 595–605. Austin, Texas, USA.
- Harris, Zellig S. 2002. The structure of science information. *Journal of Biomedical Informatics* 35 (4). 215–221.
- Ling, Wang, Chris Dyer, Alan W. Black & Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299–1304. Denver, Colorado, USA.
- Plag, Ingo. 2018. *Word-formation in English*. Cambridge: Cambridge University Press.
- Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
- Santorini, Beatrice. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). *University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-90-47*.
- Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Francesca Sangati, Veronika Vincze, Behrang Qasemi Zadeh & Alexandre Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 31–47. Valencia, Spain.
- Schmid, Hans-Jörg. 2015. A blueprint of the Entrenchment-and-Conventionalization Model. *Yearbook of the German Cognitive Linguistics Association* 3 (1). 3–26.
- Schmid, Helmut. 2013. Probabilistic part-of-speech tagging using decision trees. In D.B. Jones & H. Somers (eds.), *New methods in language processing*, 154–164. London: Routledge.
- Schneider, Gerold, Mennatallah El-Assady & Hans Martin Lehmann. 2017. Tools and methods for processing and visualizing large corpora. *Studies in Variation, Contacts and Change in English* 19. Helsinki: VARIENG, University of Helsinki. http://www.helsinki.fi/varieng/series/volumes/19/schneider_el-assady_lehmann/
- Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27 (3), 379–423.
- Simpson-Vlach, Rita & Ellis, Nick C. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31 (4). 487–512.
- Teich, Elke, Peter Fankhauser, Stefania Degaetano-Ortlieb & Yuri Bizzoni. 2021. Less is more/more diverse: On the communicative utility of linguistic conventionalization. *Frontiers in Communication* 5. 620275.
- Tourtouri, Elli N., Francesca Delogu, Les Sikos & Matthew W. Crocker. 2019. Rational over-specification in visually-situated comprehension and production. *Journal of Cultural Cognitive Science* 3 (2). 175–202.
- Universal Dependencies. 2024a. compound: compound. (<https://universaldependencies.org/en/dep/compound.html>) (Accessed 2026-02-12.)

- Universal Dependencies. 2024b. compound:prt: Phrasal verb particle. (<https://universaldependencies.org/en/dep/compound-prt.html>) (Accessed 2026-02-12.)
- Universal Dependencies. 2024c. fixed: fixed multiword expression. (<https://universaldependencies.org/en/dep/fixed.html>) (Accessed 2026-02-12.)
- Universal Dependencies. 2024d. flat: flat expression. (<https://universaldependencies.org/u/dep/flat>) (Accessed 2026-02-12.)
- Venhuizen, Noortje J., Matthew W. Crocker & Harm Brouwer. 2019. Semantic entropy in language comprehension. *Entropy* 21 (12). 1159.