# Multi-word Expressions in English Scientific Writing

**Diego Alves**
Saarland University, Germany
diego.alves@uni-saarland.de

**Stefan Fischer**
Saarland University, Germany
stefan.fischer@uni-saarland.de

**Stefania Degaetano-Ortlieb**
Saarland University, Germany
s.degaetano@mx.uni-saarland.de

**Elke Teich**
Saarland University, Germany
e.teich@mx.uni-saarland.de

## Abstract

Multi-Word Expressions (MWEs) play a pivotal role in language use overall and in register formation more specifically, e.g. encoding domain-specific terminology. Our study focuses on the identification and categorization of MWEs used in English scientific writing, considering their formal characteristics as well as their developmental trajectory over time from the mid-17th century to the present. For this, we develop an approach combining three different types of methods to identify MWEs (Universal Dependency annotation, Partitioner and the Academic Formulas List) and selected measures to characterize MWE properties (e.g., dispersion by Kullback-Leibler Divergence and several association measures). This allows us to inspect MWE types in a novel data-driven way regarding their functions and change over time in specialized discourse.

## 1 Introduction

Regularity in language does not only concern structural aspects (syntax, morphology), but also the way we combine words. Some word combinations are perceived as patterns that are associated with specific meanings or connotations, collectively referred to as multi-word expressions (MWEs). MWEs range from idioms that are formally fixed and have a figurative meaning (e.g., *kick the bucket*) to compounds (*bus ticket*) or phrasal verbs (*take a ride*) that are typically compositional and often lexically fairly productive (cf. Avgustinova and Iomdin (2019)).

MWEs are ubiquitous since they contribute to language efficiency by having highly predictable transitions from one word to the next or, if highly conventionalized, they can be retrieved from the lexicon rather than processed incrementally, thus giving them a processing advantage over other word sequences. Furthermore, MWEs play a crucial role in register formation as they provide conventional encodings of context-specific meanings. For example, MWEs such as *in no event* or *said therein* are typical of legal language and rarely encountered elsewhere, while noun-noun and adjective-noun combinations such as *iron oxide* or *sulphuric acid* are a typical type of MWEs used in scientific language forming domain-specific terminology.

In particular, we are interested in MWEs in scientific English from a diachronic perspective (mid 17th century to today). Scientific English develops into a recognizable register during the late modern period and becomes highly conventionalized in modern times. Overall, we want to better understand the process of register formation and whether processing advantages might have an impact. Specifically, we ask (i) what are the MWEs used in scientific English, (ii) which specific *types* of MWEs are used in the scientific domain, and (iii) how to characterize the diachronic development of MWEs in this domain, e.g., do specific MWEs trend in particular periods, do MWEs become more/less fixed and/or productive over time, thus contributing to conventionalization. For instance, we expect that MWEs serving domain-specific terminology (such as noun-noun compounds) will be quite agile and productive, while others, e.g., expressions of stance (e.g. *it is interesting*) will change at a lower rate and be less varied. We develop an approach to identify MWEs in scientific texts in order to be able to address these questions and better understand the role MWEs play in register formation. The scientific domain is well-suited as it encompasses different types of MWEs ranging from scientific terms up to stylistic expressions. In this paper, we take a first step towards answering the above questions, focusing on the identification of MWEs and possibilities of categorization into types by applying dispersion and association measures to our diachronic data set.

The remainder of the paper is organized as



*Proceedings of LaTeCH-CLfL 2024*, pages 67–76

follows. In Section 2 we discuss related work on MWEs in computational linguistics/NLP. Sections 3 and 4 presents our methods and results, including a preliminary diachronic analysis. We conclude with a summary and outlook (Section 5).

## 2 Related Work

From a linguistic perspective, there are numerous corpus-based accounts of MWEs in different registers, including the scientific one (e.g. Biber and Barbieri (2007); Hyland (2008); Liu (2012)). Some of these descriptions include lists of MWEs used in academic texts that are freely available. We make use of one specific list (Simpson-Vlach and Ellis, 2010) in our own approach (Section 3), but obviously such lists are always biased (time, corpus). Therefore, a sound, reusable computational method for identifying MWEs is needed, including analysis of their inherent properties (such as transparency, compositionality, fixedness; cf. also Schulte im Walde and Smolka (2019)).

From a computational point of view, MWEs have been considered "a pain in the neck" (Sag et al., 2002) because they are not trivial to identify, let alone classify, by standard language models or NLP tools. Two formal criteria of MWEs are *predictability* in a given context (e.g., register) and relative *fixedness* of the elements within the expression. In combination with relative frequency, well-established measures to assess MWE candidates are pointwise mutual information (PMI) and log-likelihood, commonly also applied to identify collocations (e.g., Evert (2008); Ramisch et al. (2010); Wahl and Gries (2018); Fabre et al. (2019)).

Regarding the identification of MWEs using machine-learning methods, Ramisch et al. (2023) conducted a survey on existing MWE corpora and evaluation methods. They showed that most of the existing tools for this specific NLP task are based either on DiMSUM (Schneider et al., 2016) or PARSEME (Savary et al., 2015) corpora and that, due to the complexity of the task and differences between approaches, results are not always comparable. PARSEME acknowledges discontinuity, variability, nesting, and overlaps and has a fine-grained MWE classification, however, it considers only verbal MWEs. On the other hand, DiMSUM corpus is annotated for most major MWE categories (i.e., nominal, verbal, adverbial, and functional), but does not include any category labels. Considering the scope of our study, the broader coverage of DiMSUM seems more relevant and aligned with our aims. The complexity of the automatic extraction of MWEs is noticeable in works such as Tanner and Hoffman (2023) where state-of-the-art tools have F1-scores below 65.

Given the complexity of the task, different approaches focus on different aspects of MWEs, so we decided to combine the state-of-the-art approaches for a more comprehensive treatment.

## 3 Methods

### 3.1 Dataset

As the main objective of this study is to investigate the role of MWEs in the development of English scientific writing, we decided to use the Royal Society Corpus (RSC) 6.0, a diachronic corpus of scientific English covering the period from 1665 until 1996. The RSC comprises 47 837 texts (295 895 749 tokens), mainly scientific articles covering a wide range of areas from both the mathematical and physical sciences and the biological sciences, and is based on the Philosophical Transactions and Proceedings of the Royal Society of London (Fischer et al., 2020).

Given its fair size and time coverage, the RSC is not only particularly relevant for diachronic linguistic analysis (e.g., Feltgen et al. (2017); Degaetano-Ortlieb and Teich (2018); Degaetano-Ortlieb and Teich (2022)), but also for historical and cultural analysis (e.g., Fyfe et al. (2015); Moxham and Fyfe (2018)).

### 3.2 Extraction of Multi-word Expressions

To identify and extract MWEs from the RSC corpus, we combined three different approaches which are schematised in Figure 1. The idea was to increase the number of identified MWEs, reducing biases related to the recall of each approach. From each method, we extract a list of MWEs which are, then, merged into the final RSC MWE list. For each method, MWEs are extracted in lowercase. Each method is described in detail in the subsections below.

#### 3.2.1 Universal Dependencies Method

The Universal Dependencies[1] (UD) guidelines for the annotation of dependency relations (De Marneffe et al., 2021) include 5 dependency labels which concern MWEs: i) compound - combinations of tokens that morphosyntactically behave as single

[1] https://universaldependencies.org/

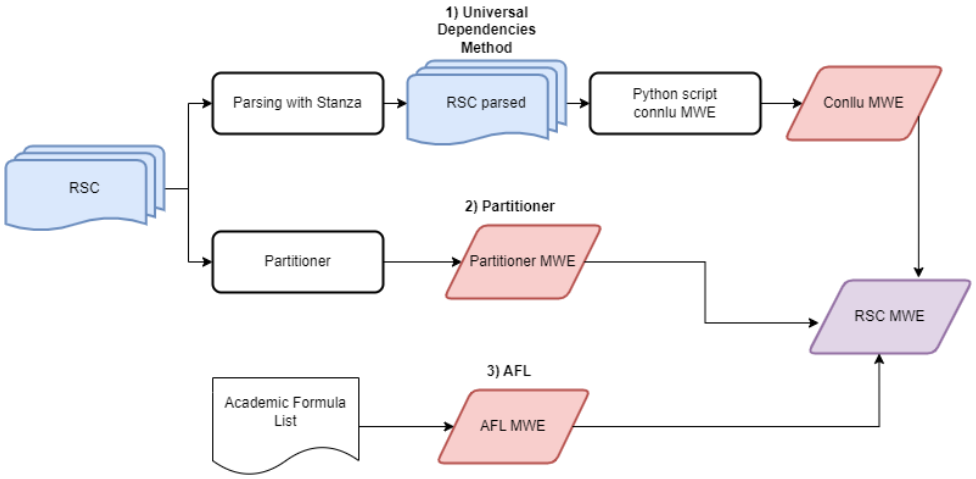


Figure 1: Methodology for extracting MWEs from the RSC corpus.

words. In English, we find most commonly nominal compounds written as separate words, for example, *orange juice*; b) compound:prt - phrasal verbs (e.g., *shut down* and *find out*); c) 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*; d) flat:foreign - sequences of foreign words; and e) fixed - used for certain fixed grammaticalized expressions which tend to behave like function words (e.g., *because of*, *in spite of*, *as well as*).

CoNNL-U is the standard format for texts containing morphosyntactic annotations following the UD guidelines. It is supported by state-of-the-art dependency parsers (e.g., Stanza) and can be easily queried for specific syntactic information. Thus, from a parsed corpus, it is possible to identify the word units composing the different types of MWE according to the UD framework.

The RSC 6.0 was parsed using Stanza tool (Qi et al., 2020) and the combined model for the English language provided by the developers which was trained with different UD corpora. Then, we developed a Python script using pyconll library[2] to identify and count the MWEs in the RSC texts per year and merged the results in a unified list of UD MWEs concerning all RSC.

The UD method for extracting MWEs depends on the accuracy of the parser, thus, although Stanza is a state-of-the-art tool for dependency annotation, some errors are inevitable. A manual evaluation of 70 sentences (10 per 50-year period of the RSC) showed that the accuracy of the parser is equal or higher than 85% for compound:prt, compound, and fixed MWEs, and equal to 75% for flat ones. The scores are quite consistent throughout the different time periods[3].

Another bias related to the UD method concerns the fact that many MWEs are not captured as they are described syntactically with dependency relations different from the 5 ones described above (e.g., *in terms of*, *so far*, *as so*). Therefore, this method alone is not enough for a global diachronic analysis of MWEs.

### 3.2.2 Partitioner

Partitioner 0.1.2[4] is a Python module that performs tokenization with generalisations into MWE segmentation using a supervised machine learning algorithm (Williams, 2016). It was presented by Tanner and Hoffman (2023) as one of the state-of-the-art tools for MWE extraction (evaluated using the DiMSUM corpus).

We applied the partitioner method to the RSC texts and, as was the case of the UD method, extracted the ensemble of MWEs in the RSC and also identified the MWEs occurring each year.

The partitioner memory overhead comes from the English Wikipedia data set, thus, it may also fail in identifying certain MWEs from the earlier periods of the RSC. Moreover, although it is a state-of-the-art tool (with better recall than the others listed by Tanner and Hoffman (2023)), it is clear that it is not possible to identify all MWEs in our corpus.

[2]https://github.com/pyconll/pyconll

[3]flat:foreign was not evaluated as this class is rare in the RSC and it did not appear in the evaluation set

[4]https://pypi.org/project/partitioner/

#### 3.2.3 Academic Formulas List

The third approach that we selected regarding MWE identification in the RSC concerns the Academic Formulas List (AFL), which is a list of the most common formulaic sequences in academic English. It is composed of a core list of 207 formulaic expressions found in written and spoken academic language, a specific list of 200 expressions from written corpora, and another one (also with 200 expressions) based on spoken academic English texts (Simpson-Vlach and Ellis, 2010). The AFL multi-word expressions were identified by the authors with a special measure of usefulness called the formula teaching worth (FTW), which combines frequency and mutual information measures.

Using a Python script, we identified and counted all AFL MWEs in the RSC. In total, 506 out of the 607 MWEs in the AFL occur in our corpus. As expected, most of the AFL expressions that do not appear in the RSC concern the ones from the AFL spoken list (e.g., *I'll talk about*, *gonna talk about*, *let's look at*).

#### 3.2.4 Merged MWEs

Once we identified and counted the MWEs in the RSC with the three methods, we merged the lists to create our final set of MWEs.

Since UD MWEs are grammatically motivated and AFL MWEs were selected using specific measures, we kept all the elements from these lists. However, regarding the partitioner method, we consider only the MWEs with frequency (in the whole corpus) >3, following the threshold defined by Gries (2022). The aim is to avoid, in our final list, syntagmas such as a determiner followed by a noun as well as other sequence of tokens which are not MWEs as they are not grammatically motivated like flat and compound structures and are not frequent in the text to be considered a collostructure. Moreover, we decided to exclude MWEs composed only of numbers.

Regarding the frequency values, if the MWE appeared in more than one list, we considered its frequency to be the highest number when comparing values from the different approaches.

### 3.3 Dimensions of Information

Several measures are described in the literature to characterize MWE properties. Gries (2022) defined eight different ones which he used to identify MWEs in the Brown corpus (Francis and Kucera, 1979) using a multi-dimensional strategy based on an information-theoretic approach.

In our case, MWEs were extracted using automatic methods, thus, our aim regarding dimensions of information is to use these metrics to describe the multi-word units identified in the RSC. Besides the MWE frequency provided by the scripts of the three approaches, we also calculated, for each MWE, its dispersion and association values across years.

#### 3.3.1 Dispersion

The dispersion measure assesses the spread of an MWE within a corpus. It is defined by Gries (2022) as a normalized version of the Kullback–Leibler divergence (KLD), which is a unidirectional measure quantifying how much in percent of a word's total occurrences in each corpus part diverges from the corpus part sizes in percent. Dispersion values vary from 0 to 1, the higher that number, the more heterogeneously distributed the MWE is. In this study, dispersion for each MWE was calculated across time by subdividing the RSC per year.

Thus, with the frequency of each MWE and the size of each corpus part (number of tokens) per year, it was possible to calculate the normalized dispersion values of all MWEs of our merged list.

#### 3.3.2 Association

The Association measures of bi-grams are defined as (i) the degree to which the first token attracts the second one, and (ii) the degree to which the second token of the MWE attracts the first. For n-grams with $n > 2$, we calculate as many association measures as necessary to describe the whole MWE, considering the whole left context. For example, for the MWE *in spite of*, we calculate: a) association of *in* and *spite*; b) Association of *spite* and *in*; c) association of *in spite* and *of*; and c) association of *of* and *in spite*.

Associations measures are also obtained using normalized KLD as described by Gries (2022). Thus, for each MWE from the merged list, we calculated the different association values considering the whole corpus and also for each 50-year period of the RSC.

## 4 Results

### 4.1 Extraction of MWEs

We present in Table 1 the details of the RSC MWE list in terms of the number of MWE types per class of MWE: (i) UD MWEs correspond to MWEs identified only with the UD method as well as the ones

identified by both UD and partitioner approaches; (ii) Other MWEs are the partitioner MWEs which do not appear in the UD list; and (iii) AFL MWEs are the MWEs provided by the AFL approach.

| Method | MWE |
|---|---|
| UD | 3 147 597 |
| Other | 181 659 |
| AFL | 506 |
| **Total** | 3 329 762 |

Table 1: Number of MWE types of each extraction approach and for the RSC MWE merged list.

It is possible to notice that the majority of the RSC MWEs (94%) come from the UD method. This is due to our decision to keep even the MWEs extracted via this method with a frequency $< 3$. Moreover, most MWEs in our list (69%) appear only once in the whole corpus.

Table 2 presents the distribution of the number of MWE types regarding UD MWEs in terms of dependency relation.

Compound and flat are the UD MWE classes with the highest number of MWE types in the RSC, however, they have a high number of types that occur only once in the corpus (hapax percentage higher than 70%). Most of these MWEs correspond to specific entities that are only mentioned in the precise context of specific articles (e.g., *oligocene regime*; *wavelength translators*; *Prince Joseph Oscar*) and did not become part of scientific terminology. The flat:foreign class is essentially composed of hapaxes, thus, of lesser interest for our study. The flat:foreign MWEs with a frequency $> 1$ (8 in total) concern mostly parsing errors (e.g., *J. McLean*, *complete collection*, *rb 27*).

Figure 2 presents the relative frequency of each MWE class per year in the RSC. It is possible to notice a clear tendency of increasing the usage of compounds in scientific English (as observed previously by Degaetano-Ortlieb and Teich (2018)), with a more pronounced slope from the second half of the nineteenth century. Moreover, flat MWEs seem to become increasingly more common specially in the second half of the twentieth century.

| UD MWE | MWE | % |
|---|---|---|
| compound | 2 523 696 | 80.2 |
| flat | 604 057 | 19.2 |
| compound:prt | 16 337 | 0.5 |
| fixed | 3 107 | 0.1 |
| flat:foreign | 400 | 0.0 |

Table 2: Distribution of UD MWEs in terms of dependency relation.

Furthermore, applying the Mann-Kendall trend test to each class (Hussain and Mahmud, 2019), with the exception of phrasal verbs (i.e., compound:prt), all the other classes present an overall increasing tendency (p-value below 0.05) even though, in some cases, decreasing periods are observed.

### 4.2 Dimensions of Information

#### 4.2.1 Dispersion and Association overview

As previously mentioned, we focus our analysis on two specific dimensions of information: dispersion and association. Thus, to have a graphic overview of the distribution of the different classes of MWEs identified in the RSC according to these metrics, we plotted the graphs presented in Figure 3.

Each graph represents the MWEs of the specific class in red, and in black, the other ones. To improve the visualisation, we plot only the types with a frequency $> 10$. For dispersion, each type has one value, while for association, the number depends on the number of words composing the MWE. Therefore, what is plotted corresponds to the mean of the different association values[5].

As expected, the different classes of MWEs behave differently in terms of distribution regarding dispersion and association metrics.

Most AFL MWEs are positioned in the lower left quadrant, thus indicating that these units are fairly well distributed within the RSC but with low mean association values. This is due to the fact that most of the AFL MWEs are composed of words that appear in many other contexts (e.g., *that is the*, *it is important*, *on the other*). The AFL elements with a mean association value around 0.5 are the ones where at least one word is more usually present in that specific construction (e.g., *in accordance with*). Few AFL MWEs are positioned in the upper left quadrant (i.e.; not homogeneously dispersed in the RSC) and usually concern MWEs with personal pronouns (mostly *I* and *you*).

Compound and flat classes are the ones with the highest number of MWE types. In both cases, most of the MWEs are positioned in the upper quadrants of the graphs. However, compound MWEs have a

[5]These graphs are available in the html format at: http://tinyurl.com/2pd8n7s8

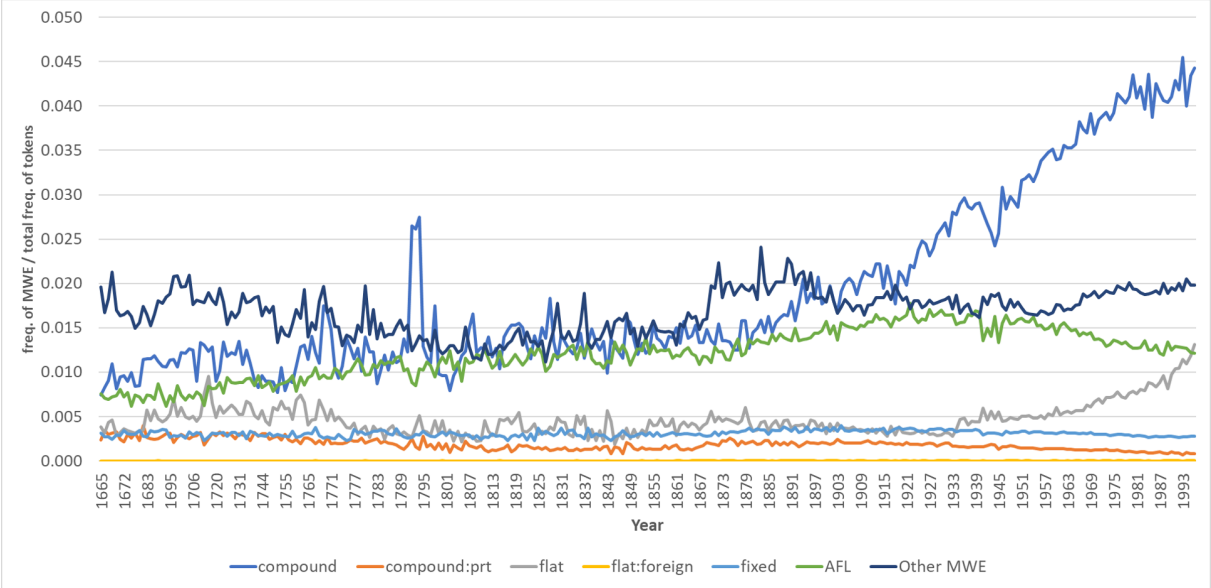


Figure 2: Relative frequency of the different classes of MWE per year in the RSC.

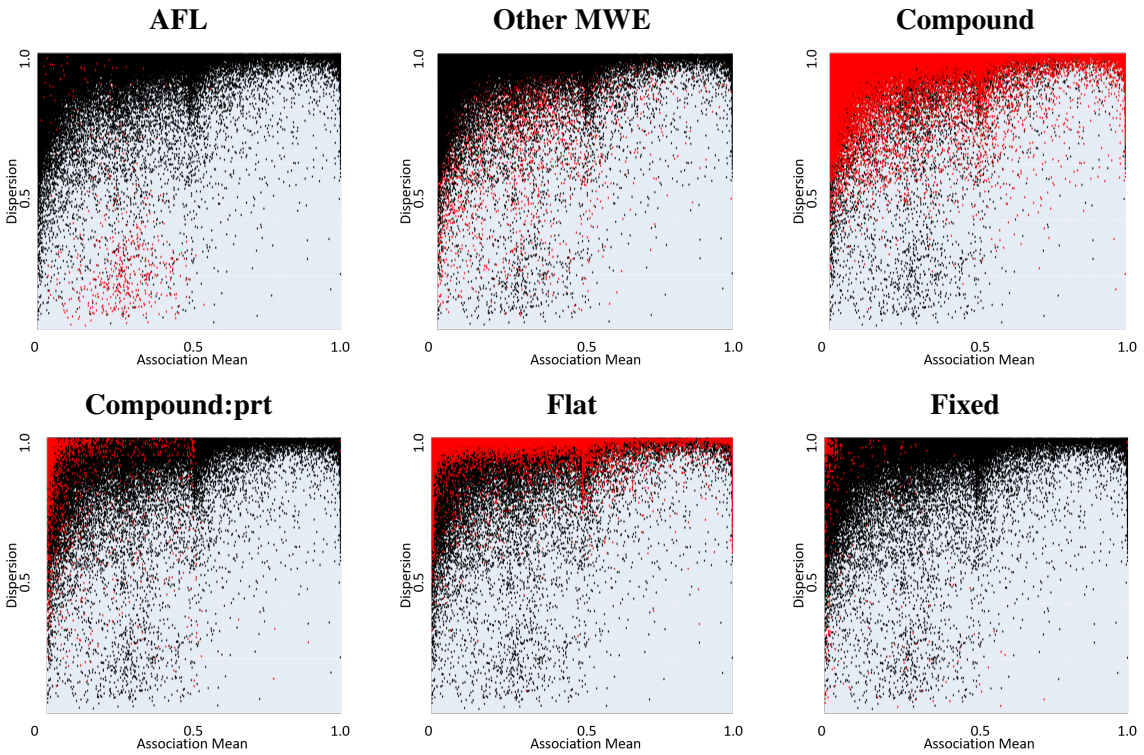


Figure 3: Distribution of the RSC MWEs in terms of Dispersion and Association. Each graph presents in red the class of MWE specified in its title.

better distribution in terms of dispersion in these quadrants. For these two classes, the elements in the upper right quadrant usually correspond to very specific scientific terms such as name of species (e.g., *Ambystoma mexicanum*), while the MWEs in the upper left side are composed of words that are more generic (e.g., *Mr Baker*, *Dr Davies*, *phase modulation*). The compound and flat MWEs in the lower right quadrant are very frequent terms that are quite homogeneously dispersed in the RSC, for example: *Royal Society*, *New York*, *refractive index*, *differential equations*, *standard deviation*. The scientific terms in the lower quadrants correspond to broad concepts usually applied in different fields.

Phrasal verbs (compound:prt) present a particular behaviour. Usually, the association value of its preposition in regard to the verb is very low as this element can appear in a large variety of other contexts. Thus, it explains the fact that most compound:prt MWEs are positioned in the left quadrants. More specifically, the vast majority of these MWEs are not well dispersed in the RSC, thus, the upper left quadrant is the most populated one with this class. Some specific phrasal verbs which are common in the scientific language are better distributed in our corpus, such as: *carried out*, *pointed out*, *depend on*. Phrasal verbs with mean association values close to 0.5 are the ones for which the verb is not encountered in other contexts in the RSC (e.g., *churned up*, *smoothes out*, *budded off*).

Regarding fixed MWEs, we observe a cluster of elements on the left upper side and many others spread over the left lower quadrant. The MWEs in the upper side correspond to unusual terms (frequency below 50) such as *according with*, *one other*, *without than*, while the ones in the lower part of the graph occur more frequently (more than 100 occurrences). Moreover, some fixed MWEs that are homogeneously distributed in the RSC have a mean association value closer to 0.5, indicating that at least one of its units is strongly attracted to the other words forming the MWE. This is the case for *due to*, *less than*, *rather than*, etc.

Finally, the MWEs from the "Other MWE" class (extracted using the partitioner tool) are mostly present on the left upper and lower parts of the graph. A qualitative analysis of these terms shows that some of them correspond to nominal phrases composed of an adjective and a noun (e.g., *electrical stimulation*, *practical applications*). Also, many of-genitive examples can be identified as "Other MWE" such as *University of Bristol* and *Department of Chemistry*. We also notice many cases of discourse markers such as *at first sight*, *for example*, etc.

#### 4.2.2 Diachronic Analysis of Association Values

To analyse the diachronic evolution of MWEs in scientific English, we focus on the examination of the mean association values throughout the 50-year periods. As described in Section 3.3.2, for each period, we calculated the association metrics for each MWE present in the sub-corpus as well as its mean value.

Our aim is to check whether the identified MWEs became more or less fixed in time, a phenomenon that can indicate possible conventionalization processes in this specific register of the English language. Therefore, we used the Mann-Kendall trend test[6] which is suited to the analysis of time series data regarding increasing or decreasing trends (Hussain and Mahmud, 2019).

As we are interested in diachronic trends, only MWEs that appear in at least 2 periods were examined (316 390 out of 3 329 762). Thus, we applied the original Mann-Kendall test proposed by the pymannkendall module to these MWEs and extracted the following results: i) trend: increasing or decreasing (if p-value $< 0.05$) and no trend (if p-value $> 0.05$) and ii) slope: value representing the rate of change (positive for increasing values of mean association and negative when decreasing).

Table 3 presents the results for each class of MWE in the RSC with detailed information regarding the number of MWEs with increasing and decreasing trends as well as the number of elements where no trend was observed. Overall, it is possible to observe that, for all classes of MWEs, the number of elements presenting no statistically valid trend is higher than in the cases where an increase or decrease is attested. Moreover, considering the percentage of MWEs with an increasing or decreasing tendency, AFL is the class with the highest number of MWEs where changes have occurred (34%), Fixed and Compound:prt classes present changes for 5 to 8%, and for the other classes, the percentage is below 1%.

Besides having the highest percentage of statistically valid trends, AFL is the only class for which the amount of MWEs with an increasing trend is

[6] pymannkendall 1.4.3 Python module available at: https://pypi.org/project/pymannkendall/

| Trend | Compound | Compound:prt | Flat | Fixed | AFL | Other MWE |
|---|---|---|---|---|---|---|
| Increasing | 60 | 15 | 20 | 7 | **149** | 61 |
| Decreasing | **437** | **181** | **84** | **40** | 13 | **365** |
| No trend | 128 007 | 3 895 | 18 781 | 630 | 309 | 48 146 |
| Total | 128 504 | 4 091 | 18 885 | 677 | 471 | 48 572 |

Table 3: Mean Association values trends for each class of MWEs in the RSC. In bold are highlighted the highest values comparing increasing and decreasing trends.

| MWE Class | Increasing Trend | Decreasing Trend |
|---|---|---|
| compound | *North Carolina*, *University College*, *Great Britain* | *os ilium*, *radius vector*, *os sacrum* |
| compound:prt | *depend upon*, *carry out*, *break down* | *set down*, *taken out*, *let loose* |
| flat | *St. Petersburgh*, *red deer*, *J. D.* | *Dr. Johnson*, *Thomas Barker*, *James Stirling* |
| fixed | *of course*, *no doubt*, *whether or not* | *as if*, *some other*, *it is* |
| AFL | *should be noted*, *the other hand*, *on the other hand* | *of the same*, *and if you*, *a kind of* |
| Other MWE | *prime minister*, *at first sight*, *give rise* | *at variance*, *inmost recesses*, *in all likelihood* |

Table 4: Top-3 MWEs per class with increasing and decreasing trends in terms of mean association value.

higher than the decreasing one. As previously explained, AFL MWEs were identified in corpora of academic English using specific metrics, therefore, they correspond to formulaic expressions specific to this register, similar to RSC.

Table 4 presents, for each class, the three MWEs with the highest rate of increase and decrease of the mean value of association measures. These results show that, in the evolution of scientific writing, specific lexical groups regarding this domain became more fixed, thus, indicating a conventionalization process. For the other classes, changes are less significant and the predominance of decreasing trends can be due to the MWEs' semantic characteristics. Compound, flat, and some Other MWEs usually refer to entities, thus related to the evolution of research topics and their terminology. On the other hand, the decrease observed in phrasal verbs and fixed MWEs could be related to a tendency of standardisation in terms of lexical choices, with some specific elements from these classes being preferred over the others.

## 5 Conclusions and Future Work

We presented a multifaceted approach to identify MWEs in scientific English for analysing their evolution from the mid-17th century. Our approach uniquely combines three distinct methods: (1) Universal Dependency annotation, which was key in uncovering syntactic structures of MWEs, (2) Partitioner, segmenting texts to detect MWEs effectively, and (3) the Academic Formulas List, which further enriched our analysis by providing a benchmark for MWEs used in the scientific context. Our methodology went beyond identification; we used tools like Kullback-Leibler Divergence for dispersion analysis and various association measures to characterise MWEs (cf. Gries (2022)). This revealed their dynamic nature and evolving roles in scientific discourse. Some MWEs adapted over time, reflecting changes in scientific language, while others remained consistent, signifying their entrenched role in scientific communication.

Our findings not only enhance understanding of MWEs in scientific English but also pave the way for future linguistic research, particularly in language evolution and specialized registers. We currently work on integrating MWEs in word embeddings to classify them semantically and model their temporal dynamics in terms of productivity. Also, we intend to compute surprisal of MWEs to link up with processing explanations (e.g. Siyanova-Chanturia et al. (2017); Bhattasali et al. (2020)).

## Acknowledgements

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

## References


Tanya Avgustinova and Leonid Iomdin. 2019. Towards a typology of microsyntactic constructions. In *Proceedings of the International Conference on Computational and Corpus-Based Phraseology*, pages 15–30.

Shohini Bhattasali, Murielle Fabre, Christophe Pallier, and John Hale. 2020. Modeling conventionalization and predictability within MWEs at the brain level. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 313–322, New York, New York. Association for Computational Linguistics.

Douglas Biber and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26:263–286.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the 2nd SIGHUM LaTeCH-CLfL workshop*, pages 22–33.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.

Stefan Evert. 2008. Corpora and collocations. In *Corpus linguistics. An international handbook*, volume 2, pages 1212–1248. Mouton de Gruyter.

Murielle Fabre, Shohini Bhattasali, Christophe Pallier, and John Hale. 2019. Modeling Conventionalization and Predictability in Multi-Word Expressions at Brain-level. Proceedings of the Society for Computation in Linguistics.

Quentin Feltgen, Benjamin Fagard, and Jean-Pierre Nadal. 2017. Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. *Royal Society Open Science*, 4(11):170830.

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.

W. Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Letters to the Editor*, 5(2):7.

Aileen Fyfe, Julie McDougall-Waters, and Noah Moxham. 2015. 350 years of scientific periodicals. *Notes and Records: the Royal Society journal of the history of science*, 69(3):227–239.

Stefan Th. Gries. 2022. Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis. Journal in English Lexicology*, (19).

Md. Manjurul Hussain and Ishtiak Mahmud. 2019. pyMannKendall: a python package for non parametric Mann Kendall family of trend tests. *Journal of Open Source Software*, 4(39):1556.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27:4–21.

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

Noah Moxham and Aileen Fyfe. 2018. The Royal Society and the prehistory of peer review, 1665–1965. *The Historical Journal*, 61(4):863–889.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. MWEtoolkit: A framework for multiword expression identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. ELRA.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

Ivan A. Sag, Tim Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *International conference on intelligent text processing and computational linguistics*, pages 1–15, Berlin. Springer.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 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)*, Poznań, Poland.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. Semeval-2016 task 10: Detecting minimal semantic units and their meanings (dimsum). In *Proceedings of the 10th International*

*Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559.

Sabine Schulte im Walde and Eva Smolka, editors. 2019. *The role of constituents in multiword expressions*. Number 4 in Phraseology and Multiword Expressions. Language Science Press, Berlin.

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

Anna Siyanova-Chanturia, Kathy Conklin, Sendy Caffarra, Edith Kaan, and Walter J.B. van Heuven. 2017. Representation and processing of multi-word expressions in the brain. *Brain and Language*, 175:111–122.

Joshua Tanner and Jacob Hoffman. 2023. MWE as WSD: Solving multiword expression identification with word sense disambiguation. *arXiv preprint arXiv:2303.06623*.

Alexander Wahl and Stefan Th. Gries. 2018. Multi-word expressions: A novel computational approach to their bottom-up statistical extraction. In Pascual Cantos-Gómez and Moisés Almela-Sánchez, editors, *Lexical Collocation Analysis: Advances and Applications*, pages 85–109. Springer International Publishing, Cham.

Jake Ryland Williams. 2016. Boundary-based MWE segmentation with text partitioning. *arXiv preprint arXiv:1608.02025*.