Diachronic Analysis of Phrasal Verbs in English Scientific Writing

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

Abstract

Phrasal verbs (PVs) are a specific type of multi-word expression and a specific feature of the English language. However, their usage in scientific prose is limited. Our study focuses on the analysis of phrasal verbs in the scientific domain using information theory methods to describe diachronic phenomena such as conventionalization and diversification regarding the usage of PVs. Thus, we analysed their developmental trajectory over time from the mid-17th century to the end of the 20th century by measuring the relative entropy (Kullback-Leibler divergence), predictability in the context of the phrasal verbs particles (surprisal), and the paradigmatic variability using word embedding spaces. We were able to identify interesting phenomena such as the process of conventionalization over the 20th century and the peaks of diversification throughout the centuries.

1 Introduction

Multi-word Expressions (MWEs) are sequences composed of two or more words that have a degree of conventionality among speakers of the language community, holding a strong relationship in communicating meaning (Siyanova-Chanturia and Sidtis, 2018). MWEs encompass idioms that are formally fixed and have a figurative meaning (e.g., *kick the bucket*), compounds (*bus ticket*), phrasal verbs (*take a ride*), and other formulaic expressions that are typically compositional and often lexically fairly productive (cf. Avgustinova and Iomdin (2019)).

MWEs contribute to language efficiency due to the highly predictable transitions from one word to the next and/or because of their high degree of conventionalization (i.e., convergence in linguistic usage over time). Also, MWEs have a strong influence on register formation, providing conventional encodings of context-specific meanings.

We are principally interested in MWEs in scientific English from a diachronic perspective (mid-17th century to today). Scientific English developed into a recognizable register during the late modern period and became highly conventionalized in modern times (cf. Degaetano-Ortlieb and Teich (2022)).

However, phrasal verbs (PVs), despite being a specific type of multi-word expression and one of the most distinctive features of the English language, are less common in academic prose, when compared to other registers. In the scientific register, usually more specialized verbs are preferred (cf. Biber et al. (2021) and Brown et al. (2015)).

The usage of PVs in English scientific writing indicates specific lexical choices influenced by contextual configurations and communicative constraints. Thus, our aim is to investigate, using information theory measures, whether phrasal verbs contribute to standardization in scientific English as other types of MWEs and grammatical constructions do. Our idea is to analyse if the effects of conventionalization of phrasal verbs can be observed over time with three different approaches: 1) analysis of PVs temporal dynamics using relative entropy, 2) study of the predictability in the context of the PVs particles using surprisal measure, and 3) examination of the paradigmatic variability of PVs using embeddings.

The remainder of the paper is organized as follows. In Section 2 we discuss related work on PVs in scientific English. Sections 3 and 4 present our methods and results, followed by a discussion of the main findings in Section 5. We conclude with a summary and outlook (Section 6).

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 8–16
March 3-4, 2025

2 Related Work

As previously mentioned, PVs are known for being less common in scientific texts when compared to other registers. (Biber et al., 2000) shows that PVs are mostly used in speech and fiction. News texts tend to use less than these two genres, but academic prose is where PVs have the least overall frequency per million words.

Regarding diachronic analysis of PVs in scientific English, Alves et al. (2024) showed that compared to other types of MWEs, PVs are the only ones presenting a decrease in its relative frequency over time (mid-17th century to end of 20th century). Moreover, PVs present a specific behaviour regarding dispersion and association measures. In terms of dispersion, most PVs are not homogeneously distributed over time, only very specific ones commonly used in academic texts such as *carried out*, *pointed out*, and *depend on*. Regarding the association measures, as the verbs and particles are also found in other contexts, in most cases, the values were quite low, except for specific cases where the verb is mostly used with its particle (e.g., *churned up*, *smoothes out*, *budded off*). Although the authors present a preliminary diachronic analysis of the evolution of the association measure, no conventionalization study was presented.

The diachronic changes of the paradigmatic variability of different parts-of-speech in scientific English using word embedding space were analysed by Teich et al. (2021). Overall, there is a reduction of paradigmatic variability over time for the different grammatical classes. However, PVs were not analysed separately to see if their behaviour is similar or discrepant when compared to other verbs.

Moreover, there are numerous corpus-based studies 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 as part of English for Academic Purposes (EAP). However, since PVs are not commonly used in scientific texts, they are usually not considered in the analysis.

Regarding computational methods for identifying PVs, the PARSEME initative[1] clearly identifies PVs or verb-particle constructions (VPCs) as one category of verbal MWEs. Multilingual corpora annotated following PARSEME guidelines are available, however, without any diachronic data.

Finally, in terms of studies regarding the cognitive processing of PVs, most studies concern L2 learners and the difficulties of learning these specific MWEs (cf. Alejo-González (2010); Mohammed (2019); Alisoy (2023). In their study, Perdomo and Kaan (2023) looked at surprisal measures to analyse the effects in priming of phrasal verb construction alternations, comparing native speakers and L2 learners, thus, focusing on learning difficulties, not in conventionalization processes as this paper.

3 Methods

3.1 Dataset

As our objective is to investigate the conventionalization processes of PVs in the development of English scientific writing, we used the Royal Society Corpus (RSC) 6.0, which is a diachronic corpus of scientific English covering the period from 1665 until 1996.

It comprises 47,837 texts (295,895,749 tokens), which are mainly scientific articles covering a wide range of areas from mathematical, physical, and biological sciences, and is based on the Philosophical Transactions and Proceedings of the Royal Society of London (Fischer et al., 2020).

The RSC 6.0 was parsed using Stanza tool (Qi et al., 2020) and the combined model for English, provided by the developers, which was trained with different Universal Dependencies[2] (UD) corpora. To extract the PVs from the RSC, we developed a Python script using pyconll library[3] to identify and count the PVs[4] in the RSC texts per year. A manual evaluation of 140 sentences (20 per 50-year period of the RSC) showed that the accuracy of the Stanza parser is 90 regarding PVs.

3.2 Information Theory Measures

To analyse the diachronic phenomena regarding PVs in Scientific English, we applied three different methods to measure the relative entropy (Kullback-Leibler divergence), the surprisal of the particles, and the paradigmatic variability. The

[1] https://gitlab.com/parseme/corpora/-/wikis/home

[2] https://universaldependencies.org/

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

[4] Phrasal verbs are easily identified in texts parsed with UD corpora as the dependency label of the PV particle is *compound:prt* and its head is the verb.

workflow is schematized in Figure 4 and further described in the following sub-sections.

3.2.1 Kullback-Leibler Divergence

To identify evolutionary trends in the use of phrasal verbs (PVs) within the RSC, we applied relative entropy, specifically the Kullback-Leibler Divergence (KLD; Kullback and Leibler (1951)). This method compares probability distributions by measuring the additional bits required to encode dataset A when using a (non-optimal) model based on dataset B for a given set of elements X, as described in Equation 1. In this study, A and B correspond to sub-sets of the RSC (e.g. time slices) and X, i.e. the ensemble of PVs.

$$D_{KL}(A\|B) = \sum_{x \in X} A(x) \log \left(\frac{A(x)}{B(x)} \right) \quad (1)$$

The KLD measure provides an indication of the degree of divergence between corpora and identifies the features that are primarily associated with a difference. Possible discrepancies regarding the vocabulary size of the subcorpora are controlled by using Jelinek-Mercer smoothing and lambda 0.05 (cf. Zhai and Lafferty (2004) and Fankhauser et al. (2014)).

To detect periods of change in the use of PVs using KLD, we adopt the methodology described in Degaetano-Ortlieb and Teich (2018)[5]. We compare 20-year windows of past and present language use sliding with a 5-year gap over the timeline (e.g. t1=1665-1685, t2=1671-1691). Then, by plotting the divergence for each comparison on the timeline, we can inspect peaks or troughs which indicate a change: a peak is an indication that the divergence of the analysed feature increases, and is thus *typical* of the future 20 years in comparison to the past 20 years.

Due to the asymmetric characteristic of the KLD, we are only interested in the direction from subsequent periods to the preceding ones as we aim to determine periodization from past to present in the development of PVs usage in English scientific writing.

In this study, we examined the KLD at two different levels: a) all PVs combined, to verify if a conventionalization process can be identified and b) each PV individually, to identify individual diachronic phenomena.

[5]Degaetano-Ortlieb and Teich (2018) make the code available at: https://stefaniadegaetano.com/code/

3.2.2 Surprisal

Surprisal is formalized as the negative log probability of a unit in context which results in bits of information (Shannon, 1948), as defined in Equation 2.

$$Surprisal(unit_i) = \log_2(unit_i|Context) \quad (2)$$

A decrease in the surprisal of a specific term can indicate a conventionalization phenomenon as showed by Degaetano-Ortlieb and Teich (2022) regarding scientific English using a four-gram language model. N-grams surprisal models have limitations, thus, Steuer et al. (2024) propose a transformer-based surprisal model trained with the RSC corpus.

Our analysis concerns the diachronic changes of the surprisal values of the PVs particles. Thus, using the transformer-based model cited above trained over the RSC divided into 10-year periods, we extracted, per year of the RSC, the surprisal values of the particles identified in the parsed version of the dataset.

3.2.3 Paradigmatic Variability

To analyse diachronic changes in the paradigmatic context of PVs, we apply a context-aware version of entropy, paradigmatic variability, based on word embeddings and the close neighbours of a word in the vector space within a given radius (Teich et al., 2021). As previously mentioned, a drop in paradigmatic variability indicates a conventionalization phenomenon.

Regarding the word embeddings model, we used structured skip-grams (Ling et al., 2015) as it presents the advantage of representing each position in the left and right context separately, not as a mere bag of words as in simple skip-gram models.

The paradigmatic variability of a word over time is calculated by comparing period-specific word embedding models (i.e., per decade, from the 1660s to the 1990s). We followed the same procedure as presented in Teich et al. (2021), with the initialisation of the first decade being done with an atemporal embeddings model trained on the complete corpus as proposed by Fankhauser and Kupietz (2017). All following decades were then initialised with the embeddings of the previous decade. In our study, as our objective is to analyse PVs, the verbs and particles were joined

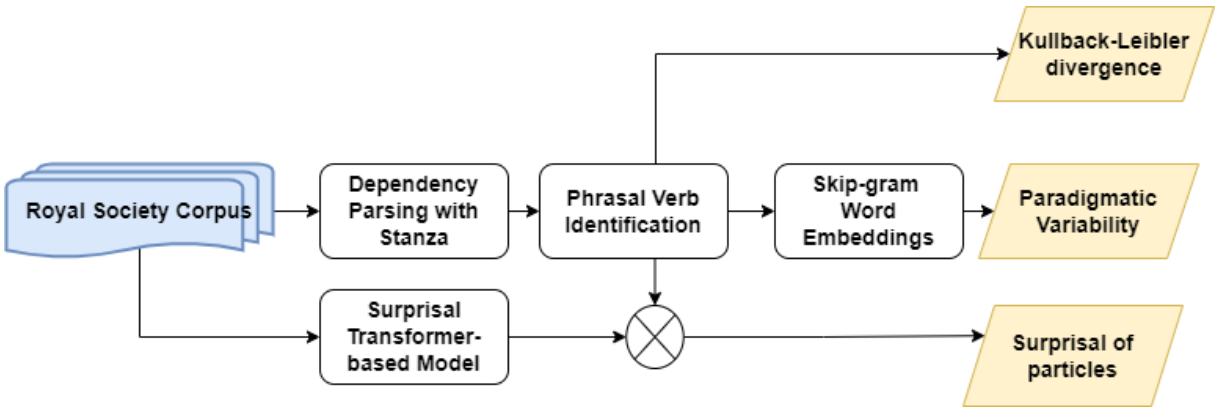

Figure 1: Experimental workflow.

with | | in the generation of the embedding space, thus allowing the differentiation between PVs and the other verbs.

This choice concerning the initialization process has the advantage of better representing low-frequency words in the embedding space, avoiding low-frequency words appearing in the centre of the space in the first few periods, and reducing bias regarding the movement in the embedding space over time. The subsequent statistical analysis of the vector space models only considers words with a frequency higher than 50[6].

Once the word embeddings for each decade are obtained, the paradigmatic variability of a word x, pvar(x), can be calculated as the entropy over a probability distribution, which is based on the probability $p(x_i|C_x)$ of a word x_i from the neighbourhood C_x being chosen instead of word x.

This is calculated using both the cosine similarity in the vector space between x_i and x and the frequency of x_i (freq(x_i)).

Thus:

$$\begin{aligned} \text{pvar}(x) &= \text{H}(\text{P}(.|C_x)) \\ &= -\sum_{\cos(x_i,x)>\theta} \text{p}(x_i|C_x)\log\left(\text{p}(x_i|C_x)\right) \end{aligned} \tag{3}$$

$$\text{with } \text{p}(x_i|C_x) = \frac{\cos(x_i,x)\,\text{freq}(x_i)}{\sum_{x_j}\cos(x_j,x)\,\text{freq}(x_j)} \tag{4}$$

The θ threshold was set to 0.6 and we considered a maximum of 30 neighbours. Thus, a word with a homogeneous distribution of neighbours has a high value of pvar(x).

[6]The other parameters used to generate the embeddings were: type 3; size 100; negative 10; hs 0; sample 1e-4; threads 4; binary 0; and iter 5.

4 Results

4.1 Kullback-Leibler Divergence

Figure 2 presents the relative entropy values (i.e., Kullback-Leibler divergence) of PVs in the RSC corpus over time as described in Section 3.2.1.

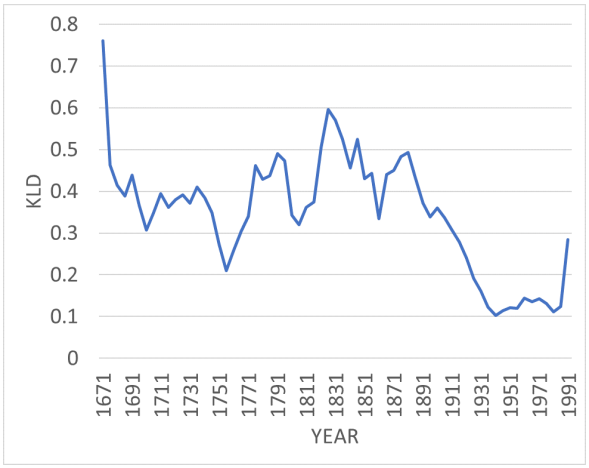

Figure 2: KLD measures for phrasal verbs in the Royal Society Corpus.

It is possible to observe peaks and troughs around the value of 0.4 from the seventeenth century to the end of the nineteenth century. On the other hand, at the beginning of the twentieth century, we clearly see a declining tendency of KLD, indicating, thus, a conventionalization in the usage of this feature, with a stabilization around 0.1 in the second half of this century.

To better understand the diachronic usage of PVs in scientific English, we also looked at the point-wise KLD, checking the relative entropy

shifts for each PV type. For each 20-year period used for the KLD calculation, we examined the number of PVs with positive values of divergence (i.e., PVs that became more typical), and the number with negative KLD (i.e., PVs that became less distinctive). Figure 3 shows the results of this analysis.

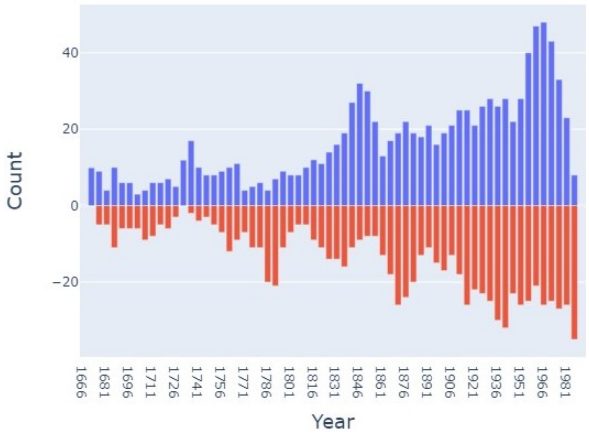

Figure 3: Number of phrasal verb types having positive (blue) or negative (red) KLD values per period of the RSC.

Besides the overall increasing trend in the number of PV types with positive KLD, it is possible to notice periods with higher increase, probably due to the specific textual needs of each period. Moreover, we can observe that the number of PVs with negative KLD also increases over time. Furthermore, the periods with more PVs having negative values of KLD, usually succeed periods where there is a peak in the number of PVs with positive values.

The increase in the number of PVs with positive KLD indicates a cyclical process of diversification (i.e., linguistic items acquiring different, more specific usages/meanings). Even though the overall relative frequency of PVs reduces over time, more different types are being used in specific periods. However, due to peaks regarding PVs with negative KLD, it seems that the usage of the new types does not become conventionalized.

4.2 Surprisal

As described in Section 3.2.2, another way of identifying possible conventionalization processes is using surprisal measures. PVs being MWEs, the surprisal of the particle is expected to be lower than the measure for the correspondent verbs. A decrease in time of the mean surprisal value of the particles indicates a conventionalization regarding the usage of these grammatical constructions.

Figure 4 presents the plot of the mean surprisal values of the phrasal verbs present in the RSC per year.

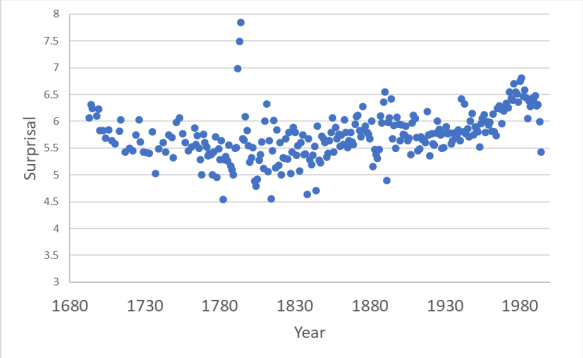

Figure 4: Mean surprisal value of phrasal verbs particles per year of the RSC.

Applying the Mann-Kendall trend test (Hussain and Mahmud, 2019), we observe that there is an overall statistically valid (i.e., a p-value below 0.00001) increasing tendency regarding the surprisal values of the particles.

This result can be correlated with the KLD observations. We observe that the conventionalization of the usage of PVs only happened in the twentieth century, moreover, throughout the centuries, we notice an increase in the usage of different PV types. Moreover, as shown by Alves et al. (2024), the relative frequency of PVs decreases over time in the RSC. All these factors contribute to an increase in the surprisal values.

In addition, it is also possible to notice that, even though there is an overall increasing tendency, there are periods with a decrease in the surprisal values and others with a more accentuated increase. When comparing Figures 3 and 4, we observe that periods with a high increase in the number of PV types (i.e., 1836-1856 and 1956-1976) also correspond to periods of accentuated increase regarding surprisal values.

Another factor that may influence the suprisal value is the distance between the verb and the particle. In the RSC, we find examples such as:

1. It suggested that development could be *broken down* into series of gene controlled chemical reactions. (*d = 1*, 1995)

2. . . . which it gently touched with little or no damage , *blowing* only *off* a few tiles. (*d = 2*,

1695)

3. … but BICHAT was continually *holding* a thing *up* by the wrong end … (*d* = *3*, 1823)

4. … that his assistance should be sought to *bring* the new edition *up* to the existing state … (*d* = *4*, 1908)

5. … and as I *wrote* many of the 336 them *down* from his own dictation … (*d* = *5*, 1840)

Thus, we decided to conduct an analysis of the diachronic evolution of the mean distance between verbs and particles in the RSC. Figure 5 shows the plot of the mean distance per 50-year period.

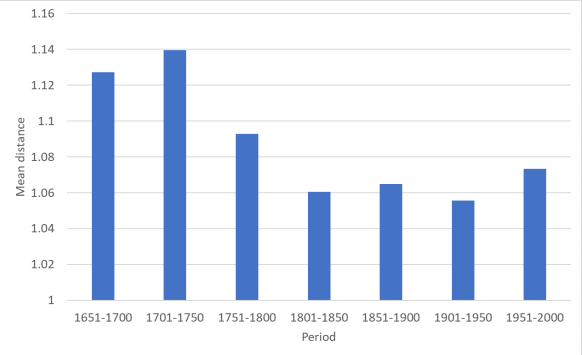

Figure 5: Mean distance between verbs and particles per 50-year of the RSC.

A statistical analysis of these results showed that p-value is below 0.001 for the following comparisons:

- 1701-1750 and 1751-1800
- 1751-1800 and 1801-1850
- 1901-1950 and 1951-2000

Thus, it is possible to notice a clear decrease in the mean value from the eighteenth century to the mid-nineteenth century, followed by a stabilization until the mid-twentieth century when a new increase is observed. The decreasing period regarding the mean distance between verbs and particles corresponds to a period with also a decrease in surprisal values (4). Moreover, the peak of surprisal (around 1970) is observed when there is also an increase in the mean distance.

4.2.1 Paradigmatic Variability

As previously explained in Section 3.2.3, using word embedding spaces, we calculated the paradigmatic variability of PVs per decade of the RSC. Figure 6 shows the results and the comparison with the variability of other verbs and all parts-of-speech in the dataset.

As shown by Teich et al. (2021), the paradigmatic variability of all words (i.e., all parts-of-speech) decreases over time as a general trend. This is due to two main mechanisms: conventionalization — a word becoming the dominant choice within its neighbourhood by frequency, possibly replacing other, alternative words — and diversification, i.e., words within a neighbourhood becoming more distant, leading to a split into two or more neighbourhoods.

Regarding non-phrasal verbs, they begin with a slightly higher paradigmatic variability than all POS, but end up with a tendency of lower variability. On the other hand, PVs start out with similar values as other verbs, but the paradigmatic variability decrease is much more accentuated, especially in the twentieth century, where the KLD measures already showed signs of conventionalization, as shown in Figure 2, and diversification (Figure 3).

Thus, it is possible to assume that the PVs have overcome a more accentuated process of conventionalization and diversification over time than other types of verbs in scientific English.

5 Discussion

In this study, our main objective was to analyse the contribution of PVs regarding the conventionalization processes happening in scientific English.

By analysing three different methods to measure linguistic shifts over time, we were able to notice that, although PVs are less common in scientific prose, they have undergone interesting diachronic phenomena.

Regarding the relative entropy measures (i.e., KLD), it was possible to notice that a conventionalization process occurred only throughout the twentieth century (Figure 2).

From the seventeenth to the twentieth century, although some small peaks and troughs can be observed, the KLD values did not change considerably. This tendency is different from what was observed by Degaetano-Ortlieb and Teich (2018) who analysed the whole lexicon (i.e., lemmas). In

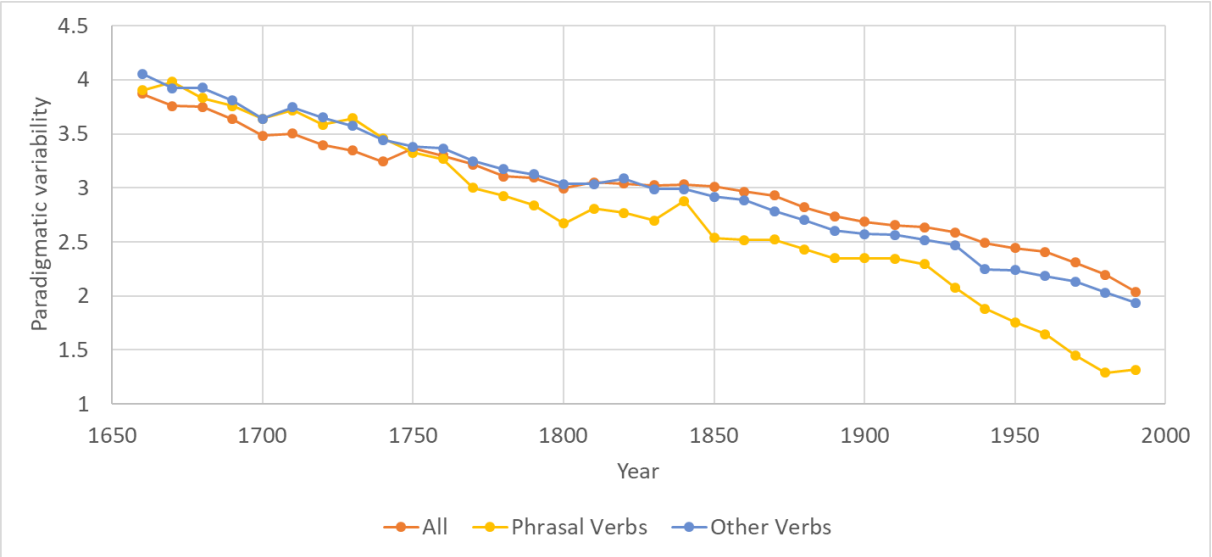

Figure 6: Paradigmatic Variability over time of phrasal Verbs compared to other verbs and all parts-of-speech in the RSC.

their study, the decreasing tendency is relatively constant throughout time.

Moreover, it was possible to verify that the conventionalization process occurs in parallel with a peak of diversification, as shown in Figure 3. This diversification has an impact on the surprisal measures, increasing the surprisal of the particles.

Both conventionalization and diversification processes are also confirmed with the paradigmatic variability analysis (Figure 6). PVs undergo a more accentuated decrease in their paradigmatic variability over time when compared to other verbs, principally during the twentieth century.

To better understand the peaks regarding the diversification of PVs, we analysed in detail the results of the KLD for each PV, per 20-year slice, present in the RSC.

What is possible to observe is that, throughout time, there are shifts regarding the PVs with peaks of KLD, i.e., verbs becoming more distinctive of specific periods.

Regarding the two main peaks of diversification identified in Figure 3, around 1846 and 1971, we can see that the PVs with the highest values of KLD in these periods differ considerably.

- 1846: *carry out, break up, filter off, bring out, split up, build up, map out, sum up, spread out, shut off.*
- 1971: *turn out, point out, rule out, end up, make up, go on, open up, break down, take on, bring together.*

A clear distinction can be made when analysing these two periods. In the nineteenth century, most PVs are linked to the description of experimental design, while in the twentieth century, there is a shift towards verbs focusing on the presentation of results, i.e., on the outcome of the research.

In Figure 3, we showed that the diversification is cyclic, new phrasal verbs are used in specific periods due to particular textual needs and, then, become less typical in future ones, however, in a more conventionalized way throughout the twentieth century.

Regarding the surprisal analysis, it is possible to notice that the shifts in the surprisal of the particles are a complex phenomenon. It is influenced not only by the peaks regarding the diversification process, but also by the changes regarding the distance between the verb and the particle, and, by the decrease regarding the relative frequency (i.e., with the verbs and particles appearing in more varied contexts, not as phrasal verbs).

6 Conclusions and Future Work

In this paper, we presented a multifaceted approach to characterize diachronic shifts regarding the usage of PVs in scientific English from the mid-seventeenth century to the end of the twentieth century by applying different information theory methods to the Royal Society Corpus.

By measuring the Kullback-Leibler divergence, we showed that the process of conventionalization

of PVs occurred mostly throughout the twentieth century. Moreover, we observed that peaks of diversification (i.e., increase in the number of PV types) happened in specific periods, followed by periods with a high number of PVs becoming less typical.

In terms of surprisal measure regarding the particles, we identified an overall tendency of increase, however, it was also possible to notice periods of accentuated increase, and some periods of decrease. These phenomena are probably correlated to the decrease regarding the relative frequency, to the peaks of diversification, and, to the distance between the verb and particles.

The analysis of the paradigmatic variability showed that PVs have a more accentuated decrease over time when compared to other verbs. This is probably due to the usage of PVs in specific contexts, where they cannot be replaced by similar terms. Moreover, the highest decrease regarding this measure was observed during the twentieth century, when a conventionalization phenomenon was detected using KLD.

Our findings not only enhance understanding of PVs in scientific English but also pave the way for future linguistic research, particularly in language evolution and specialized registers. In future work, we intend to proceed with the analysis by conducting a semantic analysis of the PVs with peaks of divergence to better understand their usage throughout time. Moreover, as part of a larger study, these results will be integrated with other types of MWEs (e.g., compounds, fixed expressions) to better understand the impact of the usage of these formulaic expressions in scientific texts throughout time.

Acknowledgements

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

References

Rafael Alejo-González. 2010. Making sense of phrasal verbs: A cognitive linguistic account of l2 learning. *AILA review*, 23(1):50–71.

Hasan Alisoy. 2023. Enhancing understanding of english phrasal verbs in first-year elt students through cognitive-linguistic methods.

Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024. 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.

Tanya Avgustinova and Leonid Iomdin. 2019. https://link.springer.com/chapter/10.1007/978-3-030-30135-4_2 Towards a typology of microsyntactic constructions. In *Proceedings of the International Conference on Computational and Corpus-Based Phraseology*, pages 15–30.

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 Leech, Susan Conrad, and Edward Finegan. 2000. Longman grammar of spoken and written english.

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.

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.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In *LREC*, pages 4125–4128.

Peter Fankhauser and Marc Kupietz. 2017. Visual correlation for detecting patterns in language change. In *Visualisierungsprozesse in den Humanities. Linguistische Perspektiven auf Prägungen, Praktiken, Positionen (VisuHu 2017). Tagung vom 17. bis 19. Juli 2017, Universität Zürich*. Universität Zürich.

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.

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(1):4–21.

Solomon Kullback and Richard A Leibler. 1951. On Information and Sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Wang Ling, Chris Dyer, Alan W Black, and 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*, pages 1299–1304.

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.

Al-Otaibi Ghuzayyil Mohammed. 2019. A cognitive approach to the instruction of phrasal verbs: Rudzka-ostyn's model. *Journal of Language and Education*, 5(2 (18)):10–25.

Michelle Perdomo and Edith Kaan. 2023. Surprisal effects in priming of phrasal verb construction alternations in native speakers and l2 learners.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. https://nlp.stanford.edu/pubs/qi2020stanza.pdf 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*.

Anna Siyanova-Chanturia and Diana Van Lancker Sidtis. 2018. What online processing tells us about formulaic language. *Understanding formulaic language*, pages 38–61.

Julius Steuer, Marie-Pauline Krielke, Stefan Fischer, Stefania Degaetano-Ortlieb, Marius Mosbach, and Dietrich Klakow. 2024. Modeling diachronic change in english scientific writing over 300+ years with transformer-based language model surprisal. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, pages 12–23.

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*, 5:620275.

Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.