Diachronic Analysis of Multi-word Expression Functional Categories in Scientific English

Diego Alves, Stefania Degaetano-Ortlieb, Elena Schmidt, Elke Teich
Saarland University
Saarbrücken, Germany
diego.alves@uni-saarland.de, s.degaetano@mx.uni-saarland.de
elsc00001@stud.uni-saarland.de , e.teich@mx.uni-saarland.de

Abstract

We present a diachronic analysis of multi-word expressions (MWEs) in English based on the Royal Society Corpus, a dataset containing 300+ years of the scientific publications of the Royal Society of London. Specifically, we investigate the functions of MWEs, such as stance markers ("it is interesting") or discourse organizers ("in this section"), and their development over time. Our approach is multi-disciplinary: to detect MWEs we use Universal Dependencies, to classify them functionally we use an approach from register theory, and to assess their role in diachronic development we use an information-theoretic measure, relative entropy.

Keywords: multi-word expressions, universal dependencies, relative entropy, discourse functions, diachronic analysis

1. Introduction

In this paper, we analyze multi-word expressions (MWEs) and the functions they fulfill in scientific writing, inspecting diachronic changes from the mid 17th century to today. From a communicative perspective, MWEs contribute to language efficiency as they constitute highly predictable linguistic material with a clear processing advantage for language users. Their use in scientific writing is particularly interesting due to the high informational load encountered within the scientific domain, where MWEs can act as devices to smooth the informational load in the signal (Conklin and Schmitt, 2012).

There has been a long-standing tradition to identify and analyze MWEs in scientific text and academic writing more widely, most prominently in research on English for Academic Purposes (EAP) (cf. Oakey (2020)). We combine this approach considering the Academic Formula List (AFL) with a UD-based approach, were we use the dependency relation label `fixed` to identify further MWEs not included in the AFL list. As it has been shown that scientific writing becomes increasingly conventionalized over time (see e.g.Degaetano-Ortlieb and Teich (2019)), the `fixed` MWEs are particularly important for a diachronic analysis aimed at investigating communicative efficiency. In this study, we focus on the most frequent grammaticalized fixed expressions identified from the RSC combined with a set of formulaic expressions commonly used in the scientific domain that can be considered as MWEs due to the statistical criteria defined by Simpson-Vlach and Ellis (2010).

Moreover, we label each identified MWE with functional categories to assess (a) the functions MWEs have fulfilled in scientific writing across 300 years, and (b) whether there have been changes in their usage over time. We derive the functions *stance expressions, discourse organizers*, and *referential expressions* from extensive previous work based on Hallidayan register theory (Halliday and Matthiessen, 2014) and widely used by EAP researchers (Biber et al., 2004; Simpson-Vlach and Ellis, 2010; Liu, 2012). Finally, to assess change regarding MWEs, we employ a method from language modeling, relative entropy (Kullback-Leibler Divergence).

The remainder of the paper is organized as follows. In Section 2 we discuss related work on functional categories of MWEs. Sections 3 and 4 present our methods and results. We conclude with a summary of our findings and perspectives for future work (Section 5).

2. Related Work

There are numerous corpus-based accounts regarding the usage of MWEs in different registers, including the scientific one (e.g. Biber and Barbieri (2007); Hyland (2008); Liu (2012)), considering also their classification in terms of functions (see Biber et al. (2004); Simpson-Vlach and Ellis (2010) and Oakey (2020) for an overview). These studies are usually based on strategies for identifying formulaic, pre-fabricated, chunk-like and otherwise phraseological linguistic items considering frequency-based measures (such as MPI) derived from corpora (see work on lexical bundles (Biber and Barbieri, 2007; Hyland, 2008), academic formulas (Simpson-Vlach and Ellis, 2010), and multi-word constructions (Liu, 2012)).

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@LREC-COLING-2024, pages 81–87. May 25, 2024.

Computational linguistic accounts usually focus on techniques to identify and describe patterns of co-occurrence of linguistic units (e.g.Evert (2005); Gries (2022)). To identify potential MWE candidates different measures are applied. Gries (2022) proposes a strategy based on eight different dimensions of information, while Simpson-Vlach and Ellis (2010) define a formula teaching worth (FTW) score based on frequency and mutual information. The identification of MWEs using machine-learning methods are typically based either on DiMSUM (Schneider et al., 2016) or PARSEME (Savary et al., 2015) corpora and the complexity of this task can be attested by the low F1-scores of the state-of-the-art tools (i.e., below 65 as presented by Tanner and Hoffman (2023)). PARSEME corpus divides MWEs into different categories, but they are based on structural properties, not on their functions. Moreover, these datasets are not composed of scientific texts, and thus not totally suitable to address our research question.

Although the study of MWEs is a very active field, both from a linguistic and a computational point of view, the diachronic development of MWEs and their functions remains under-researched. While Biber et al. (2004) and Simpson-Vlach and Ellis (2010) propose a classification of MWEs in terms of discourse functions, these categories have not been examined diachronically. Alves et al. (2024) presented a study concerning the development of MWEs association metrics in scientific English, however, MWE functions were not the main focus of the analysis. Consequently, there are hardly any ready-to-use methodological approaches. With our work, we intend to fill these gaps.

3. Data and Methods

3.1. Data

As a data source, we use the Royal Society Corpus (RSC) 6.0[1], a diachronic corpus of scientific English covering the period from 1665 until 1996. This resource comprises 47,837 texts (295,895,749 tokens), mainly scientific articles covering a wide range of areas from mathematics to physical and biological sciences, and is based on the Philosophical Transactions and Proceedings of the Royal Society of London (Fischer et al., 2020). Table 1 shows a detailed overview of the distribution of texts and tokens over time.

There has been extensive work on the proceedings and transactions of the Royal Society based on the RSC, showing how the scientific register has evolved from an involved verbal style of writing (papers were read out aloud by fellows at the Royal Society of London in the beginning of the society) to a highly informational style of writing meant for purely written expert-to-expert communication. This development is specific to scientific writing and not observed in a register-mixed corpus (cf. Degaetano-Ortlieb and Teich (2019)). Also, we observe diversification in linguistic usage reflecting disciplinary specialization (e.g. modern chemistry emerges in the 18th c.) (Bizzoni et al., 2020) and a general conventionalization trend (Teich et al., 2021). Together, linguistic diversification and conventionalization address the communicative demands of modern science communication. In this paper, we expand this research by specifically investigating MWEs in the RSC since they are highly conventionalized structures.

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

Table 1: Size of the Royal Society Corpus 6.0 over time

3.2. Identifying MWEs in the RSC

In the present study, we focus on two specific kinds of MWEs that were extracted from the RSC using two different approaches: (a) fixed MWEs extracted from the UD-parsed version of the RSC, and (b) ensemble of MWEs provided by the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010).

Fixed Multi-word Expressions The Universal Dependencies[2] (UD) guidelines for morphosyntactic annotations (De Marneffe et al., 2021) encompass the relation label *fixed* for certain fixed grammaticalized expressions which tend to behave like function words (e.g. *because of*, *in spite of*, *as well as*) with distinct functions.

To extract the fixed MWEs, we parsed the RSC 6.0 using Stanza tool (Qi et al., 2020) with the combined model for the English language trained on different UD corpora (i.e., EWT, GUM, GUMReddit, PUD, and Pronouns). Using a Python script with the pyconll library[3], we identified and counted the fixed MWEs in the RSC texts per year.[4]

From the list of fixed MWEs, we identified the 100 most frequent ones and manually annotated

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

[2] https://universaldependencies.org/

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

[4] A manual evaluation of 70 sentences (10 per 50-year period of the RSC) showed that the labelled attachment score of the parser is equal or higher than 85% for fixed MWEs in the different time periods.

Function	Type	MWEs	Examples
Stance	epistemic	84	*it is important*, *according to*
	attitudinal/modality	24	*we have to*, *needs to be*
	intention/prediction	11	*if you want to*, *to do so*
	ability	34	*can be found*, *it is possible to*
Discourse	topic introduction/focus	31	*in this article*, *for example in*
	topic elaboration/clarification	70	*due to the fact*, *the reason for*
Reference	identification/focus	61	*such as the*, *as can be seen in*
	imprecision	3	*and so on*, *and so forth*
	specification of attributes	177	*a form of*, *on the basis of*
	time/place/text reference	57	*at the end of*, *in between*

Table 2: Functional categories and types (cf. Biber et al. (2004)).

them according to the taxonomy in Section 3.2.[5] Since we consider only the fixed MWEs with high frequency in the RSC and conducted a manual evaluation of the identified expressions, we assume that the parsing errors have been minimized in this study.

AFL Multi-word Expressions The Academic Formulas List is an inventory of the most common formulaic sequences in academic English. It is composed of: a) a core list of 207 formulaic expressions found in written and spoken academic language (e.g.*in terms of* and *at the same time*; b) 200 expressions from written corpora (e.g. *on the other hand* and *it should be noted*); and c) 200 MWEs extracted from spoken academic English texts (e.g. *be able to* and *if you look at*) (Simpson-Vlach and Ellis, 2010). The AFL MWEs were identified by the authors with a special measure of usefulness called the formula teaching worth (FTW), which combines frequency and mutual information measures. Thus, the classification of the formulaic expressions from the AFL as MWEs is done due this statistical criterion.

3.3. MWE Functional Categories

We follow the taxonomy proposed by Biber et al. (2004), which captures the major functions of MWEs with three primary categories: (a) stance expressions, which express attitudes or assessments of certainty, framing other propositions; (b) discourse organizers that reflect relationships between parts of the discourse; and (c) referential expressions that refer to physical or abstract entities, or to the textual context, identifying a specific entity or pointing out to a specific attribute of it.

Table 2 presents a summarized version of the taxonomy established by Biber et al. (2004) (i.e., functions and types) together with the number of MWEs per type and examples observed in the RSC.

[5]The annotation was made by a linguistics student and verified by two specialists

Note that Simpson-Vlach and Ellis (2010) classified most of the AFL MWEs according to a taxonomy similar to the one proposed by Biber et al. (2004). We selected these categorised MWEs to be examined in this study, adjusting the taxonomy according to Table 2.

3.4. Modeling Change with Relative Entropy

To analyse the diachronic development of the different MWE functional categories, first, we examined the relative frequency per year.

To detect evolutionary trends, we applied relative entropy, specifically Kullback-Leibler Divergence (KLD; Kullback and Leibler (1951)), a method for comparing probability distributions measuring the number of additional bits needed to encode a given data set A when a (non-optimal) model based on a data set B is used for a set of elements X. In our case, A and B correspond to sub-sets of the RSC (e.g. time slices) and X, i.e. the ensemble of MWEs of each function.

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

KLD provides an indication of the degree of divergence between corpora and identifies the features that are primarily associated with a difference.[6]

To detect periods of change using KLD given each functional category (stance, discourse, and reference), we adopt the methodology described in Degaetano-Ortlieb and Teich (2018).[7] Basically, we compare 20-year windows of past and present language use sliding with a 5-year gap over the time line (e.g. t1=1665-1685, t2=1691-1711). By plotting the divergence for each comparison on the time line, we can inspect peaks or troughs which

[6]Discrepancies regarding vocabulary size are controlled by applying Jelinek-Mercer smoothing with lambda 0.05 (cf. Zhai and Lafferty (2004) and Fankhauser et al. (2014)).

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

indicates a change. A peak indicates that the divergence of that features increases, and is thus *typical* of the future 20 years in comparison to the past 20 years. In particular, we consider the pointwise KLD, i.e. the individual KLD of each feature (here: either functions or types), in order to determine a feature's rise or decrease in typicality.

4. Results

4.1. Frequency-based Trends

Figure 1 presents the evolution of each main functional category per year by relative frequency (i.e., MWEs occurrence/no. of tokens of each period).

In general, all three functions present an increasing tendency across time until the beginning of the twentieth century. The usage of referential expressions (black line) has a considerable increase in the second half of the eighteenth century. Moreover, from 1925 on, while both discourse (blue) and reference MWEs (red) present a decreasing tendency, the use of stance expressions seems to steadily increase even though these expressions remain relatively low in frequency.

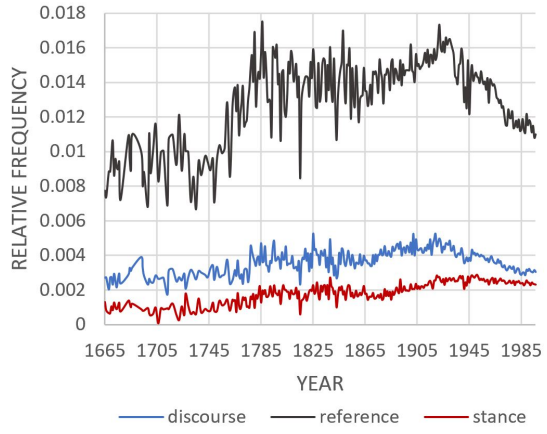

Figure 1: Relative frequency for each function.

4.2. Diachronic Trends by Divergence

While relative frequencies pinpoint the rise or decline of specific linguistic features over time, KLD provides a detailed quantification of the overall linguistic shift from one period to another, identifying even those changes that do not correspond to simple increases or decreases in usage frequency. Thus, KLD provides insights into the degree of linguistic change and allows to identify more subtle patterns of linguistic evolution that relative frequencies alone may not discern. Figure 2 presents the overall results per category for all the MWEs (AFL and fixed). We can observe that from the 17th to

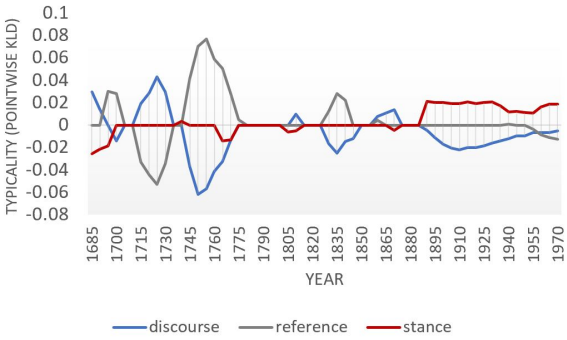

Figure 2: KLD measures for each function.

the beginning of 20th century, reference and discourse MWEs tend to behave in opposite directions, i.e. when reference becomes typical, discourse goes down in typicality and vice versa, while stance MWEs present less change. The scenario changes in the 20th century when the presence of stance expressions in the corpus becomes more typical.

To better understand these diachronic trends, we also applied KLD considering the types of each function (see Figure 3). The main trends observed for discourse and referential expressions are due to the function types 'topic elaboration/clarification' and 'specification of attributes types', respectively. While the topic elaboration/clarification function is used to signal further explication providing a clearer understanding or additional information related to the topic being discussed (e.g. *in order to*, *as a result*, *the reason for*), the specification of attributes function type serves as a way to provide framing information (e.g. *the way which*, *the level of*, *these two*), i.e. essentially specifying or detailing characteristics, qualities, or attributes of a subject. These trends may be influenced by a variety of factors. Historical and cultural contexts that value explicit reasoning may lead to a preference for elaborate discourse, while changes in academic standards and expectations could necessitate a more precise specification of attributes. The rise of particular disciplines and interdisciplinary research, along with technological advancements that shape information dissemination, could also play significant roles.

Considering the increase in divergence for stance expression in the more contemporary period, we can observe that the peak is indicated by three out of four types for stance expressions. By 1825, ability becomes more typical showing an increased distinctive use (e.g. *can be used/found/expressed*), followed by attitudinal expressions until almost 100 years later where they decrease in divergence around the 1930s, when epistemic expressions (e.g. *according to, at least*) become typical. Around that period, also identification and focus reference expressions (e.g. *there has been, can be seen*) increase in typicality as well as topic and introduc-

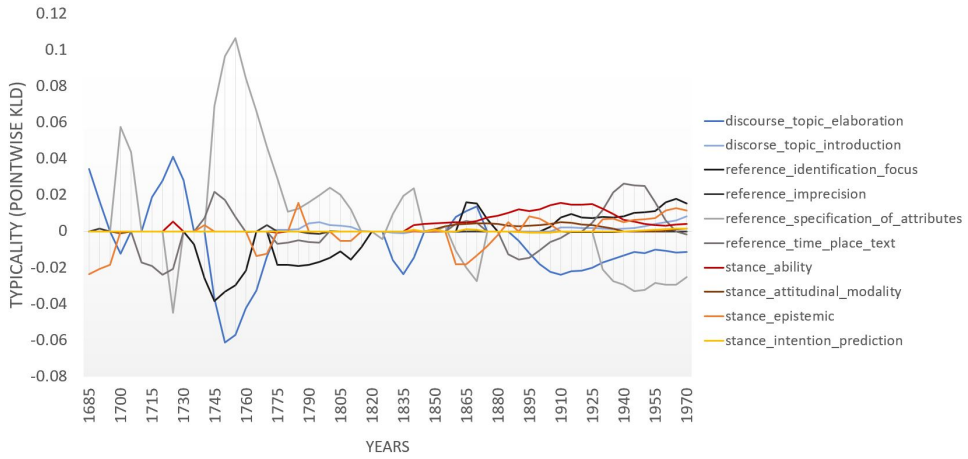

Figure 3: KLD measures for each function (colour) and its types (shades of a colour)

tion discourse organizers (e.g. *first of all, in this paper we*). During that period, there is also a peak in time, place and textual reference (e.g. *as shown in, shown in figure*). Overall, there is a trend towards a more varied distinctive use of MWE function types towards the more contemporary period. These trends seem to signal a use of MWEs to be increasingly inclined to articulate evidence-based reasoning as shown by MWEs such as *according to* or *as shown in*. These expressions serve to direct the reader's attention to evidence or examples that support the argument being made, which is a fundamental aspect of scholarly work.

5. Conclusion and Future Work

In this paper we have presented an analysis of MWEs in scientific writing, tracing the evolution of their functions over a span of three centuries. Our investigation reveals a dynamic landscape of MWE usage, marked by significant shifts in function that reflect changing priorities and practices within the scientific community over time. In the initial stages, we observed a competitive relationship between discourse and reference functions of MWEs. This competition underscores the evolving nature of scientific discourse, as authors sought to balance the needs for clarity and precision with the demands of argumentation and discourse structuring. Towards the recent 100 years, our findings indicate a diversification in MWE functions, with stance expressions taking on a leading role. The shift towards epistemic stance, reference of identification/focus, of place/time/textual and discourse organizers of topic and introduction seems to be a means of directing the reader's attention to evidence-based information.

Combining the AFL list with a UD-based approach to identify MWEs not covered by the AFL, allowed us to capture a broader range of conventionalized expressions that contribute to the diachronic trend of increasing conventionalization in scientific writing. The application of relative entropy as a methodological tool has further enriched our understanding of change over time, offering a quantitative measure of the shifts in MWE usage.

The functional categorization of MWEs, grounded in Hallidayan register theory, provides a solid theoretical framework for our analysis of functions and types. A limitation of our study is the uneven distribution of data across periods, with more material from recent periods, which may skew perceptions of MWE functionality and its evolution over time. Also, the diachrony of our data might present gaps within the AFL list. In future work, we aim to expand our research in three ways: (1) increase the number of MWEs related to the different functions and compare the obtained results with analysis of other domains; (2) model MWEs at the paradigmatic level by word embeddings to further increase coverage of items; (3) apply probabilistic measures of processing (e.g. surprisal) to gain insights on processing effects of conventionalization of MWEs. Overall, we aim to work towards gaining further insights into the complex ways MWEs serve the communicative needs of scientific writers and compare their usage across scientific domains and other registers.

Acknowledgements

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

6. Bibliographical References

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, St. Julians, Malta. Association for Computational Linguistics.

Douglas Biber and Federica Barbieri. 2007. Lexical Bundles in University Spoken and Written Registers. *English for specific purposes*, 26(3):263–286.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at. . . : Lexical Bundles in University Teaching and Textbooks. *Applied linguistics*, 25(3):371–405.

Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. 2020. Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach. *Frontiers in Artificial Intelligence*, 3.

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

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 Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at COLING2018*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

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

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Universität Stuttgart.

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

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.

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).

MAK Halliday and CMIM Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*, volume 17. Routledge.

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.

Dilin Liu. 2012. The Most Frequently-used Multiword Constructions in Academic Written English: A Multi-corpus Study. *English for Specific Purposes*, 31(1):25–35.

David Oakey. 2020. Phrases in EAP Academic Writing Pedagogy: Illuminating Halliday's Influence on Research and Practice, journal = Journal of English for Academic Purposes. 44:100829.

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*.

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.

Rita Simpson-Vlach and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied linguistics*, 31(4):487–512.

Joshua Tanner and Jacob Hoffman. 2023. MWE as WSD: Solving Multiword Expression Identification with Word Sense Disambiguation. *arXiv preprint arXiv:2303.06623*.

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.

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.