Multi-word expressions in biomedical abstracts and their plain English adaptations

Sergei Bagdasarov
Saarland University, Germany
sergeiba@lst.uni-saarland.de

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

Abstract

This study analyzes the use of multi-word expressions (MWEs), prefabricated sequences of words (e.g. *in this case, this means that, healthcare service, follow up*) in biomedical abstracts and their plain language adaptations. While English academic writing became highly specialized and complex from the late 19th century onwards, recent decades have seen a rising demand for a lay-friendly language in scientific content, especially in the health domain, to bridge a communication gap between experts and laypersons. Based on previous research showing that MWEs are easier to process than non-formulaic word sequences of comparable length, we hypothesize that they can potentially be used to create a more reader-friendly language. Our preliminary results suggest some significant differences between complex and plain abstracts when it comes to the usage patterns and informational load of MWEs.

1 Introduction

Previous diachronic research has shown that English scientific writing developed a compressed code of communication that is efficient for its primary users (i.e. scientists) (Halliday, 1988; Biber and Gray, 2016; Degaetano-Ortlieb and Teich, 2016, 2022). However, the consequence of this process was that academic papers became almost incomprehensible for a general audience, which poses a considerable problem as the need to draw knowledge directly from scientific publications is growing among laypersons, particularly in the health domain.

To address this issue, many scholars and journals encourage scientists to use plain language in their papers or at least include plain language summaries of their work (Hauck, 2019; Sedgwick et al., 2021). While writing recommendations on plain language abound, they seem to pay little attention to multi-word expressions (MWEs), i.e. prefabricated sequences of several words that are argued to foster the fluency of language use (Sinclair, 1991; Pawley and Syder, 1983).

Our goal is to investigate whether, and if so to what extent MWEs can ease the processing of plain language texts. To this end, we analyze the use of MWEs in abstracts from biomedical papers ("complex abstracts") and their plain language adaptations ("plain abstracts"). We pose the following questions: (i) Does the use of MWEs differ in complex and plain abstracts and, more specifically, do plain abstracts use more MWEs? (ii) Are MWEs in plain abstracts easier to process?

In general, we expect to see more MWEs in plain abstracts. In terms of MWE types, we anticipate that nominal MWEs (e.g. compound nouns used as terms), typically associated with technical scientific writing, will be less characteristic of plain abstracts. Moreover, we expect MWEs in plain abstracts to be less informationally loaded on average (and therefore easier to process).

The remainder of the paper is structured as follows. Section 2 is dedicated to MWE processing. Section 3 describes our data and methodology. In Section 4, we present our analysis results. In Section 5, we provide a summary and prospects of future work.

2 Background and Related Work

Linguistic studies in recent decades have revealed that MWEs make up a large proportion of language use and that they are less costly in processing than other sequences of words (Erman and Warren, 2000; Foster, 2001). For instance, Conklin and Schmitt (2008) prove that MWEs have shorter reading times in comparison to non-formulaic expressions. Li et al. (2021) and Siyanova-Chanturia et al. (2011) arrive at similar conclusions using eye-tracking. The assumption about a processing advantage of MWEs has also been corroborated by EEG studies (cf. Tremblay et al. (2011); Siyanova-Chanturia et al. (2017)). Further evidence is pro-

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 483–488
November 16, 2024. ©2024 Association for Computational Linguistics

vided from speech processing: formulaic expressions are produced faster and more fluently than comparable, non-formulaic expressions and recognized better (e.g. under acoustic degradation; (Rammell et al., 2017)).

While it now seems increasingly clear that MWEs are faster and easier to process than non-formulaic language, what still remains open is whether the use of MWEs is influenced by other factors. In register theory, it is widely assumed that speakers adjust their language according to the particular communicative situation (Biber, 2012; Biber and Conrad, 2019; Conrad and Biber, 2005). One of the parameters describing the communicative situation is the relationship between the speaker and the recipient. For instance, in case of complex abstracts both speaker and recipient have professional knowledge of the subject. In contrast, plain abstracts are written by well-versed speakers for lay recipients. Hence, it is plausible to suppose that this shift in the level of expertise should be reflected in the use of MWEs, i.e. MWEs should be employed in plain abstracts in such a way that reduces the processing cost for the recipient.

3 Methodology

3.1 Data

We use the Plain Language Adaptation of Biomedical Abstracts dataset (PLABA) (Attal et al., 2023). The biomedical abstracts come from PubMed and were transformed into plain language by human writers on a sentence basis, with sometimes multiple plain language adaptations being written for one complex abstract. Some relevant corpus statistics is summarized in Table 1.[1]

	#Abstracts	#Tokens	#Types
Complex	749	199,851	17,425
Plain	919	249,301	13,117

Table 1: PLABA corpus data

We performed tokenization and sentence segmentation with TreeTagger (Schmid, 1994, 1995). The pretokenized abstracts were then parsed with the state-of-the-art Stanza parser (Qi et al., 2020).

3.2 MWE Identification

Following Alves et al. (2024a,b), we use Universal Dependencies (UD) and the Academic Formulas List (AFL) to identify MWEs in our corpus.

The UD framework (de Marneffe et al., 2021) contains five MWE-related labels: 1. compound — combinations of tokens that morphosyntactically behave as single words; in English this label refers mostly to nominal compounds (e.g. *muscle cramps*), 2. compound:prt — phrasal verbs (e.g *follow up*), 3. fixed — certain grammaticized expressions normally acting as function words (e.g. *according to*), 4. flat — sequences where none of the words can be identified as the head, in our case these are mostly proper names (e.g. *Moderna mRNA-1273*), 5. flat:foreign — sequences of foreign words[2].

The identification of MWEs according to the UD method was performed using a Python script that extracted all words labelled with the above mentioned tags and their corresponding heads (if any). For instance, some occurrences of the word *muscle* were labelled with the *compound* tag during parsing, with the word *cramps* being identified as their head. So, the resulting MWE is *muscle cramps*.

The AFL (Simpson-Vlach and Ellis, 2010) includes 207 core formulaic expressions common for both written and spoken academic English, 200 expressions common for written academic English and 200 expressions common for spoken academic English. The authors selected the MWEs based on a measure called "formula teaching worth", which combines frequency and mutual information. For this study, we relied only on core and written MWEs. Using a Python script, we iterated through both lists and extracted all MWEs that appear at least once in our data.

After applying the UD and AFL methods, we merged all extracted MWEs into one final list. No frequency thresholds were used since the UD labels are grammatically motivated and the AFL MWEs had already been predefined based on specific measures.

3.3 Relative Entropy

We use the asymmetric variant of relative entropy, known as Kullback-Leibler Divergence (KLD) (Kullback and Leibler, 1951), to investigate the use of MWEs in complex and plain abstracts. KLD allows us to compare two probability distributions

[1] The number of abstracts available in PLABA at the time of our study differs from the number of abstracts stated in the original publication by Attal et al. (2023). Table 1 contains statistics on the actual data employed in our study.

[2] This category is not attested in our data.

A and B (here, MWEs in complex and plain abstracts) by showing the number of additional bits of information needed to encode one distribution using the other one. The formal representation of KLD is shown in equation 1:

$$D(A||B) = \sum_i p(feature_i|A) \log_2 \frac{p(feature_i|A)}{p(feature_i|B)} \quad (1)$$

A KLD value of 0 would mean that the usage patterns of MWEs are exactly the same in complex and plain abstracts, while a value greater than 0 would indicate a divergence. Moreover, KLD shows the contributions of individual features to the overall divergence, allowing us to generate a list of the most relevant features (i.e. MWEs).

3.4 Surprisal

To quantify the informativity of MWEs, we use surprisal, a measure that shows how much information (in bits) a word carries in a given context (Shannon, 1948):

$$S(word) = -\log_2 p(word|context) \quad (2)$$

As shown by reading time or specific EEG signals, surprisal is proportional to cognitive effort. Hence, a high surprisal of a MWE would be indicative of its high processing cost and vice versa.

In this study, we estimated surprisal of a given word *n* based on the four-gram model where words *n-1*, *n-2* and *n-3* are taken as context (cf. Genzel and Charniak (2002)). Additionally, we computed average surprisal for each MWE. For this, we first estimated average surprisal of each individual instance of a MWE, then summed all values and divided them by the number of occurrences of a MWE in the corpus.

4 Results

Contrary to what we expected, plain abstracts employ fewer MWEs, both in terms of unique occurrences (6,155 vs 6,700) and total frequency (62,976.08 vs 63,802.53 occurrences per million words). Compounds are the most common MWE type in both abstract categories as reflected in Table 2 and Figure 1. The most notorious differences in frequencies were observed for the proper nouns (flat) and phrasal verbs (compound:prt).

Type	Description	Complex	Plain
compound	compounds	6,309	5,702
compound:prt	phrasal verbs	26	114
fixed	fixed expressions	30	35
flat	proper names	67	29
afl	academic formulas	268	275

Table 2: Unique MWEs identified in complex and plain abstracts

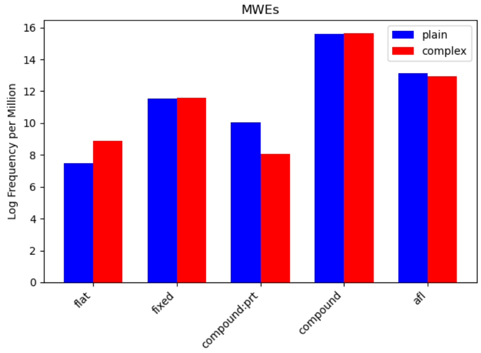

Figure 1: MWE frequency in complex and plain abstracts.

However, going beyond mere frequency estimations, our KLD analysis revealed a considerable difference in the use of MWEs in both directions of comparison: 3.35 bits and 3.18 bits for complex VS plain and plain VS complex, respectively. A look at the most distinctive features (see Figures 2 and 3) also offers interesting insights. While both types of abstracts are characterized by compound MWEs to a great extent, we see, for instance, that complex abstracts have more statistical terminology (e.g. *confidence interval, mean age, odds ratio* etc.) and different research design terms (e.g. *cohort study, crossover study, control group* etc.).

In contrast, such MWEs are not encountered among the features distinctive of plain abstracts. This is probably due to the fact that such statistical and methodological information is not relevant for a lay person and, therefore, can be left out to enhance readability.

Moreover, we see numerous examples of specialized terminology denoting biological and medical phenomena (e.g. *dopamine receptor, plasma concentration* etc.). Since it is impossible to just delete such terms without loosing information relevant to the reader, plain abstracts try to use more common equivalents (e.g. *blood sugar levels* instead of *blood glucose levels*). Sometimes such transformations lead to the creation of MWEs in cases where no MWE is used in complex abstracts. For instance, a one-word term *placebo* turns into a compound *dummy treatment*. Or an adjective-noun term *neurodegenerative disease* is replaced with a

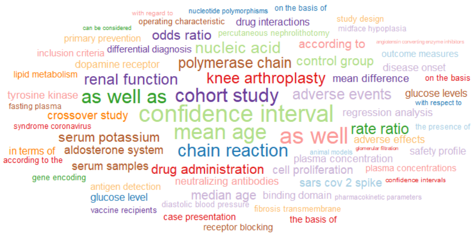

Figure 2: 70 most distinctive MWEs in complex abstracts.

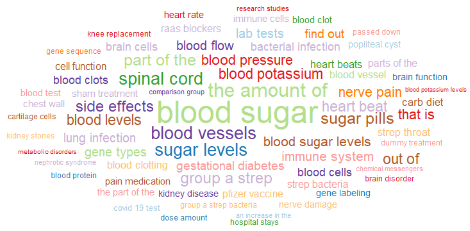

Figure 3: 70 most distinctive MWEs in plain abstracts.

noun compound *brain disorder*.

Other MWE types are also attested among the most characteristic features, albeit marginally. For instance, phrasal verbs seem to be more distinctive of plain abstracts (e.g. *find out, make up*). Fixed and AFL MWEs are present in both lists, however complex abstracts seem to employ more sophisticated expressions typical of elaborated writing (e.g. *as well as, in terms of, according to the* etc.). Flat MWEs are not attested among the most distinctive features.

In terms of informativity, as measured by surprisal, we observed significant differences between complex and plain abstracts for compound and AFL categories, while phrasal verbs showed a marginally significant difference.[3] All of these three MWE types have lower surprisal for plain abstracts (see Figure 4), which is, in principle, in line with our expectations although we anticipated a more pronounced trend.

Lower surprisal values in plain compound MWEs might be indicative of MWEs being used to effectively reduce processing effort. Consider, for instance, the MWE *blood glucose levels* which is typically used in complex abstracts and has a surprisal of 4.73 bits. Its plain language equivalent *blood sugar levels*, however, transmits 3.24 bits of information on average, thus being easier to process.

The same seems to hold for cases where plain abstracts use an MWE instead of a noun with an adjective premodifier, which is a very common terminology formation pattern in scientific texts. Compare, for example, the following terms from complex abstracts and their plain language equivalents: *renal cancer* (7.41 bits) vs *kidney cancer* (5.91 bits) and *neurodegenerative disease* (12.20 bits) vs *brain disorder* (3.65 bits).

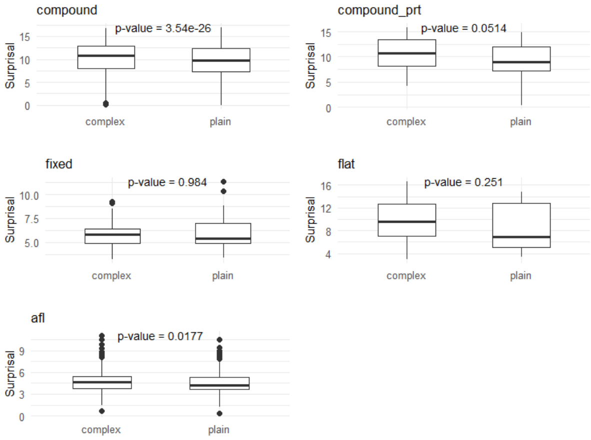

Figure 4: Comparison of average MWE surprisal in complex and plain abstracts across different MWE types.

A similar mechanism seems to apply to phrasal verb MWEs that can be used as an alternative to more complex verbs. Consider, for instance, Example (1) extracted from a complex abstract and its plain language adaptation shown in Example (2) (values in parenthesis indicate surprisal in the corresponding sentence)

(1) ... *as the risk of detrimental outcomes* ***increases*** (12.46) *with delayed surgical intervention.*

(2) ...*since the risk of harmful effects* ***goes up*** (8.69) *with delayed surgery.*

While in general our findings do suggest that some types of MWEs *per se* seem to be easier to process in plain language abstracts, a more in-depth analysis is needed to investigate how rewrit-

[3] For determining statistical significance we used the Wilcoxon rank sum test available in R (R Development Core Team).

ing strategies like those described above affect processing complexity on the sentence and text level.

5 Conclusion and Future Work

In this study, we investigated MWEs in biomedical abstracts and their plain language adaptations. We were able to establish some differences in the use of MWEs (e.g. more prominent use of statistical and methodological terms in complex abstracts, greater reliance on phrasal verbs in plain abstracts).

Furthermore, we found that the informational load of compound, phrasal verb and AFL MWEs is lower in plain abstracts, suggesting that the use of MWEs might play a role in decreasing processing cost in the transition from complex to plain language.

In future studies, we will focus on the MWE types individually to investigate why plain MWEs have lower surprisal. Additionally, we are planning to expand our methodology to account for factors that might be correlated with the MWE processing cost (e.g. association strength among the component parts of an MWE).

Limitations

Our study is based on a relatively small corpus: roughly 200,000 words for complex abstracts and 250,000 words for plain abstracts. Moreover, we are not aware which journals the abstracts come from and whether the authors of abstracts are native speakers of English. These factors might also influence the use of MWEs. The creation of a larger dataset with detailed meta-information may be addressed in future research.

Ethics Statement

This does not apply to our research since we did not perform any experiments nor collected personal data.

Acknowledgements

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

References

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

Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024b. Multi-word expressions in English scientific writing. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 67–76, St. Julians, Malta. Association for Computational Linguistics.

Kelly Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1):9–37.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2nd edition. Cambridge University Press, Cambridge.

Douglas Biber and Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Studies in English Language. Cambridge University Press, Cambridge, UK.

Kathy Conklin and Norbert Schmitt. 2008. Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers? *Applied Linguistics*, 29(1):72–89.

Susan Conrad and Douglas Biber. 2005. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, 20(2004):56–71.

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. 2016. Information-based modeling of diachronic linguistic change: From typicality to productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 165–173, Berlin, Germany. Association for Computational Linguistics.

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.

Britt Erman and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text & Talk*, 20(1):29–62.

Pauline Foster. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In Martin

Bygate, Peter Skehan, and Merrill Swain, editors, *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*, pages 75–93. Longman, Harlow.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th ACL*, pages 199–206, Philadelphia, PA, USA.

M.A.K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of written English: Situational factors and linguistic features*, pages 162–177. Pinter, London.

Steven A. Hauck. 2019. Sharing planetary science in plain language. *Journal of Geophysical Research: Planets*, 124(10):2462–2464.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Hui Li, Kate L. Warrington, Ascensión Pagán, Kevin B. Paterson, and Xinchun Wang. 2021. Independent effects of collocation strength and contextual predictability on eye movements in reading. *Language, Cognition and Neuroscience*, 36(8):1001–1009.

Andrew Pawley and Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards and Richard W. Schmidt, editors, *Language and communication*, pages 191–225. Longman, London, UK.

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*, pages 101–108, Online. Association for Computational Linguistics.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. ISBN 3-900051-07-0.

C. Sophie Rammell, Diana Van Lancker Sidtis, and David B. Pisoni. 2017. Perception of formulaic and novel expressions under acoustic degradation. *Ment Lex*, 12(2):234–262.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Kyoto, Japan.

Cassie Sedgwick, Laura Belmonte, Amanda Margolis, Patricia Osborn Shafer, Jennifer Pitterle, and Barry E. Gidal. 2021. Extending the reach of science – talk in plain language. *Epilepsy Behavior Reports*, 16:100493.

Claude Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656.

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

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

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.

Anna Siyanova-Chanturia, Kathy Conklin, and Norbert Schmitt. 2011. Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27(2):251–272.

Antoine Tremblay, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language learning*, 61(2):569–613.