Analyzing Multi-word Expressions in Complex and Plain Language Biomedical Texts: An Information-theoretic Approach

Sergei Bagdasarov*[1], Elke Teich[1], and Diego Alves[1]

[1]Saarland University – Allemagne

Résumé

We focus on multi-word expressions (MWEs) – formulaic sequences of at least two words – in complex and plain English biomedical abstracts. Starting from the premise that language production varies depending on the situational context of communication (Biber, 2012), we formulate the following research question: How does the shift in the target audience background knowledge (specialist-to-specialist vs specialist-to-layperson communication) affect the MWE usage?

To answer this question, we use the Plain Language Adaptation of Biomedical Abstracts (PLABA) dataset (Attal et al., 2023) – a parallel corpus of English biomedical abstracts and their human-created plain language adaptations. Following (Alves et al., 2024), we use Universal Dependencies (de Marneffe et al., 2021), Academic Formulas List (Simpson-Vlach and Ellis, 2011) and a supervised machine learning algorithm (Williams, 2017) to extract MWEs. This combined approach allows us to cover a broad spectrum of MWEs ranging from complex function words (*due to*, *in terms of*) and phrasal verbs (*follow up*, *carry out*) to formulaic noun phrases (*muscle cramps*, *logistic regression*) and common academic formulas (*it is interesting to*).

For the comparative analysis of plain vs. complex abstracts, we employed the asymmetric variant of relative entropy, also known as Kullback-Leibler Divergence (KLD) (Kullback and Leibler, 1951; Fankhauser et al., 2014). KLD is an information-theoretic measure that allows us to quantify the difference (in bits) between two probability distributions (in our case, MWEs in complex and plain abstracts) and identify features contributing to the divergence.

Furthermore, we used normalized KLD and surprisal for a more fine-grained analysis of MWEs. Normalized KLD measures the internal association strength of MWE components (Gries, 2022). Surprisal (Shannon, 1948) is another information-theoretic measure that quantifies (in bits) the amount of information carried by a word in a certain context and is known to correlate with processing effort (Demberg and Keller, 2008).

Our analysis revealed that the most notorious trend concerns terminology patterns which undergo different manipulations in plain abstracts: e.g. omission of statistical and methodological terminology (*confidence interval*) and terminology transformations (*brain disorder* instead of *neurodegenerative disease*). Moreover, plain language texts contain more verbal

*Intervenant

MWEs – possibly a sign of colloquialization. We also show that MWEs in complex abstracts exhibit higher association scores and higher surprisal in some categories. The former can be interpreted as evidence of a more consistent use of formulaic language in complex texts. The latter is indicative of a lower processing cost of MWEs in plain language abstracts.

References

- Diego Alves, Stefania Degaetano-Ortlieb, Elena Schmidt, and Elke Teich. 2024. 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.
- 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.
- Vera Demberg and Frank Keller. 2008. Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity. Cognition. 2008 Nov;109(2):193-210. Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), pages 4125–4128, Reykjavik, Iceland. ELRA.
- Stefan Th. Gries. 2022. What do (some of) our association measures measure (most)? Association? Journal of Second Language Studies, 5(1): 1-33.
- Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. The Annals of Mathematical Statistics, 22(1): 79–86.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. Computational Linguistics, 47(2): 255–308.
- 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.
- Jake Williams. 2017. Boundary-based MWE segmentation with text partitioning. In Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.