Domain expertise reduces reading times of multi-word expressions in academic texts

Sergei Bagdasarov (Saarland University), Marie-Pauline Krielke (Saarland University), Diego Alves (Saarland University)
sergeiba@lst.uni-saarland.de

Multi-word expressions (MWEs) are frequently co-occurring word combinations (Wahl & Gries, 2018) and represent a special case in cognitive processing. Due to their frequency and predictability, they are known to be processed faster than matched control phrases (Siyanova-Chanturia, 2013), while processing effort is also known to depend on intra-subject factors such as language proficiency: natives read MWEs faster than non-natives (Underwood et al., n.d.). Following rational communication principles assuming highest communicative efficiency with the lowest effort possible, MWEs contribute to language efficiency by representing highly predictable linguistic material with a clear processing advantage (Conklin & Schmitt, 2008). This processing advantage is especially relevant in scientific writing posing several cognitive challenges such as high information density, abstractness, constant lexical innovation, etc.

Considering the processing advantage of highly predictable linguistic material, our hypothesis is that domain-specific MWEs should be read faster by in-domain experts than by out-of-domain experts due to background knowledge and frequent exposure to certain types of expressions. We also expect to find different effects for different types of MWEs (e.g. discourse structure markers vs. multi-word terminology) and measures associated with processing effort such as reading times (RTs) should reflect this.

For the present study, we use the Potsdam Textbook Corpus (POTEC, Jäger et al., 2021), a naturalistic eye-tracking-while-reading corpus comprising eye-movement data from domain experts (physics and biology) and novices reading 12 German scientific texts. It follows a 2×2×2 fully-crossed factorial design, with the level of study and discipline as between-subject factors, and text domain as a within-subject factor.

We extract MWEs from the corpus using a novel 8-dimensional method proposed by Gries (2022). The method leverages both traditional frequency-based parameters and different information-theoretical measures like normalized and relative entropy. After manually filtering noisy output (e.g. chunks that only contain grammatical words or span phrase boudaries like *welche die* or *wird in der*), we obtained a list of 99 MWEs, e.g. *in der Nähe zum*, *extensiv ausgebildetes Wurzelsystem*, *qualitativ und quantitativ* etc.

To analyze the total fixation times of MWEs, we fitted a mixed-effects linear model with logged total fixation times as response and expertise level as well as the presence of general and domain-specific terminology as predictors, while controlling for variability due to differences among readers, texts, trials and MWEs. For this, we used the lmerTest package (Kuznetsova et al., 2017) available in RStudio (RStudio Team, 2020). Preliminary findings show that MWEs in general are read faster by in-domain experts (estimate $= -0.19$, SE $= 0.018$, $p < 2 \times 10^{-16}$, see Figure 1). If a MWE contains a domain-specific term, the fixation time increases (estimate $= 0.21$, SE $= 0.07$, $p = 0.0078$. However, as expected, in-domain experts read MWEs with domain-specific terminology slightly faster (estimate $= -0.058$, SE $= 0.0262$, $p = 0.0259$).

Due to the small size of the corpus, the quality of the extracted MWEs is rather limited. We thus aim to refine our MWE extraction method to improve the fit of the model. Moreover, we intend to analyse other reading time measures, comparing different classes of MWEs.

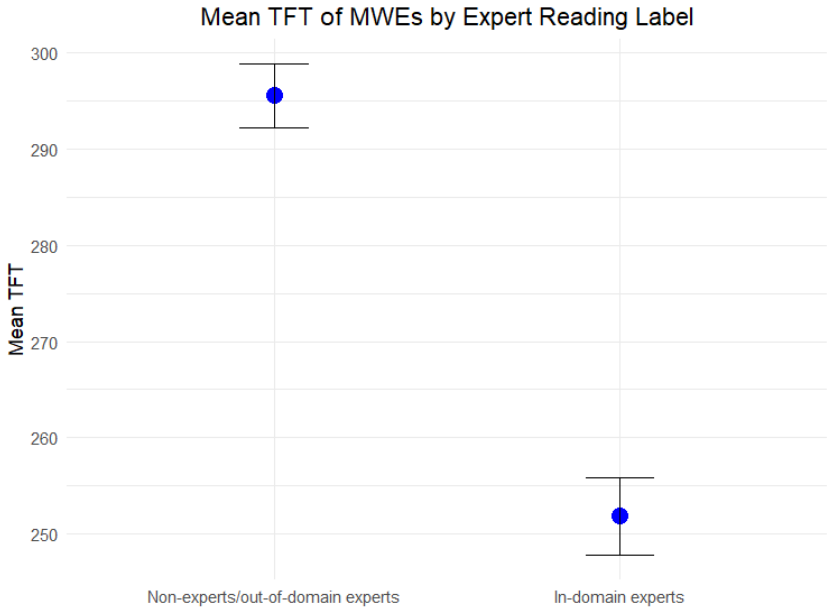

Figure 1: Mean total fixation time for in-domain experts and non-experts/out-of-domain experts (bars indicate standard error)

References

Conklin, Kathy & Norbert Schmitt (Mar. 2008). "Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?" In: *Applied Linguistics* 29.1, pp. 72–89.

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

Jäger, Lena A. et al. (Jan. 22, 2021). "Potsdam Textbook Corpus (PoTeC)". In: *Jäger, Lena A; Kern, Thomas; Haller, Patrick (2021). Potsdam Textbook Corpus (PoTeC). OSF: Open Science Framework.* Place: OSF Publisher: Open Science Framework.

Kuznetsova, Alexandra et al. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13, pp. 1–26.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA.

Siyanova-Chanturia, Anna (2013). "Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings". In: *The Mental Lexicon* 8.2. Publisher: John Benjamins, pp. 245–268.

Underwood, Geoffrey et al. (n.d.). "An eye-movement study into the processing of formulaic sequences". In: ().

Wahl, Alexander & Stefan Th. Gries (2018). "Multi-word Expressions: A Novel Computational Approach to Their Bottom-Up Statistical Extraction". In: *Lexical Collocation Analysis*. Ed. by Pascual Cantos-Gómez & Moisés Almela-Sánchez. Series Title: Quantitative Methods in the Humanities and Social Sciences. Cham: Springer International Publishing, pp. 85–109.