Multi-word expressions and language efficiency: an information-theoretic account

Stefan Fischer[1], Peter Fankhauser[2] & Elke Teich[1]
[1]Universität des Saarlandes, [2]Leibniz-Institut für Deutsche Sprache
stefan.fischer@uni-saarland.de, fankhauser@ids-mannheim.de,
e.teich@mx.uni-saarland.de

Multi-word expressions (MWEs) are a cornerstone in conventionalized language use and vital for the perceived fluency of a message (Fillmore 1979). From a processing perspective, MWEs seem to have an advantage over arbitrary word sequences due to highly predictable transitions from one word to the next, or they may be perceived as wholes (see e.g. Siyanova-Chanturia et al. 2017).

The emergence and use of specific MWEs is typically context-dependent and register-specific. In our work, we investigate MWEs in the scientific domain from a diachronic perspective, asking what is the contribution of MWEs in the development of "scientific language" (here: English)? We assume that over time scientific English develops an optimal code for scientific expert communication characterized by high information density (Halliday 2004; Teich et al. 2021).

Using a large diachronic corpus of English scientific texts (Fischer et al. 2020), we work in a data-driven fashion using various established word association measures (e.g. log-likelihood, PMI) to identify and classify MWEs by time periods (e.g. 50-year periods). In a complementary step, we account for the environments of words using selected computational language models (statistical models, embeddings; cf. Fankhauser & Kupietz 2022). On this basis, we then analyse the informational characteristics of MWEs diachronically: The more conventionalized an MWE becomes, the lower its surprisal (higher predictability of the MWE) and the lower the uncertainty about an upcoming word within the MWE (entropy). We expect to see that while specific MWEs come and go over time, during their life cycles they will exhibit surprisal/entropy reduction, thus contributing to language efficiency.

References: • Fankhauser, P. & Kupietz, M. (2022). Count-Based and Predictive Language Models for Exploring DeReKo. *Proceedings of LREC 2022 Workshop CMLC-10 2022*, 27–31. • Fillmore, C. J. (1979). On fluency. In Fillmore, C. J., Kempler, D. & Wang, W. S-Y. (eds.), *Individual Differences in Language Ability and Language Behavior*, 85–101. Academic Press. • Fischer, S., Knappen, J., Menzel, K. & Teich, E. (2020). The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study. *Proceedings of LREC 2020* (online). • Halliday, MAK (2004). The Language of Science. In Webster, J. (ed.), *The Collected Works of M. A. K. Halliday*. London: Continuum. • Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E. & van Heuven, W. J. (2017). Representation and processing of multi-word expressions in the brain. *Brain and Language*, 175:111–122. • Teich, E., Fankhauser, P., Degaetano-Ortlieb, S. & Bizzoni, Y. (2021). Less is More/More Diverse: On the Communicative Utility of Linguistic Conventionalization. *Frontiers in Communication, Topic: Rational Approaches in Language Science*. 5:620275 (online).