More complex or just more diverse? Capturing diachronic linguistic variation

Stefan Fischer, Elke Teich
Universität des Saarlandes
stefan.fischer@uni-saarland.de, e.teich@mx.uni-saarland.de

We present a diachronic comparison of general (register-mixed) and scientific English in the late modern period (1700–1900). For our analysis we use two corpora which are comparable in size and time-span: the Corpus of Late Modern English (CLMET; De Smet et al. 2015) and the Royal Society Corpus (RSC; Kermes et al. 2016).

Previous studies of scientific English found a diachronic tendency from a verbal, involved to a more nominal, abstract style compared to other discourse types (cf. Halliday 1988; Biber & Gray 2011). The features reported include type-token ratio, lexical density, number of words per sentence and relative frequency of nominal vs. verbal categories—all potential indicators of linguistic complexity at a shallow level.

We present results for these common measures on our data set as well as for selected information-theoretic measures, notably relative entropy (Kullback–Leibler divergence: KLD) and surprisal. For instance, using KLD, we observe a continuous divergence between general and scientific language based on word unigrams as well as part-of-speech trigrams. Lexical density increases over time for both scientific language and general language. In both corpora, sentence length decreases by roughly 25%, with scientific sentences being longer on average. On the other hand, mean sentence surprisal remains stable over time.

The poster will give an overview of our results using the selected measures and discuss possible interpretations. Moreover, we will assess their utility for capturing linguistic diversification, showing that the information-theoretic measures are fairly fine-tuned, robust and link up well to explanations in terms of linguistic complexity and rational communication (cf. Hale 2016; Crocker, Demberg, & Teich 2016).

References: • Biber, D., & Gray, B. (2011). The historical shift of scientific academic prose in English towards less explicit styles of expression: Writing without Verbs. Bhatia, V., Sánchez, P., & Pérez-Paredes, P. (Eds.), *Researching Specialized Languages*, 11–24. Amsterdam: John Benjamins. • Crocker, M. W., Demberg, V., & Teich, E. (2016). Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz*, *30*(1), 77–81. • De Smet, H., Flach, S., Tyrkkö, J., & Diller, H.-J. (2015). The Corpus of Late Modern English (CLMET), version 3.1: Improved tokenization and linguistic annotation. Retrieved 1 November 2018, from http://fedora.clarin-d.uni-saarland.de/clmet/clmet.html • Hale, J. (2016). Information-theoretical Complexity Metrics. *Language and Linguistics Compass*, *10*(9), 397–412 • Halliday, M. A. K. (1988). On the Language of Physical Science. Ghadessy, M. (Ed.), *Registers of Written English: Situational Factors and Linguistic Features*, 162–178. London: Pinter. • Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J., & Teich, E. (2016). The Royal Society Corpus: From Uncharted Data to Corpus. Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., … Piperidis, S. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).