**Computational linguistic modelling of the temporal dynamics of scientific communication: a quantitative corpus study on the journal Nature**

Gard Jenset[1], Isabell Landwehr[2], Barbara McGillivray[3] and Stefania Degaetano-Ortlieb[2]
[1]Springer Nature Group, [2] Saarland University, [3]King's College London and The Alan Turing Institute

We trace the linguistic evolution of English written scientific communication within the journal Nature, one of the world's leading multidisciplinary science journals, published since 1869. Our study applies computational models for diachronic linguistic analysis to investigate the statistical distribution of lexical and lexical-semantic features in a collection consisting of over 230,000 titles and abstracts from articles published in the journal Nature between 1869 and 2022, accessed via the Dimensions database (Hook et al. 2018).

We dynamically model changes in scientific language use over time. This overcomes the limitations of working with raw frequencies which tend to highlight only high-frequency features, disregarding low-frequency items (e.g. Biber and Gray 2016; Moskowich and Crespo 2012; Rissanen et al. 1997; Teich et al. 2016). We compare changes in probability distributions of individual lexical, grammatical, and semantic features with relative entropy as a measure of divergence for entire sets of features (e.g. all lemmas, parts of speech etc.), allowing for a comprehensive coverage of frequency bands. The dynamicity of the model is achieved by sliding over the timeline and continuously comparing adjacent time spans. The more a distribution of a feature changes over time, the higher the divergence will be, indicating changes in use. The sum of all features' divergence at a particular point in time gives an overall estimate of how much current language use is distinct from past practices, i.e. if a large number of features shows an increase in divergence over a time span, this will indicate a period of change. In terms of interpretability of the model, we are not only able to detect periods of change in a data-driven fashion, but can attribute these changes to sets of linguistic features that contribute to them. In addition, drawing on title and abstract embeddings for Nature articles using Google's Universal Sentence Encoder, we measure the trends in similarity between articles over time.

Previous work on the publications of The Royal Society of London (Degaetano-Ortlieb and Teich 2019, Degaetano-Ortlieb 2021) has proven the adaptability of applying dynamic divergence models to investigate change in scientific language use, showing specialisation trends at the lexical level and at the same time grammatical conventionalization trends. Sun et al. (2021) show similar results employing word embeddings methods. Research using embedding technologies applied to the labels of scientific disciplines rather than to the linguistic content has also found evidence for disciplines undergoing a process of global convergence combined with local specialisation (McGillivray et al. 2022). Previous work on Nature (Monastersky and Van Noorden 2019a) has shown specialisation of particular keywords in individual titles and abstracts. Our overarching question is whether these trends can be found for the journal Nature at scale, indicating general mechanisms of change in language use which contribute to the formation of the English scientific register. In addition, we are interested in changes that might be an indication of journal-specific linguistic features, especially considering the leading position of Nature in the scientific research landscape, as well as the journal's shift in focus over time (Monastersky and Van Noorden 2019a). We investigate the following sub-questions: (a) Can we observe similar/diverging diachronic trends between Nature and The Royal Society corpus, i.e. can we detect lexical and lexical-semantic diversification and grammatical conventionalization in Nature? (b) While we would assume similar diverging trends at the lexical level (new discoveries and technical advancement call for new linguistic expressions), do we encounter journal-specific trends at the grammatical and semantic level, and if so, are these disparate trends or do some trends start off in one journal and are picked up later in the other? Here we assume, besides grammatical trends indicating terminology formation processes, also changes in grammatical features that indicate text structuring functions (e.g. introductory linguistic

material such as prepositional phrases or discourse markers) and those that meet expressive needs given extra-linguistic pressures, such as passive voice usage during periods of increased experimental work).

**References**


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

Degaetano-Ortlieb, S. (2021). Measuring informativity: The rise of compounds as informationally dense structures in 20th century Scientific English. In Elena Soave and Douglas Biber (eds.), Corpus Approaches to Register Variation, chapter 11, John Benjamins Publishing Company, pp. 291-312.

Degaetano-Ortlieb, Stefania and Teich, Elke. "Toward an optimal code for communication: The case of scientific English" Corpus Linguistics and Linguistic Theory, vol. 18, no. 1, 2022, pp. 175-207.

Hook, D.W., Porter, S.J. and Herzog, C., 2018. Dimensions: building context for search and evaluation. Frontiers in Research Metrics and Analytics, 3, p.23.

McGillivray, B., Jenset, G.B., Salama, K. and Schut, D. 2022. Investigating patterns of change, stability, and interaction among scientific disciplines using embeddings. Humanities and Social Sciences Communications 9, 285. https://doi.org/10.1057/s41599-022-01267-5

Monastersky, Richard & Van Noorden, Richard. 2019a. 150 years of Nature: a data graphic charts our evolution. Nature: 575, 22-23. https://doi.org/10.1038/d41586-019-03305-w

Monastersky, Richard & Van Noorden, Richard. 2019b. 150 years of Nature: a data graphic charts our evolution. Supplementary information: Methodology. Nature: 575.
https://www.nature.com/magazine-assets/d41586-019-03305-w/17345736 (last accessed date: 23 December 2022).

Moskowich, Isabel & Begona Crespo (eds.). 2012. Astronomy Playne and simple: The writing of science between 1700 and 1900. Amsterdam/Philadelphia: John Benjamins.

Rissanen, Matti, Merja Kytö & Kirsi Heikkonen (eds.). 1997. English in transition: Corpus-based studies in linguistic variation and genre analysis. Berlin: Mouton de Gruyter.

Sun, K., Liu, H. and Xiong, W., 2021. The evolutionary pattern of language in scientific writings: A case study of Philosophical Transactions of Royal Society (1665–1869). Scientometrics, 126(2), pp.1695-1724.

Teich, Elke, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes & Ekaterina Lapshinova-Koltunski. 2016. The linguistic construal of disciplinarity: A data mining approach using register features. Journal of the Association for Information Science and Technology (JASIST) 67(7). 1668–1678.