Noun Phrases in English Scientific Writing: The Diachronic Evolution of Information Density

Isabell Landwehr (Saarland University) • Coauthors: Arianna Bienati

Complex noun phrases (NPs) are a key feature of English scientific writing and typically used for encoding technical and specialized concepts (Halliday 1988; Banks 2008). Over time, scientific writing has moved from an emphasis on clausal structures towards an emphasis on phrasal structures, allowing the transmission of information in a highly compressed way (Biber & Gray 2011). In this way, scientific English has evolved to be an optimized code for expert-to-expert communication (Degaetano-Ortlieb & Teich 2022).

In this study, we analyze these diachronic optimization mechanisms using the information-theoretic notion of Uniform Information Density (UID; Jaeger 2010). The UID hypothesis is based on the possibility of modeling language using surprisal (i.e. predictability in context, Shannon 1948) and predicts that language users prefer structures that are informationally more uniform if several encoding options exist (Jaeger 2010). Evidence for this theory has been found cross-linguistically for many linguistic phenomena, especially syntactic ones (Jaeger 2010; Clark et al. 2023; Liang et al. 2024). We aim to add a diachronic and register-informed perspective, focusing on NPs due to their significance for scientific writing. Our hypothesis is that NPs exhibit increased UID over time, as the register becomes increasingly optimized.

UID has been operationalized in various ways (for an overview, see Meister et al. 2021). Since we assess the information profiles of noun phrases, we choose measures of local variability. Local variability can be conceptualized as the magnitude of information gains and losses in the stream of communication. It is commonly operationalized by computing the delta of adjacent tokens' surprisals and then applying a descriptive statistical summary: Collins (2014) proposes taking the average of the differences and names the resulting measure UIDev, while Bates and Shepard (1993) use the standard deviation, resulting in a measure known as Information Fluctuation Complexity (IFC).

Our dataset is the Royal Society Corpus (Fischer et al. 2020; Menzel et al. 2021), a diachronic corpus of scientific writing. We use a version enriched with Universal Dependencies (De Marneffe et al. 2021) and surprisal annotation, the latter based on a 4-gram language model trained on the corpus. Furthermore, we focus on the articles in the Series A and Series B journals of the corpus, which contain texts from biology, physics and mathematics and span the time from 1887 to 1990. After extracting all subject and direct object NPs, we apply both UIDev and IFC as measures of UID. Using a linear-mixed effects model, we analyze the diachronic development of UIDev and IFC. We expect a positive effect of time on UID, indicating a trend towards increased optimization of the register over time.

References

David Banks. 2008. The Development of Scientific Writing: Linguistic Features and Historical Context. University of Toronto Press.

John E. Bates and Harvey K. Shepard. 1993. Measuring complexity using information fluctuation. Physics Letters A 172(6): 416-425.

Douglas Biber and Bethany Gray. 2011. Grammatical change in the noun phrase: The influence of written language use. English Language & Linguistics, 15(2): 223–250.

Thomas H. Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell and Roger Levy. 2023. A Cross-Linguistic Pressure for Uniform Information Density in Word Order. Transactions of the Association for Computational Linguistics, 11: 1048–65.

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. Journal of Psycholinguistic Research, 43(5): 651–681.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific English. Corpus Linguistics and Linguistic Theory, 18(1): 175–207.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre and Daniel Zeman. 2021. Universal Dependencies. Computational Linguistics, 47(2): 255–308.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 794-802. European Language Resources Association.

Michael A. K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy (ed.), Registers of written English: Situational factors and linguistic features, pp. 162-177.

Florian T. Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology, 61(1): 23-62.

Yiming Liang, Pascal Amsili, Heather Burnett and Vera Demberg. 2024. Uniform Information Density Explains Subject Doubling in French. Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0).

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell and Roger Levy. 2021. Revisiting the Uniform Information Density Hypothesis. In Proceedings of the 2021

Conference on Empirical Methods in Natural Language Processing, pp. 963–80. Association for Computational Linguistics.

Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. Generating linguistically relevant metadata for the Royal Society Corpus. Research in Corpus Linguistics, 9(1):1–18.

Claude E. Shannon. 1948. A Mathematical Theory of Communication. The Bell System Technical Journal, 27(3):379–423.