Information Density in Scientific Writing: A Diachronic Perspective

We report on a project investigating the development of scientific writing in English from the mid-17th century to present. While scientific discourse is a much researched topic, including its historical development (see e.g. Banks (2008) in the context of Systemic Functional Grammar), it has so far not been modeled from the perspective of information density.

Our starting assumption is that as science develops to be an established sociocultural domain, it becomes more specialized, on the one hand, and more conventionalized, on the other hand. Thus, denser linguistic encodings are required for specialist communication to be functional, potentially increasing the information density of scientific texts (Halliday and Martin, 1993:54-68). More specifically, we pursue the following hypotheses:

- As an effect of specialization, scientific texts will exhibit greater encoding density over time, i.e. denser linguistic forms will be increasingly used.
- As an effect of conventionalization, scientific texts will exhibit greater linguistic uniformity over time, i.e. the linguistic forms used will be less varied.

We further assume that these effects are measurable in the linguistic signal in terms of information density (see below).

We have built a diachronic corpus of scientific texts from the Philosophical Transactions and Proceedings of the Royal Society of London. We have chosen these materials due to the prominent role of the Royal Society in forming English scientific discourse (cf. Atkinson, 1998). At the time of writing, the corpus comprises 23 million tokens for the time period between 1776 and 1869 (other time periods will follow). The corpus has been normalized, tokenized and part-of-speech tagged.

For analysis, we combine methods from register theory (Halliday and Hasan, 1985) and computational language modeling (Manning et al., 2009:237-240). The former provides us with features that are potentially register-forming (cf. also Ure, 1971; 1982); the latter provides us with models with which we can measure information density. We thus pursue two complementary methodological approaches:

- Pattern-based extraction and quantification of linguistic constructions that are potentially involved in manipulating information density. Here, all linguistic levels are relevant (cf. Harris, 1991), from lexis and grammar to cohesion and generic structure. We have started with the level of lexico-grammar, inspecting for instance morphological compression (derivational processes such as conversion, compounding) and syntactic reduction (e.g. reduced vs full relative clauses).
- Measuring information density using information-theoretic models (Shannon, 1949). In current practice, information density is measured based on the probability of an item conditioned on context as $ID(item) = -\log_2 P(item \mid \text{Context})$. For our purposes, we need to compare such probabilities to assess the *relative* information density (cross-entropy) of texts along a time line. Here, we apply various probability distance measures, notably Kullback-Leibler divergence (Fankhauser et al., 2014).

The ultimate goal is to test hypotheses about which kinds of linguistic patterns contribute to relative information density and to what extent: e.g. if reduced relative clauses increase over time, what is the effect on information density, if any, and how big is the effect?

The present research is an extension of our previous work on register formation in contemporary scientific English on the basis of the SciTex corpus (Kermes and Teich, 2012; Degaetano-Ortlieb et al., 2014; Teich et al., to appear) to which we are now adding a diachronic perspective.

Acknowledgment. This work is supported by *Deutsche Forschungsgemeinschaft* (DFG) under grant SFB 1102 “Information density and linguistic encoding” (www.sfb1102.uni-saarland.de).

References

Atkinson, Dwight, 1998. *Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. Routledge, New York.

Banks, David, 2008. *The development of scientific writing: Linguistic features and historical context*. Equinox, London.

Degaetano-Ortlieb, Stefania, Peter Fankhauser, Hannah Kermes, Ekaterina Lapshinova-Koltunski, Noam Ordan and Elke Teich, 2014. Data Mining with Shallow vs. Linguistic Features to Study Diversification of Scientific Registers. *Proceedings of the of the Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland.

Fankhauser, Peter, Hannah Kermes and Elke Teich, 2014 Combining Macro- and Microanalysis for Exploring the Construal of Scientific Disciplinarity. *Proceedings of Digital Humanities 2014.* Lausanne, Switzerland.

Halliday, M.A.K. and Ruqaiya Hasan, 1985. *Language, context and text: Aspects of language in a social semiotic perspective*. Deakin University Press, Geelong.

Halliday, M.A.K. and James R. Martin, 1993. *Writing science: Literacy and discursive power*. Falmer Press, London.

Harris, Zellig, 1991. *A theory of language and information: A mathematical approach*. Clarendon Press, Oxford.

Kermes, Hannah and Elke Teich, 2012 "Formulaic expressions in scientific texts: Corpus design, extraction and exploration." *Lexicographica* 28.1: 99-120.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze, 2009. *An introduction to information retrieval*. Cambridge University Press, Cambridge UK.

Shannon, Claude E., 1949. *The mathematical theory of communication*. University of Illinois Press, Urbana and Chicago (1983 edition).

Teich, Elke, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes and Ekaterina Lapshinova-Koltunski,, to appear. “The Linguistic Construal of Disciplinarity: A Data Mining Approach Using Register Features. “ *Journal of the Association for Information Science and Technology (JASIST)*.

Ure, Jean, 1971. Lexical density and register differentiation. In G. E. Perren and J. L. M. Trim, editors, *Applications of linguistics. Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969*, pp. 443–452. Cambridge University Press, Cambridge UK.

Ure, Jean, 1982. Introduction: Approaches to the study of register range. *International Journal of the Sociology of Language*, 35:5–23.