Information density in scientific writing: Exploring the SciTex corpus

Stefania Degaetano-Ortlieb (Saarland University)
Hannah Kermes (Saarland University)
Ashraf Khamis (Saarland University)
Elke Teich (Saarland University)

The linguistic evolution of scientific writing is characterized by two major motifs: *specialization* and *conventionalization*. The assumption is that as scientific domains become more specialized, particular meanings become more predictable in these domains and call for denser encodings that minimize redundancy while maintaining accuracy in transmission. Specialization is manifested linguistically by densification in encoding, something observed, for example, on single text instances from the language of physical science (Halliday 1988). Balancing the effects of specialization, conventionalization leads to greater linguistic uniformity, i.e. over time scientific texts show greater resemblance to one another and are more clearly distinguishable as scientific. Our main hypothesis is that the linguistic features realizing specialization and conventionalization serve to optimize information density in scientific writing. This hypothesis is based on recent work in psycholinguistics, which suggests that there is a correlation between variation in linguistic encoding and information density (see e.g. Aylett and Turk (2004); Levy (2008)). It is assumed that highly informative (i.e. informationally dense) parts of an utterance are less predictable and thus realized by more expanded linguistic forms, while less informative parts are realized by shorter, more reduced forms.

To empirically investigate information density in scientific writing, we use the SciTex corpus (see Teich and Fankhauser 2010; Degaetano et al. 2013) which covers nine scientific disciplines (computer science, computational linguistics, linguistics, bioinformatics, biology, digital construction, mechanical engineering, microelectronics, and electrical engineering). The corpus is annotated for structural information (such as sections (Abstract, Introduction, etc.), paragraphs, and sentences) as well as positional information (such as lemma and part of speech). To compare informationally dense vs. less informationally dense text, we consider *abstracts* vs. *research articles without their abstracts*, assuming that abstracts are more informationally dense than their research articles.

In terms of methods, we use (1) text classification and (2) calculation of cross-entropy rate. Text classification is performed by considering linguistic features possibly involved in optimizing information density (e.g. high/low standardized type-token ratio, high/low lexical density, complex/simple NPs, complex/simple clause structure, use/omission of relativizer, etc.), looking at how well abstracts can be distinguished from research articles by these features and which features mainly contribute to the distinction. Calculation of cross-entropy rate is based on Genzel and Charniak (2002), considering entropy at each token position. Particular tokens have higher entropy rates, showing peaks in entropy (e.g. lexical words), while others have lower entropy rates, pointing at troughs (e.g. function words). This helps to explore whether abstracts have a higher cross-entropy rate than research articles (i.e. show a higher amount of peaks in entropy, being thus more informationally dense). Here, we also consider the variation among scientific disciplines.

In the talk, we will present the methodology applied and the results from (1) text classification, which show that abstracts are indeed distinct from research articles in terms of linguistic features involved in information density, as well as from (2) cross-entropy calculation, which also points to distinctions between abstracts and research articles and differences across disciplines.