Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, Noam Ordan, Elke Teich

Information Density and Scientific Literacy in English

In the evolution of English scientific writing, we hypothesize a process of linguistic densification, predicting that as scientific activity becomes more specialized, particular meanings become more predictable and call for denser encodings that optimize efficiency in communication.

To investigate this hypothesis, we employ (a) corpus-based register analysis (linguistic variation according to situation), (b) data mining (especially text classification to compare text/text classes in terms of differential linguistic encodings), and (c) computational language modeling which provides us with measures of information density in terms of entropy. These techniques are applied to observe synchronic as well as diachronic differences in scientific text productions. For the synchronic part, we investigate differences between abstracts vs. research article bodies, assuming that abstracts will show a higher encoding density than research articles. For the diachronic part, we consider research articles from the Proceedings of the Royal Society of London from the mid 17th century to present, assuming that earlier productions will show lower encoding density than contemporary ones.

In the talk, we will show our research design as well as selected preliminary analyses.