Abs-150

Hannah Kermes

Usage and function of formulaic expressions in scientific texts

Formulaic expressions are commonplace in scientific riting. Scientists use patterns such as 'based on the', 'the/a number of', 'in other words' consciously or unconsciously to convey research interests, the theoretical / data bases of studies, results of experiments, scientific findings, conclusions and as discourse organizers.

Research studies such as Biber et al. (2004); Biber (2006) and Simpson (2004) use frequencies to identify lexical bundles typical for academic language. Biber et al. (2004) differentiate structural types and distinguish among three primary functions (stance expressions, discourse organizers and referential expressions) each with several subcategories. They show the distribution of the different structural types and categories across different types of academic language (conversation, classroom teaching, textbooks and academic prose). Simpson-Vlach and Ellis (2010) in addition argue for the use of statistical measures such as mutual information for the identification of academic formula. Mutual information allows to find typical and salient but low-frequent patterns, which cannot be identified on a simple frequency basis. They classify their list of formula according to the same primary categories used by Biber et al. (2004), adjusting the scheme of subcategories where their data makes it necessary.

In our approach we want to go a step further, investigating the linguistic usage of formulaic expressions in a mostly automatic fashion. The questions we are interested in are: How are formulaic expressions used in scientific text, when are they used and what function do they have? We extract frequency distributions not only with respect to text type but also with regard to the occurrence of formula within the texts. Do they occur in the abstract, the introduction, the main part or to the end of a text. Are they distributed evenly throughout the text? Are there formula dense text areas, with a relative high frequency of formulaic expressions, and areas with a low density of formulas? We are also interested in differences and commonalities with respect to the frequency distribution of these features across different scientific disciplines. Taking the extracted frequency distribution of the features and statistical measures as basis, we want to group the formulas and find out, whether it is possible to draw conclusions about their function. The results are compared to the lists of Biber et al. (2004) and Simpson-Vlach and Ellis (2010).

As a data basis for our study we use the DaSciTex (Darmstadt Scientific Text) corpus. A corpus with approximately 17 million words (around 2000 texts) from nine scientific disciplines: four interdisciplinarydomains (computational linguistics, bioinformatics, computer-aided design, micro-electronics) and the corresponding “pure” disciplines (computer science, linguistics, biology, mechanical engineering, electricalengineering) (Teich and Holtz, 2009; Teich and Fankhauser, 2010).

References

Biber, Douglas (2006). University language. John Benjamins, Amsterdam.

Biber, Douglas, Conrad, Susan, and Cortes, Viviana (2004). ”If you look at ...:” Lexical Bundles in University Teaching and Textbooks. Applied Linguistics, 25(3), 371–405.

Simpson, Rita (2004). Stylistic features of academic speech: The role of formulaic expressions. In T. Upton and U. Connor, editors, Discourse in the professions: Perspectives from corpus linguistics. John Benjamins, Amsterdam.

Simpson-Vlach, Rita and Ellis, Nick C. (2010). An Academic Formulas List (AFL). Applied Linguistics, 31(4), 487–512.

Teich, Elke and Fankhauser, Peter (2010). Exploring a corpus of scientific text using data mining. In S. T. Grief, S. Wulf, and M. Davies, editors, Corpus-linguistic applications: Current studies, new directions. Rodopi.

Teich, Elke and Holtz, Mônica (2009). Scientific registers in contact. An exploration of the lexicogrammatical properties of interdisciplinary discourses. International Journal of Corpus Linguistics, 14(4), 524–548.