Peter Uhrig (University of Erlangen-Nürnberg)

Title

Collo-Phenomena, Lexicography and AI – Leveraging Computational Advances for more Efficient Collocation Candidate Extraction

Abstract

The talk explores the integration of linguistic research and artificial intelligence to improve the extraction and analysis of collocations. The use of dependency-annotated corpora (e.g. Uhrig/Proisl 2012) will be discussed, highlighting the role of syntactic information for identifying collocational patterns with high precision. Following this, the interaction of dependency parsing with association measures and frequency thresholds will be discussed with respect to two different gold standards (Evert et al. 2017) to identify the most effective statistical methods for recognizing significant collocations in large text corpora.

Different parsers and parsing schemes (Uhrig et al. 2018) will be examined, assessing their impact on the accuracy and reliability of syntactic annotation, which is essential for precise collocation extraction. A short excursus to larger structures (Proisl 2018) and crossmodal collostructions (Uhrig 2021) will be included, demonstrating how integrating multiple modes of linguistic data can add relevant information and create new challenges for methodology and lexicography.

The potential of large language models (LLMs) in collocation lexicography will also be discussed. Specifically, the use of LLMs for creating lists of collocation candidates, grouping these candidates semantically, and generating entire collocation dictionary entries will be examined.

The talk discusses efficient and effective practices in collocation candidate extraction for lexicography that leverage computational tools and methodologies.

Works Cited

Evert, Stefan/Peter Uhrig/Sabine Bartsch/Thomas Proisl (2017): "E-VIEW-alation – a large-scale evaluation study of association measures for collocation identification." In *Electronic lexicography in the 21st century. Proceedings of the eLex 2017 conference*, Leiden, The Netherlands.

Proisl, Thomas (2019): *The cooccurrence of linguistic structures.* Erlangen: FAU University Press.

Uhrig, Peter (2021): *Large-Scale Multimodal Corpus Linguistics – The Big Data Turn.* Habilitation Thesis, FAU Erlangen-Nürnberg.

Uhrig, Peter/Stefan Evert/Thomas Proisl (2018): "Collocation candidate extraction from dependency-annotated corpora: Exploring the differences between parsers and dependency annotation schemes." In: Pascual Cantos Gómez and Moisés Almela Sánchez (eds.) *Lexical Collocation Analysis: Advances and Applications.* Berlin: Springer.

Uhrig, Peter/Thomas Proisl (2012): “Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates.” *Lexicographica* 28.