Enhancing National Corpus Infrastructure with Multidimensional Model of Register Variation

This presentation explores the role of multidimensional analysis (MDA) in enhancing the infrastructure of national corpora, specifically through the lens of the Czech National Corpus (CNC, www.korpus.cz). Devised by Douglas Biber (Biber, 1988, 1995; Biber & Conrad, 2009), MDA offers a corpus-based methodology for describing register variation, focusing on the systematic and pervasive co-occurrence of linguistic features (grammatical as well as lexical) across different texts.

The talk will present findings from (Cvrček et al., 2018; Cvrček, Laubeová, et al., 2020), which applied MDA to the CNC, demonstrating its utility in identifying major communication registers and dimensions of variation. The implications of these findings are multifaceted: not only do they contribute to the enhancement of corpus infrastructure by introducing a new level of metadata (such as text positioning on specific dimensions or prevailing register of texts), but they also inform broader linguistic research.

The presentation will explore how MDA aids in understanding language variability and establishing criteria for corpus representativeness. This includes, among other things, a comparative analysis of traditional, carefully designed corpora and opportunistic web-crawled corpora (Cvrček, Komrsková, et al., 2020). The application of MDA extends beyond corpus construction; it is instrumental in related research areas, including analysis of specific genres (e.g., parliamentary speeches), the evolution of text types (such as journalistic texts and fiction) over time, and the examination of elicited texts (e.g. in psycholinguistics).

References

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press.

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018). From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2018-0020

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., Zasina, A. J., & Benko, V. (2020). Comparing web-crawled and traditional corpora. *Language Resources and Evaluation*, *54*, 713–745. https://doi.org/10.1007/s10579-020-09487-4

Cvrček, V., Laubeová, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2020). *Registry v češtině*. Nakladatelství Lidové noviny.