Exploring text-internal register organization using machine learning

Erik Henriksson, Saara Hellström, Antti Kanner & Veronika Laippala
TurkuNLP, University of Turku, Finland
[eihenr, sherik, aokann, mavela]@utu.fi

In the text-linguistic approach to register variation (Biber 2019), registers are seen as culturally recognized text varieties associated with a communicative situation. Registers are analyzed through pervasive linguistic features. Although text-internal variation has attracted some attention in register studies as well, the internal structure of texts is much more central in discourse-analytic approaches to genre (Swales 1990) and studies on discourse organization (Braud et al. 2023).

In a recent study, however, Henriksson et al. (2025) presented a sub-document register segmentation method, showing that text-internal passages featuring distinct registers can be found and automatically identified in web data. In this presentation, we examine the inner structure of texts as reflected by the predicted sub-document registers. How frequent are texts with different registers assigned to different text segments? What typical register sequences can we find, and how do these relate to the overall register of the text?

For our analysis, we use the Corpus of Online Registers of English featuring the unrestricted web (Egbert et al. 2015). The register labels are created with ModernBERT and the register segmentation method by Henriksson et al. (2025), achieving an 0.86 F1 for entire documents and 0.84 IAA for the segments. Our first findings indicate that 26% of the corpus documents have several registers assigned to their subparts, the most frequent combinations being *Narrative* + *Informational*. In the presentation, we will examine the linguistic characteristics of these combinations and the text-internal register organization they reflect.

References: • Biber, D. (2019). Text-linguistic approaches to register variation. *Register Studies* 1(1), 42–75. • Braud, C., et al. (2023). The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of DISRPT 2023*. Toronto, Canada. • Egbert, J., Biber, D., & M. Davies (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology.* • Henriksson, E., Hellström, S. & V. Laippala (2025). Analyzing register variation in web texts through automatic segmentation. In *Proceedings of NLP4DH.* ACL. • Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.