Outliers, novel registers or plain errors? Analysing texts with no register label

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

Within the register framework, registers are defined as culturally recognized, situationally bound language use with different purposes (Biber & Egbert 2023). Hybrids are texts mixing features of various registers, lacking a register category and thus considered not culturally recognized (Biber & Egbert 2023). Automatic text classification predicts the likelihood of a text belonging to a register, making it possible not only to recognize single-class texts (registers) and multi-class texts (hybrids) but also texts that are neither, i.e., they do not correspond to any register or a combination thereof. Such texts have often been excluded from previous research as outliers. Using French and Swedish HPLT web data (Burchell et al. 2025), we explore the linguistic features and functions of the unlabelled texts, and their cross-textual (dis)similarities.

Adapting from Santini (2007), we study four hypotheses for unlabelled texts: they belong to 1) outliers, 2) yet uncovered registers, 3) still only emerging register(s), or 4) prediction errors. We test the hypotheses by applying multi-dimensional analysis (MDA; Biber 1988) and document embeddings (see, e.g., Grootendorst 2022). In MDA, cooccurring linguistic features form dimensions based on which the texts can be characterized and compared. Document embeddings generated by a language model are dense numerical representations of texts that capture semantic and contextual relationships in multidimensional vector space. We believe that these two methods complement each other in situating the unlabelled texts in the continuous space of register variation (Biber et al. 2020).

References: • Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press. • Biber, D. & J. Egbert (2023). What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies* 5(1), 1–22. • Biber, D., Egbert, J. & D. Keller (2020). Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory* 16(3), 581–616. • Burchell, L. et al. (2025). *An expanded massive multilingual dataset for high-performance language technologies (HPLT)*. arXiv. https://doi.org/10.48550/arXiv.2503.10267 • Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv. https://arxiv.org/pdf/2203.05794 • Santini, M. (2007). Characterizing genres of web pages: Genre hybridism and individualization. *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, USA, 1–10.