Analysing text macro structures with automatic register identification

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

Broadly, registers describe the variety of language use appropriate for a particular social context (Biber 1995). Genres in linguistic text analysis (Swales 1990; Martin & Rose 2008) are categories of texts sharing mutual social purposes and structural parameters derived from those purposes. Theoretically, these notions can interact unproblematically, as a genre's structural parameters may dictate which registers are possible in which segments of texts. On the other hand, register identification is thought to handle complete documents, as pervasive linguistic features of a document resonate throughout full documents. In practice, the reconciliation of genres and registers has equally proven challenging, as it is for example somewhat typical that a category of texts could be identified both as a genre and a register.

Against this background, it is interesting that in automatic register identification, the beginnings of documents are shown to provide better identification accuracy than other document parts (Laippala et al. 2023). It thus seems that information about text macro structure is useful for register identification and, further, that for machine-learning models the division between genre and register may not be clear-cut. In our paper, we engage with the question of what kind of information automatic register identification systems base their assessments on and, more specifically, whether they might actually carry genre-related information.

In our paper, we will investigate this phenomenon by analysing the performance of automatic register identification in the HPLT 2.0 dataset (Burchell et al. 2025) for English, French and Finnish. We will also experiment with identification systems based on different types of linguistic structures: mainstream embeddings based on sub-lexical units ("tokens"), alongside with more bespoke approaches based on functional and grammatical features coming from parser output and syntactic constructions based on dependency biarchs and triarchs. The aim is to see how the performance of the automatic register identification varies across text segments, different sets of linguistic features and different languages.

References Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press. • Burchell, L et al. (2025). *An expanded massive multilingual dataset for high-performance language technologies (HPLT)*. arXiv. • Martin, J. R., & D. Rose (2008). *Genre relations: Mapping culture*. Equinox Publishing. • Laippala, V., et al (2023). Register identification from the unrestricted open web using the Corpus of Online Registers of English. *Language Resources and Evaluation* 57, 1045–1079. • Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.