Using deep learning to examine cross-linguistic similarities of registers

Erik Henriksson, Amanda Myntti, Veronika Laippala

University of Turku, Finland

This study examines the use of multilingual deep learning to analyze cross-linguistic similarities of registers – situationally defined text varieties such as news or reviews (Biber 1988). Register studies have repeatedly shown that differences in the situational context of a text are reflected in its linguistic characteristics. However, little is known about register variation across languages (see, however, Biber 2014; Li et al. 2023). One of the reasons for this is the lack of methods enabling the analysis of registers without the manual interpretation of register characteristics in each language. In this study, we apply multilingual deep learning to fill this gap.

We examine cross-linguistic similarities of registers using the deep learning model XLM-R (Conneau et al. 2020). Specifically, we target eight registers and eight languages: English, French, Swedish, Finnish, Turkish, Urdu, Chinese, and Farsi. First, using the multilingual CORE corpora (Laippala et al. 2022) and XLM-R, we train a multilingual register identification model. The model learns to classify documents to register classes and creates document vectors that represent the documents in one multilingual vector space. This allows us to examine registers and their similarities across languages by calculating document similarities in the vector space. Second, we extract keywords for the registers using the trained model and the model explanation method SACX (Rönnqvist et al. 2022). This enables the analysis of the linguistic motivation behind the learnt model. We group the keywords using semantic and grammatical criteria and analyze the registers and their similarities across languages based on these groupings. Furthermore, we compare our findings to previous studies based on more frequently applied statistical methods, such as multi-dimensional analysis.

Preliminary results show that the model learns to identify the registers at a nearly human-level performance. In the vector space, the documents are structured to language-independent and register-specific groupings. This shows that the model has learnt language-independent representations of the registers. Furthermore, the analysis of the keywords shows that the learning is based on linguistically motivated features. For instance, the keywords feature semantic properties such as stance and functional features such as reporting verbs that characterize registers across languages – and have been identified as register characteristics in previous studies focusing on individual languages. Thus, our findings support the existence of register universals (Biber 2014) and encourage the use of multilingual deep learning for cross-linguistic corpus analyses.

Keywords: multilingual machine learning, web-as-corpus, web registers, register universals, keyword extraction

References:

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D. (2013). In Gray, Bethany. (2013). Interview with Douglas Biber. *Journal of English Linguistics* 41(4), 359–379. https://doi.org/10.1177/0075424213502237.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contras*t, 14(1), 7-34. https://doi.org/10.1075/lic.14.1.02bib

Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440-8451. Association for Computational Linguistics.

Laippala, V., Salmela, A., Rönnqvist, S., Aji, A. F., Chang, L.-H., Dhifallah, A., Goulart, L., Kortelainen, H., Pàmies, M., Prina Dutra, D., Skantsi, V., Sutawika, L., & Pyysalo, S. (2022). Towards better structured and less noisy Web data: Oscar with Register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, 215-221. Association for Computational Linguistics.

Li, H., Dunn, J. & Nini, A. (2022). Register variation remains stable across 60 languages. *Corpus Linguistics and Linguistic Theory,* 19(3), 397-426. https://doi.org/10.1515/cllt-2021-0090

Rönnqvist, S., Kyröläinen, A.-J., Myntti, A., & Laippala, V. (2022). Explaining Classes through Stable Word Attributions. In *Findings of the Association for Computational Linguistics,* 1063-1074. Association for Computational Linguistics.