Comparing French and Swedish web registers using multilingual word vectors

Saara Hellström (University of Turku)

The web features a wide variety of registers (Biber, 1988), i.e., situationally defined language use with different purposes (e.g., blogs, news, recipes), in numerous languages. Yet online language use in other languages than English (Biber & Egbert, 2018) remains largely unexplored. Moreover, comparisons across languages are manually conducted which is time-consuming and prone to subjective interpretations. Our study expands web register research to French and Swedish and examines the register characteristics using multilingual word vectors allowing the analysis of registers in one multilingual space without manual comparison. Our research aims at answering the following questions: 1) What kind of keyword groupings does the clustering of the word embeddings reveal? and 2) What (dis)similarities does the clustering of the word embeddings reveal about the languages and registers?

Our data consists of the newly established FreCORE and SweCORE corpora including similarly register-annotated web documents. In our analysis, we first extract the keywords, i.e., the statistically overrepresented words indicating what the texts are about (Scott & Tribble, 2006, pp. 55-59), from the corpora using text dispersion keyness (Egbert & Biber, 2019) to get the language specific characteristics for the registers. Then, using the fastText tools, we transform the keywords into word vectors, i.e., linguistically motivated, numerical representations of words derived from a language model. The word vectors present words in one multilingual space where semantically similar words are represented by similar vectors even across languages. Finally, to examine the cross-lingual similarities of the keywords and what they tell about the registers, we cluster the word vectors with KMeans.

Nineteen clusters offer the best fit to the data. Our analysis shows that the clusters group keywords based on their topical or grammatical features: e.g., the cluster POLITICS/POWER (topic) includes *pouvoir* – *makt* (power; authority), *people* – *folket* (people) while the cluster STANCE (grammar) features *pense* – *tänker* (thinks), *vrai* – *sant* (true). Moreover, the keywords in each cluster tend to belong to certain dominant registers, and these prominent registers and clusters are often the same in both French and Swedish. The keywords within a register group coherently which suggests that clustering could be a viable method to group keywords computationally instead of the laborious manual grouping. These findings suggest that there are more cross-linguistic similarities than dissimilarities between the French and Swedish web registers.

References

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

Biber, D. & Egbert, J. (2018). *Register Variation Online*. Cambridge University Press.

Egbert, J. & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77–104.

Scott, M. & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. John Benjamins Publishing Company.