A Glimpse into Terminology Research with R: Two Experiments Exploring Diastratic Variation in a Large Specialized Corpus

Nicolás Gonzalez Granado*†[1], Aurélie Picton‡[2], and Patrick Drouin§[3]

[1]Département de traitement informatique multilingue [Genève] (TIM) – Suisse
[2]Département de traitement informatique multilingue [Genève] (TIM) – Suisse
[3]Observatoire de linguistique Sens-Texte (OLST) – Département de linguistique et de traduction - Université de Montréal - C.P. 6128, succ. Centre-ville - Montréal (Québec) H3C 3J7 - Canada, Canada

Résumé

The increasing possibilities for the study of specialized discourse have seen terminologists dealing with large volumes of more heterogenous corpus data. In this context, researchers are faced with the prospect that ready-made tools might fall short, which emphasizes the need for programming languages. R has become one of the most popular choices among linguists as a tool for both data extraction and data evaluation tasks (e.g. Desagulier, 2017). For common operations such as concordances, the benefits of developing customized scripts may not compensate for a steep learning curve (Anthony, 2013). However, when it comes to advanced techniques, writing code opens doors that otherwise remain closed.
Given that R is free and open-source, it has attracted a huge community that allows it to keep up with new methods in statistical analysis and machine learning (Wickham, 2019). As an illustration, we propose an experiment in each of these two areas. Both aim to examine diastratic variation, understood as the coexistence of different language uses within groups of experts in the same field. The corpus that we have chosen for these tests, created for the Humanitarian Encyclopedia project, contains over 70 million occurrences and can be subdivided based on various criteria. In this case, we center on the eleven types of humanitarian organizations and their subcorpora, all of very disparate sizes.
In each experiment, we focus on a different phenomenon of diastratic variation, providing a step-by-step description of the approach adopted to investigate it. First, we compare the terminologies of the organization types and establish whether certain communities of humanitarian actors favor specific concepts. To remedy the imbalanced sizes of the subcorpora, we turn to correspondence analysis, an exploratory technique to reveal patterns of association in categorical data and display them in two-dimensional plots (e.g. Glynn, 2014). Second, we represent humanitarian terms as more or less distant points in space by capturing their meanings as vectors. Building on the assumption that similarity in meaning correlates with similarity in distribution, the word2vec algorithms rely on deep learning technology to infer the meaning of a lexical unit from its contexts in a corpus (Mikolov et al., 2013). Together, the two experiments lead us to discuss key perspectives and limitations for R in terminology

*Intervenant
†Auteur correspondant: Nicolas.Gonzalez@etu.unige.ch
‡Auteur correspondant: aurelie.picton@unige.ch
§Auteur correspondant: patrick.drouin@umontreal.ca

studies.

Anthony, L. (2013). A Critical Look at Software Tools in Corpus Linguistics. *Linguistic Research, 30*(2), 141–161. https://doi.org/10.17250/khisli.30.2

Desagulier, G. (2017). *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Springer International Publishing.

Glynn, D. (2014). Correspondence Analysis: Exploring Data and Identifying Patterns. In D. Glynn & J. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy* (pp. 443–485). John Benjamins Publishing Company.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. ICLR Workshop Track 2013: Scottsdale, USA. https://arxiv.org/abs/1301.3781

Wickham, H. (2019). *Advanced R (2nd ed.)*. CRC Press.

Mots-Clés: specialized corpora, large corpora, R, diastratic variation