Introducing MMDA: An interactive toolkit for CDA

Stefan Evert, Philipp Heinrich
Computational Corpus Linguistics, Friedrich-Alexander-Universität Erlangen-Nürnberg

Corpus-based discourse analysis (CDA) is a popular and highly successful technique for the investigation of socio-political research questions (see e.g. Baker 2006; McEnery et al. 2015). The CDA procedure starts from collocation analyses for selected subcorpora and/or keyword analyses of suitable (sub-)corpora. Collocates (or keywords) are then grouped into categories that are supposed to reflect discursive positions, i.e. attitudes towards the topic. These interpretations are verified and refined by careful inspection of the corresponding KWIC concordances.

Given the wide-spread use of CDA and recent advances in computational linguistic techniques, it is surprising that researchers still have to rely on general-purpose corpus tools such as CQPweb (Hardie 2012). There is no dedicated software solution that combines the necessary steps of CDA into a single consistent and efficient working environment. As a result, the CDA workflow remains unidirectional: the researcher carries out one or more collocation analyses with fixed parameters (such as association measure and span size) and a pre-defined topic node; concordances for salient collocates are viewed in the corpus tool; then the lists of collocates are exported to a text editor or similar program where they can be grouped into discursive positions. Even though the CDA procedure builds on a manual categorization of lexical items, there is no clear pre-determined classifcation scheme; hence CDA has not been benefitted from the recent development of powerful end-to-end classifiers based on deep learning techniques.

In our contribution, we present an interactive software toolkit called MMDA (for “mixed-methods discourse analysis”), which enables the user to carry out multiple collocation analyses in parallel and which visualizes the results in an intuitive way. The user can try out different parameter settings in real time, which provides a more comprehensive understanding of the semantic space of the discourse. From a Digital Humanities perspective, our approach can be understood as an attempt to blend “close” and “distant” reading techniques. Our visualization is a two-dimensional semantically structured map of the discourse, based on word embeddings (cf. Mikolov et al. 2018), which we created for the respective linguistic registers. The MMDA toolkit represents a first step towards a more sophisticated CDA methodology, in which the manual categorization procedure is operationalized in terms of so-called *discoursemes*, groups of related collocates that form building blocks of discursive positions. Furthermore, our toolkit supports the interactive exploration of second-order collocates, i.e. co-occurrences of discoursemes in the context of the selected topic node.

We demonstrate the usefulness of our toolkit with two use cases: the Bavarian parliamentary elections (Landtagswahlen) of 2018 and the discourse around the topic *austerity*. Since both case studies are part of a larger research agenda aimed at understanding political discourse in the transnational algorithmic public sphere (cf. Heinrich et al. 2018), we analyze newspaper texts as well as Twitter corpora. The MMDA approach, combined with the triangulation of discourseme semantics via second-order collocates, allows us e.g. to analyze the interplay of ideologies against the backdrop of topics such as *migration* or *austerity*. Last but not least, we also demonstrate how the removal of noise (social bots and duplicate tweets, cf. Schäfer et al. 2017) impacts the collocational profiles.

References

- Baker, P. (2006). *Using Corpora in Discourse Analysis.* London: Continuum Books.
- Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380–409.
- Heinrich, P., Adrian, C., Kalashnikova, O., Schäfer, F., and Evert, S. (2018). A Transnational Analysis of News and Tweets about Nuclear Phase-Out in the Aftermath of the Fukushima Incident. In *Proceedings of the LREC 2018 Workshop on Computational Impact Detection from Text Data*, pages 8–16. Miyazaki, Japan.
- McEnery, T., McGlashan, M. and Love, R. (2015). Press and Social Media Reaction to Ideologically Inspired Murder: The Case of Lee Rigby. *Discourse and Communication* 9(2): 1–23.
- Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC 2018), pages 52–55, Miyazaki, Japan.
- Schäfer, F., Evert, S., and Heinrich, P. (2017). Japan's 2014 general election: Political bots, right-wing internet activism and PM Abe Shinzō's hidden nationalist agenda. *Big Data* 5(4): 294–309.