Using DiaCollo for historical research

Bryan Jurish
Berlin-Brandenburgische Akademie der Wissenschaften
Berlin, Germany
jurish@bbaw.de

Maret Nieländer
Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung
Braunschweig, Germany
nielaender@leibniz-gei.de

Abstract

This article presents some applications of the open-source software tool DiaCollo for historical research. Developed in a cooperation between computational linguists and historians within the framework of CLARIN-D's discipline-specific working groups, DiaCollo can be used to explore and visualize diachronic collocation phenomena in large text corpora. In this paper, we briefly discuss the constitution and aims of the CLARIN-D discipline-specific working groups, and then introduce and demonstrate DiaCollo in more detail from a user perspective, providing concrete examples from the newspaper "*Die Grenzboten*" ("messengers from the borders") and other historical text corpora. Our goal is to demonstrate the utility of the software tool for historical research, and to raise awareness regarding the need for well-curated data and solutions for specific scientific interests.

1 Introduction

Ever since their establishment in 2011, German CLARIN centers have worked together with discipline-specific working groups to develop and improve their services in close dialogue with the needs of philologies, history, social science, etc. The German CLARIN initiative, CLARIN-D, has strong roots in computational linguistics. With the help of the working groups, it has been possible to curate and integrate corpus data that is important to different fields of the humanities and social sciences, as well as to disseminate knowledge of the usefulness of computational linguistic methods for other disciplines.

Developed in a collaboration between historians and computational linguistics within the context of the CLARIN-D working groups, DiaCollo (Jurish, 2015) is an open-source software tool for exploration and interactive visualization of diachronic change with respect to collocation behavior in large collections of (historical) text. In addition to the technical, implementation-oriented issues common to all software development projects on the one hand and the various challenges of source criticism characteristic for historical research on the other, interdisciplinary collaborations of this kind present challenges all their own, ranging from lack of established shared terminology (e.g. "term", "concept", "query", "type/token", "collocant/collocate", "relevance") to fundamentally different approaches to what constitutes "research activity" as such (analytic/stipulative vs. hermeneutic/interpretive). Over the course of the collaboration, DiaCollo underwent several iterations of the software development lifecycle phases of "planning", "implementation", and "evaluation" – in the latter case relying on extensive feedback from the working group's historians to identify missing functionality and potentially useful new features.

After its initial release, DiaCollo was integrated into the corpus administration framework of the CLARIN service center at the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW). At the time of writing (April, 2019), DiaCollo indices for 70 distinct curated text corpora comprising a total of over 15,000,000,000 (15G) source tokens have been indexed and deployed at the BBAW, where they enjoy a modicum of popularity with an average of about 500 queries per day over the past

12 months. Of these curated corpora, 18 indices are publicly accessible and 3 more can be queried using CLARIN credentials.[1]

2 Background

In linguistics, collocations are sets of words or terms that frequently occur in one another's vicinity, presumably because they belong to the same "semantic field" and thus shape their respective meanings, as suggested by J.R. Firth's well-known assertion that "you shall know a word by the company it keeps" and Wittgenstein's famous *"die Bedeutung eines Wortes ist sein Gebrauch in der Sprache"* ("the meaning of a word is its use in the language"). For example, the fact that the words "smoke" and "fire" tend to occur near one another in a text corpus suggests that there is indeed a semantic relation between them – in this case, a causal one.

Previous work in computational linguistics has established a number of methods for unsupervised discovery of collocations in text corpora, based on distributional properties of the collocated terms alone (see e.g. Evert, 2008). Informally, distributional collocation discovery procedures identify those word-pairs as potential collocations which occur together substantially more often than would be expected under "chance" conditions. Collocation profiling is a related technique which requires the user to provide one or more search terms of interest (the "collocant"), and searches for those terms in the corpus which associate most strongly with the collocant (i.e. the "collocates"). The association strength of a particular candidate collocate is estimated with regard to its own independent frequency in the corpus as well as that of the collocant, and should provide a quantitative approximation of the "relevance" of the respective collocate for the given collocant. To illustrate, a simple collocation profiling procedure would investigate all words in a pre-defined neighborhood of the search term, e.g. within a window of 5 words to the left and right. The more often one of these words occurs together with the collocant, compared to its frequency in the corpus overall, the stronger its association with the search term will be.

Synchronic collocation analysis has long been employed to provide evidence for typical usage(s) of words/concepts in the corpus as a whole, that is, semantics. It is also possible to compare collocation-profiles of different words, to look at differences and similarities in usage (e.g. for lexicography). "Ready-to-use" implementations include both the DWDS "*Wortprofil*" database[2] and Cyril Belica's co-occurrence database "CCDB"[3]. More complex user queries are possible (and familiarity with the associated software tools and interfaces required) when using the *Deutsches Referenzkorpus* (DeReKo) with the "COSMAS II" interface.

When analyzing historical text, synchronic collocation analysis can be a point of departure for comparing the usage of certain terms in historic source material with their use in the contemporary reference corpora. In order to be truly useful for historical research, collocation analysis should also provide methods that reveal changes in language use over time (in specific corpora), allowing users to trace phenomena such as semantic shifts, discourse trends, history of concepts, introduction of neologisms, etc. DiaCollo has been specifically developed for this purpose. As a free, open-source, language-agnostic software package[4], it can also be integrated into other project contexts and corpus infrastructures.

DiaCollo corpus data must be pre-tokenized and each document must be assigned a characteristic date (e.g. year of publication) to represent the diachronic axis. If provided by the corpus, DiaCollo can also make use of additional token-level attributes such as lemmata or part-of-speech tags as well as document-level metadata such as author or genre to enable finer-grained queries and aggregation of result profiles (Jurish, 2018). As with any other data-driven procedure, DiaCollo is subject to "garbage-in / garbage-out" phenomena: "messy" corpora containing abundant OCR or annotation errors, mistokenizations, and/or incorrect document metadata are less likely to produce satisfying results for humanities researchers than "tidy", well-curated corpora with accurate metadata and reliable linguistic annotations.

[1] https://kaskade.dwds.de/~jurish/diacollo/corpora/ (this and subsequent URLs last accessed 16 April, 2019)

[2] https://www.dwds.de/d/ressources#wortprofil

[3] http://corpora.ids-mannheim.de/ccdb/

[4] https://metacpan.org/release/DiaColloDB/

3 Use Case: Debates on Education in *Die Grenzboten*

As an introduction to DiaCollo’s functionality, we will consider the collocates of a simple search term *Schule* (“school”) in the largest historical corpus available at the BBAW, the *Deutsches Textarchiv*[5] (“German Text Archive”, DTA). Presentation of results in HTML format displays up to the specified number (kbest) of collocates (e.g. 10) for *Schule* discovered within the chosen time slice (e.g. a decade) in the form of a table. Each row of the table includes a color-code indicating the strength of the collocate’s association preference as well as links to (close approximations of) the underlying corpus evidence for the corresponding collocation pair as Keywords-in-Context (KWIC), allowing the user to focus her attention more closely on the original text source. Additional visualization modes such as the “bubble” and “cloud” formats display changes in the collocates on an interactive timeline. For the example query, the collocates give quite obvious evidence that the term *Schule* is associated with words within the semantic field of the institution of the church in the earliest documents queried (e.g. in the 1560s: *Kloster* (“cloister”), *Pfarrherr* (“pastor”), and *Kirche* (“church”)). The findings imply that the influence of this institution on the school system begins to mingle with worldy institutions in texts from the 1710s, where collocates include *Kirche* (“church”), as well as *Inspektor* (“inspector”), *preußisch* (“prussian”), and *Universität* (“university”); the term “church” disappears from the lists of top-10 collocates from the 1770s onwards (but re-occurs in the 1840s and 1890s).

We will now further demonstrate the use of DiaCollo by looking at German education policy as discussed in a historical periodical. *Die Grenzboten* was a German-language national-liberal magazine published from 1841 to 1922. Originally published as a (bi-)weekly periodical, the 311 volumes (roughly 180,000 pages) of *Die Grenzboten* were first digitized by the Staats- und Universiätsbibliothek Bremen[6] with funding from the German Research Association (DFG), and have been integrated into the BBAW CLARIN service center’s corpus infrastructure. *Die Grenzboten* covered a wide range of subjects in politics, literature, and the arts throughout the ‘long’ nineteenth century, and over the course of its publication was witness to several changes and attempted reforms of school systems in German-speaking territories. Using DiaCollo, we will now explore *Die Grenzboten*’s stance on education policy.

3.1 Is the corpus a source for research into the history of education?

A time series analysis of the absolute frequency of selected relevant terms such as *Schule* (“school”), *Schulgesetz* (“school law”), *Schulbuch* (“textbook”), and other terms denoting various types of German schools shows that the lemma “school” was indeed mentioned in every year of *Die Grenzboten*'s publication. Its raw frequency peaked at more than 500 tokens in 1890.[7] Its relative frequency in the *Die Grenzboten*-corpus is twice as high[8] as in the corresponding texts (1840–1920) from the aggregated DTA and *Digitales Wörterbuch der deutschen Sprache* (DWDS)[9] “core” corpus. A DiaCollo search[10] for collocates of *Schule* in ten-year epochs beginning at 1840 provides ample results from which to explore the school-related topics discussed in *Die Grenzboten*. Of the top-10 collocates per decade, most are nouns, some adjectives and one a finite verb (*gehören,* “to belong”).

3.2 Are all findings relevant? Disambiguation by targeted close reading

DiaCollo’s KWIC facility allows one to quickly check whether the results are applicable to a particular research question. In this case, strong adjective collocates of *Schule* are often associated with the sense of “school” as “doctrine”, e.g. an artistic school or school of thought, which is irrelevant when looking at education policy. Another adjective collocate of interest is the lemma *hoch* (“high”). In DiaCollo's ‘cloud’ visualization for this query, it becomes evident that the adjective already appeared among the ten best collocates per epoch after 1870. Examination of the corresponding KWIC hits reveals that these collocates refer almost exclusively to secondary (“higher”) schools, supporting the impression

[5] https://kaskade.dwds.de/~jurish/cac2019/Schule-dta
[6] University of Bremen: Grenzboten project, https://www.suub.uni-bremen.de/ueber-uns/projekte/grenzboten/
[7] https://kaskade.dwds.de/~jurish/cac2019/Schule-ts
[8] https://kaskade.dwds.de/~jurish/cac2019/hist-gb
[9] https://kaskade.dwds.de/~jurish/cac2019/hist-dta+dwds
[10] https://kaskade.dwds.de/~jurish/cac2019/Schule-gb

that *Die Grenzboten* was on the whole more concerned with higher education than with the *Volksschule* which provided basic primary (rural) education.

3.3 Do the findings offer tracks to specific discourses/debates?

Finally, the adjectives *konfessionell* (“denominational”) and *öffentlich* (“public”) were examined. These collocates appear among the top ten between 1860 and 1879, as do the nouns *Gemeinde* (“parish”/”congregation”) and *Kirche* (“church”) – the latter being as prominent and persistent as more expected noun collocates such as *Kind* (“child”) or *Lehrer* (“teacher”). Using DiaCollo’s on-the-fly filtering function to restrict our attention to adjective collocates only[11], the 1860s and 1870s documents reveal the adjectives *protestantisch* (“protestant”) and *evangelisch* (“evangelical”) as well as *katholisch* (“catholic”) as strong collocates of *Schule*.

We may assume that the prominence of this terminology involving religious denominations at that particular time was caused by the contemporary debates – since referred to as the *Kulturkampf* (“cultural struggle”) – concerning the rights and spheres of influence of state (Prussia) and church (Pope Pius IX) which started in some German territories in the 1860s and reached their peak in the 1870s. The debates involved the issue of who should be in charge of education and curricula, and how to deal with different religious denominations in schools. Loyal supporters of the Roman Catholic Church were referred to as *ultramontan* (“ultramontane”) during this period. A simple frequency query[12] shows that this kind of terminology is indeed present in the *Grenzboten* corpus, the former peaking and the latter beginning in the 1870s. Among the strong collocates of *Kulturkampf* and *ultramonan* are no terms that would hint at education though.

The connection only becomes clear if we turn our attention to all co-occurrences of *ultramontan* and GermaNet (Hamp & Feldweg, 1997) hyponyms of the synset *Bildungseinrichtung* (“educational institution”) or compounds matching a simple regular expression and using a rather broad paragraph-wide search window. Through closer reading of the corpus hits, we find evidence for anti-Catholic opinions in debates about education emanating from various sources[13].

4 Conclusion

DiaCollo serves as an effective automatic tool for the analysis of semantic change with respect to terms and concepts in diachronic perspective. Designed and optimized for the needs of humanities researchers, DiaCollo’s expressive query language and flexibility make it a powerful tool for corpus research. Thorough documentation, tutorials, and references to previous work as well as user-oriented dissemination in the form of workshops and lectures by the CLARIN-D working group “history” make it easy for the inexperienced to learn and provide a useful resource for more experienced users. Actively maintained and supported, DiaCollo continues to evolve and adapt in response to and in co-operation with its user community. Our use cases have shown the necessity of constant shifts between close and distant reading methods, which DiaCollo facilitates. Although ensuring interoperability between tools and maintaining the high standards of data curation necessary for reliable results requires considerable effort across all disciplines, we believe the prospective gain for the scientific community will justify the endeavor.

References

Evert, S. “Corpora and collocations.” In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, Berlin, Mouton de Gruyter, pp. 1212–1248, 2008.

Hamp, B., Feldweg, H. “GermaNet – a lexical-semantic net for German.” In Proceedings of the ACL workshop *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.

Jurish, B. “DiaCollo: On the trail of diachronic collocations.” In K. De Smedt (ed.), *CLARIN Annual Conference 2015* (Wrocław, Poland), 2015.

Jurish, B. “Diachronic Collocations, Genre, and DiaCollo.” In Whitt, R. J. (ed.), *Diachronic Corpora, Genre, and Language Change*. Amsterdam, John Benjamins, pp. 42–64, 2018.

[11] https://kaskade.dwds.de/~jurish/cac2019/Schule-gb-adj

[12] https://kaskade.dwds.de/~jurish/cac2019/ultramontan-freq

[13] https://kaskade.dwds.de/~jurish/cac2019/ultramontan-germanet