

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 bi-weekly German-language newspaper “Die Grenzboten” (1841-1922) (“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.¹ 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, 2018) 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 (March, 2020), DiaCollo indices for 78 distinct curated text corpora comprising a total of over 23,000,000,000 (23G) source tokens have been indexed and deployed at the BBAW,

¹<https://www.clarin-d.net/en/disciplines>

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, 3 require registration with the www.dwds.de web platform, and 3 more can be queried using CLARIN credentials.² The BBAW corpus infrastructure is strongly focused on German-language sources – both historical and contemporary – but also includes DiaCollo indices for several English and French corpora, developed in cooperation with the respective corpus providers. DiaCollo itself is language-agnostic and can be applied to any source corpus, provided that it has been appropriately pre-processed. Documentation and demos are available in German and English.

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 (1957) well-known assertion that "you shall know a word by the company it keeps" and Wittgenstein's (1953) 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, i.e. (distributional) 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*" database³ and Cyril Belica's co-occurrence database "CCDB"⁴. 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 part of departure for comparing the usage of certain terms in historic source material with their use in the contemporary reference corpora. Historical corpora (if existent and accessible in sufficient quality and quantity for the culture/time period of interest) can provide empirical data regarding the use of specific terms during a certain epoch. By comparing historical and synchronic reference corpora, researchers can identify linguistic divergences, which in turn may lead to hypotheses (and perhaps even conclusions) about an author's or issuing institution's specific intentions as realized by their "linguistic framing" of the phenomena under discussion.

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 package⁵, it can also be integrated into other project contexts and corpus infrastructures.

²<http://kaskade.dwds.de/~jurish/diacollo/corpora/>

³<http://www.dwds.de/d/ressourcen#wortprofil>

⁴<http://corpora.ids-mannheim.de/ccdb/>

⁵<http://metacpan.org/release/DiaColloDB/>

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 (Nieländer & Weiß, 2018).

3 Output, Visualization, and Usability

The usefulness of computer-aided text analysis for historical research and the acceptance of such methods in the discipline are influenced by a number of factors. For this reason, historians working in the CLARIN-D working group "Contemporary History" (2014-2016) and "History" (2016-present) were involved in the development of DiaCollo from the very beginning. Historians traditionally had to investigate any diachronic and synchronic differences in an individual's or a group's language use by personally “learning” these different usages, as one would learn different languages, and then comparing them. This method – as large parts of historic methodology in general – focussed on finding meaningful anomalies (as opposed to overall patterns), and required careful and close reading of original source texts or true-to-text reproductions.

The goal was to design and implement the software tool in such a way that “distant reading” – a bird’s-eye view of large digital text corpora – would be possible. At the same time, the algorithm’s functions should be made transparent and its parameters should remain flexible enough to accommodate the requirements of the users’ respective research interests. The result displays should be linked to the sources in order to enable differentiated interpretation with the help of direct source study.

3.1 Output

The output of DiaCollo should be as complete, correct, and reproducible as possible. These requirements derive from a projection of quality standards of analog historical research onto computer-aided research on the one hand, and from generic criteria for software development on the other. Completeness of output is favored by historians because their source materials are usually sparse, fragmentary, and biased. Every detail is important in order to keep the gaps (which have to be reconstructed) as small as possible, so as to be able to draw as well-founded a conclusion as possible from the particular evidence to the bigger picture. So ideally, digital corpora should encompass all sources relevant to a given subject, and an analysis tool should find every instance of a sought-after linguistic phenomenon in the corpus, and include that instance in subsequent calculations. In contrast to techniques like topic modeling, collocation analysis can yield complete, correct and reproducible results in this sense, provided that the corpora are of high data quality and are have undergone appropriate and sufficiently accurate NLP preprocessing (e.g. tokenization, lemmatization, etc.).

Paradoxically, this preprocessing both guarantees and compromises the correctness-criterium. Individual- or time-specific peculiarities such as orthographic features, foreign-language insertions, illegible typescript or handwriting, the choice of words, a gap, an addition, a typographical or spelling error – can be decisive in assessing the authenticity and significance of a source, and might indeed be precisely the phenomenon sought after in a specific research design. Consequently, this type of information should not be lost when working with digital tools. But in order to support search functions in the text, the peculiarities and “quirks” must be leveled and the text normalized, which is diametrically opposed to the requirements of source criticism. The *Basisformat*⁶ (“base format”) of the *Deutsches Textarchiv* (“German Text Archive”, DTA) provides a set of effective and field-tested guidelines for corpus annotation recommended by the *Deutsche Forschungsgemeinschaft* (“German Research Foundation”, DFG). Digital source criticism must encompass an understanding of these labor-intensive data curation efforts, both of their limits and of the opportunities they provide by enabling complex queries with search engines such as DDC.⁷

⁶<http://www.deutschestextarchiv.de/doku/basisformat/>

⁷<http://www.deutschestextarchiv.de/doku/software#ddc> , <http://kaskade.dwds.de/dstar/dta/diacollo/help.perl#queries>

N: Total number of words in selected time slice
f1: Total frequency of the query term ("Macht") in selected slice
f2: Total frequency of collocate ("Kraft") in time slice
label: Time slice label ("1610s")
pos: Part of speech ("NN" noun)
KWIC: Link to Keyword in context
score: Strength of collocation (score/color coded)
f12: Frequency of the collocation

Collocates for "Macht" ("power") in the 1610s and 1620s, DTA corpus

N	f1	f2	f12	score	label	lemma	pos	
6862864	1535	3958	25	7.2205	1610	Gewalt	NN	KWIC
6862864	1535	5148	10	4.2958	1610	König	NN	KWIC
6862864	1535	8486	5	4.0312	1610	Kraft	NN	KWIC
6862864	1535	14774	5	3.285	1610	Land	NN	KWIC
6862864	1535	27484	8	3.1753	1610	Tag	NN	KWIC
6862864	1535	21166	5	2.8515	1610	Sünde	NN	KWIC
6862864	1535	34046	5	2.1790	1610	groß	ADJ	KWIC
6862864	1535	3028	5	2.1182	1610	Leben	NN	KWIC
6862864	1535	154341	17	1.8374	1610	Gott	NN	KWIC
6862864	1535	56867	5	1.4882	1610	Mensch	NN	KWIC
3564750	150	150	6	9.3561	1620	Macht	NN	KWIC
3564750	150	659	6	7.9250	1620	Kraft	NN	KWIC
3564750	150	8900	7	4.6637	1620	Gott	NN	KWIC

Figure 1: Annotated screenshot of DiaCollo's tabular HTML display format

3.2 Visualization

In addition to the citation forms of the k -best collocates discovered per epoch, DiaCollo's default tabular HTML output (Figure 1) includes a number of additional data columns for each collocate row. This supplementary data represents the empirical basis by means of which DiaCollo computes the association strength between the user's search term(s) and the collocate item in question. In particular, each collocate row contains the minimum date ("label") for the current corpus epoch, the total size of the epoch ("N"), the total frequencies of both the user's search term ("f1") and the current collocate item ("f2") in the current epoch, and the final association score itself ("score") as computed by the selected scoring function.

DiaCollo also offers several interactive visualization formats for diachronic collocation data, including animated bubble charts and tag-clouds, in which each collocate item's association score is mapped directly to visually salient properties (size and color) of the corresponding display element, and the "raw" empirical corpus frequencies are not displayed by default. In general, such visualizations can facilitate comprehension by reducing the cognitive workload involved in interpreting search results, but may fail to adequately capture all relevant aspects of the underlying data, leading to oversimplification and the danger of "jumping to conclusions" (Jurish, 2016b). DiaCollo's visualization formats were developed in close collaboration with and in response to the needs of the contributing humanities researchers. Color-coding and differences in font-size allow "intuitive" comparisons of large datasets at a single glance. In the interactive visualizations, detailed information (f1, f2, f12, etc.) about individual collocation pairs can be displayed in a popup window by clicking on a collocate item of interest. At the same time, some of the pitfalls of distant reading (e.g. extractions of single words "solely" on the basis of distributional properties alone) are ameliorated by establishing a direct connection to the underlying text sources. A KWIC ("Keywords in Context") view of corpus hits for a given collocation pair as well as links to the original texts are offered (provided the requisite preprocessing has been done). DiaCollo thus attempts to allow smooth transitions from distant to close reading in all supported visualization formats.

3.3 Usability

The statistics and algorithms underlying tools such as DiaCollo, as well as corpus query languages, are only starting to be included in the academic curriculum of historians. They are neither trivial nor easy to grasp without this kind of specialized training. DiaCollo allows for free choice of up to 15 different parameters influencing the final result and its method of presentation. The flexibility offered by this large selection of choices is what makes it useful for historians – but not necessarily easy to use. Lifting the obligation to explicitly specify all parameter values by providing sensible defaults and suppressing "raw" frequency data (f1, f2 etc.) from the output presentation might increase users' initial comfort with the tool, but hinder understanding and interpretation of the results obtained. Thorough documentation, examples, and use-cases are needed for users with different levels of experience. DiaCollo comes with documentation, use cases and tutorials in German and English that aid digital source

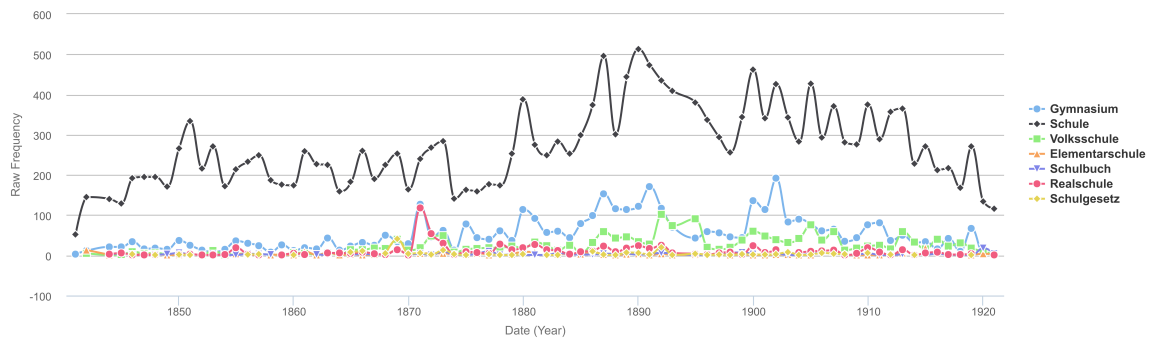


Figure 2: *Grenzboten* corpus frequencies for school-related terms

criticism⁸. Examples range from simple single-word queries to more elaborate constructions using regular expressions, thesaurus expansion, Boolean operators, and document metadata filters. Even so, there is further demand for finer-grained tutorials, for introductory workshops as provided by members of the CLARIN-D working group “History”⁹, and for one-to-one feedback and support as provided by the experts at the BBAW.

4 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*¹⁰ (“German Text Archive”, DTA). Presentation of results in HTML format displays up to the specified number (kbest) of collocates (by default 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 “school” 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 worldly 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 further demonstrate the use of DiaCollo by looking at German education policy as discussed in a historical periodical. This is a typical use case for historical research: determining to what extent this particular corpus is relevant for specific research questions, e.g. for historical textbook research as conducted by researchers at institutions like the Georg-Eckert-Institute (GEI). *Die Grenzboten* was a German-language national-liberal magazine published from 1841 to 1922, covering a wide range of subjects in politics, literature, and the arts throughout the ‘long’ nineteenth century (Werner, 1922). Its “messengers from the borders” did not limit their reports to German-speaking territories, but also addressed issues in other European countries and the rest of the world. Its coverage of civic life, opinions, and debates before and after the revolution of 1848, the restoration period, industrialization, the German Empire (*Kaiserreich*), and the First World War makes this periodical a valuable source for a broad range of disciplines.

Originally published as a (bi-)weekly periodical, the 311 volumes (roughly 180,000 pages) of *Die Grenzboten* were first digitized by the Staats- und Universitätsbibliothek Bremen¹¹ with funding from

⁸Tutorial in German: <http://kaskade.dwds.de/diacollo-tutorial>; DiaCollo help in English:

<http://kaskade.dwds.de/dstar/dta/diacollo/help.perl>; Short guide in English and German: <https://www.clarin-d.net/en/diacollo-en>, <https://www.clarin-d.net/de/kollokationsanalyse-in-diachroner-perspektive>

⁹<https://www.clarin-d.net/en/disciplines/history>

¹⁰<http://kaskade.dwds.de/~jurish/cac2019/Schule-dta>

¹¹University of Bremen: Grenzboten project, <http://www.suub.uni-bremen.de/ueber-uns/projekte/grenzboten/>

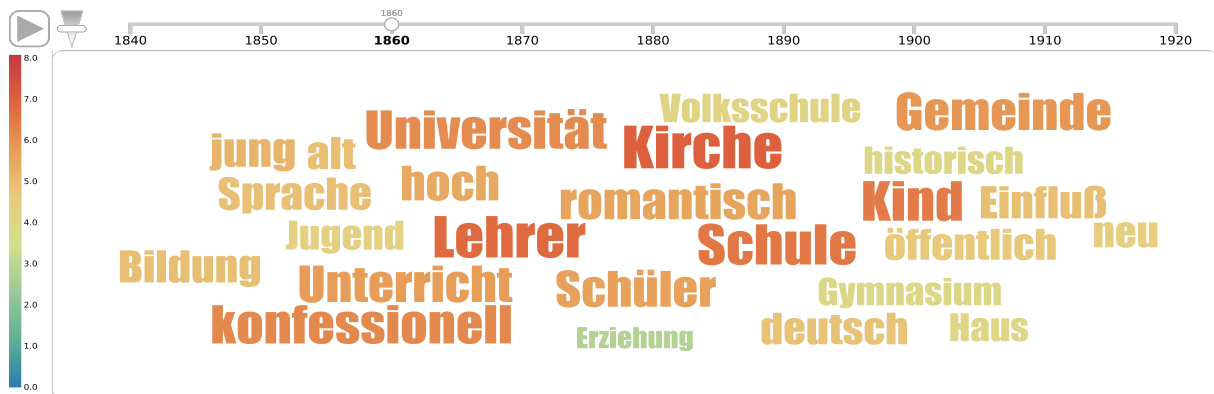


Figure 3: DiaCollo ‘cloud’ visualization of “Schule” collocates in the *Grenzboten* corpus (1860-1869)

the German Research Association (DFG), and have been integrated into the BBAW CLARIN service center’s corpus infrastructure. Over the course of its publication, *Die Grenzboten* was witness to several changes and attempted reforms of school systems in German-speaking territories. Its political agenda changed over time as well: the publication’s original focus on Austria shifted to Prussia when its founder Ignaz Kuranda (1811-1884) left the editorial team in 1848. After 1870, *Die Grenzboten* increasingly became a voice for conservative views. But throughout its lifetime, the magazine self-identified as part of the liberal opposition which favored German unity under Prussian leadership, the so-called “small German solution” (von Wurzbach-Tannenberg, 1865; Werner, 1922). Using DiaCollo, we will now explore *Die Grenzboten*’s stance on education policy.

4.1 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 publication¹² (Figure 2). Its raw frequency peaked at more than 500 tokens in 1890, and its relative frequency in the *Die Grenzboten* corpus is twice as high¹³ as in the corresponding texts (1840–1920) from the aggregated DTA and *Digitales Wörterbuch der deutschen Sprache* (DWDS)¹⁴ “core” corpus. A DiaCollo search¹⁵ 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”).

4.2 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 interactive ‘cloud’ visualization for this query¹⁶, it becomes evident that the adjective already appeared among the ten best collocates per epoch after 1870, and was strongly associated with *Schule* throughout the entire corpus (Figure 3). Examination of the corresponding KWIC hits reveals that these collocates refer almost exclusively to secondary (“higher”) schools, both historic and contemporary, in German-speaking countries and elsewhere. Quite often, the encompassing articles deal with access granted or denied to higher education. The query results lack an antonym – this may be a result of it being distributed over a larger field of words (such as “basic”, “elementary”, or “primary”), but also supports the impression that *Die Grenzboten* was on the whole more concerned with higher education than with the *Volksschule* which provided basic primary (rural) education.

¹²<http://kaskade.dwds.de/~jurish/cac2019/Schule-ts>

¹³<http://kaskade.dwds.de/~jurish/cac2019/hist-gb>

¹⁴<http://kaskade.dwds.de/~jurish/cac2019/hist-dta+dwds>

¹⁵<http://kaskade.dwds.de/~jurish/cac2019/Schule-gb>

¹⁶<http://kaskade.dwds.de/~jurish/cac2019/Schule-cloud>

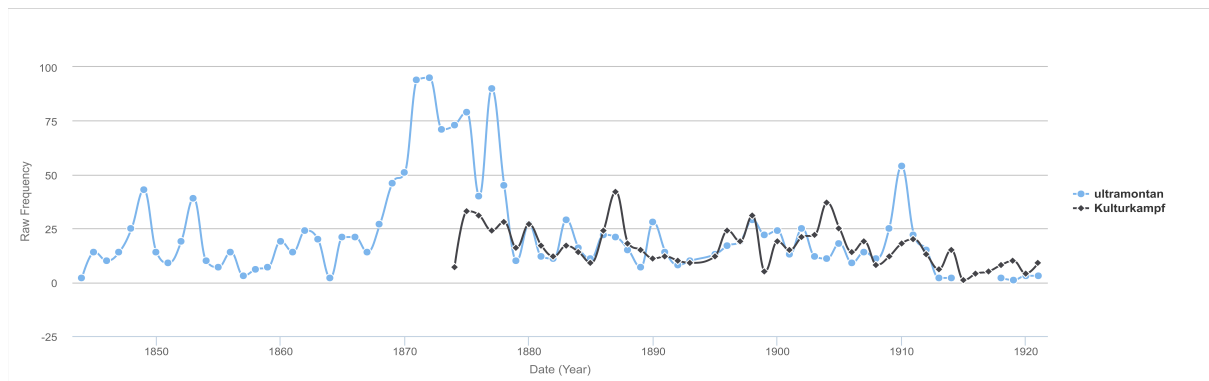


Figure 4: *Grenzboten* corpus frequencies for "ultramontan" and "Kulturkampf"

4.3 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 only¹⁷, 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 query¹⁸ (Figure 4) 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*¹⁹ and *ultramontan*²⁰ are no terms that would hint at education, though. A manual check of the sources can be time consuming; in this case they do indeed yield results relevant for debates on education. Even in cases where the interpretation seems straightforward – as in *Kulturkampf ... entbrannt* ("cultural struggle ... erupted") – a look at the sources is not superfluous. In this case, a parliamentary debate is commented upon. According to the author, the cultural struggle erupted "in the master's mansion" (*im Herrenhaus*) too, after a heated parliamentary discussion of several ultramontan petitions concerning religious education in different school types and a discussion of the slogan "*Trennung von Kirche und Schule*" ("separation of church and school"; *Die Grenzboten*, "Vom Preußischen Landtage", p. 237f).

The connection of the cultural struggle with debates on education only becomes clear at the level of collocates if we turn our attention to all co-occurrences of *ultramontan* and GermaNet (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010) 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.²¹ In an article celebrating 30 years of the Gustav-Adolf Verein (today 'Gustav-Adolf-Werk') – which aided protestants living away from larger congregations (in diaspora) – we find for example the following "activity report":

"Tausende von Kindern evangelischer Eltern wurden durch ihn [den Verein] der katholischen Schule, in die sie nothgedrungen gehen mußten, entnommen und so vor den Nachstellungen der ultramontanen Propaganda bewahrt." [*Die Grenzboten*, "Der Gustav Adolf Verein", p. 503f]

¹⁷<http://kaskade.dwds.de/~jurish/cac2019/Schule-gb-adj>

¹⁸<http://kaskade.dwds.de/~jurish/cac2019/ultramontan-freq>

¹⁹<http://kaskade.dwds.de/~jurish/cac2019/Kulturkampf-collocates>

²⁰<http://kaskade.dwds.de/~jurish/cac2019/ultramontan-collocates>

²¹<http://kaskade.dwds.de/~jurish/cac2019/ultramontan-germanet>

(“Thousands of protestant parents' children were taken [by the Gustav-Adolf-Society] from the Catholic schools they had been forced to attend, and thus spared from the harassments of ultramontane propaganda.”)

So even if this important part of the debates on education policy was not immediately apparent in the results of our initial DiaCollo queries, the subsequent indications combined with informed curiosity and further investigation (by means of focused queries and close reading of sources) produces more satisfying results.

5 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 useful aid for corpus exploration and 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 easier for the inexperienced to learn and provide a useful resource for more experienced users. Actively maintained and supported as part of the ongoing development cycle, 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

- Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin.
- John Rupert Firth. 1957. *Papers in Linguistics 1934–1951*. Oxford University Press, London.
1862. Der Gustav-Adolf-Verein. *Die Grenzboten*, 21:502–515. <http://brema.suub.uni-bremen.de/grenzboten/periodical/titleinfo/114291>
- Birgit Hamp and Helmut Feldweg. 1997. 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*, Madrid, Spain.
- Verena Henrich and Erhard Hinrichs. 2010. GernEdiT – the GermaNet editing tool. In *Proceedings LREC 2010*, pages 2228–2235, Valletta, Malta.
- Bryan Jurish, Alexander Geyken, and Thomas Werneke. 2016. DiaCollo: diachronen Kollokationen auf der Spur. In *Proceedings DHd 2016: Modellierung – Vernetzung – Visualisierung*, pages 172–175.
- Bryan Jurish. 2015. DiaCollo: On the trail of diachronic collocations. In Koenraad De Smedt, editor, *Proceedings CLARIN Annual Conference 2015*, pages 28–31, Wrocław, Poland.
- Bryan Jurish. 2018. Diachronic collocations, genre, and DiaCollo. In Richard Jason Whitt, editor, *Diachronic Corpora, Genre, and Language Change*, pages 42–64.
- Maret Nieländer and Andreas Weiß. 2018. »Schönere Daten« – Nachnutzung und Aufbereitung für die Verwendung in Digital-Humanities-Projekten. In Maret Nieländer and Ernesto William De Luca, editors, *Digital Humanities in der internationalen Schulbuchforschung*, pages 91–116. V&R unipress, Göttingen.
1878. Vom preußischen Landtage – Berlin, 27. Januar. *Die Grenzboten*, 37:234–239. <http://brema.suub.uni-bremen.de/grenzboten/periodical/pageview/139535>
- Constantin von Wurzbach-Tannenberg. 1865. Kuranda, Ignaz. In *Biographisches Lexikon des Kaiserthums Oesterreich*, volume 13, pages 407–416. Staatsdruckerei, Vienna.
- Fritz Werner. 1922. Die Grenzboten: aus der Geschichte einer achtzigjährigen Zeitschrift nationaler Bedeutung. *Die Grenzboten*, 81:448–452. <http://brema.suub.uni-bremen.de/grenzboten/periodical/titleinfo/178709>