DiaCollo: On the trail of diachronic collocations

Bryan Jurish
Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstrasse 22-23 · 10117 Berlin · Germany
jurish@bbaw.de

Abstract

This paper presents DiaCollo, a software tool developed in the context of CLARIN for the efficient extraction, comparison, and interactive visualization of collocations from a diachronic text corpus. Unlike other conventional collocation extractors, DiaCollo is suitable for extraction and analysis of diachronic collocation data, i.e. collocate pairs whose association strength depends on the date of their occurrence. By tracking changes in a word's typical collocates over time, DiaCollo can help to provide a clearer picture of diachronic changes in the word's usage, in particular those related to semantic shift. Beyond the domain of linguistics, DiaCollo profiles can be used to provide humanities researchers with an overview of the discourse topics commonly associated with a particular query term and their variation over time or corpus subset, while comparison or "diff" profiles highlight the most prominent differences between two independent target queries. In addition to traditional static tabular display formats, a web-service plugin also offers a number of intuitive interactive online visualizations for diachronic profile data for non-technical users.

1 Introduction

In recent years, an increasing number of large diachronic text corpora have become available for linguistic and humanities research, including the *Deutsches Textarchiv*[1] (Geyken et al., 2011) and the Corpus of Historical American English[2] (Davies, 2012). While the broad time spans represented by these corpora offer unique research opportunities, they also present numerous challenges for conventional natural language processing techniques, which in turn often rely on implicit assumptions of corpus homogeneity – in particular with respect to the temporal axis. Indeed, even putatively synchronic newspaper corpora have a nontrivial temporal extension and can reveal date-dependent phenomena if queried appropriately (Scharloth et al., 2013). This paper addresses the problem of automatic *collocation profiling* (Church and Hanks, 1990; Evert, 2005) in such diachronic corpora by introducing a new software tool "DiaCollo" explicitly designed for this purpose which allows users to choose the granularity of the diachronic axis on a per-query basis.

DiaCollo is a modular software package for the efficient extraction, comparison, and interactive visualization of collocations from a diachronic text corpus. Unlike conventional collocation extractors such as DWDS Wortprofil[3] (Didakowski and Geyken, 2013), Sketch Engine[4] (Kilgarriff and Tugwell, 2002; Rychlý, 2008), or the UCS toolkit[5], DiaCollo is suitable for extraction and analysis of diachronic collocation data, i.e. collocate pairs whose association strength depends on the date of their occurrence and/or other document-level properties. By tracking changes in a word's typical collocates over time and applying J. R. Firth's famous principle that "you shall know a word by the company it keeps" (Firth, 1957), DiaCollo can help to provide a clearer picture of diachronic changes in the word's usage.

[1] http://www.deutschestextarchiv.de

[2] http://corpus.byu.edu/coha

[3] http://zwei.dwds.de/wp

[4] http://www.sketchengine.co.uk

[5] http://www.collocations.de/software.html

DiaCollo was developed in the context of CLARIN in order to aid historians participating in the CLARIN Working Groups in their analysis of the changes in discourse topics associated with selected terms as manifested by changes in those terms' context distributions, and has been successfully applied to both mid-sized and large corpus archives, including the *Deutsches Textarchiv* (1600–1900, ca. 2.6K documents, 173M tokens) and a large heterogeneous newspaper corpus (1946–2015, ca. 10M documents, 4G tokens).

2 Implementation

DiaCollo is implemented as a Perl library, including efficient re-usable classes for dealing with native index structures such as $(string \leftrightarrow integer)$ mappings, n-tuple inventories and component-wise equivalence classes, or hierarchical tuple-pair frequency databases. DiaCollo indices are suitable for use in a high-load environment, since no persistent server process is required and all runtime access to native index data structures occurs either via direct file I/O or (optionally) via the `mmap()` system call for efficient kernel-managed page caching, relying on the underlying filesystem cache to optimize access speed.

In addition to the programmatic API provided by the Perl modules, DiaCollo provides both a command-line interface as well as a modular plugin for the D* corpus administration framework which includes a publicly accessible RESTful web service (Fielding, 2000) and a form-based user interface for evaluation of runtime database queries and interactive visualization of query results. The remainder of this paper describes the DiaCollo web service, whose architecture closely mirrors the independently documented command-line and Perl APIs. A publicly accessible web front-end for the *Deutsches Textarchiv* corpus can be found at `http://kaskade.dwds.de/dstar/dta/diacollo/`, and the source code is available via CPAN at `http://metacpan.org/release/DiaCollоDB`.

Requests & Parameters DiaCollo is a request-oriented service: it accepts a user request as a set of *parameter=value* pairs and returns a corresponding *profile* for the term(s) queried. Parameters are passed to the service RESTfully via the URL query string or HTTP POST request as for a standard web form. Each request must contains at least a *query* parameter specifying the target term(s) to be profiled. The date-range to be profiled can be specified with the *date* parameter, while the *slice* parameter can be used to alter the granularity of the returned profile data by specifying the size in years of a single profile epoch. Candidate collocates can be filtered by means of the *groupby* parameter, and result-set pruning is controlled by the *score*, *kbest*, and *global* parameters.

Profiles & Diffs The results of a simple DiaCollo user request are returned as a tabular *profile* of the k-best collocates for the queried word(s) or phrase(s) in each of the requested date sub-intervals ("epochs" or "slices", e.g. decades) specified by the `date` and `slice` parameters. Alternatively, the user may request a comparison or "diff" profile in order to highlight the most prominent differences between two independent queries, e.g. between two different words or between occurrences of the same word in different date intervals, corpus subsets, or lexical environments.

Indices, Attributes & Aggregation For maximum efficiency, DiaCollo uses an internal "native" index structure over the input corpus content words to compute collocation profiles. Each indexed word is treated as an n-tuple of linguistically salient token- and/or document-attributes selected at compile-time, in addition to the document date. User `query` and `groupby` request parameters are interpreted as logical conjunctions of restrictions over these attributes, selecting the precise token tuple(s) to be profiled. For finer-grained selection of profiling targets, DiaCollo supports the full range of the DDC[6] query language (Sokirko, 2003; Jurish et al., 2014) via the `ddc` and `diff-ddc` profile types whenever the DiaCollo instance is associated with an underlying DDC server back-end.

Scoring & Pruning DiaCollo assigns each collocate w_2 in a unary profile for a target term w_1 a real-valued score by means of a user-specified *score function*. Supported score functions include absolute raw- and log-frequency (*f, lf*), normalized raw- and log-frequency per million tokens (*fm, lfm*), pointwise

[6] `http://www.ddc-concordance.org`

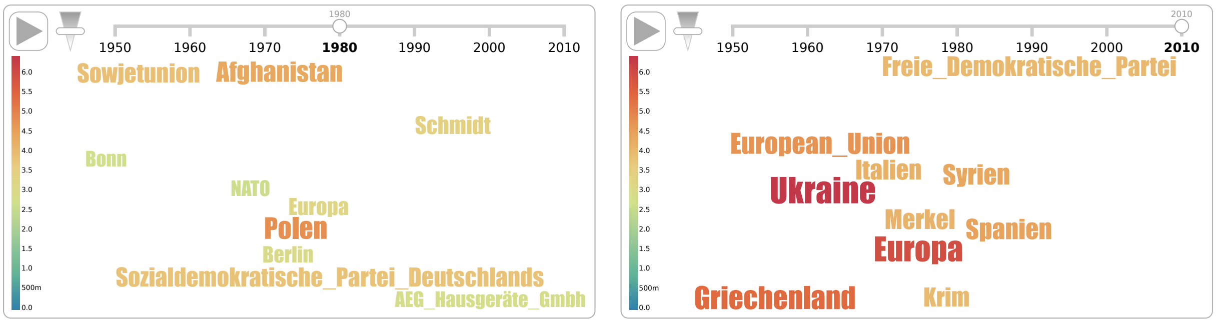

Figure 1: DiaCollo interactive tag-cloud visualization of the ten best proper name collocates of the noun *Krise* ("crisis") in the German weekly newspaper *DIE ZEIT* for the epochs 1980–1989 (left) and 2010–2014 (right).

mutual information × log-frequency product (*mi*), and the scaled log-Dice coefficient (*ld*) as proposed by Rychlý (2008). Candidate collocates are ranked in descending order by the associated scores, and the k-best candidates in each epoch are selected and returned. "Diff" queries compute independent profiles p_{a} and p_{b} for the *query* and *bquery* parameters, respectively. After ranking according to the selected score function, a comparison-profile $p_{\mathrm{a-b}}$ is computed as $p_{\mathrm{a-b}} : w_2 \mapsto p_{\mathrm{a}}(w_2) - p_{\mathrm{b}}(w_2)$ for each of the up to $2k$ collocates $w_2 \in k\text{-best}(p_{\mathrm{a}}) \cup k\text{-best}(p_{\mathrm{b}})$ and the k-best of these with the greatest absolute differences $|p_{\mathrm{a-b}}(w_2)|$ are selected and returned.

Output & Visualization DiaCollo supports a number of different output formats for returned profile data, including TAB-separated plain text suitable spreadsheet import, native JSON for further automated processing, and a simple tabular HTML format. In addition to the static tabular formats, the D* web-service plugin also offers several interactive online visualizations for diachronic profile data, including two-dimensional time series plots using the Highcharts JavaScript library, flash-based motion charts using the Google Motion Chart library, and interactive tag-cloud and bubble-chart visualizations using the D3.js library. The HTML and interactive D3-based display formats provide an intuitive color-coded representation of the association score (rsp. score-difference for "diff" profiles) associated with each collocation pair, as well as hyperlinks to underlying corpus hits ("KWIC-links") for each data point displayed.

3 Example

Figure 1 contains example tag-cloud visualizations for a unary DiaCollo profile of proper name collocates for the noun *Krise* ("crisis") in 10-year epochs over an archive of the German weekly newspaper *DIE ZEIT* spanning the interval 1950–2014. Since the term "crisis" usually refers to a short-lived and inherently unstable situation, its typical collocates can be expected to vary widely over time, reflecting changes in the discourse environment which in the case of a newspaper corpus can themselves be assumed to refer to events in the world at large. Indeed, the data in Figure 1 can easily be traced to prominent political events of the associated epochs. Conspicuous *Krise*-collocates and associated events in the 1980s include *Afghanistan* and *Sowjetunion* ("Soviet Union") for the Soviet-Afghan war (1979–1989); *Polen* ("Poland") due to the *Solidarność* movement and declaration of martial law in 1981; *Schmidt* and *Sozialdemokratische Partei Deutschlands* (SPD) referring to the collapse of the SPD-led German government coalition under Helmut Schmidt in 1982; and *NATO*, *Bonn*, and *Berlin* in the context of NATO's deployment of mid-range missiles in western Europe. The foreshortened final epoch (2010–2014) can be traced to civil wars in the Ukraine and Syria (*Syrien*), the Greek government-debt crisis and its repercussions in the European Union (*Griechenland, Italien, Spanien*), the Russian annexation of Crimea (*Krim*), and the German FDP (*Freie Demokratische Partei*) party's loss in the 2013 federal elections.

4 Summary & Outlook

This paper introduced DiaCollo, a new software tool for the efficient extraction, comparison, and interactive online visualization of collocations specially tailored to the unique demands of diachronic text corpora. In its top-level incarnation as a modular web service plugin for the D* corpus administration framework, DiaCollo provides a simple and intuitive interface for assisting linguists, lexicographers, and humanities researchers to acquire a clearer picture of diachronic variation in a word's usage over time or corpus subset. Future work will focus on implementing new visualization techniques for DiaCollo profile data, as well as extending the profiling back-end to handle other corpus input formats or online search APIs such as CLARIN Federated Content Search.

References

[Church and Hanks1990] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

[Davies2012] Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.

[Didakowski and Geyken2013] Jörg Didakowski and Alexander Geyken. 2013. From DWDS corpora to a German word profile – methodological problems and solutions. In Andrea Abel and Lothar Lemnitzer, editors, *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information*, (OPAL X/2012). IDS, Mannheim.

[Evert2005] Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

[Fielding2000] Roy Thomas Fielding. 2000. *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, University of California, Irvine.

[Firth1957] John Rupert Firth. 1957. *Papers in Linguistics 1934–1951*. Oxford University Press, London.

[Geyken et al.2011] Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. Das deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In Silke Schomburg, Claus Leggewie, Henning Lobin, and Cornelius Puschmann, editors, *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, pages 157–161.

[Jurish et al.2014] Bryan Jurish, Christian Thomas, and Frank Wiegand. 2014. Querying the deutsches Textarchiv. In Udo Kruschwitz, Frank Hopfgartner, and Cathal Gurrin, editors, *Proceedings of the Workshop "Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities" (MindTheGap 2014)*, pages 25–30, Berlin, Germany, 4th March.

[Kilgarriff and Tugwell2002] Adam Kilgarriff and David Tugwell. 2002. Sketching words. In Marie-Hélène Corréard, editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, EURALEX, pages 125–137.

[Rychlý2008] Pavel Rychlý. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9.

[Scharloth et al.2013] Joachim Scharloth, David Eugster, and Noah Bubenhofer. 2013. Das Wuchern der Rhizome. linguistische Diskursanalyse und Data-driven Turn. In Dietrich Busse and Wolfgang Teubert, editors, *Linguistische Diskursanalyse. Neue Perspektiven*, pages 345–380. VS Verlag, Wiesbaden.

[Sokirko2003] Alexey Sokirko. 2003. A technical overview of DWDS/Dialing Concordance. Talk delivered at the meeting *Computational linguistics and intellectual technologies*, Protvino, Russia.