Diachronic Collocations and Genre

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

Abstract

This paper outlines some potential applications of the open-source software tool "DiaCollo" to multi-genre diachronic corpora. Explicitly developed for the efficient extraction, comparison, and interactive visualization of collocations from a diachronic text corpus, DiaCollo – unlike conventional collocation extractors – is suitable for processing collocation 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, especially those related to semantic shift or discourse environment. Use of the flexible DDC search engine back-end allows user queries to make explicit reference to genre and other document-level metadata, thus allowing e.g. independent genre-local profiles or cross-genre comparisons. 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 immediate inspection.

1 Introduction

DiaCollo is a software tool for automatic *collocation profiling* (Church and Hanks, 1990; Evert, 2005) in diachronic corpora such as the *Deutsches Textarchiv*[1] (Geyken et al., 2011) or the Corpus of Historical American English[2] (Davies, 2012) which allows users to choose the granularity of the diachronic axis on a per-query basis (Jurish, 2015). Unlike conventional collocation extractors such as DWDS Wortprofil[3] (Didakowski and Geyken, 2013) or Sketch Engine[4] (Kilgarriff and Tugwell, 2002), 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 such as author or genre.

2 Implementation

DiaCollo is implemented as a Perl library, and provides both a command-line interface as well as a modular RESTful web service plugin (Fielding, 2000) with a form-based user interface for evaluation of runtime database queries and interactive visualization of query results. For finer-grained selection of profiling targets, DiaCollo supports the full range of the DDC[5] query

[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.ddc-concordance.org

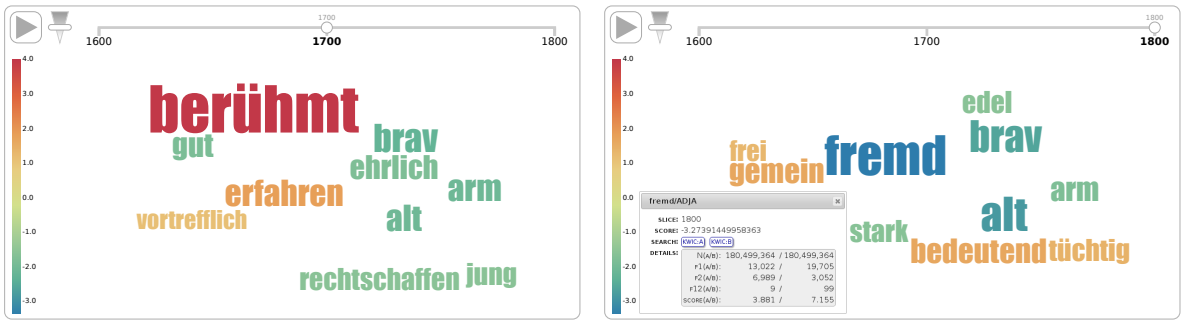

Figure 1: DiaCollo interactive tag-cloud visualization of the $k = 10$ most strongly divergent adjectives immediately preceding the noun *Mann* ("man") in the genres "science" (warm colors) and "*belles lettres*" (cool colors) over the *Deutsches Textarchiv* corpus for the epochs 1700–1799 (left) and 1800–1899 (right).

language (Sokirko, 2003; Jurish et al., 2014) whenever the DiaCollo instance is associated with an underlying DDC server back-end. In particular, use of the DDC back-end allows explicit reference to all document-level metadata encoded in the corpus, including e.g. text genre for the *Deutsches Textarchiv* DiaCollo index accessible at `http://kaskade.dwds.de/dstar/dta/diacollo/`.[6]

3 Example

Figure 1 contains example tag-cloud visualizations for a cross-genre comparison over the *Deutsches Textarchiv* corpus. The DDC back-end was used to acquire raw frequency counts over 100-year epochs for all adjectives in immediately preceding the noun *Mann* ("man") in the genres *Wissenschaft* ("science") and *Belletristik* ("belles lettres"), respectively. After computing association scores (Evert, 2008; Rychlý, 2008) for each such candidate collocate, the DiaCollo engine extracts and returns the k collocates in each epoch with the greatest absolute score differences. In the tag-cloud visualization mode, absolute score differences are mapped to tag font-size, and the signs of the score differences are mapped to an intuitive color-scale, with warm tones indicating a relative preference for the "science" genre and cool tones indicating a preference for "belles lettres". As Figure 1 shows, men in scientific texts are more likely to be described as *berühmt* ("famous"), *erfahren* ("experienced"), *bedeutend* ("significant"), or *tüchtig* ("capable"); while men in belles lettres are more likely to be designated *brav* ("well-behaved"), *rechtschaffen* ("righteous"), *arm* ("poor"), *alt* ("old") – assumedly reflecting the properties considered most salient in the context of the respective genres.

4 Conclusion

A new software tool "DiaCollo" for the efficient extraction, comparison, and interactive online visualization of collocations was introduced. In its top-level incarnation as a modular web service plugin, DiaCollo provides a simple and intuitive interface for assisting linguists,

[6]For faster processing, arbitrary token- and/or document-attributes from the source corpus can be selected for inclusion in DiaCollo's native index structure at index compilation time. The default configuration includes only the token attributes *Lemma* and *Pos* ("part-of-speech") in its native index.

lexicographers, and humanities researchers to acquire a clearer picture of variation in a word's usage over time and/or corpus subset. Use of either the flexible DDC search engine back-end or compile-time index-attribute selection allows user queries to make explicit reference to genre and other document-level metadata, thus allowing cross-genre diachronic comparisons, as demonstrated on the basis of a simple example. Publicly accessible DiaCollo web-service instances exist for a number of German corpora hosted by the DWDS project at the Berlin-Brandenburg Academy of Sciences, and the DiaCollo source code itself is available via CPAN at `http://metacpan.org/release/DiaColloDB`.

References

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

M. Davies. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157, 2012. URL `http://davies-linguistics.byu.edu/ling450/davies_corpora_2011.pdf`.

J. Didakowski and A. Geyken. From DWDS corpora to a German word profile – methodological problems and solutions. In A. Abel and L. Lemnitzer, editors, *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information*, (OPAL X/2012). IDS, Mannheim, 2013. URL `http://www.dwds.de/static/website/publications/pdf/didakowski_geyken_internetlexikografie_2012_final.pdf`.

S. Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2005. URL `http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/`.

S. Evert. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin, 2008. URL `http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_manuscript.pdf`.

R. T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000. URL `https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm`.

A. Geyken, S. Haaf, B. Jurish, M. Schulz, J. Steinmann, C. Thomas, and F. Wiegand. Das deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In S. Schomburg, C. Leggewie, H. Lobin, and C. Puschmann, editors, *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, pages 157–161, 2011. URL `http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf#page=159`.

B. Jurish. DiaCollo: On the trail of diachronic collocations. In K. De Smedt, editor, *CLARIN Annual Conference 2015 (Wroław, Poland, October 14–16 2015)*, pages 28–31, 2015. URL `http://www.clarin.eu/sites/default/files/book%20of%20abstracts%202015.pdf`.

B. Jurish, C. Thomas, and F. Wiegand. Querying the deutsches Textarchiv. In U. Kruschwitz, F. Hopfgartner, and C. 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 2014. URL http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf.

A. Kilgarriff and D. Tugwell. Sketching words. In M.-H. Corréard, editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, EURALEX, pages 125–137, 2002. URL http://www.kilgarriff.co.uk/Publications/2002-KilgTugwell-AtkinsFest.pdf.

P. Rychlý. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9, 2008. URL http://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf.

A. Sokirko. A technical overview of DWDS/Dialing Concordance. Talk delivered at the meeting *Computational linguistics and intellectual technologies*, Protvino, Russia, 2003. URL http://www.aot.ru/docs/OverviewOfConcordance.htm.