Visualizing Semantic Change with DiaCollo

Bryan Jurish  
Berlin-Brandenburg Academy of Sciences and Humanities, Berlin  
jurish@bbaw.de

Maret Nielaender  
Georg Eckert Institute for International Textbook Research, Braunschweig  
nielaender@leibniz-gei.de

Thomas Werneke  
Centre for Contemporary History, Potsdam  
werneke@zzf-potsdam.de

Genealogies of Knowledge:  
Translating Political and Scientific Thought across Time and Space  
University of Manchester

8th December, 2017
Overview

The Situation
- Diachronic Text Corpora
- Collocation Profiling
- Diachronic Collocation Profiling

DiaCollo
- Requests & Parameters
- Profiles, Diffs & Indices

Use Cases
- *Die Grenzboten* corpus
- *Die Grenzboten* & anti-semitism
- *Die Grenzboten* & education policy

Summary & Conclusion
The Situation: Diachronic Text Corpora

- heterogeneous text collections, especially with respect to *date of origin*
  - other partitionings potentially relevant too, e.g. by author, text class, etc.

- increasing number available for linguistic & humanities research, e.g.
  - *Deutsches Textarchiv (DTA)* (Geyken 2013)
  - *Royal Society Corpus (RSC)* (Kermes et al. 2016)
  - *Corpus of Historical American English (COHA)* (Davies 2012)

- ...but even putatively “synchronic” corpora have a temporal extension, e.g.
  - *DWDS/ZEIT (“Kohl”)* (1946–2016)
  - *DDR Presseportal (“Ausreise”)* (1945–1993)

- should expose temporal effects of e.g. *semantic shift, discourse trends*

- problematic for conventional natural language processing tools
  - implicit assumptions of *homogeneity*
The Situation: Collocation Profiling

“Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache”
(‘The meaning of a word is its use in the language’) — L. Wittgenstein

“You shall know a word by the company it keeps” — J. R. Firth

Basic Idea
(Church & Hanks 1990; Manning & Schütze 1999; Evert 2005)
- **lookup** all candidate collocates \(w_2\) occurring with the target term \(w_1\)
- **rank** candidates by association score
  - “chance” co-occurrences with high-frequency items must be **filtered out**!
  - statistical methods require **large data sample**

What for?
- computational lexicography (Kilgarriff & Tugwell 2002; Didakowski & Geyken 2013)
- neologism detection (Kilgarriff et al. 2015)
- “text mining” / “distant reading” (Heyer et al. 2006; Moretti 2013)
Diachronic Collocation Profiling

The Problem: (temporal) heterogeneity
- conventional collocation extractors assume **corpus homogeneity**
- co-occurrence frequencies are computed only for **word-pairs** \( (w_1, w_2) \)
- influence of **occurrence date** (and other document properties) is irrevocably lost

A Solution (sketch)
- represent terms as \( n \)-tuples of independent attributes, **including occurrence date**
  - alternative: “document” level co-occurrences over sparse TDF matrix
- partition corpus **on-the-fly** into **user-specified intervals** (“date slices”, “epochs”)
- collect independent slice-wise profiles into final result set

Advantages
- full support for diachronic axis
- variable query-level granularity
- flexible attribute selection
- multiple association scores

Drawbacks
- sparse data requires larger corpora
- computationally expensive
- large index size
- no syntactic relations (yet)
DiaCollo: Overview

General Background
- developed to aid CLARIN historians in analyzing discourse topic trends
- successfully applied to mid-sized and large corpora, including:
  - Royal Society *Philosophical Transactions* (1665–1869, 9.8K documents, 35M tokens)
  - J. G. Dingler’s *Polytechnisches Journal* (1820–1931, 19K documents, 35M tokens)
  - *Deutsches Textarchiv* (1600–1900, 3.6K documents, 205M tokens)
  - *DDR-Presseportal* (1945–1994, 4.1M documents, 1.3G tokens)
  - *DWDS Zeitungen* (1946–2016, 10M documents, 4.7G tokens)

Implementation
- Perl API, command-line, & RESTful DDC/D* web-service plugin + GUI
- fast native indices over n-tuple inventories, equivalence classes, etc.
- **scalable** even in a high-load environment
  - no persistent server process is required
  - native index access via direct file I/O or `mmap()` system call
- various output & visualization formats, e.g. TSV, JSON, HTML, d3-cloud
DiaCollo: Profiles, Diffs & Indices

Profiles & Diffs

- simple request → unary profile for collocant(s)
  - filtered & projected to selected attribute(s)
  - aggregated into independent slice-wise sub-intervals
  - trimmed to \(k\)-best collocates for target word(s)

- diff request → comparison of two independent targets
  - highlights differences or similarities of target queries
  - can be used to compare different words
    ...or different corpus subsets w.r.t. a given word

Indices & Attributes

- compile-time filtering of native indices: frequency thresholds, PoS-tags
- default index attributes: Lemma \((l)\), Pos \(p\)
- finer-grained queries possible with TDF or DDC back-ends
- “live” KWIC-links to underlying corpus hits ⇒ DDC search engine
- batteries not included: corpus preprocessing, analysis, & full-text search index
  - see e.g. Jurish (2003); Geyken & Hanneforth (2006); Jurish et al. (2014), ...
DiaCollo: Scoring & Comparison Functions

Selected Score Functions

- **f**  
  raw collocation frequency  
  \[ f_{12} \]

- **lf**  
  collocation log-frequency  
  \[ \log_2(f_{12} + \varepsilon) \]

- **mi**  
  pointwise MI \( \times \) log-frequency  
  \[ \approx \log_2 \frac{f_{12} \times N}{f_1 \times f_2} \times \log_2 f_{12} \]

- **ll**  
  log-likelihood (Dunning 1993)  
  \[ \approx \operatorname{sgn}(f_{12}|f_1, f_2) \times \log \frac{L(H_0)}{L(H_1)} \]

- **ld**  
  log-Dice coefficient (Rychlé 2008)  
  \[ \approx 14 + \log_2 \frac{2 \times f_{12}}{f_1 + f_2} \]

Selected Diff Operations

- **diff**  
  raw score difference  
  \[ s_a - s_b \]

- **adiff**  
  absolute score difference  
  \[ |s_a - s_b| \]

- **avg**  
  arithmetic average  
  \[ \frac{s_a + s_b}{2} \]

- **max**  
  maximum  
  \[ \max\{s_a, s_b\} \]

- **min**  
  minimum  
  \[ \min\{s_a, s_b\} \]

- **havg**  
  harmonic average  
  \[ \approx \frac{2s_a s_b}{s_a + s_b} \]
Use Cases
Die Grenzboten Corpus

- *Die Grenzboten* ("the messengers from the border(s)") was a bi-weekly national-liberal German language periodical published 1841–1922.

- covered a wide range of politics, literature, and the arts throughout the ‘long’ nineteenth Century:
  - coverage of civic life, opinions, and debates surrounding the revolution of 1848, the restoration period, industrialization, the German Empire (*Kaiserreich*), and the First World War ∼ valuable source for a broad range of disciplines.

- 270 volumes (ca. 187,000 pages) digitized, OCR’ed, and structured by the SuUB Bremen in the context of a DFG-Project:
  - integrated into the corpus research infrastructure of the *Deutsches Textarchiv* at the BBAW CLARIN Service Center.
Use Cases

Basic Idea

- explore the corpus of national-liberal cultural history in order to...
  - test and demonstrate the utility of DiaCollo itself,
  - and that of other NLP software tools provided by CLARIN-D
  ...for the field of historical semantics.

- method: “blended reading”

Use Case 1

- *Die Grenzboten’s* stance towards Jewish people (Thomas Werneke)

Use Case 2

- *Die Grenzboten’s* content with respect to education policy (Maret Nieländer)
“Unter dem Begriff des ‘blended reading’ schlagen wir eine Strategie im Sinne einer Best Practice vor, die semiautomatische Analyseverfahren mit klassischer Textlektüre so integriert, dass sozialwissenschaftliche Erkenntnispotentiale, die sich auf die Auswertung großer Textdatenmengen stützen, optimal ausgeschöpft werden.”

“With the concept of ‘blended reading’ we propose a best-practice strategy which integrates semi-automatic analysis techniques with traditional (hermeneutic, ‘close’) reading such that the potential for socioscientific insights supported by the evaluation of large text data-sets is optimally exploited.”

— Stulpe & Lemke (2016)
Blended Reading: anti-semitism

Searching for anti-semitic ressentiments within the corpus

Step 1
- isolate a promising time slice
  - choosing the 1880s, covering the ‘Antisemitismusstreit’

Step 2
- identify collocates of the term Jude (“jew”)
  - select unusual collocates typically relating to ressentiments

Step 3
- analyze text passages via key-word-in-context (KWIC)
  - conventional source exegesis
Query: 50 strongest common noun collocates of the term ‘Jude’, 1880–1889

Four collocates evoked particular curiosity:

- Zahl ("number")
- Vermehrung ("reproduction, proliferation")
- Wucher ("usury")
- Handel ("trade")
"But, apart from these circumstances, the proliferation [Vermehrung] of the Jews still appears too significant to be attributed only to particularly favorable biotic conditions, to greater fertility and longer lifespan, which by the way should not be denied them."

“The most capable farmer often cannot measure up to the simplest Jew when it comes to money and trade, to certificate and mortgage and exchange.”


**Use Case 1: Conclusions**

- We have found evidence of stereotypizations, pointing to a cultural racism.
- The authors in part convey their arguments together with typical cultural anti-semitic resentments.
Blended Reading: *Schule* ("school")

**Step 1: query corpus vocabulary database (LexDB)**

- identify relevant terms within the corpus (e.g. *Schule*, 1840–1899)
  - ...in the *Deutsches Textarchiv*: 98.79 per million tokens
  - ...in *Die Grenzboten*: 237.86 per million tokens
- select interesting terms
  - identify high-frequency compounds containing *Schule*

**Step 2: query DiaCollo**

- identify strong adjective, common noun, & proper name collocates
- identify possible debates in the corpus via query results
- close reading in the texts via “key-word-in-context” (KWIC) hyperlinks
## Blended Reading: Frequencies & Time Series

### Frequency (absolute) vs. Date

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schule</td>
<td>20960</td>
</tr>
<tr>
<td>Hochschule</td>
<td>2156</td>
</tr>
<tr>
<td>Volksschule</td>
<td>1860</td>
</tr>
<tr>
<td>Realschule</td>
<td>762</td>
</tr>
<tr>
<td>Mädchenschule</td>
<td>450</td>
</tr>
<tr>
<td>Mittelschule</td>
<td>416</td>
</tr>
<tr>
<td>Bürgerschule</td>
<td>280</td>
</tr>
<tr>
<td>Fortbildungsschule</td>
<td>269</td>
</tr>
<tr>
<td>Fachschule</td>
<td>248</td>
</tr>
<tr>
<td>Vorschule</td>
<td>237</td>
</tr>
<tr>
<td>Elementarschule</td>
<td>234</td>
</tr>
<tr>
<td>Malerschule</td>
<td>186</td>
</tr>
<tr>
<td>Privatschule</td>
<td>184</td>
</tr>
<tr>
<td>Einheitsschule</td>
<td>183</td>
</tr>
<tr>
<td>Töchterschule</td>
<td>171</td>
</tr>
<tr>
<td>Kunstschule</td>
<td>168</td>
</tr>
<tr>
<td>Gelehrtschule</td>
<td>159</td>
</tr>
<tr>
<td>Klosterschule</td>
<td>141</td>
</tr>
<tr>
<td>Dorfschule</td>
<td>126</td>
</tr>
<tr>
<td>Gewerbeschule</td>
<td>124</td>
</tr>
<tr>
<td>Kriegsschule</td>
<td>116</td>
</tr>
<tr>
<td>Lateinschule</td>
<td>110</td>
</tr>
<tr>
<td>Militärschule</td>
<td>95</td>
</tr>
<tr>
<td>Knabenschule</td>
<td>92</td>
</tr>
<tr>
<td>Oberrealschule</td>
<td>88</td>
</tr>
<tr>
<td>Staatsschule</td>
<td>85</td>
</tr>
<tr>
<td>Musikschule</td>
<td>83</td>
</tr>
</tbody>
</table>
‘Gymnasium’: DiaCollo Collocates

1850s:
- Gymnasium
- Universität
- Schüler
- Lehrer
- Realgymnasium
- Griechische
- humanistisch

1870s:
- Gymnasium
- Universität
- Schüler
- Lehrer
- Realschule
- Griechische
- humanistisch
‘Erziehungen’: DiaCollo Collocates
(“education, upbringing”)

Attributive Adjectives (ADJA)

- **1850s**: häuslich, sittlich (“domestic, moral”)
- **1870s**: weiblich, ästhetisch (“feminine, aesthetic”)
- **1910s**: national, later staatsbürgerlich (“national, civic”)

Common Nouns (NN)

- **1860s**: Jugend, Kind (“youth, child”)
- **1870s**: Unterricht, Schule & Ausbildung appear (“instruction, school, training”)

Proper Names (NE)

- **1840s**: Frankreich (“France”)
- **1860–1880s**: Hegel, Lessing, Fichte, Schiller, Rousseau
- **1890–1910s**: Paul Güßfeldt, Georg Kerschensteiner
Collocate ‘Kirche’ ("church")

- persistently prominent throughout the entire corpus
- 1850s–1880s: konfessionell ("confessional")
- 1890s–1910s: Religionsunterricht ("religious education")

Refining the Search

- restrict to attributive adjective collocates (GROUPBY: 1,p=ADJA)
  - protestantisch ("protestant") 1860s
  - katholisch ("Catholic") 1860s-1870s
  - evangelisch ("Protestant, Evangelical") 1860s-1870s
  - konfessionell ("confessional") 1860s-1880s
  - kirchlich ("churchly") 1870s

- collocates related to church & religious confession peak in the 1860s–1870s
- also prominent: öffentlich ("public"); 1840s, 1870s–1900s
  - KWIC ⇝ stance of publicly funded schools w.r.t. church influence in education
Education Policy: *Kulturkampf*

*Kulturkampf* ("cultural struggle")
- rights & influences of state (Prussia) vs. church (Pope Pius IX)
- *ultramontan* ("ultramontane") ⇔ staunch supporters of the Catholic Church

Refining the Search: GermaNet thesaurus + paragraph search window
(Hamp & Feldweg 1997; Henrich & Hinrichs 2010)
- corpus hits show evidence for anti-Catholic opinions in debates on education
  - who should be in charge of education and curricula?
  - how to deal with different religious denominations in schools?

Use Case 2: Conclusions
- some important aspects of debate not apparent from initial naïve DiaCollo queries
- informed curiosity & focused investigation leads to very satisfying results
Summary & Conclusion I

Diachronic Collocation Profiling

- “meaning is use” \(\leadsto\) ... in temporal context
- diachronic text corpora \(\leadsto\) semantic shift, discourse trends
- conventional tools \(\leadsto\) implicit assumptions of homogeneity
- diachronic profiling \(\leadsto\) date-dependent lexemes

DiaCollo

- on-the-fly corpus partitioning \(\leadsto\) arbitrary query-level granularity
- DDC/D* integration \(\leadsto\) fine-grained queries, corpus KWIC links
- RESTful web service \(\leadsto\) external API, online visualization
Summary & Conclusion II

DiaCollo’s Application in the field of Historical Semantics

Advantages

- support for semiasological methods
- inductive and explorative methods
- fluent “blending” from distant to close reading (and back)

Problems & Issues

- onomasiological properties are not (currently) recoverable
  - e.g. by comparing word clouds
- digital corpora (sources):
  - quantity (or lack thereof)
  - quality (metadata, format, preprocessing)
  - legal aspects, copyright issues

Mildly Pontifical Remarks (“Outlook”)

- distant reading will eventually require us to read more rather than less
- in the end, we may have to read more sources than ever before!
Thank you for listening!

http://kaskade.dwds.de/~jurish/diacollo2017
http://metacpan.org/release/DiaColloDB
References
References


References


References


