Using DIACOLLO for Historical Research

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

CLARIN Annual Conference 2019
Leipzig, Germany

1st October, 2019
Overview

- Collaborative software development
- Corpora & collocations
- DiaCollo: diachronic collocation profiling
- Use case: Education policy in *Die Grenzboten*
- Summary & conclusion
Software Development Cycle

Planning
- identify desiderata & bugs
- sketch next steps

Implementation
- coding & documentation
- release & deployment

Evaluation
- testing “in the wild”
- user feedback
Corpora & Collocations

Diachronic Text Corpora

- heterogeneous with respect to date of origin
- should expose temporal effects of e.g. semantic shift, discourse trends
- problematic for conventional NLP tools (which assume homogeneity)

Collocation Profiling  
(Church & Hanks 1990; Manning & Schütze 1999; Evert 2005)

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

- prompt user for target collocant term(s) of interest ($w_1$)
- lookup all candidate collocates ($w_2$) co-occurring with $w_1$
- rank candidates by association score
  - score function $\varphi(f_1, f_2, f_{12}, N)$ approximates relevance of $w_2$ to $w_1$
  - “chance” co-occurrences with high-frequency $w_2$ should be filtered out!
  - statistical method requires large data sample
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
- partition corpus on-the-fly into user-specified intervals ("date slices", "epochs")
- collect independent epoch-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: Development

Planning & Evaluation
- in collaboration with DWDS lexicographers & CLARIN-D historians

Implementation
- Perl+PDL API, CLI, client/server
  - RESTful D* web-service + GUI
- various output & visualization formats, e.g.
  - TSV, JSON, HTML, Highcharts, d3-cloud, ...
- batteries not included
  - tokenization, annotation, full-text search, ...
- garbage in ⇛ garbage out
  - “messy” corpora ⇛ unsatisfying results

Deployment
- successfully applied to 70 distinct curated corpora at the BBAW, including:
  - Royal Society *Philosophical Transactions* (1665–1869, 9.8K documents, 35M tokens)
  - Deutsches Textarchiv (1600–1900, 3.6K documents, 205M tokens)
  - DWDS Zeitungen (1946–2019, 16M documents, 6.3G tokens)
DiaCollo: Scoring & Comparison Functions

Selected Score Functions

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

- **If**  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)  
  \[ \approx \text{sgn}(f_{12}|f_1, f_2) \times \log \frac{L(H_0)}{L(H_1)} \]

- **ld**  log-Dice coefficient (Rychlý 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 Case: Education Policy in *Die Grenzboten*
‘Schule’: DiaCollo Query (DTA)

Target as:
- LEMMA(s), e.g. Maske
- /REGEX/, e.g. /\^Masken.*$/

DDC QUERY

Target date(s):
- DATE(s), e.g. 1900:1999 or *:* or 1900:* /REGEX/, e.g. /\^18\[345]/
‘Schule’: DiaCollo Collocates (DTA: HTML)

1560–1569

<table>
<thead>
<tr>
<th>N</th>
<th>f1</th>
<th>f2</th>
<th>f12</th>
<th>score</th>
<th>label</th>
<th>lemma</th>
<th>pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>592882</td>
<td>1630</td>
<td>1152</td>
<td>40</td>
<td>8.8800</td>
<td>1560</td>
<td>Kloster</td>
<td>NIN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>1038</td>
<td>21</td>
<td>8.0108</td>
<td>1560</td>
<td>Knabe</td>
<td>NIN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>412</td>
<td>15</td>
<td>7.9111</td>
<td>1560</td>
<td>Schuemeister</td>
<td>NIN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>1630</td>
<td>22</td>
<td>7.7888</td>
<td>1560</td>
<td>Schule</td>
<td>NIN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>1987</td>
<td>23</td>
<td>7.7030</td>
<td>1560</td>
<td>Ordnung</td>
<td>NIN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>54</td>
<td>9</td>
<td>7.4522</td>
<td>1560</td>
<td>partikular</td>
<td>ADJA</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>1370</td>
<td>15</td>
<td>7.3561</td>
<td>1560</td>
<td>Fleiß</td>
<td>NIN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>382</td>
<td>10</td>
<td>7.3475</td>
<td>1560</td>
<td>Pfarrherr</td>
<td>NIN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>5791</td>
<td>35</td>
<td>7.2719</td>
<td>1560</td>
<td>Kirche</td>
<td>NIN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>425</td>
<td>9</td>
<td>7.1650</td>
<td>1560</td>
<td>Flecken</td>
<td>NIN</td>
</tr>
</tbody>
</table>

- association with religious institutions
  - *Kloster* (“cloister”)
  - *Pfarrherr* (“pastor”)
  - *Kirche* (“church”)
### ‘Schule’: DiaCollo Collocates (DTA: HTML)

#### 1560–1569

<table>
<thead>
<tr>
<th>N</th>
<th>f1</th>
<th>f2</th>
<th>f12</th>
<th>score</th>
<th>label</th>
<th>lemma</th>
<th>pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>592882</td>
<td>1630</td>
<td>1152</td>
<td>40</td>
<td>8.8800</td>
<td>1560</td>
<td>Kloster</td>
<td>NN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>1038</td>
<td>21</td>
<td>8.0108</td>
<td>1560</td>
<td>Knabe</td>
<td>NN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>412</td>
<td>15</td>
<td>7.9111</td>
<td>1560</td>
<td>Schumieister</td>
<td>NN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>1630</td>
<td>22</td>
<td>7.7888</td>
<td>1560</td>
<td>Schule</td>
<td>NN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>1987</td>
<td>23</td>
<td>7.7030</td>
<td>1560</td>
<td>Ordnung</td>
<td>NN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>54</td>
<td>9</td>
<td>7.4522</td>
<td>1560</td>
<td>partikular</td>
<td>ADJA</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>1370</td>
<td>15</td>
<td>7.3561</td>
<td>1560</td>
<td>Fleiß</td>
<td>NN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>382</td>
<td>10</td>
<td>7.3475</td>
<td>1560</td>
<td>Pfarrherr</td>
<td>NN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>5791</td>
<td>35</td>
<td>7.2719</td>
<td>1560</td>
<td>Kirche</td>
<td>NN</td>
</tr>
<tr>
<td>592882</td>
<td>1630</td>
<td>425</td>
<td>9</td>
<td>7.1650</td>
<td>1560</td>
<td>Flecken</td>
<td>NN</td>
</tr>
</tbody>
</table>

#### 1710–1719

<table>
<thead>
<tr>
<th>N</th>
<th>f1</th>
<th>f2</th>
<th>f12</th>
<th>score</th>
<th>label</th>
<th>lemma</th>
<th>pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>13801428</td>
<td>2241</td>
<td>2241</td>
<td>16</td>
<td>6.8701</td>
<td>1710</td>
<td>Schule</td>
<td>NN</td>
</tr>
<tr>
<td>13801428</td>
<td>2241</td>
<td>14479</td>
<td>44</td>
<td>6.4301</td>
<td>1710</td>
<td>Kirche</td>
<td>NN</td>
</tr>
<tr>
<td>13801428</td>
<td>2241</td>
<td>227</td>
<td>6</td>
<td>6.3158</td>
<td>1710</td>
<td>Inspektor</td>
<td>NN</td>
</tr>
<tr>
<td>13801428</td>
<td>2241</td>
<td>206</td>
<td>5</td>
<td>6.0651</td>
<td>1710</td>
<td>mechanisch</td>
<td>ADJA</td>
</tr>
<tr>
<td>13801428</td>
<td>2241</td>
<td>335</td>
<td>5</td>
<td>5.9910</td>
<td>1710</td>
<td>Besuchung</td>
<td>NN</td>
</tr>
<tr>
<td>13801428</td>
<td>2241</td>
<td>818</td>
<td>5</td>
<td>5.7431</td>
<td>1710</td>
<td>preußisch</td>
<td>ADJA</td>
</tr>
<tr>
<td>13801428</td>
<td>2241</td>
<td>1266</td>
<td>5</td>
<td>5.5459</td>
<td>1710</td>
<td>Besserung</td>
<td>NN</td>
</tr>
<tr>
<td>13801428</td>
<td>2241</td>
<td>1969</td>
<td>6</td>
<td>5.5454</td>
<td>1710</td>
<td>Universität</td>
<td>NN</td>
</tr>
<tr>
<td>13801428</td>
<td>2241</td>
<td>2462</td>
<td>6</td>
<td>5.3856</td>
<td>1710</td>
<td>Lehrer</td>
<td>NN</td>
</tr>
<tr>
<td>13801428</td>
<td>2241</td>
<td>3418</td>
<td>6</td>
<td>5.1186</td>
<td>1710</td>
<td>Jugend</td>
<td>NN</td>
</tr>
</tbody>
</table>

- association with religious institutions
  - Kloster (“cloister”)
  - Pfarrherr (“pastor”)
  - Kirche (“church”)
- stronger secular associations
  - Inspektor (“inspector”)
  - preußisch (“prussian”)
  - Universität (“university”)
- trend continues as time progresses
‘Schule’: DiaCollo Collocates (DTA: lemma-cloud)

1560s:
- Pfarrherr
- partikular
- Ordnung
- Schule
- Flecken
- Schulmeister
- Knabe
- Kloster

1710s:
- mechanisch
- Schule
- Lehrer
- Jugend
- Universität
- Kirche
- Inspektor
- Besuchung
- preußisch
- Besserung
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.

- 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.

http://brema.suub.uni-bremen.de/grenzboten
http://www.deutschestextarchiv.de/doku/textquellen#grenzboten
Are *Die Grenzboten* concerned with education?

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

- identify relevant terms in the corpus, e.g. *Schule* ("school"), 1840–1899
  - ...in the *Deutsches Textarchiv*: 101.52 per million tokens
  - ...in *Die Grenzboten*: **237.29** per million tokens

**Step 2: query DiaCollo**

- identify strong collocates for *Schule* ("school")
- identify possible debates in the corpus via query results
- close reading in the texts via "keyword-in-context" (KWIC) hyperlinks
**Collocate ‘Kirche’ (“church”)**

- persistently prominent throughout the entire *Grenzboten* 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
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

- 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?

Upshot

- some important aspects of debate are not apparent from initial naïve DiaCollo queries
- informed curiosity & focused investigation leads to very satisfying results
Summary & Conclusion

Collaborative Development

- cyclic process
- elusive common ground

\[ \Rightarrow \text{feedback loop} \]

\[ \Rightarrow \text{terminology, research methodology} \]

DiaCollo

- diachronic text corpora
- conventional tools
- diachronic profiling

\[ \Rightarrow \text{semantic shift, discourse trends} \]

\[ \Rightarrow \text{implicit assumptions of homogeneity} \]

\[ \Rightarrow \text{date-dependent lexemes} \]

...as a tool for historical research

- fluent “blended”/“scalable” reading
- digital corpora (sources)

\[ \Rightarrow \text{distant } \leftrightarrow \text{close reading} \]

\[ \Rightarrow \text{quantity, quality, legal issues} \]
Thank you for listening!

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


References
