Visualisierung diachroner Kollokationen mit DiaCollo

oder:

“warum noch so‘n komisches buntiklicki DH-Werkzeug?”

Bryan Jurish
jurish@bbaw.de

Die geisteswissenschaftliche Perspektive: Welche Forschungsergebnisse lassen Digital Humanities erwarten?
Akademie der Wissenschaften und der Literatur, Mainz
19. Februar, 2016
Overview

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

DiaCollo
- Requests & Parameters
- Profile, Diffs & Indices
- Scoring & Comparison Functions

Applications & Examples

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 et al. 2011)
  - *Referenzkorpus Altdeutsch (DDD)* (Richling 2011)
  - Corpus of Historical American English (COHA) (Davies 2012)

- ... but even putatively “synchronic” corpora have a temporal extension, e.g.
  - DWDS/ZEIT (“Kohl”) (1946–2015)
  - 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

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

**Basic Idea**

*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)*
- distributional semantics  *(Schütze 1992; Sahlgren 2006)*
- 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:
  - J. G. Dingler’s *Polytechnisches Journal* (1820–1931, 19K documents, 35M tokens)
  - *Deutsches Textarchiv* (1600–1900, 2.6K documents, 173M tokens)
  - *DDR-PP Neues Deutschland* (1946–1990, 1.5M documents, 443M tokens)
  - *DWDS Zeitungen* (1946–2015, 10M documents, 4.3G 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: Requests & Parameters

- request-oriented RESTful service
- accepts user requests as set of \textit{parameter=value} pairs
- parameter passing via URL query string or HTTP POST request
- common parameters:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>query</td>
<td>target lemma(ta), regular expression, or DDC query</td>
</tr>
<tr>
<td>date</td>
<td>target date(s), interval, or regular expression</td>
</tr>
<tr>
<td>slice</td>
<td>aggregation granularity or “0” (zero) for a global profile</td>
</tr>
<tr>
<td>groupby</td>
<td>aggregation attributes with optional restrictions</td>
</tr>
<tr>
<td>score</td>
<td>score function for collocate ranking</td>
</tr>
<tr>
<td>kbest</td>
<td>maximum number of items to return per date-slice</td>
</tr>
<tr>
<td>diff</td>
<td>score aggregation function for diff profiles</td>
</tr>
<tr>
<td>global</td>
<td>request global profile pruning (vs. default slice-local pruning)</td>
</tr>
<tr>
<td>profile</td>
<td>profile type to be computed ({native,tdf,ddc} × {unary,diff})</td>
</tr>
<tr>
<td>format</td>
<td>output format or visualization mode</td>
</tr>
</tbody>
</table>
DiaCollo: Profiles, Diffs & Indices

Profiles & Diffs
- simple request → unary profile for target term(s)
  - filtered & projected to selected attribute(s)
  - trimmed to $k$-best collocates for target word(s)
  - aggregated into independent slice-wise sub-intervals
- 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
- 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 \( = f_{12} \)
- \( \text{l}_f \) collocation log-frequency \( = \log_2(f_{12} + \varepsilon) \)
- \( \text{mi} \) pointwise MI \( \times \) log-frequency \( \approx \log_2 \frac{f_{12} \times N}{f_1 \times f_2} \times \log_2 f_{12} \)
- \( \text{ll} \) log-likelihood (Dunning 1993) \( \approx \text{sgn}(f_{12}|f_1, f_2) \times \log(1 + \log \lambda) \)
- \( \text{ld} \) log-Dice coefficient (Rychlý 2008) \( \approx 14 + \log_2 \frac{2 \times f_{12}}{f_1 + f_2} \)

Selected Diff Operations
- \( \text{diff} \) raw score difference \( = s_a - s_b \)
- \( \text{adiff} \) absolute score difference \( = |s_a - s_b| \)
- \( \text{avg} \) arithmetic average \( = \frac{s_a + s_b}{2} \)
- \( \text{max} \) maximum \( = \max\{s_a, s_b\} \)
- \( \text{min} \) minimum \( = \min\{s_a, s_b\} \)
- \( \text{havg} \) harmonic average \( \approx \frac{2s_a s_b}{s_a + s_b} \)
Application Scenarios

Exploration & Discovery
- detection of temporal (historical) phenomena
- fast & flexible “distant reading” over large digital text collections
- various intuitive visualizations of query results

Analysis & Investigation
- data acquisition for hypothesis testing
- first step towards “close” hermeneutic analysis (e.g. via KWIC-links)

(Quantitative) Evaluation & Assessment
- diachronic linguistics
- historical semantics
- history of concepts
Example 1: ‘Crises’ in the News

‘Krise’ in DIE ZEIT (west) and Neues Deutschland (east)


1950–1959
- Berlin blockade aftermath

1960–1969
- anti-government protests & strikes in France

1970–1979
- Nixon & Brandt resignations; Iranian revolution

1980–1989
- Solidarność in Poland; Soviet war in Afghanistan; Schmidt coalition collapses

1990–1999
- wars in ex-Yugoslavia, Kosovo & Chechnya; financial crises in Asia & Mexico

2000–2009
- global financial crisis

2010–2015
- civil wars in Syria & the Ukraine; Greek bankruptcy

Compare:
- Krise: DDR-PP Neues Deutschland: 3-year slices, proper name collocates (NE)
- Krise: DDR-PP Neues Deutschland: 5-year slices, common noun collocates (NN)
Example 1: Selected Lemma-Clouds

1980–1989:
- Sowjetunion
- Polen
- Europa
- NATO
- Afghanistan
- AEG_Hausgeräte_GmbH
- Bonn
- Schmidt
- Berlin
- Sozialdemokratische_Partei_Deutschlands

2010–2014:
- Kiew
- European_Union
- Merkel
- Spanien
- Griechenland
- Syrien
- Italien
- Ukraine
- Krim
Example 2: Lexicography
‘autofrei’ (automobile-free)

Lexicography & Collocations
- collocation preferences correlate strongly with word meanings
- new senses (“neosemantemes”) ⇒ new collocates
  - *Maus* ("mouse"): rodent vs. input device
  - *Ampel* ("traffic light"): traffic signal vs. political coalition

The case of *autofrei* ("automobile-free")
- Duden: *keinen Autoverkehr aufweisend* ("lacking automobile traffic")
- DWDS corpora reveal ***two sub-senses***:
  - 1970–1989: ... by ordinance (⇝ *Sonntag, Innenstadt*)
  - 1990–present: ... voluntary (⇝ *Wohnanlage, Siedlung*)

http://kaskade.dwds.de/dstar/zeitungen/diacollo/?q=autofrei&ds=5&f=bub
Example 2: Selected Bubble-Charts

1985–1989

- Innenstadt
- Sonntag
- AL
- Berlin
- Stadt

1990–1994

- Aktionswoche
- Farmsen
- Innenstadt
- Wohngebiet
- Wohnen
- autoarm
- Zone
- Modellversuch
- Siedlung
- City
Example 3: Gender & Cultural Bias
‘Mann’ vs. ‘Frau’ in the Deutsches Textarchiv (1600–1900)

http://kaskade.dwds.de/dstar/dta/diacollo/?q=Mann&bq=Frau&d=1600:1899&ds=25&gb=1,p%3DADJA&f=cld&p=d2

Disclaimer
- historical corpus data can reveal persistent cultural biases
- linked collocation data does not reflect the opinions of this author or the BBAW!

Observations
- biological fact: schwangere Frau (only appears 1675–1724)
- fixed & formulaic expressions very prominent
  - gnädige Frau (masculine variant: gnädiger Herr)
  - Frau X geborene Y (birth- vs. married surname)
  - der gemeine Mann (masculine generic)
- pretty much exclusively cultural bias:
  - Mann ⇛ berühmt, ehrlich, gelehrt, tapfer, weise, …
  - Frau ⇛ betrübt, lieb, schön, tugendreich, verwitwet, …
- differences grow less pronounced in late 18th & 19th centuries
Example 3: Selected Lemma-Clouds

1725–1749:
lieb
groß
ander
gemein
gebären
weise
glelehrt
gnädig
eigen

1825–1849:
edel
jung
ander
gnädig
lieb
schön
gut
grau
deutsch
Example 4: Genealogy of Terminology

Habermas vs. Cassirer in the DWDS Kernkorpus

http://kaskade.dwds.de/dstar/kern/diacollo/?ds=0&bds=0&k=20&p=diff-tdf&f=cld&diff=adiff

QUERY: * #has[author, /Habermas/]
~QUERY: * #has[author, /Cassirer/]
GROUPBY: 1, p=NN

Remarks
- uses TDF (term × document) matrix back-end for bibliographic meta-data queries
- sets slice=0 parameter to acquire date-independent profiles
- groupby clause selects only common noun lemmata (STTS tag NN)
- modest sample size (Habermas: 516k tokens, Cassirer: 130k tokens)
- Habermas himself openly acknowledges Cassirer’s influence

Differences (diff=adiff)
- Habermas ⇝ Handeln, Gesellschaft, Öffentlichkeit, Meinung, Norm, . . .
- Cassirer ⇝ Anschauung, Bestimmung, Bezeichnung, Erkenntnis, Sein, . . .

Similarities (diff=havg, diff=min)
- Analyse, Ausdruck, Begriff, Beziehung, Funktion, Sinn, Sprache, . . .
Example 4: Lemma-Clouds

differences
(diff=adiff)

similarities
(diff=havg)
Example 5: 400 Years of Potables

‘[GETRÄNK] trinken’ in aggregated DTA+DWDS (1600–2000)

http://kaskade.dwds.de/dstar/dta+dwds/diacollo/?d=1600%3A1999&ds=50&k=20&p=ddc&f=cld&g=1&G=1
QUERY: "(Getränk|gn-sub WITH $p=NN)=2 (trinken WITH $p=/VV[IP]/)" #FMIN 1

Remarks
- uses DDC back-end for fine-grained data acquisition
- uses GermaNet thesaurus-based lexical expansion for Getränk (“beverage”)
- considers only those target terms immediately preceding verb trinken (“to drink”)
- “global” profile uses shared target-set to avoid visual clutter

Observations
- near-constants: Bier, Milch, Wasser, Wein (“beer, milk, water, wine”)
- 1650–1750: Tee, Kaffee, Schokolade (“tea, coffee, chocolate”) appear
- 1800–1900: Schnaps displaces Branntwein; Champagner appears
- 1850–1900: Alkohol (“alcohol”) as category of beverages
- 1900–2000: Kognak, Saft, Sekt, Whisky (“cognac, juice, sparkling wine, whisky”)
Example 5: Time Series \( (k = 10) \)

DiaCollo Profile

\[ (\text{Getränk}\_\text{gn-sub WITH } p=\text{NN})=2 \text{ (trinken WITH } p=\text{VV[IP]})) \#\text{FMIN 1} \]

<table>
<thead>
<tr>
<th>Date (slice)</th>
<th>Score (log Dice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1600</td>
<td>0</td>
</tr>
<tr>
<td>1700</td>
<td>5</td>
</tr>
<tr>
<td>1800</td>
<td>2.5</td>
</tr>
<tr>
<td>1900</td>
<td>-2.5</td>
</tr>
</tbody>
</table>

Graph showing the DiaCollo Profile over time with different drinks such as Alkohol, Bier, Branntwein, Kaffee, Milch, Schnaps, Sekt, Tee, and Wasser.
Summary & Conclusion

Diachronic Collocation Profiling
- diachronic text corpora
- conventional tools
- diachronic profiling
  - semantic shift, discourse trends
  - implicit assumptions of homogeneity
  - date-dependent lexemes

DiaCollo
- on-the-fly corpus partitioning
- DDC/D* integration
- RESTful web service
  - arbitrary query granularity
  - fine-grained queries, corpus KWIC links
  - external API, online visualization

Applications
- exploration & discovery
- analysis & investigation
- evaluation & assessment
  - large source collections
  - data acquisition for hypothesis testing
  - historical semantics, history of concepts, &c.
Thank you for listening!

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