DiaCollo – computergestützte Analyse von Kollokationen im diachronen Verlauf

Bryan Jurish
jurish@bbaw.de

Thomas Werneke
werneke@zzf-potsdam.de

Digitale Geschichtswissenschaft – neue Tools für neue Fragen?
Tagung der CLARIN-D Facharbeitsgruppen "Neuere Geschichte" und "Zeitgeschichte"
Berlin-Brandenburgische Akademie der Wissenschaften
8th–9th February, 2016

Overview

The Situation

- Diachronic Text Corpora
- Collocation Profiling
- Diachronic Collocation Profiling

DiaCollo

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

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)*
  - DWDS/Blogs ("Browser") *(1994–2014)*
- 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

(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)*
- 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 term vocabulary **on-the-fly** into **user-specified intervals** ("date slices")
- 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 *(Fielding 2000)*
- accepts user requests as set of *parameter=value* pairs
- parameter passing via URL query string or HTTP POST request
- common parameters:

Parameter	Description
`query`	target lemma(ta), regular expression, or DDC query
`date`	target date(s), interval, or regular expression
`slice`	aggregation granularity or "0" (zero) for a global profile
`groupby`	aggregation attributes with optional restrictions
`score`	score function for collocate ranking
`kbest`	maximum number of items to return per date-slice
`diff`	score aggregation function for diff profiles
`global`	request global profile pruning (vs. default slice-local pruning)
`profile`	profile type to be computed ({native,tdf,ddc} × {unary,diff})
`format`	output format or visualization mode

DiaCollo: Profiles, Diffs & Indices

Profiles & Diffs

- simple request → unary **profile** for target term(s) *(profile, query)*
  - **filtered** & **projected** to selected attribute(s) *(groupby)*
  - **trimmed** to k-best collocates for target word(s) *(score, kbest, global)*
  - **aggregated** into independent slice-wise sub-intervals *(date, slice)*
- diff request → **comparison** of two independent targets *(profile, bquery, ...)*
  - highlights **differences** or **similarities** of target queries *(diff)*
  - can be used to compare different words *(query ≠ bquery)*

    ...or different corpus subsets w.r.t. a given word *(e.g. date ≠ bdate)*

Indices & Attributes

- compile-time filtering of native indices: frequency threshholds, PoS-tags
- default index attributes: *Lemma (l)*, *Pos (p)*
- finer-grained queries possible with TDF or DDC back-ends

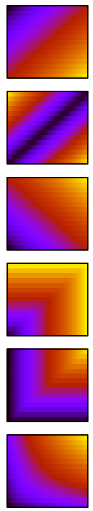
DiaCollo: Scoring & Comparison Functions

Selected Score Functions

- **f** raw collocation frequency $= f_{12}$
- **lf** collocation log-frequency $= \log_2(f_{12} + \varepsilon)$
- **mi** pointwise MI × 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) $\approx \mathrm{sgn}(f_{12} | f_1, f_2) \times \log(1 + \log \lambda)$
- **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 $\approx \frac{2 s_a s_b}{s_a + s_b}$



Examples

Fallbeispiele

Anwendungsszenarien in den historisch arbeitenden Fachdisziplinen

Mehrwert von DiaCollo

- macht historische Dimension der Sprache in digitalen Quellensammlungen sichtbar
- Untersuchung diachronen semantischen Wandels innerhalb großer Textmengen
- vielseitige Visualisierungsoptionen der Abfrageergebnisse

Anwendungsbereiche

- als Analysewerkzeug für Historische Semantik / Diskurs- und Begriffsgeschichte
- als Auswertungswerkzeug in der Kulturgeschichte (des Politischen)
- als allgemeines Hilfsmittel bei der Analyse großer Quellenkorpora

Darstellung des Potentials anhand von Fallbeispielen

- Eigennamen in der "Krise" in west- und ostdeutschen Printmedien
- Semantische Differenzierung bei Vergleichsabfrage – *Mann* vs. *Frau*
- DDC-Abfrage für "[GETRÄNK] *trinken*" im Deutschen Textarchiv

Example 1: *Krise* ("crisis") in the weekly *DIE ZEIT*

http://kaskade.dwds.de/dstar/zeit/diacollo/?q=Krise&d=1950:2014&gb=l,p%3DNE

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–present
- 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:

2010–2014:

Example 2: *Mann* vs. *Frau* in the DTA

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

Disclaimer

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

Observations

- 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* $\rightsquigarrow$ *berühmt, ehrlich, gelehrt, tapfer, weise, . . .*
  - *Frau* $\rightsquigarrow$ *betrübt, lieb, schön, tugendreich, verwitwet, . . .*
- differences grow less pronounced in late 18th & 19th centuries

Example 2: Selected Lemma-Clouds

1725–1749:

1825–1849:

Example 3: 400 Years of Potables

http://kaskade.dwds.de/dstar/dta+dwds/diacollo/?d=1600%3A1999&ds=50&k=20&p=ddc&f=cld&g=l&G=1

QUERY: "(Getränk|gn-sub WITH $p=NN)=2 (trinken WITH $p=/VV[IP]/)" #FMIN 1

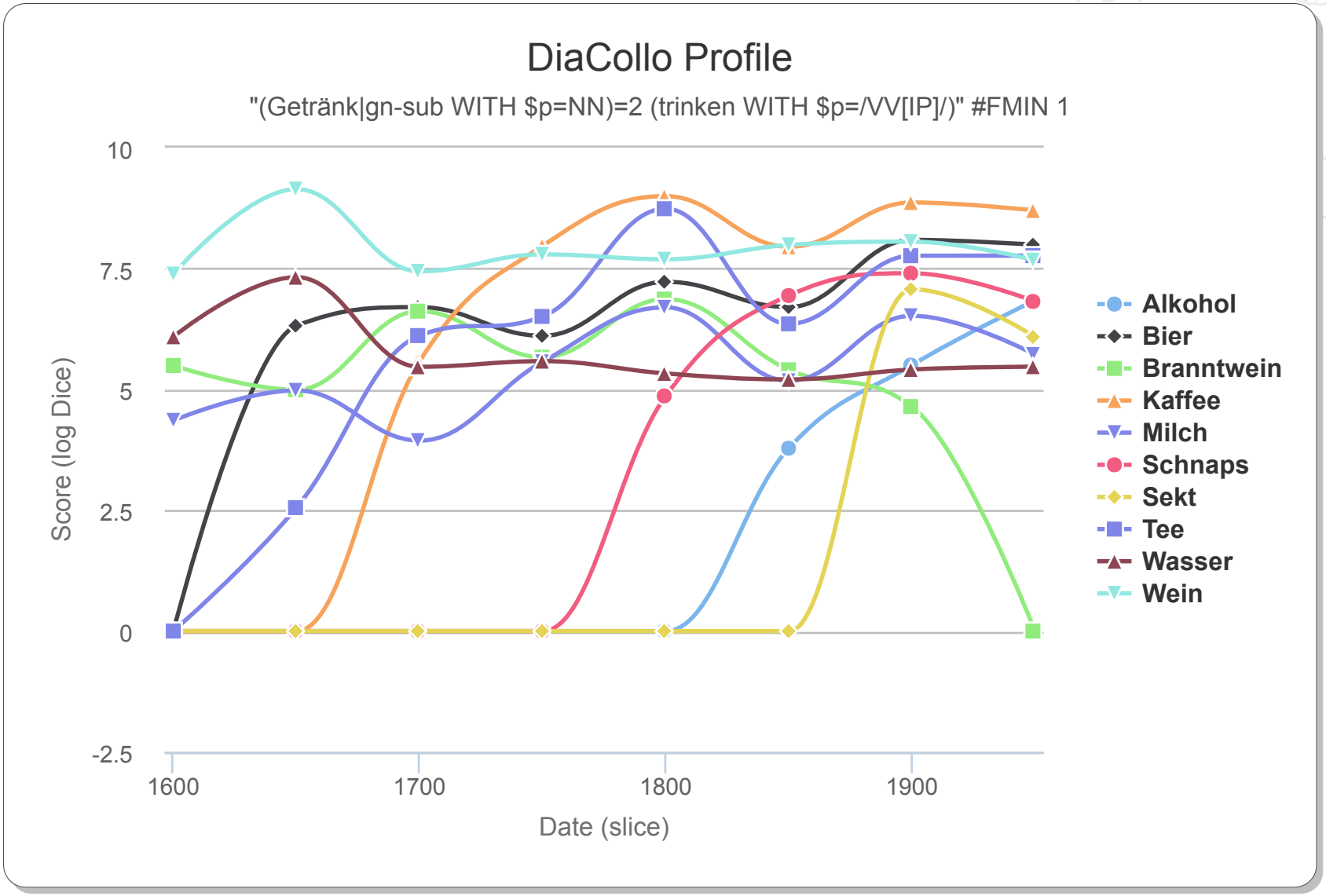
Remarks

- 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 3: Time Series $(k = 10)$



Summary & Conclusion

Diachronic Collocation Profiling

- diachronic text corpora ⇝ *semantic shift, discourse trends*
- conventional tools ⇝ *implicit assumptions of homogeneity*
- diachronic profiling ⇝ *date-dependent lexemes*

DiaCollo

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

Applications

- exploration & discovery ⇝ *large source collections*
- analysis & investigation ⇝ *data acquisition for hypothesis testing*
- interpretation & assessment ⇝ *historical semantics, history of concepts, &c.*

— The End —

Thank you for listening!

http://kaskade.dwds.de/diacollo/

http://metacpan.org/release/DiaColloDB