Neue Entwicklungen und Wege bei der Erstellung, Erweiterung und Nutzung von Korpora am Zentrum Sprache

Alexander Geyken, Bryan Jurish
{geyken,jurish}@bbaw.de

Neue Wege in der Nutzung von Korpora – Data-Mining für die textorientierten Geisteswissenschaften
Abschlusstagung des BMBF-Projekts KobRA
Berlin-Brandenburgische Akademie der Wissenschaften
30th September, 2015
Overview

(New) Corpora at the Zentrum Sprache

Some Current Uses of Corpora in the DWDS Project

Diachronic Collocations

- Diachronic Text Corpora
- Collocation Profiling
- Diachronic Collocation Profiling

DiaCollo

- Requests & Parameters
- Profile, Diffs & Indices
- Association Score Functions

Examples

Summary & Outlook
(New) Corpora at the Zentrum Sprache

Corpora embedded in a new platform

- public: 1.5G Tokens
  - Reference corpora: DWDS-Kernkorpus, Deutsches Textarchiv
  - Newspaper corpora: ZEIT, Berliner Zeitung, Tagesspiegel
  - Special corpora: blogs, film subtitles

- internal: 7G Tokens
  - Newspaper corpora: Bild, FAZ, NZZ, Spiegel, Der Standard, Welt
  - Under construction:
    - extended blog corpus (2G Tokens)
    - newspapers (“DDR-Presseportal”): Neues Deutschland, Berliner Zeitung

http://zwei.dwds.de (beta)
Some Current Uses of Corpora in the DWDS Project

- Concordance search
  - Used in university courses and in general by users interested in word use

- Extraction of “Good Examples” *(cf. Lemnitzer this morning)*
  - Used to restrict the result sets of corpus queries for lexicographic work

- Synchronic collocation profiling *(DWDS-Wortprofil)*
  - Used to extract collocation information for the DWDS dictionary

- Diachronic collocation profiling *(this talk!)*
  - Used to investigate temporal shift in word meaning
  - Useful for lexicographers, historians, *etc.*
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 (1946–1994)

- should reveal temporal phenomena such as *semantic shift*

- 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
- distributional semantics
- text mining / “distant reading”  (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
- 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, e.g.
  - J. G. Dingler’s *Polytechnisches Journal* (1820–1931, 19K documents, 35M tokens)
  - *Deutsches Textarchiv* (1600–1900, 2.6K documents, 173M 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 `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 (<code>{native,ddc} \times \{unary,diff\}</code>)</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 threshholds, PoS-tags
- default index attributes: **Lemma** ($l$), **Pos** ($p$)
- finer-grained queries possible with [DDC] back-end
DiaCollo: Scoring Functions

**Supported Score Functions**

- **f**  raw collocation frequency  \(= f_{12} \)
- **lf**  collocation log-frequency  \(= \log_2(f_{12} + \varepsilon) \)
- **mi**  pointwise MI \(\times\) log-frequency  \(\approx \log_2 \frac{f_{12} \times N}{f_1 \times f_2} \times \log_2 f_{12} \)
- **ld**  log-Dice coefficient (Rychlý 2008)  \(\approx 14 + \log_2 \frac{2 \times f_{12}}{f_1 + f_2} \)

**Supported 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  \(\approx \frac{2s_a s_b}{s_a + s_b} \)
- **gavg**  geometric average  \(\approx \sqrt{s_a s_b} \)
Example 1: *Krise* ("crisis") in the weekly *DIE ZEIT*

http://kaskade.dwds.de/dstar/zeit/diacollo/?q=Krise&d=1950:2014&gb=1,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
Example 1: Selected Lemma-Clouds

1980–1989:
- Sowjetunion
- Polen
- Europa
- NATO
- Afghanistan
- AEG_Hausgeräte_GmbH
- Bonn
- 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*)
Example 2: Selected Bubble-Charts

1985–1989

Innenstadt
Sonntag
AL
Berlin
Stadt

1990–1994

Aktionswoche
Innenstadt
Wohngebiet
Wohnen
autoarm
Zone
Siedlung
Farmsen
City
Modellversuch
Example 3: *Mann* vs. *Frau* in the DTA

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 the authors or the BBAW!

**Observations**

- fixed & formulaic expressions very prominent
  - *gnädige Frau*  
  - *Frau X geborene Y*  
  - *der gemeine Mann*  
- pretty much exclusively cultural bias:
  - *Mann* $\leadsto$ berühmt, ehrlich, gelehrt, tapfer, weise, ...  
  - *Frau* $\leadsto$ betrübt, lieb, schön, tugendreich, verwitwet, ...
- differences grow less pronounced in late 18\textsuperscript{th} & 19\textsuperscript{th} centuries
Example 3: Selected Lemma-Clouds

1725–1749:

lieb, groß, ander, gnädig, gemein, gebären, weise, gelehrt

1825–1849:

lieb, schön, edel, jung, groß, ander, gnädig, grau, gut, deutsch
Example 4: 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]/)"

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

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 4: Time Series \( (k = 10) \)

DiaCollo Profile

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

- Alkohol
- Bier
- Branntwein
- Kaffee
- Milch
- Schnaps
- Sekt
- Tee
- Wasser
- Wein
Summary & Outlook

Diachronic Collocation Profiling
- diachronic text corpora
- conventional tools
- diachronic profiling

DiaCollo
- on-the-fly corpus partitioning
- attribute-wise term indices
- “diff” profile mode
- DDC/D* integration
- RESTful web service

Future Work
- distributional semantic profiles
- cross-product visualizations
- ...and more!

〜 semantic shift, discourse trends
〜 implicit assumptions of homogeneity
〜 date-dependent lexemes

〜 arbitrary query granularity
〜 flexible result filtering
〜 direct comparison
〜 fine-grained queries, corpus KWIC links
〜 external API, online visualization

(Berry et al. 1995; Blei et al., 2003)
(Barnes & Hut 1986)
Thank you for listening!

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