Exploring the internal heterogeneity of a corpus of Classical French with DiaCollo

Bryan Jurish
Berlin-Brandenburgische Akademie der Wissenschaften
jurish@bbaw.de

Annette Gerstenberg
Freie Universität Berlin
annette.gerstenberg@fu-berlin.de

Global Philology Open Conference
Universität Leipzig
22nd February, 2016
Overview

The Situation
- Diachronic (heterogeneous) Text Corpora
- Collocation Profiling
- Diachronic Collocation Profiling

DiaCollo
- Requests & Parameters
- Profile, Diffs & Indices

APWCF Corpus
- Background
- Subcorpora
- Sources & Enrichments

Examples

Summary & Conclusion
The Situation: Diachronic Text Corpora

- heterogeneous text collections, especially with respect to *date of origin*
  - other partitionings may be relevant too, e.g. by genre, *location*, etc.

- increasing number available for linguistic & humanities research, e.g.
  - *Deutsches Textarchiv (DTA)* (Geyken 2013)
  - *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/ZET (Kohl ∼ politician vs. “cabbage”) (1946–2016)
  - DDR Presseportal (Ausreise ∼ “departure”) (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: Requests & Parameters

- Perl API, RESTful web-service *(Fielding 2000)* + web-form GUI
- 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 collocant 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>epoch granularity or “0” (zero) for a date-independent profile</td>
</tr>
<tr>
<td>groupby</td>
<td>projected collocate attributes with optional restrictions</td>
</tr>
<tr>
<td>score</td>
<td>association score function for collocate ranking</td>
</tr>
<tr>
<td>kbest</td>
<td>maximum number of collocate items to return per epoch</td>
</tr>
<tr>
<td>diff</td>
<td>binary score comparison operation for diff profiles</td>
</tr>
<tr>
<td>global</td>
<td>request global profile pruning (vs. default epoch-local pruning)</td>
</tr>
<tr>
<td>profile</td>
<td>profile type to be computed (<em>{native,tdf,ddc}</em> × <em>{unary,diff}</em> )</td>
</tr>
<tr>
<td>format</td>
<td>output format or visualization mode (e.g. TSV, JSON, HTML, d3-cloud, ...)</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 epoch-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 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 Association Score Functions

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

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

- **mi1** pointwise mutual information
  \[ \approx \log_2 \frac{f_{12} \times N}{f_1 \times f_2} \]

- **milf** 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)
  \[ \approx \text{sgn}(f_{12} | f_1, f_2) \times \log(1 + \log \lambda) \]

- **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
  \[ = \frac{1}{2}(s_a + s_b) \]

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

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

- **havg** harmonic average
  \[ \approx \frac{2 \times s_a \times s_b}{s_a + s_b} \]
**Classical French**: same structures as modern French, but
- linguistic norm is not yet stable
- variation and patterns of usage of grammatical features
- linguistic change on the levels of semantics and pragmatics

**Acta Pacis Westphalicae** (1643–1648): The French correspondence
- Diplomatic letters between Paris (government) and diplomats at Münster
- Ambassadors are committed to achieving diplomatic goals
  - convincing the government to adapt the instructions
- Diplomatic letters: formal constraints *versus* expressive needs

**Linguistic interest**
- **Diachronic variation**: comparison with existing resources of Classical French
- **Synchronic variation**: genre-internal heterogeneity

**Genre-internal heterogeneity**: hypothesis of different levels of formality
- Two subcorpora: ”government” (Paris) and ”ambassadors” (Münster)
- Register-variation reflected in the use of linguistic variables
APWCF: Correspondence

French Delegation: Longueville, D'Avaux, Servien

APWCF
Letters between Paris and Münster, 1643–1648

Queen Anne, Cardinal Mazarin, Brienne

10 days
APWCF: Subcorpora

GOVERNMENT

Monsieur
Monsieur sg., pl.

AMBASSADORS

Madame
Votre Majesté
Monseigneur
Votre Éminence

APWCF
Letters between Paris and Münster, 1643–1648
2 Subcorpora
"gov." vs. "amb."
APWCF: Data

- Letters mostly conserved in French Archives
  - Archives du Ministère des Affaires Etrangères
  - microforms: Zentrum für Historische Forschung, Bonn

- Digital edition (PDF, XML)
  - Bayerische Staatsbibliothek
  - Zentrum für historische Forschung Bonn

- Linguistic corpus: AG
  - Part-of-Speech Tagging (PRESTO, Cologne/Lyon)
  - XML / TXM (Lyon)

- Corpus size: 8 volumes of French edition, 2.4M Tokens
APWCF: Transcription

Diplomatic transcription, spelling variants preserved
- *traittés* vs. *traittéz* vs. *traittez*
- *estat* (old) or *état* (mod.) as appearing in the manuscript
- Punctuation almost preserved, but...

Modernized
- Some adaptations of punctuation
- *u/v, i/j* modernized
- Capitalization of proper names and titles
- Diacritics normalized (*lavis* → *l’avis*, *francais* → *français*)
- Abbreviated titles/words: full form
Examples
Example 1: *ledict* (chancellery style)

http://kaskade.dwds.de/dstar/apwcf/diacollo/?as=0&bs=0&p=d1&sf=mi1&f=html ...

**QUERY**: doc.loc=*Paris*               **SLICE**: 0

**QUERY**: doc.loc=*Münster*        **SLICE**: 0

**SCORE**: mi1                       **GROUPBY**: w=*ledict*

<table>
<thead>
<tr>
<th></th>
<th>Paris</th>
<th>Münster</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>2,462,443</td>
<td>2,462,443</td>
</tr>
<tr>
<td>$f_1$</td>
<td>746,786</td>
<td>1,153,939</td>
</tr>
<tr>
<td>$f_2 = f_{12}$</td>
<td>284</td>
<td>414</td>
</tr>
<tr>
<td><strong>score (mi1)</strong></td>
<td>1.721</td>
<td>1.093</td>
</tr>
<tr>
<td><strong>diff (Paris - Münster)</strong></td>
<td><strong>0.6278</strong></td>
<td></td>
</tr>
</tbody>
</table>

- simple example uses “unigrams” comparison profile ($f_2 = f_{12}$)
- pointwise mutual information (mi1) score function
- “Paris” shows definite preference the archaic form *ledict* (chancellery style)
Example 2: **PLAIRE** (situational context)

```
http://kaskade.dwds.de/dstar/apwcf/diacollo/?as=0&bs=0&p=d1&sf=ll&f=cloud ...
QUERY: doc.loc=Paris SLICE: 0
∼QUERY: doc.loc=Münster KBEST: 25
SCORE: 11 GROUPBY: w,l=PLAIRE
```

- situational context: lemma *PLAIRE*, e.g. *s’il vous plaît* (“please”)
- log-likelihood association scores (robust)
- attribute-cloud visualization: warm colors ∼ Paris, cool colors ∼ Münster
Example 3: *plaît* vs. *plaist* (orthographic variation)

http://kaskade.dwds.de/dstar/apwcf/diacollo/?as=0&bs=0&p=d1&sf=ll&f=cloud ... 
QUERY: doc.loc=Paris  SLICE: 0
~QUERY: doc.loc=Münster  SCORE: 11
GROUPBY: w=plaît|plaist

- zoom: “*plaist/plaît de ’s’il vous plaît*”
- more frequent in Münster
- relatively more frequent use of the archaic variant *plaist*
- transcription in general respects orthographic variation
  - typically not transparent in historical editions
Example 4: **POUVOIR** (request & response)

- **Paris**: request “you could” (*puissiez/pourrez*)
- **Münster**: response “we could/will be able” (*pouvions/pourrons*)

---

http://kaskade.dwds.de/dstar/apwcf/diacollo/?as=0&bs=0&p=d1&sf=ll&f=cloud ...  
QUERY: doc.loc=Paris  SLICE: 0  
~QUERY: doc.loc=Münster  KBEST: 25  
SCORE: 11  GROUPBY: w,1=POUVOIR
Example 5: Speech Acts

- Diplomatic negotiations: overt speech act verbs
- **Paris**: discuter, discuté, escrivez, escrivois, ...
- **Münster**: comuniquer, escrivons, escrivismes, parlasmes, parlèrent, ...

http://kaskade.dwds.de/dstar/apwcf/diacollo/?as=0&bs=0&p=d1&sf=11&f=cloud ... 

QUERY: doc.loc=Paris KBEST: 50 
~QUERY: doc.loc=Münster SCORE: 11 
GROUPBY: w,l={ESCRIRE|ÉCRIRE|PARLER|COMMUNICQUER|COMMUNIQUER|DISCUTER}
Summary & Conclusion

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

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

APWCF + DiaCollo
- metadata-based filtering
  \(\sim\) location-specific profiles
- “diff” profile mode
  \(\sim\) inter-subcorpus comparisons
- metadata-based aggregation
  \(\sim\) subcorpus preference profiles
Thank you for listening!

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

APWCF: http://wikis.fu-berlin.de/pages/viewpage.action?pageId=594936338