DiaCollo: on the trail of diachronic collocations
Bryan Jurish
Berlin-Brandenburg Academy of Sciences and Humanities

Abstract
DiaCollo is a new software tool for the efficient extraction, comparison, and interactive visualization of collocations from a diachronic text corpus. Unlike other conventional collocation extractors, DiaCollo is suitable for extraction and analysis of diachronic collocation data: to colocate pairs whose association strength depends on the date of their occurrence. By tracking changes in a word’s typical collocates over time, DiaCollo can help to provide a clearer picture of diachronic changes in the word’s usage, especially those related to semantic shift or discourse environment.

The Situation
Diachronic Text Corpora
- heterogeneous text collections
  - especially with respect to date of origin
- increasing number available, e.g.
  - Deutsches Textarchiv (DTA) [4]
  - Historical American English (COHA) [2]
- even putatively “synchronous” corpora have a nontrivial temporal extension [8]

Collocation Profiling
“Shall you know a word by the company it keeps” – J. R. Firth
- find “significant” collocates of a target term
- rank candidates by association score
- filter out “chance” co-occurrences
- statistical methods require large sample
- existing methods [1, 3, 7] implicitly assume corpus homogeneity

Diachronic Profiling

Idea
- represent terms as attribute n-tuples
  - including document date!
- partition term vocabulary on-the-fly
  - user-specified epochs
- collect epoch profiles into final result-set

Advantages
- full support for diachronic axis
- variable query-level granularity
- flexible attribute selection

Drawbacks
- sparse data requires larger corpora
- computationally expensive
- large index size

Implementation
Interfaces
- Perl API & command-line utilities
- RESTful web-service plugin + GUI

Features
- scalable even in a high-load environment
  - no persistent server process required
- index access via file I/O or mmap () syscall
- supports both unary and “diff” profiles
- full DDC query support via ddc back-end

Output & Visualization
- TSV, JSON, HTML, Highcharts, d3-cloud, …

References

Example 1: Krise (“crisis”) in the weekly Die Zeit (1946–2014)

- 1990: NATO Pershing-II missiles in western Europe
- 1980–1989: Solidarność & martial law in Poland
- 1990: Collapse of Helmut Schmidt (SPD) coalition
- 1990s: AEG sells consumer electronics division
  - subsequent takeover by Daimler-Benz AG
- 2010–2014: civil wars in Ukraine & Syria
- 2010–2014: German FDP loses Bundestag presence

Example 2: 400 Years of Potables (1600–1999)

Remarks
- DDC back-end + GermaNet [5, 6] expansion
- fine-grained search for beverages in object position of verb trinken (“to drink”)

Observations
- staples ~ constants, e.g.
  - Bier, Milch, Wasser (“beer, milk, water”)
- 1650–1750: Tee, Kaffee (“tea, coffee”) appear
- 1800–1900: Schnaps displaces Branntwein
- 1850–1900: Alkohol (“alcohol”) as a beverage

Example 3: Gender Bias (1600–1900)

- comparison profile: Mann (“man”) vs. Frau (“woman”)
  - node size indicates absolute association score difference
- fixed & formulaic expressions very prominent
  - gnädige Frau (“milady”) → masculine: gnädiger Herr
  - Frau X geborene Y (“born”) → birth- vs. married surname
  - der gemeine Mann (“common”) → masculine generic
- historical corpus data can reveal persistent cultural biases
  - Mann ~ berühmt, ehrlich, gelehrt, . . . ("famous, honest, learned, . . .")
  - Frau ~ lieb, schön, verwitwet, . . . ("dear, beautiful, widowed, . . .")
- differences grow less pronounced in late 18th & 19th centuries
- political discourse: deutsch, eigen, frei ("German, own, free")

References