

Visualizing Semantic Change with DiaCollo

Bryan Jurish

Berlin-Brandenburg. Akademie der Wissenschaften
jurish@bbaw.de

Thomas Werneke

Zentrum für Zeithistorische Forschung
werneke@zzf-potsdam.de

Concepts in the World: Politics, Knowledge, and Time
20th International Conference on Conceptual History
University of Oslo

23rd September, 2017



Overview

The Situation

- Diachronic Text Corpora
- Collocation Profiling
- Diachronic Collocation Profiling

DiaCollo

- Requests & Parameters
- Profiles, Diffs & Indices

Applications & Examples

- Application Scenarios
- Modular Language in the GDR Press

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 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/ZEIT (“Kohl”) (1946–2016)
 - ▶ DDR Presseportal (“Ausreise”) (1945–1993)
 - ▶ DWDS/Blogs (“Browser”) (1994–2016)
- 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

*“Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache”
 (‘The meaning of a word is its use in the language’)*
 — L. Wittgenstein

“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)
- “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)
 - ▶ *Die Grenzboten* (1841–1921, 308 volumes, 88M tokens)
 - ▶ *Deutsches Textarchiv* (1600–1900, 3.6K documents, 205M tokens)
 - ▶ *DDR-Presseportal* (1945–1994, 4.1M documents, 1.3G tokens)
 - ▶ *DWDS Zeitungen* (1946–2016, 10M documents, 4.7G 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: Profiles, Diffs & Indices

Profiles & Diffs

- simple request → unary **profile** for collocant(s)
 - ▶ **filtered** & **projected** to selected attribute(s)
 - ▶ **aggregated** into independent slice-wise sub-intervals
 - ▶ **trimmed** to k -best collocates for target word(s)

*(profile, query)
(groupby)
(date, slice)
(score, kbest, global)*
- 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

*(profile, bquery, ...)
(diff)
(query \neq bquery)
(e.g. date \neq bdate)*

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
- “live” KWIC-links to underlying corpus hits ⇒ **DDC search engine**
- **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	$= f_{12}$	
■ If	collocation log-frequency	$= \log_2(f_{12} + \varepsilon)$	
■ mi	pointwise MI \times log-frequency	$\approx \log_2 \frac{f_{12} \times N}{f_1 \times f_2} \times \log_2 f_{12}$	
■ ll	log-likelihood (Dunning 1993)	$\approx \text{sgn}(f_{12} f_1, f_2) \times \log \frac{L(H_0)}{L(H_1)}$	
■ 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	$\approx \frac{2s_a s_b}{s_a + s_b}$	

Application Scenarios

Exploration & Discovery

- detection of temporal (historical) phenomena
- fast & flexible “distant reading” over **large digital text collections**
- various intuitive visualizations of query results
 - ▶ e.g. interactive tag-clouds, bubble charts, time series plots, ...

Analysis & Investigation

- data acquisition for hypothesis testing
 - ▶ TSV format ↵ spreadsheet import, R's `read.delim()`
- first step towards “close” hermeneutic analysis
 - ▶ KWIC-links ↵ corpus hits (approximation) for a given collocation pair

(Quantitative) Evaluation & Assessment

- at least to the extent that “the numbers can speak for themselves”
 - ▶ beware corpus- and preprocessing-artefacts!

Examples

Example 1: Newsworthy Crises

'Krise' in DIE ZEIT (west) and Neues Deutschland (east)

http://kaskade.dwds.de/dstar/zeit/diacollo/?q=Krise&d=1950:*&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

2010–2016

- 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:



2010–2016:



Example 2: Gender & Cultural Bias

'Mann' vs. 'Frau' in the Deutsches Textarchiv (1600–1900)

<http://kaskade.dwds.de/dstar/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 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 2: Selected Lemma-Clouds

1725–1749:



1825–1849:



Example 3: Modular Language in the GDR Press

'Wohl' in the Neues Deutschland (ND) and the Berliner Zeitung (BZ)

(*Neues Deutschland*) <http://kaskade.dwds.de/dstar/nd/diacollo/?q=Wohl...>

(*Berliner Zeitung*) http://kaskade.dwds.de/dstar/bz_pp/diacollo/?q=Wohl...

KBEST: 25

GROUPBY: 1, p=NN

Typical Noun-Lemma Collocates of *Wohl* ("welfare", "well-being")

- **Familie** ("family")
- **Frieden** ("peace")
- **Glück** ("happiness")
- **Mensch** ("person", "human being", "humankind")
- **Volk** ("people", "nation")
- **Welt** ("world")



Example 3: Selected Lemma-Clouds

(BZ vs. ND, 1980–1989)

ND:



BZ:



Example 3: Selected Lemma-Clouds

(aggregated BZ+ND, 1950s vs. 1980s)



Example 3: Selected Corpus Hits

(Neues Deutschland)

- “Wir schlugen zum **Wohl** der **Menschen**, für das **Glück** des **Volkes**, im Interesse der Arbeiterklasse und aller Werktätigen den Kurs der Einheit von Wirtschafts- und Sozialpolitik ein.” (ND, 1989-08-12)

*“We set a course for the **well-being** of **humankind**, for the **happiness** of the **people**, . . .”*
- “Beide Seiten unterstrichen, daß ein gesicherter **Frieden** eine entscheidende Voraussetzung für soziale Sicherheit, für das **Glück** und das **Wohl** der arbeitenden **Menschen** in der **Welt** ist.” methodical (ND, 1989-06-28)

*“. . . **peace** . . . for the **happiness** and **well-being** of the working **persons** in the **world**”*
- “. . . das planvolle Wirken für das **Glück** der **Familie** und das **Wohl** unserer ganzen Gesellschaft; die Sicherheit, sozial geborgen zu sein” (ND, 1988-09-24)

*“. . . work for the **happiness** of the **family** and the **well-being** of our whole society . . .”*
- “Sie können gewiß sein, daß mit den wachsenden Anforderungen an den militärischen Dienst alles, was jeweils notwendig und möglich ist für das **Wohl** der Soldaten des **Volkes** und für das **Glück** ihrer **Familien**, für das **Glück** ihrer Kinder und Kindeskinder getan wird.” (ND, 1987-10-27)

*“. . . the **well-being** of the soldiers of the **people** and for the **happiness** of their **families**, for the **happiness** of their children, and their childrens' children”*



Example 3: Selected Corpus Hits

- “So dient unser bewährter Kurs der Einheit von Wirtschafts- und Sozialpolitik auch in dieser Hinsicht dem **Wohl** und dem **Glück** der arbeitenden **Menschen** und ist bis in die **Familie** spürbar.” (BZ, 1988-03-24)

*“Our course serves . . . the **well-being** and **happiness** of working **persons** and can even be felt in the **family**”*

- “Ich bitte Sie, mit mir das Glas zu erheben und zu trinken auf weitere Fortschritte bei der Schaffung einer **Welt** ohne Kernwaffen, auf den **Frieden** und das **Glück** aller **Völker**, auf Ihr **Wohl**, verehrte Gäste, liebe Freunde und Genossen.” (BZ, 1988-06-23)

*“Please raise your glasses . . . to the creation of a **world** without nuclear weapons, to **peace** and the **happiness** of all **peoples**, to your **well-being**, honored guests, dear friends and comrades.”*

- “Unsere gemeinsamen Ziele — und dafür setzen wir täglich unser Wissen und Können ein — sind das **Wohl** des **Volkes**, das **Glück** der **Menschen**, eine Zukunft in **Frieden** auf der **Welt**.” (BZ, 1988-10-08)

*“Our mutual goals . . . are the **well-being** of the **people**, the **happiness** of **humankind**, a future in **peace** on the **world**”*

Example 4: 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")
(Hamp & Feldweg 1997; Lemnitzer & Kunze 2007; Henrich & Hinrichs 2010)
- considers only those target terms immediately preceding verb **trinken** ("to drink")
- "global" profile uses shared target-set to avoid visual clutter

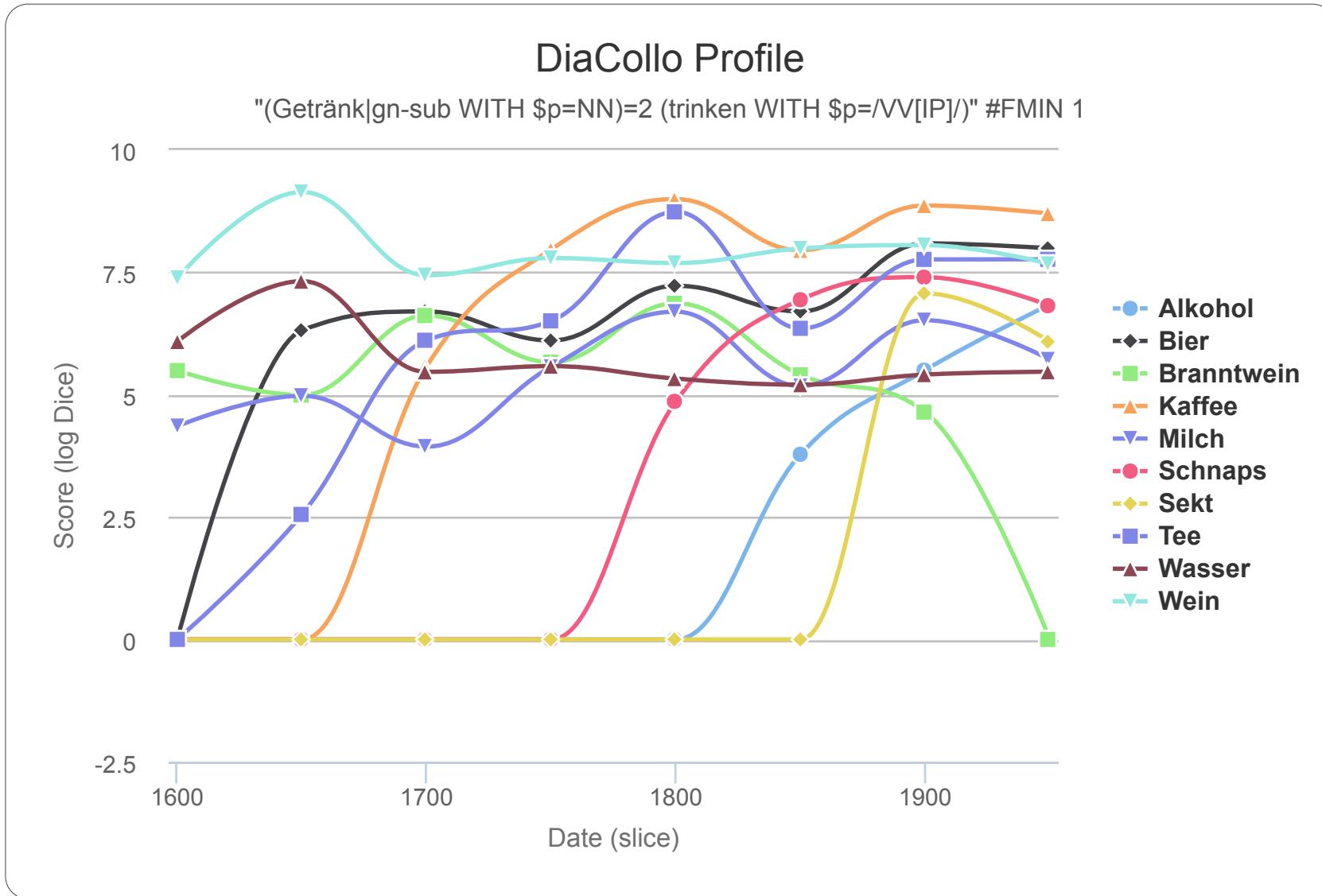
Observations

- 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")

[inspiration: C. Thomas]



Example 4: Time Series ($k = 10$)



Summary & Conclusion I

Diachronic Collocation Profiling with DiaCollo

Diachronic Collocation Profiling

- “meaning is use” $\rightsquigarrow \dots \text{in temporal context}$
- diachronic text corpora $\rightsquigarrow \text{semantic shift, discourse trends}$
- conventional tools $\rightsquigarrow \text{implicit assumptions of homogeneity}$
- diachronic profiling $\rightsquigarrow \text{date-dependent lexemes}$

DiaCollo

- on-the-fly corpus partitioning $\rightsquigarrow \text{arbitrary query-level granularity}$
- DDC/D* integration $\rightsquigarrow \text{fine-grained queries, corpus KWIC links}$
- RESTful web service $\rightsquigarrow \text{external API, online visualization}$

Summary & Conclusion II

DiaCollo's Application in the field of Historical Semantics

Advantages

- support for semiasiological methods
- inductive and explorative methods
- fluent shift from distant reading to close reading (and back)
- new digital approaches for conceptual history

Problems & Issues

- onomasiological properties are not (currently) recoverable
 - ▶ e.g. by comparing word clouds
- digital corpora (sources):
 - ▶ quantity (or lack thereof)
 - ▶ quality (metadata, format, preprocessing)
 - ▶ legal aspects, copyright issues

Mildly Pontifical Remarks (“Outlook”)

- distant reading will eventually require us to read **more** rather than **less**
- in the end, we may have to read more sources than ever before!



— *The End* —



Thank you for listening!

<http://kaskade.dwds.de/~jurish/diacollo2017>

<http://metacpan.org/release/DiaColloDB>

More Examples

Example 5: Lexicography

'autofrei' (automobile-free)

<http://kaskade.dwds.de/dstar/zeitungen/diacollo/?q=autofrei&ds=5&f=bub>

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")

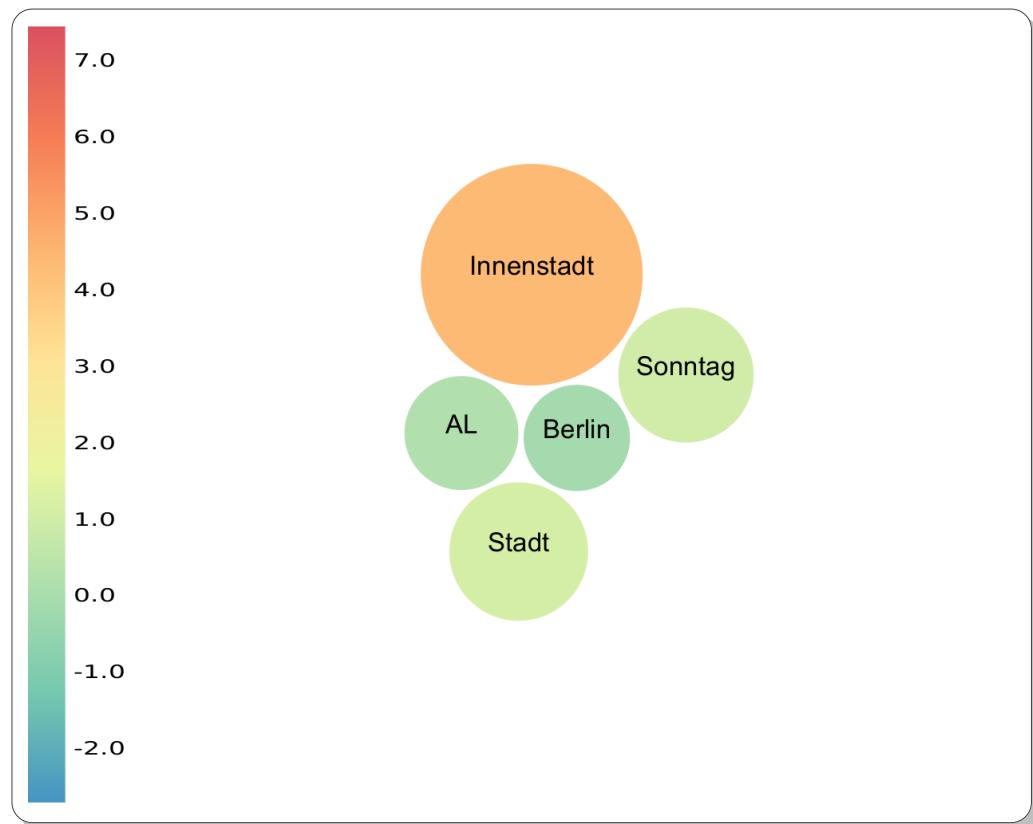
- Duden: *keinen Autoverkehr aufweisend* ("lacking automobile traffic")
- DWDS corpora reveal ***two sub-senses***:
 - ▶ **1970–1989**: ... by ordinance (↔ *Sonntag, Innenstadt*)
 - ▶ **1990–present**: ... voluntary (↔ *Wohnanlage, Siedlung*)

[source: A. Geyken]

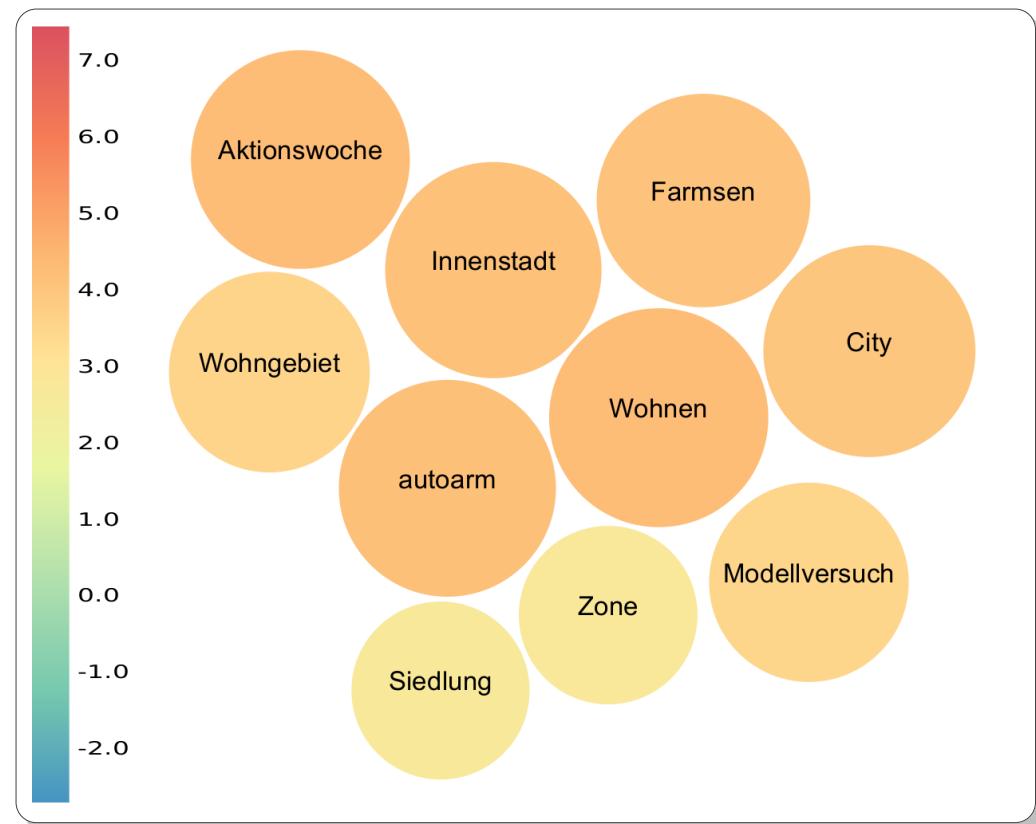


Example 5: Selected Bubble-Charts

1985–1989



1990–1994



well, you know . . .

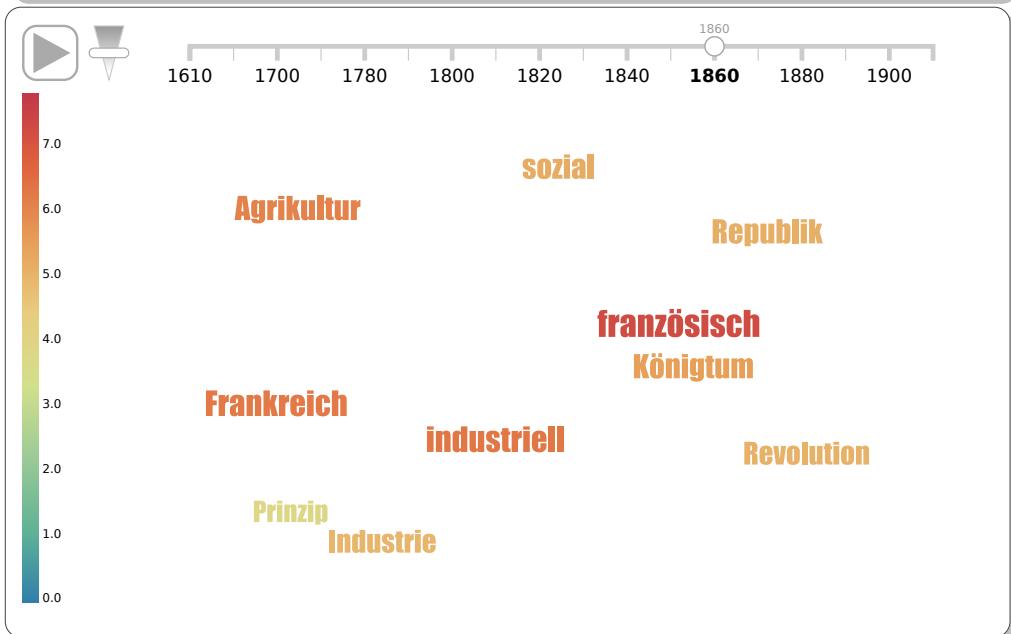
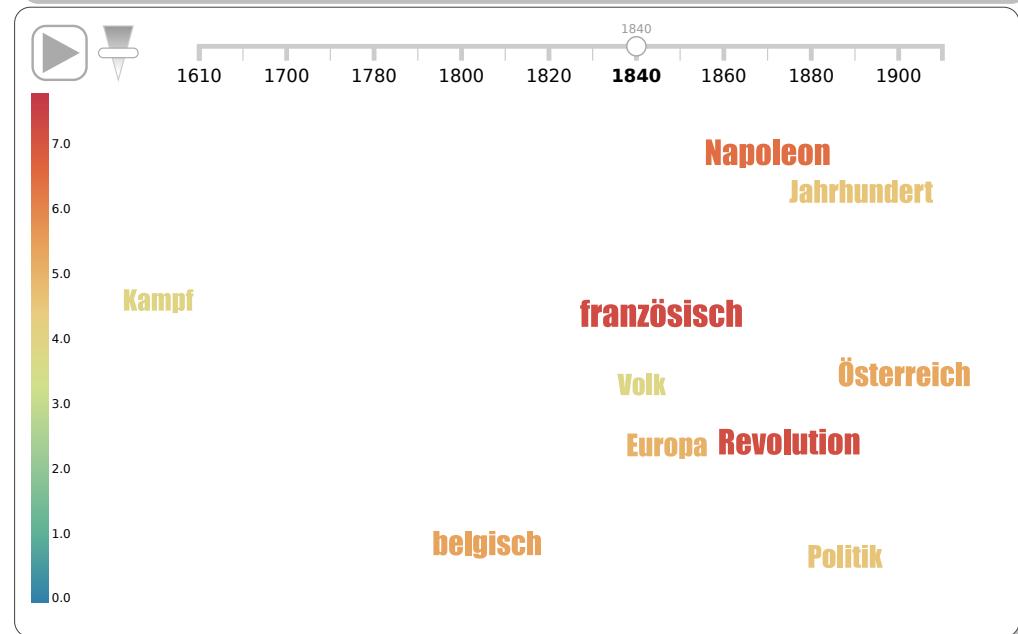
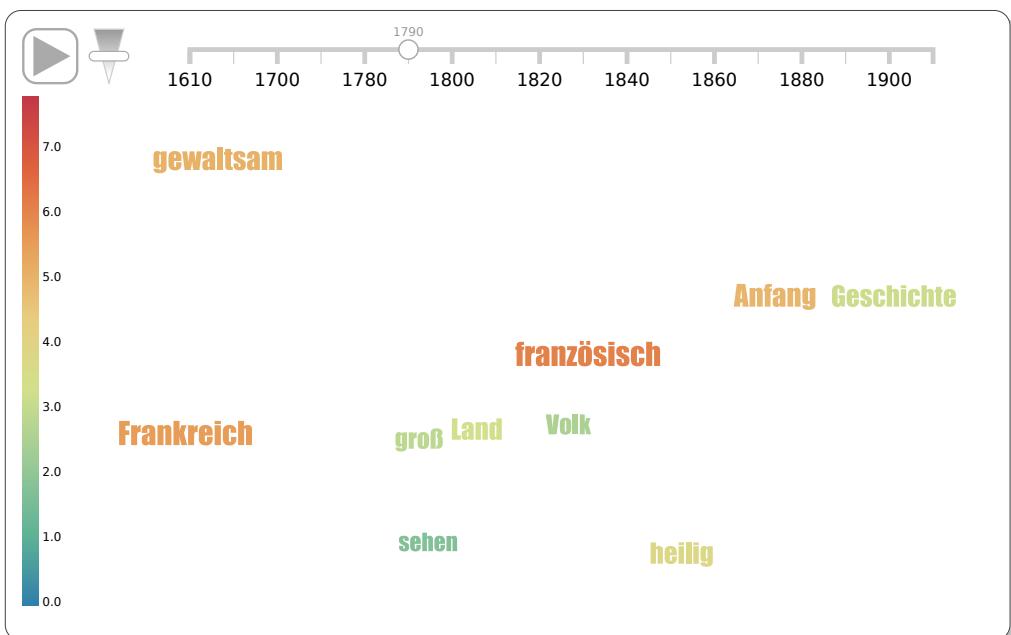
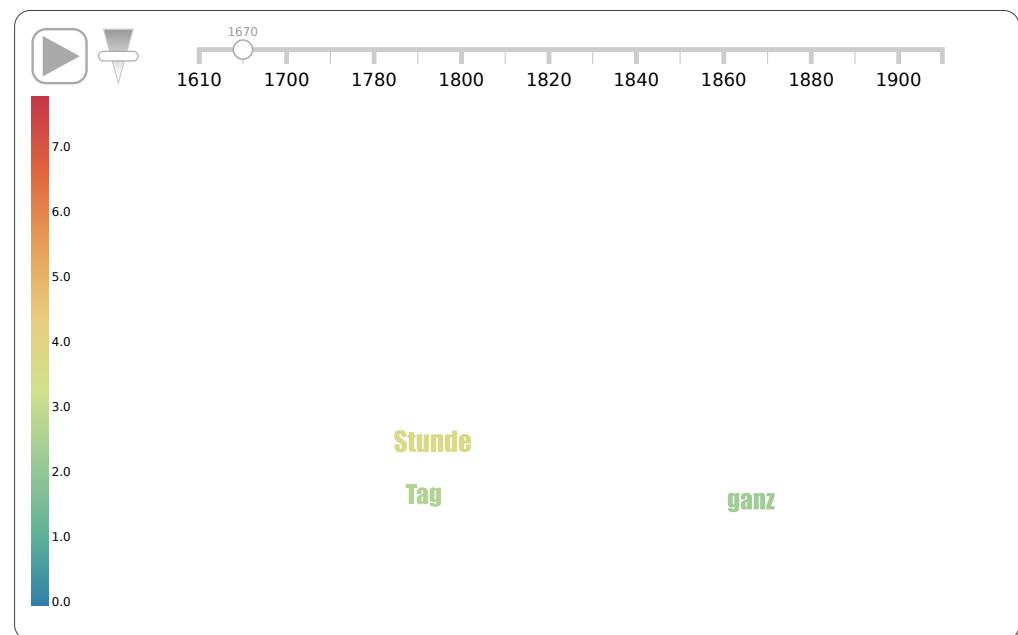
<http://kaskade.dwds.de/dstar/dta/diacollo/?q=Revolution&ds=10&f=cloud>

- < 1770: only ‘rotation’ sense
 - ▶ *ganz, Stunde, Tag* (“entire, hour, day”)
- ≥ 1770: ‘dramatic change’
 - ▶ *menschlich* (“human”)
- ≥ 1790: French Revolution
 - ▶ *französisch, Frankreich* (“French, France”)
- ≥ 1840: violent political upheaval (Napoleonic era in Europe)
 - ▶ *Napoleon, Europa*
- ≥ 1860: industrial revolution
 - ▶ *Industrie, industriell* (“industry, industrial”)
- **Bonus:** “colorful”  (DDC + GermaNet)

[source: L. Lemnitzer, A. Geyken, J. Lennon, P. McCartney]



Example 6: Selected Lemma-Clouds



Example 7: What Makes a ‘Man’?

[ADJA] Mann' in the Deutsches Textarchiv (1600–2000)

<http://kaskade.dwds.de/dstar/dta/diacollo/?profile=diff-ddc&k=25&f=cloud> ...

```
QUERY: "*=2 Mann" #has [textClass, Wissenschaft*]
~QUERY: "*=2 Mann" #has [textClass, Belletristik*]
GROUPBY: l,p=ADJA
```

Remarks

- ‘diff’ profile provides direct comparison of genres **science** vs. **belles lettres**
- uses **DDC** back-end for fine-grained data acquisition

Differences (diff=adiff)

- **Science** ↪ *berühmt, scharfsinnig, tüchtig* (“famous, astute, capable”)
- **Belles Lettres** ↪ *brav, grau, rechtschaffen* (“well-behaved, gray, righteous”)

Similarities (diff=min)

- *groß, gelehrt, gemein, jung, alt* (“great, learned, common, young, old”)

Example 7: Selected Lemma-Clouds

1700–1799
(diff=adiff)



1800–1899
(diff=adiff)



Example 8: 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: l, 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 8: Lemma-Clouds

differences
(diff=adiff)



similarities
(diff=havg)



Example 9: Pronominal Adverbs by Genre

[PAV]’ in aggregated DTA+DWDS (1600–2000)

<http://kaskade.dwds.de/dstar/dta+dwds/diacollo/?p=diff-ddc&k=50&f=cld&G=1> ...

QUERY: \$p=PAV=2 #has [textClass, **Wissenschaft***]
 ~QUERY: \$p=PAV=2 #has [textClass, **Belletristik***]

Remarks

- ‘diff’ profile provides direct comparison of genres **science** vs. **belles lettres**
- uses **DDC** back-end for querying functional category

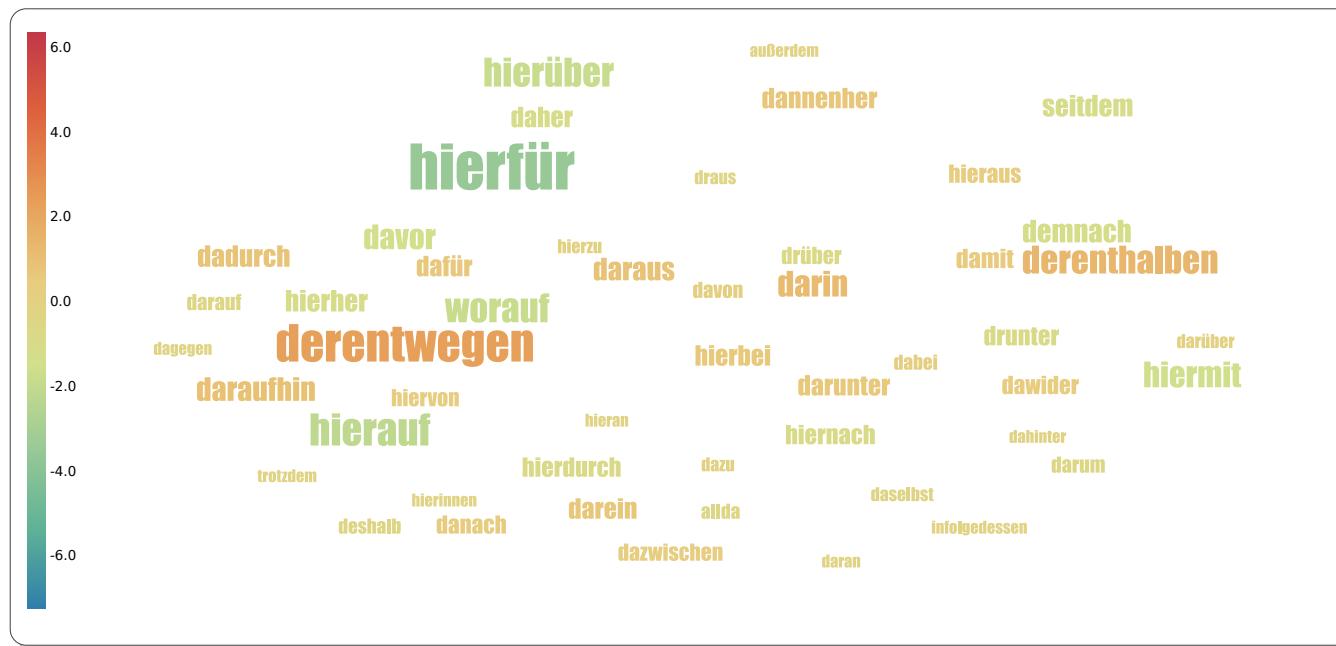
Observations

- divergent: differences grow more pronounced over time
- **Science**
 - ▶ *hier-* anaphorics ↵ *hierbei*, *hieraus*, *hierzu* (“hereby, out of which, to which”)
 - ▶ causal/logical ↵ *dennach*, *infolgedessen*, *daher* (“therefore”)
- **Belles Lettres**
 - ▶ fixed expression *drunter [und] drüber* (“higgledy-piggeldy, at sixes and sevens”)
 - ▶ spatial & temporal ↵ *dahinter*, *worauf* (“behind which, upon which”)
 - ▶ concessive & adversative ↵ *dawider*, *trotzdem* (“against which, despite which”)



Example 9: Selected Lemma-Clouds

1650–1699:



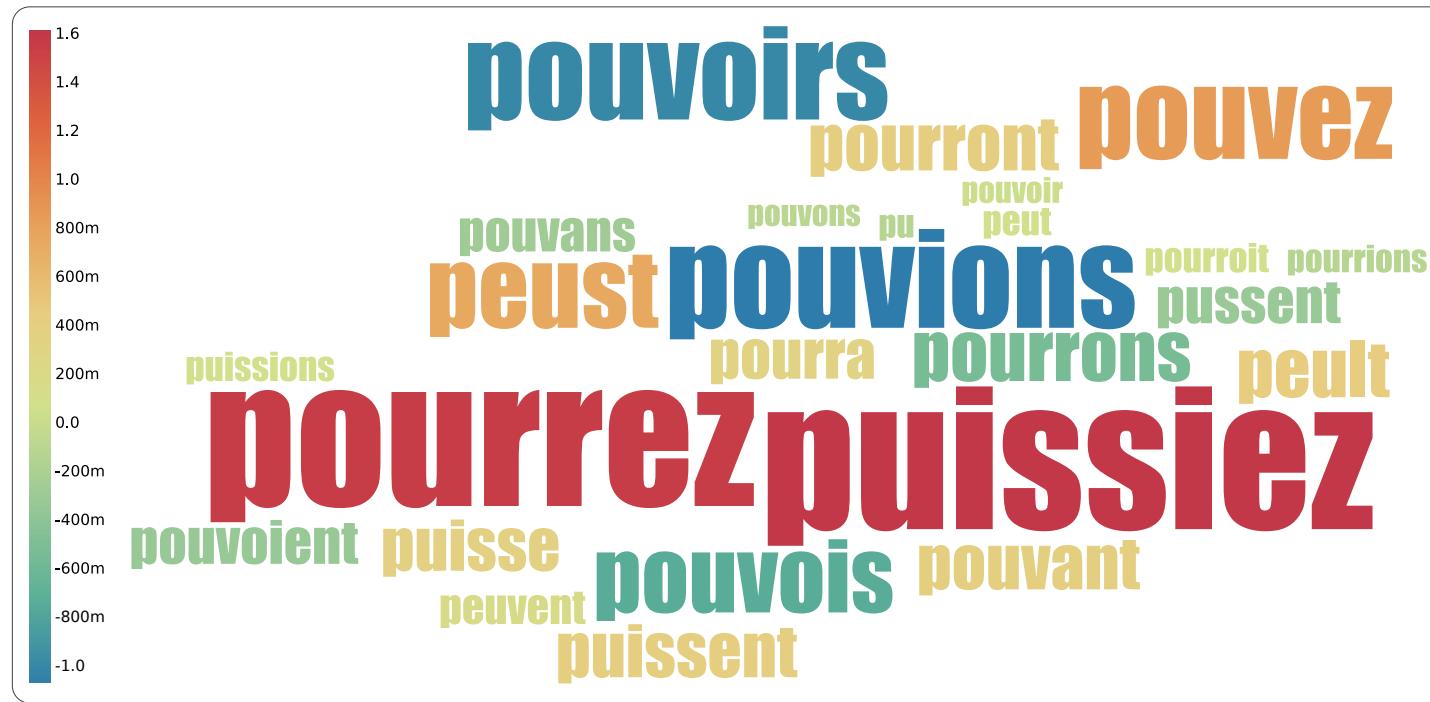
1950–1999:



Example 10: APWCF ‘POUVOIR’ (request & response) CLARIN-D

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

QUERY: doc.loc=Paris SLICE: 0
~QUERY: doc.loc=Münster KBEST: 25
SCORE: 11 GROUPBY: w,l=POUVOIR



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

[source: A. Gerstenberg]



Example 11: Speech Acts

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

QUERY: doc.loc=**Paris** KBEST: 50

~QUERY: doc.loc=**Münster** SCORE: 11

GROUPBY: w,l=ESCRIRE|ÉCRIRE|PARLER|COMMUNICQUER|COMMUNIQUER|DISCUTER



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

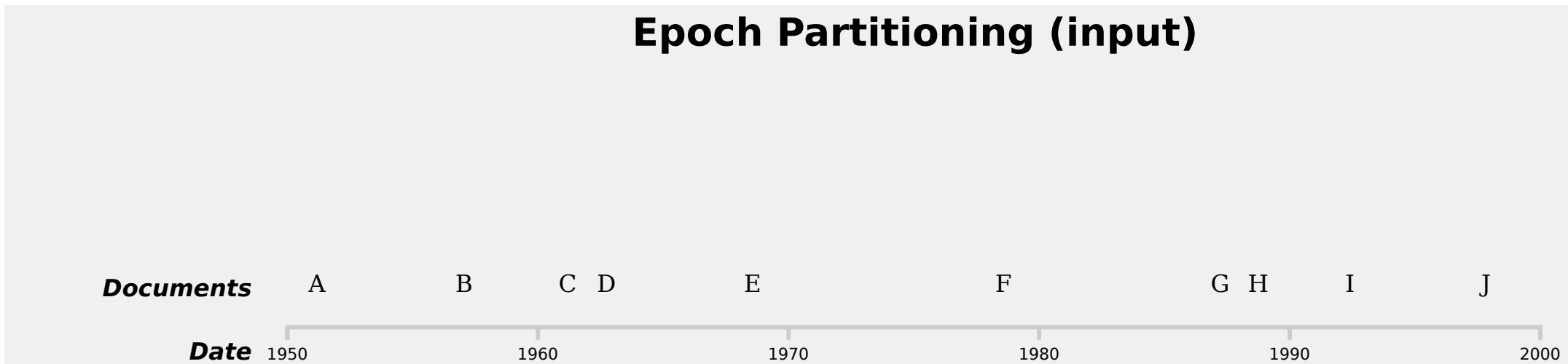
[source: A. Gerstenberg]



Gory Details

Manual Corpus Partitioning

Epoch Partitioning (input)

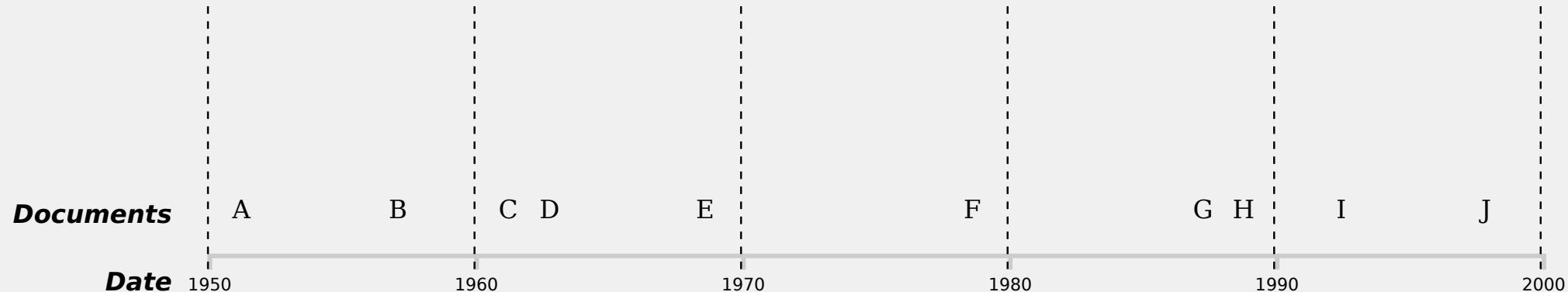


- input corpus with documents $\{A, B, \dots, J\}$ over date range (1950–1999)



Manual Corpus Partitioning

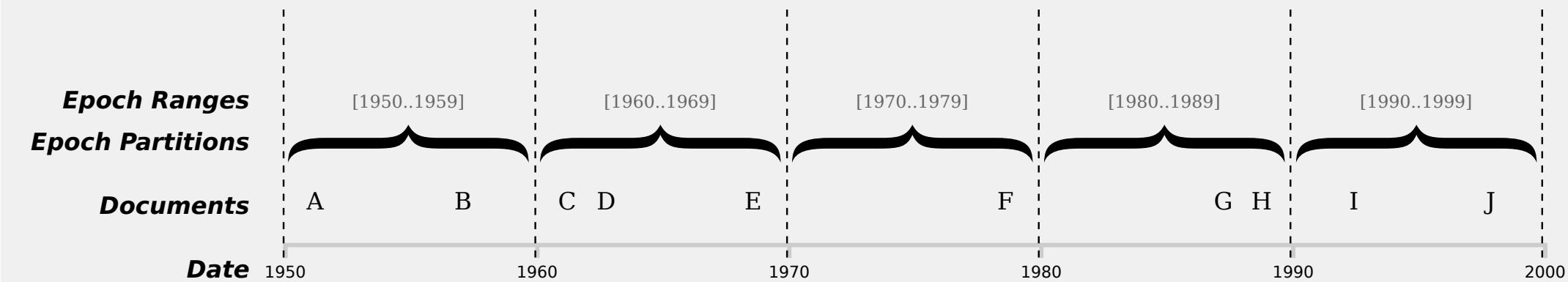
Epoch Partitioning ($E=10$)



- input corpus with documents $\{A, B, \dots, J\}$ over date range (1950–1999)
- partition by decade ($E = 10$)

Manual Corpus Partitioning

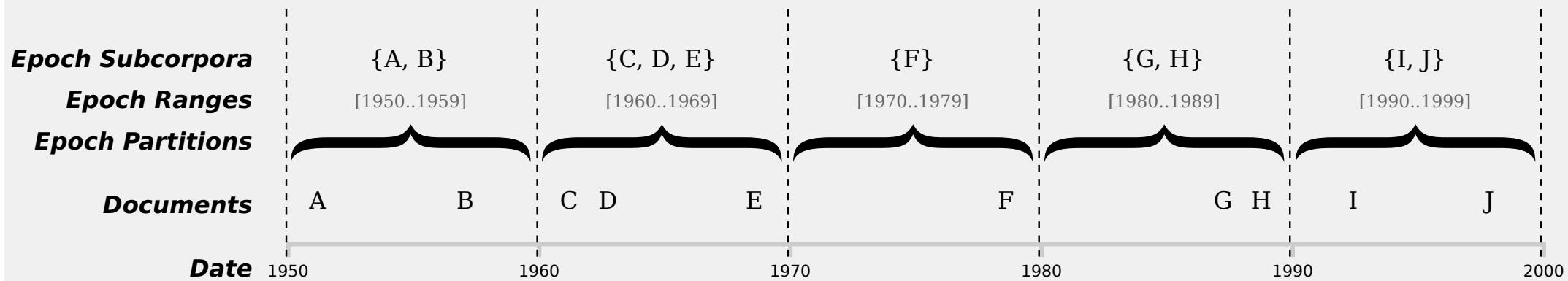
Epoch Partitioning ($E=10$)



- input corpus with documents $\{A, B, \dots, J\}$ over date range (1950–1999)
- partition by decade ($E = 10$)

Manual Corpus Partitioning

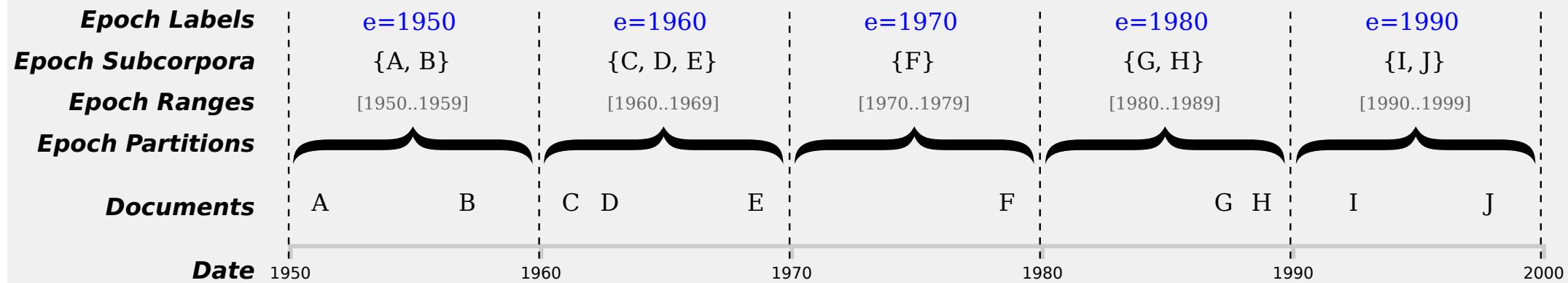
Epoch Partitioning ($E=10$)



- input corpus with documents $\{A, B, \dots, J\}$ over date range (1950–1999)
- partition by decade ($E = 10$)
- collect epoch-wise subcorpora

Manual Corpus Partitioning

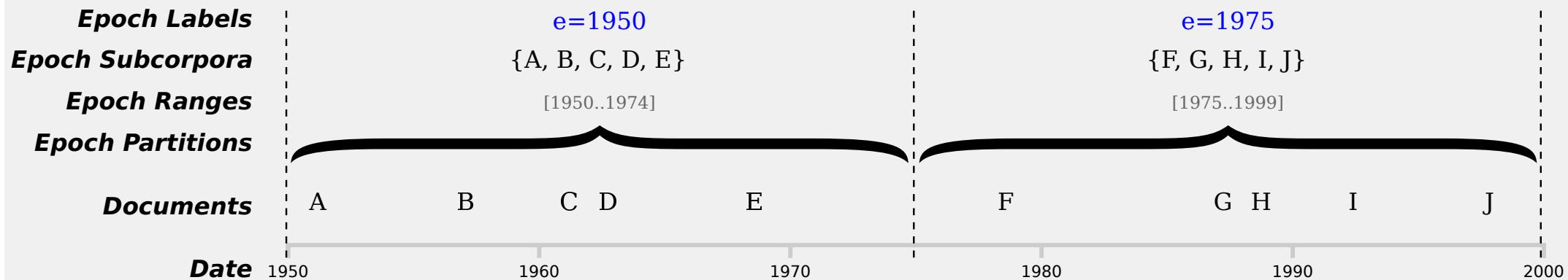
Epoch Partitioning ($E=10$)



- input corpus with documents $\{A, B, \dots, J\}$ over date range (1950–1999)
- partition by decade ($E = 10$)
- collect epoch-wise subcorpora
- label sub-corpora (e.g. by minimum date) and analyze independently

Manual Corpus Partitioning

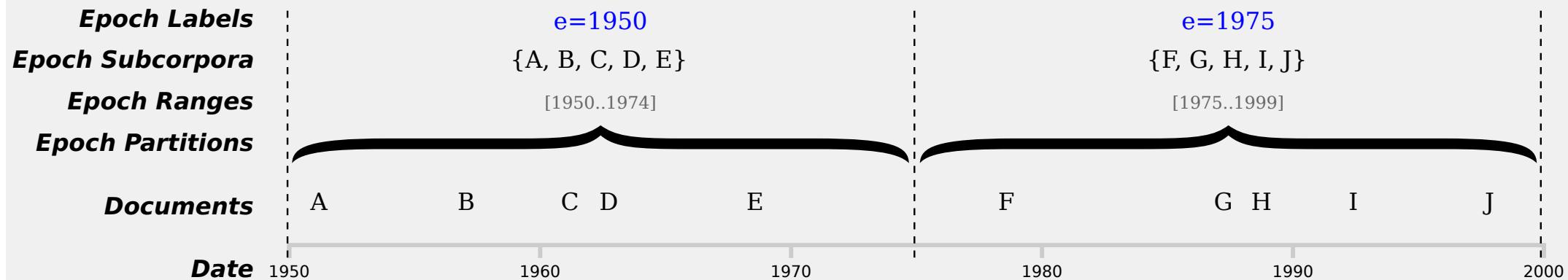
Epoch Partitioning ($E=25$)



- input corpus with documents $\{A, B, \dots, J\}$ over date range (1950–1999)
- partition by ~~decade~~ **quarter-century** ($E = 25$)
- collect epoch-wise subcorpora
- label sub-corpora (e.g. by minimum date) and analyze independently
- **Problems:**
 - ▶ static partitioning \rightsquigarrow labor-intensive, inflexible, & often inaccessible
 - ▶ “good” epoch granularity (partition size) depends on research question
- *can we generalize this?*

Manual Corpus Partitioning

Epoch Partitioning ($E=25$)



- input corpus with documents $\{A, B, \dots, J\}$ over date range (1950–1999)
- partition by ~~decade~~ **quarter-century** ($E = 25$)
- collect epoch-wise subcorpora
- label sub-corpora (e.g. by minimum date) and analyze independently
- **Problems:**
 - ▶ static partitioning \rightsquigarrow labor-intensive, inflexible, & often inaccessible
 - ▶ “good” epoch granularity (partition size) depends on research question
- *can we generalize this?*



...

Corpus Indexing

Input Corpus

- abstract input class `DiaColloDB::Document`
 - ▶ currently supported sub-classes: `DDCTabs`, `JSON`, `TCF`, `TEI`
- input corpus must be **pre-tokenized** and **pre-annotated**
 - ▶ user-defined token-attribute selection
 - ▶ D* project uses attributes Lemma and PoS (“part-of-speech”)
- may include user-defined **break markers**
 - ▶ e.g. clause-, sentence-, page-, and/or paragraph-boundaries

Content Filtering

- not all corpus types are “interesting”
 - ▶ e.g. closed classes, *hapax legomena*, etc.
- Regular expression & frequency filters used to pre-prune corpus, e.g.
 - ▶ `-0=wbad=REGEX` : surface form blacklist regex
 - ▶ `-0=pgood=REGEX` : PoS whitelist regex
 - ▶ `-tfmin=REQ` : minimum global term-tuple frequency
 - ▶ `-lfmin=REQ` : minimum global lemma frequency
 - ▶ `-cfmin=REQ` : minimum co-occurrence frequency

Basic Definitions

Corpus Data

- a corpus \mathcal{C} is list of N tokens t_i
- each token is an n_A -tuple of attribute values
- each token is associated with a unique non-negative integer date (year)

$$\mathcal{C} = t_1 t_2 \dots t_N$$

$$t_i \in \mathcal{A}_1 \times \dots \times \mathcal{A}_{n_A}$$

$$Y(t_i) \in \mathbb{N}$$

Corpus Domain

- lexical domain (term vocabulary)
- temporal domain (dates)

$$\mathcal{W} = \bigcup_{i=1}^N \{t_i\} \subseteq \mathcal{A}_1 \times \dots \times \mathcal{A}_{n_A}$$

$$\mathcal{Y} = \bigcup_{i=1}^N \{Y(t_i)\} \subset \mathbb{N}$$

Common Notation

- attribute projection
- ... for attribute-lists
- equivalence classes

$$t[j] = a_j \text{ for } t = \langle a_1, \dots, a_n \rangle$$

$$t[J] = \langle t_{j_1}, \dots, t_{j_{n_J}} \rangle \text{ for } J = \langle j_1, \dots, j_{n_J} \rangle$$

$$[u]_{T/J} = \{t \in T \mid t[J] = u\} \subseteq T$$

Runtime Data: Requests and Profiles

DiaCollo Request

runtime user input parameters:

- q a collocant selection expression (query)
- $E \in \mathbb{N}$ the target epoch size (slice)
- $G \in \langle g_1, g_2, \dots, g_{n_G} \rangle$ the collocate attributes to project (groupby)
- $H : \mathcal{Y} \times \mathcal{W}[G] \rightarrow \{0, 1\}$ a filter function (date, groupby)
- $\varphi : \mathbb{R}^4 \rightarrow \mathbb{R}$ an association score function (score)
- $k \in \mathbb{N}$ the maximum number of collocates per epoch (kbest)

$$Q = \langle q, E, G, H, \varphi, k \rangle$$

Raw Co-occurrence Frequency Profile

computation basis, for $\mathcal{E} \subset \mathbb{N}$ a finite set of corpus epochs:

$$R_Q = \langle r_N, r_1, r_2, r_{12} \rangle$$

- $r_N : \mathcal{E} \rightarrow \mathbb{N}$ the total number of corpus co-occurrences by epoch (N)
- $r_1 : \mathcal{E} \rightarrow \mathbb{N}$ independent collocant frequency by epoch (f1)
- $r_2 : \mathcal{E} \times \mathcal{W}[G] \rightarrow \mathbb{N}$ independent collocate frequency by epoch (f2)
- $r_{12} : \mathcal{E} \times \mathcal{W}[G] \rightarrow \mathbb{N}$ co-occurrence frequencies by epoch (f12)



Native Co-occurrence Relation: Indexing

(“*collocations*” profile type)

- “co-occurrence” \rightsquigarrow moving window over $\ell \in \mathbb{N}$ content tokens
- window never crosses selected **break boundaries** (e.g. sentences)
- 3-level index maps “lexical” tuple pairs to date-dependent co-frequencies for (filtered) corpus $C = s_1 \dots s_{n_S}$ of break-units (“sentences”) $s_i = t_{i1} \dots t_{in_{s_i}}$,

$$I_{12} : \mathcal{W} \rightarrow (\mathcal{W} \rightarrow (\mathcal{Y} \rightarrow \mathbb{N}))$$

$$: \langle w, v, y \rangle \mapsto \sum_{i=1}^{n_S} \sum_{j=1}^{n_{s_i}} \sum_{d=-\ell}^{\ell} \mathbf{1}[d \neq 0 \ \& \ t_{ij} = w \ \& \ t_{i(j+d)} = v \ \& \ Y(t_{ij}) = y]$$

- **Beware:** compile-time filters (pgood, tfmin, etc.) influence index content!

► **cfmin option** prunes by co-frequency

$$f(w, v, y) < f_{\text{cfmin}} \Rightarrow I_{12}(w, v, y) = 0$$

- independent “frequencies” $I_1(w, y)$, $I_N(y)$ computed as true marginals:

$$I_1 : \mathcal{W} \times \mathcal{Y} \rightarrow \mathbb{N} : \langle w, y \rangle \mapsto \sum_{v \in \mathcal{W}} I_{12}(w, v, y)$$

$$I_N : \mathcal{Y} \rightarrow \mathbb{N} : y \mapsto \sum_{w \in \mathcal{W}} I_1(w, y)$$

$$\ell = 3$$

Native Co-occurrence Relation: Context Window

Input Text The fat cat sat on the fuzzy cat .

Input Text	The	fat	cat	sat	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.

Input Text	The	fat	cat	sit	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-

Input Text	The	fat	cat	sit	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	

Input Text	The	fat	cat	sit	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		t_1	t_2	t_3			t_4	t_5	

Input Text	The	fat	cat	sat	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		t_1	$\langle t_2 \rangle$	t_3			t_4	t_5	

$$\begin{array}{l} j=1 \\ d=1 \end{array} \rightsquigarrow I_{12} = \left\{ \langle \text{fat}, \text{cat} \rangle \mapsto 1 \right\}$$

Input Text	The	fat	cat	sat	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		t_1	t_2	$\langle t_3 \rangle$			t_4	t_5	

$$\begin{array}{l} j=1 \\ d=2 \end{array} \rightsquigarrow I_{12} = \left\{ \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1 \right\}$$

Input Text	The	fat	cat	sit	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		t_1	t_2	t_3			$\langle t_4 \rangle$	t_5	

$j=1$
 $d=3$

 $\rightsquigarrow I_{12} = \left\{ \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1, \langle \text{fat}, \text{fuzzy} \rangle \mapsto 1 \right\}$

Input Text	The	fat	cat	sit	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		$\langle t_1 \rangle$	t_2	t_3			t_4	t_5	

$j=2$
 $d=-1$

 $\rightsquigarrow I_{12} = \left\{ \begin{array}{l} \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1, \langle \text{fat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{fat} \rangle \mapsto 1 \end{array} \right\}$

Input Text	The	fat	cat	sit	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		t_1	t_2	$\langle t_3 \rangle$			t_4	t_5	

$j=2$
 $d=1$

 $\rightsquigarrow I_{12} = \left\{ \begin{array}{l} \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1, \langle \text{fat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{fat} \rangle \mapsto 1, \langle \text{cat}, \text{sit} \rangle \mapsto 1 \end{array} \right\}$

Input Text	The	fat	cat	sat	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		t_1	t_2	t_3			$\langle t_4 \rangle$	t_5	

$j=2$
 $d=2$

 $\rightsquigarrow I_{12} = \left\{ \begin{array}{l} \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1, \langle \text{fat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{fat} \rangle \mapsto 1, \langle \text{cat}, \text{sit} \rangle \mapsto 1, \langle \text{cat}, \text{fuzzy} \rangle \mapsto 1 \end{array} \right\}$

Input Text	The	fat	cat	sat	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		t_1	t_2	t_3			t_4	$\langle t_5 \rangle$	

$$\begin{array}{l} j=2 \\ d=3 \end{array}$$

 \rightsquigarrow

$$I_{12} = \left\{ \begin{array}{l} \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1, \langle \text{fat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{fat} \rangle \mapsto 1, \langle \text{cat}, \text{sit} \rangle \mapsto 1, \langle \text{cat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{cat} \rangle \mapsto 1 \end{array} \right\}$$

Input Text	The	fat	cat	sat	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		$\langle t_1 \rangle$	$\langle t_2 \rangle$	t_3			$\langle t_4 \rangle$	$\langle t_5 \rangle$	

$j=3$
 $d=*$

$$I_{12} = \left\{ \begin{array}{l} \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1, \langle \text{fat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{fat} \rangle \mapsto 1, \langle \text{cat}, \text{sit} \rangle \mapsto 1, \langle \text{cat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{cat} \rangle \mapsto 1, \langle \text{sit}, \text{fat} \rangle \mapsto 1, \underline{\langle \text{sit}, \text{cat} \rangle \mapsto 2}, \\ \langle \text{sit}, \text{fuzzy} \rangle \mapsto 1 \end{array} \right\}$$

Input Text	The	fat	cat	sat	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		$\langle t_1 \rangle$	$\langle \underline{t}_2 \rangle$	$\langle t_3 \rangle$			$\boxed{t_4}$	$\langle \underline{t}_5 \rangle$	

$j=4$
 $d=*$

\rightsquigarrow

$$I_{12} = \left\{ \begin{array}{l} \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1, \langle \text{fat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{fat} \rangle \mapsto 1, \langle \text{cat}, \text{sit} \rangle \mapsto 1, \langle \text{cat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{cat} \rangle \mapsto 1, \langle \text{sit}, \text{fat} \rangle \mapsto 1, \langle \text{sit}, \text{cat} \rangle \mapsto 2, \\ \langle \text{sit}, \text{fuzzy} \rangle \mapsto 1, \langle \text{fuzzy}, \text{fat} \rangle \mapsto 1, \langle \text{fuzzy}, \text{cat} \rangle \mapsto 2, \\ \langle \text{fuzzy}, \text{sit} \rangle \mapsto 1 \end{array} \right\}$$

Input Text	The	fat	cat	sat	on	the	fuzzy	cat	.
Input Lemma	the	fat	cat	sit	on	the	fuzzy	cat	.
Filter	the	fat	cat	sit	on	the	fuzzy	cat	-
Content		fat	cat	sit			fuzzy	cat	
Tokens		t_1	$\langle t_2 \rangle$	$\langle t_3 \rangle$			$\langle t_4 \rangle$	t_5	

$j=5$
 $d=*$

\rightsquigarrow

$$I_{12} = \left\{ \begin{array}{l} \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1, \langle \text{fat}, \text{fuzzy} \rangle \mapsto 1, \\ \langle \text{cat}, \text{fat} \rangle \mapsto 1, \langle \text{cat}, \text{sit} \rangle \mapsto 2, \langle \text{cat}, \text{fuzzy} \rangle \mapsto 2, \\ \langle \text{cat}, \text{cat} \rangle \mapsto 2, \langle \text{sit}, \text{fat} \rangle \mapsto 1, \langle \text{sit}, \text{cat} \rangle \mapsto 2, \\ \langle \text{sit}, \text{fuzzy} \rangle \mapsto 1, \langle \text{fuzzy}, \text{fat} \rangle \mapsto 1, \langle \text{fuzzy}, \text{cat} \rangle \mapsto 2, \\ \langle \text{fuzzy}, \text{sit} \rangle \mapsto 1 \end{array} \right\}$$

Native Co-occurrence Relation: Runtime

Given a user-supplied query request $Q = \langle q, E, G, H, \varphi, k \rangle$

- find collocant tuple(s) $\llbracket q \rrbracket$, e.g. $\llbracket \$\text{lemma}=\text{love} \rrbracket = [\text{love}]_{\mathcal{W}/a_{\text{lemma}}}$

$$\llbracket q \rrbracket \subseteq \mathcal{W}$$

- raw index lookup:

$$\hat{I}_q : \mathcal{Y} \times \mathcal{W} \rightarrow \mathbb{N} : \langle y, v \rangle \mapsto \sum_{w \in \llbracket q \rrbracket} I_{12}(w, v, y)$$

- group collocates by attributes G :

$$\hat{I}_{q,G} : \mathcal{Y} \times \mathcal{W}[G] \rightarrow \mathbb{N} : \langle y, g \rangle \mapsto \sum_{v \in [g]_{\mathcal{W}/G}} \hat{I}_q(y, v)$$

- apply request filter restrictions H :

$$\hat{I}_{q,G,H} = \hat{I}_{q,G} \upharpoonright \text{ext}(H) : \mathcal{Y} \times \mathcal{W}[G] \rightarrow \mathbb{N}$$

- aggregate by epoch E :

$$\hat{I}_{q,G,H,E} : \mathcal{E}_E \times \mathcal{W}[G] \rightarrow \mathbb{N} : \langle e, g \rangle \mapsto \sum_{y \in [e]_E} \hat{I}_{q,G,H}(y, g)$$

► where $\tilde{E} : \mathcal{Y} \rightarrow \mathbb{N} : y \mapsto E \lfloor \frac{y}{E} \rfloor$; $\mathcal{E}_E = \tilde{E}(\mathcal{Y})$; $[e]_E = \tilde{E}^{-1}(e)$

- finalize raw frequency profile $R_Q = \langle r_N, r_1, r_2, r_{12} \rangle$

$$r_N(e) = \sum_{y \in [e]_E} I_N(y)$$

$$r_1(e) = \sum_{y \in [e]_E} \sum_{w \in \llbracket q \rrbracket} I_1(w, y)$$

$$r_2(e, g) = \sum_{y \in [e]_E} \sum_{v \in [g]_{\mathcal{W}/G}} I_1(v, y)$$

$$r_{12}(e, g) = \hat{I}_{q,G,H,E}(e, g)$$

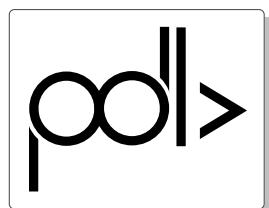
► 2-pass lookup strategy required for accurate independent collocate frequencies r_2



TDF Co-occurrence Relation: Indexing

(“*term* × *document matrix*” profile type)

- “co-occurrence” ↪ anywhere within the selected break unit (“document”)
- relatively coarse index granularity (no proximity constraints)
- for corpus partitioned into documents $\text{Doc} = \{d_1, \dots, d_{n_D}\}$, store:
 - ▶ sparse term-document frequency matrix tdf : $\mathcal{W} \times \text{Doc} \rightarrow \mathbb{N}$
 - ▶ date counts $\text{yf} : \mathcal{Y} \rightarrow \mathbb{N} : y \mapsto \sum_{w \in \mathcal{W}} \sum_{d \in \text{dy}^{-1}(y)} \text{tdf}(w, d)$
 - ▶ document dates and bibliographic metadata dy : Doc → Y
- occurrence date, bibliographic metadata stored as ***document properties***
- index uses `mmap()` on sparse matrix **PDL** via **PDL::CCS::Nd**
 - ▶ transparent on-demand paging from disk
 - ▶ fast numerical manipulation of large N-dimensional data arrays
- optimized lookup using Harwell-Boeing offset vectors
- supports Boolean query expressions and document metadata attributes



TDF Co-occurrence Relation: Runtime

- interpret collocant query $\textcolor{red}{q}$ independently as:

▶ set of terms $[\![\textcolor{red}{q}]\!]_{\mathcal{W}}$

$$[\![\textcolor{red}{q}]\!]_{\mathcal{W}} \subseteq \mathcal{W}$$

▶ set of documents $[\![\textcolor{red}{q}]\!]_{\text{Doc}}$

$$[\![\textcolor{red}{q}]\!]_{\text{Doc}} \subseteq \text{Doc}$$

- index lookup with collocate grouping:

$$\hat{I}_{\text{tdf:}\textcolor{red}{q}, \textcolor{blue}{G}} : \mathcal{Y} \times \mathcal{W}[\textcolor{blue}{G}] \rightarrow \mathbb{N}$$

$$: \langle \textcolor{green}{y}, \textcolor{blue}{g} \rangle \mapsto \sum_{d \in [\![\textcolor{red}{q}]\!]/\textcolor{green}{y}} \min \left\{ \left(\sum_{\textcolor{red}{w} \in [\![\textcolor{red}{q}]\!]_{\mathcal{W}}} \text{tdf}(\textcolor{red}{w}, d) \right), \left(\sum_{\textcolor{blue}{v} \in [\![\textcolor{blue}{g}]\!]_{\mathcal{W}/\textcolor{blue}{G}}} \text{tdf}(\textcolor{blue}{v}, d) \right) \right\}$$

▶ where $[\![\textcolor{red}{q}]\!]/\textcolor{green}{y} = [\![\textcolor{red}{q}]\!]_{\text{Doc}} \cap \text{dy}^{-1}(\textcolor{green}{y})$

- candidate filtering and epoch aggregation as for native index

- final raw frequency profile $R_Q = \langle r_N, r_1, r_2, r_{12} \rangle$

$$r_N(\textcolor{green}{e}) = \sum_{\textcolor{green}{y} \in [\![\textcolor{green}{e}]\!]_E} \text{yf}(\textcolor{green}{y})$$

$$r_1(\textcolor{green}{e}) = \sum_{\textcolor{green}{y} \in [\![\textcolor{green}{e}]\!]_E} \sum_{\textcolor{red}{w} \in [\![\textcolor{red}{q}]\!]_{\mathcal{W}}} \sum_{d \in [\![\textcolor{red}{q}]\!]/\textcolor{green}{y}} \text{tdf}(\textcolor{red}{w}, d)$$

$$r_2(\textcolor{green}{e}, \textcolor{blue}{g}) = \sum_{\textcolor{green}{y} \in [\![\textcolor{green}{e}]\!]_E} \sum_{\textcolor{blue}{v} \in [\![\textcolor{blue}{g}]\!]_{\mathcal{W}/\textcolor{blue}{G}}} \sum_{d \in \text{dy}^{-1}(\textcolor{green}{y})} \text{tdf}(\textcolor{blue}{v}, d)$$

$$r_{12}(\textcolor{green}{e}, \textcolor{blue}{g}) = \hat{I}_{\text{tdf:}\textcolor{red}{q}, \textcolor{blue}{G}, H, \textcolor{green}{E}}(\textcolor{green}{e}, \textcolor{blue}{g})$$



DDC Co-occurrence Relation

- “co-occurrence” \rightsquigarrow as returned by a **DDC** search engine query (“*ddc*” profile type)
 - ▶ requires a running DDC search engine server for the appropriate corpus
- query subscripts (“match-IDs”) identify collocant (=1) and collocates (=2)
- supports full range of the **DDC query language**, including:
 - ▶ user-specified **break collections** (e.g. sentence, file, paragraph)
 - ▶ **break-** and **token-level** Boolean query expressions
 - ▶ **phrase-** and **proximity-queries**
 - ▶ **bibliographic metadata filters**
 - ▶ server-side **term expansion pipelines**
- ***most flexible*** back-end yet implemented, but ***comparatively slow***
- generated raw frequency profile $R_Q = \langle r_N, r_1, r_2, r_{12} \rangle$

$$r_N = \lambda_q \times \text{COUNT}(* \ #\text{SEP}) \ #\text{BY}[\text{date}/E]$$

$$r_1 = \lambda_q \times \text{COUNT}(\text{KEYS}([[q \& H]] \ #\text{SEP} \ #\text{BY}[G=1]) \ #\text{SEP}) \ #\text{BY}[\text{date}/E]$$

$$r_2 = \lambda_q \times \text{COUNT}(\text{KEYS}([[q \& H]] \ #\text{SEP} \ #\text{BY}[G=2]) \ #\text{SEP}) \ #\text{BY}[\text{date}/E, G=2]$$

$$r_{12} = \text{COUNT}([[q \& H]] \ #\text{SEP} \ #\text{BY}[\text{date}/E, G=2])$$

- ▶ $[[q \& H]]$ a DDC query with optional collocate restrictions
- ▶ $\lambda_q \in \mathbb{N}$ a query-dependent scaling coefficient
- ▶ **server-side pre-pruning** via optional **#FMIN f_{cfmin}** query operator

Scoring & Pruning: Basics

- φ maps raw frequency profiles to scalar association scores

$$\varphi : \mathbb{R}^4 \rightarrow \mathbb{R}$$

- score profiles $p_{Q,e}$ computed independently for each epoch $e \in \mathcal{E}_E$:

$$p_{Q,e} : \mathcal{W}[G] \rightarrow \mathbb{R} : g \mapsto \varphi(r_N(e), r_1(e), r_2(e, g), r_{12}(e, g))$$

- k -best pruning within each epoch:

$$\hat{p}_{Q,e} = p_{Q,e} \upharpoonright \text{best}_k(p_{Q,e})$$

- ▶ “global” profiles prune by global corpus association score:

$$\hat{p}_{Q,e:\text{global}} = p_{Q,e} \upharpoonright \text{best}_k(p_{Q_{[0/E]}, e})$$

- ▶ alternative: user-specified cutoff threshold

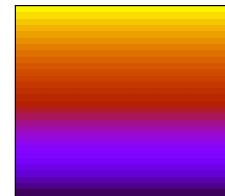
- final diachronic profile maps epoch-labels to epoch-local profiles:

$$\hat{P}_Q : \mathcal{E}_E \rightarrow \mathbb{R}^{\mathcal{W}[G]} : e \mapsto \hat{p}_{Q,e}$$

Score Functions: f (raw frequency)

Variable	Description
w_1	target tuple (“collocant”) matching the user query request (q)
w_2	collocate tuple matching the user groupby request (G)
N	total number of co-occurrences in the epoch $r_N(e)$
f_1	epoch-local frequency of the collocant term: $r_1(e)$
f_2	epoch-local frequency of the collocate term: $r_2(e, w_2)$
f_{12}	epoch-local frequency of the collocation pair: $r_{12}(e, w_2)$
ε	smoothing constant, by default $\frac{1}{2}$

$$\varphi_f(w_1, w_2) = f_{12}$$

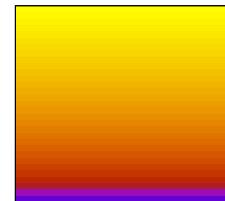


- immediately interpretable, but not very robust
- Zipf distribution leads to “lopsided” visualizations
- values may not be comparable across slices (e.g. for non-balanced corpora)
- many false positives with high-frequency collocates
- **not** generally a good measure of collocate affinity

Score Functions: If (log frequency)

Variable	Description
w_1	target tuple (“collocant”) matching the user query request (q)
w_2	collocate tuple matching the user groupby request (G)
N	total number of co-occurrences in the epoch $r_N(e)$
f_1	epoch-local frequency of the collocant term: $r_1(e)$
f_2	epoch-local frequency of the collocate term: $r_2(e, w_2)$
f_{12}	epoch-local frequency of the collocation pair: $r_{12}(e, w_2)$
ε	smoothing constant, by default $\frac{1}{2}$

$$\varphi_{\text{If}}(w_1, w_2) = \log_2(f_{12} + \varepsilon)$$

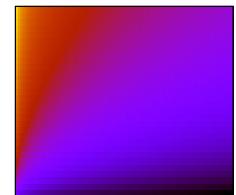


- better visual scaling than raw frequency
- otherwise shares raw frequency's shortcomings

Score Functions: mi (pointwise MI \times log-frequency)

Variable	Description
w_1	target tuple (“collocant”) matching the user query request (q)
w_2	collocate tuple matching the user groupby request (G)
N	total number of co-occurrences in the epoch $r_N(e)$
f_1	epoch-local frequency of the collocant term: $r_1(e)$
f_2	epoch-local frequency of the collocate term: $r_2(e, w_2)$
f_{12}	epoch-local frequency of the collocation pair: $r_{12}(e, w_2)$
ε	smoothing constant, by default $\frac{1}{2}$

$$\varphi_{\text{mi}}(w_1, w_2) = \log_2 \frac{(f_{12} + \varepsilon) \times (N + \varepsilon)}{(f_1 + \varepsilon) \times (f_2 + \varepsilon)} \times \log_2 (f_{12} + \varepsilon)$$

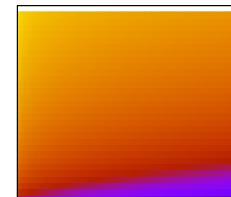


- used by first version of Sketch Engine *(Kilgarriff et al. 2004)*
- PMI gives code-length change for (optimal) joint vs. independent encodings
- PMI alone is very sensitive to low-frequency items (\rightsquigarrow longer codes)
 - ▶ *post-hoc* workaround: include log-frequency coefficient
- some preference for low-frequency collocates remains

Score Functions: II (log-likelihood)

Variable	Description
w_1	target tuple (“collocant”) matching the user query request (q)
w_2	collocate tuple matching the user groupby request (G)
N	total number of co-occurrences in the epoch $r_N(e)$
f_1	epoch-local frequency of the collocant term: $r_1(e)$
f_2	epoch-local frequency of the collocate term: $r_2(e, w_2)$
f_{12}	epoch-local frequency of the collocation pair: $r_{12}(e, w_2)$
ε	smoothing constant, by default $\frac{1}{2}$

$$\varphi_{ll}(w_1, w_2) = \text{sgn}(f_{12}|f_1, f_2) \times \log(1 + \log \lambda)$$



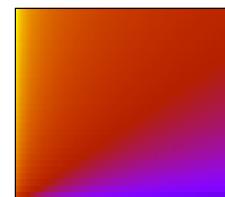
$$\log \lambda = \log \frac{L(H_0)}{L(H_1)} = f_{12} \log \frac{f_{12}^N}{f_1 f_2} + f_{12} \log \frac{f_{12}^N}{f_1 f_2} + f_{12} \log \frac{f_{12}^N}{f_1 f_2} + f_{12} \log \frac{f_{12}^N}{f_1 f_2}$$

- 1-sided variant of the binomial log likelihood ratio (*Dunning 1993; Evert 2008*)
 - ▶ only “attracting” collocate pairs are assigned positive values
- null hypothesis H_0 filters out “uninteresting” high-frequency collocates
- very sensitive to fixed & formulaic expressions \rightsquigarrow **poor visual scaling**
 - ▶ workaround: report & scale using $\log(1 + \log \lambda)$ rather than “pure” $\log \lambda$

Score Functions: Id (log-Dice coefficient)

Variable	Description
w_1	target tuple (“collocant”) matching the user query request (q)
w_2	collocate tuple matching the user groupby request (G)
N	total number of co-occurrences in the epoch $r_N(e)$
f_1	epoch-local frequency of the collocant term: $r_1(e)$
f_2	epoch-local frequency of the collocate term: $r_2(e, w_2)$
f_{12}	epoch-local frequency of the collocation pair: $r_{12}(e, w_2)$
ε	smoothing constant, by default $\frac{1}{2}$

$$\varphi_{\text{Id}}(w_1, w_2) = 14 + \log_2 \frac{2(f_{12} + \varepsilon)}{(f_1 + \varepsilon) + (f_2 + \varepsilon)}$$



- “lexicographer-friendly” association score *(Rychlý 2008)*
- less susceptible to low-frequency outliers than $\text{PMI} \times \log\text{-frequency product}$
- good filtering of “uninteresting” high-frequency collocates
- “intuitive” visual scaling (consistent with human perceptual givens)
- default score used by DiaCollo

Comparison Profiles (“Diffs”)

- **Idea:** compare two independently acquired queries Q_a and Q_b
 - ▶ comparison operation (diff)

$$\ominus : \mathbb{R}^2 \rightarrow \mathbb{R}$$

- epoch alignment (1:1, $n:1$, or $1:m$)

$$\mathcal{E}_{a \bowtie b} \subseteq \mathcal{E}_{E_a} \times \mathcal{E}_{E_b}$$

- apply by epoch $p_{Q_a \ominus Q_b, e_{ab}} : \text{Dom}_{Q_a \ominus Q_b / e_{ab}} \rightarrow \mathbb{R} : g \mapsto p_{Q_a, e_a}(g) \ominus p_{Q_b, e_b}(g)$

▶ $e_{ab} = \langle e_a, e_b \rangle \in \mathcal{E}_{a \bowtie b}$ an aligned epoch pair

▶ $\text{Dom}_{Q_a \ominus Q_b / e_{ab}} \subseteq \text{dom}(p_{Q_a, e_a}) \cup \text{dom}(p_{Q_b, e_b})$ characteristic for \ominus at e_{ab} :

- “pre-trimmed” operations

$$= \text{dom}(\hat{p}_{Q_a, e_a}) \cup \text{dom}(\hat{p}_{Q_b, e_b})$$

- “restricted” operations

$$= \text{dom}(p_{Q_a, e_a}) \cap \text{dom}(p_{Q_b, e_b})$$

- prune and collect

$$\hat{p}_{Q_a \ominus Q_b, e_{ab}} = p_{Q_a \boxminus Q_b, e_{ab}} \upharpoonright \text{best}_k(p_{Q_a \ominus Q_b, e_{ab}})$$

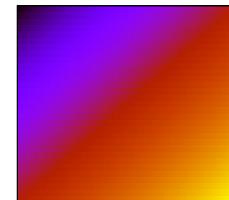
$$\hat{P}_{Q_a \ominus Q_b} : \mathcal{E}_{a \bowtie b} \rightarrow \mathbb{R}^{\mathcal{W}[G]} : e_{ab} \mapsto \hat{p}_{Q_a \ominus Q_b, e_{ab}}$$

- ▶ companion operation \boxminus (usually $= \ominus$) provides final return values
- ▶ otherwise as for unary profiles

Diff Operations: diff (raw difference)

Variable	Description
Q_a	1st profile query (query, date, slice)
Q_b	2nd profile query (bquery, bdate, bslice)
s_a	1st score value operand given collocate g : $s_a = p_{Q_a, e_a}(g)$
s_b	2nd score value operand given collocate g : $s_b = p_{Q_b, e_b}(g)$

$$s_a \ominus_{\text{diff}} s_b := s_a - s_b$$

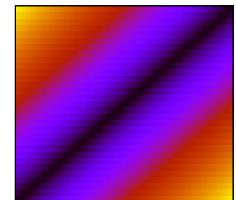


- pre-trimmed
- asymmetric
- selects collocates strongly associated only with Q_a

Diff Operations: adiff (absolute difference)

Variable	Description
Q_a	1st profile query (query, date, slice)
Q_b	2nd profile query (bquery, bdate, bslice)
s_a	1st score value operand given collocate g : $s_a = p_{Q_a, e_a}(g)$
s_b	2nd score value operand given collocate g : $s_b = p_{Q_b, e_b}(g)$

$$s_a \ominus_{\text{adiff}} s_b := |s_a - s_b| \quad ; \quad \boxminus_{\text{adiff}} := \ominus_{\text{diff}}$$

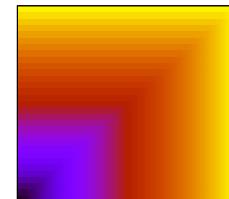


- pre-trimmed
- symmetric
- selects based on $|s_a - s_b|$, but reports raw difference $s_a - s_b$
- returns most extreme differences among strong collocates of Q_a and Q_b
- sign of returned score indicates association preference for Q_a (+) or Q_b (-)

Diff Operations: max (maximum)

Variable	Description
Q_a	1st profile query (query, date, slice)
Q_b	2nd profile query (bquery, bdate, bslice)
s_a	1st score value operand given collocate g : $s_a = p_{Q_a, e_a}(g)$
s_b	2nd score value operand given collocate g : $s_b = p_{Q_b, e_b}(g)$

$$s_a \ominus_{\max} s_b := \max\{s_a, s_b\}$$

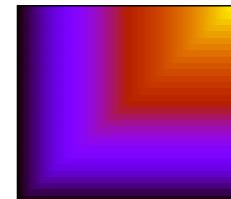


- pre-trimmed
- symmetric
- selects only stronger of the operand association scores
- potentially useful for discovering collocates deserving further investigation

Diff Operations: min (minimum)

Variable	Description
Q_a	1st profile query (query, date, slice)
Q_b	2nd profile query (bquery, bdate, bslice)
s_a	1st score value operand given collocate g : $s_a = p_{Q_a, e_a}(g)$
s_b	2nd score value operand given collocate g : $s_b = p_{Q_b, e_b}(g)$

$$s_a \ominus_{\min} s_b := \min\{s_a, s_b\}$$

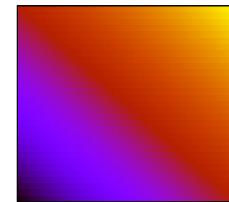


- restricted
- symmetric
- selects only weaker of the operand association scores
- high scores indicate similar strong association preferences
- very sensitive to sparse data problems (missing data \rightsquigarrow zeroes)

Diff Operations: avg (arithmetic average)

Variable	Description
Q_a	1st profile query (query, date, slice)
Q_b	2nd profile query (bquery, bdate, bslice)
s_a	1st score value operand given collocate g : $s_a = p_{Q_a, e_a}(g)$
s_b	2nd score value operand given collocate g : $s_b = p_{Q_b, e_b}(g)$

$$s_a \ominus_{\text{avg}} s_b := \frac{s_a + s_b}{2}$$

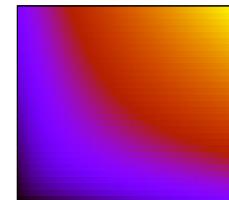


- restricted
- symmetric
- selects strong associations for either Q_a or Q_b , preferring shared associations
- **not** very sensitive to non-uniform operand values
 - ▶ high scores do not necessarily indicate similar collocation behavior

Diff Operations: havg (harmonic average)

Variable	Description
Q_a	1st profile query (query, date, slice)
Q_b	2nd profile query (bquery, bdate, bslice)
s_a	1st score value operand given collocate g : $s_a = p_{Q_a, e_a}(g)$
s_b	2nd score value operand given collocate g : $s_b = p_{Q_b, e_b}(g)$

$$s_a \ominus_{\text{havg}} s_b := \frac{2s_a s_b}{s_a + s_b}$$



- restricted
- symmetric
- selects uniformly strong associations for both Q_a and Q_b
- to avoid singularities, actually computed as:

$$\begin{aligned} \text{havg}(s_a, s_b) &:= \begin{cases} 0 & \text{if } s_a \leq 0 \text{ or } s_b \leq 0 \\ \frac{2s_a s_b}{s_a + s_b} & \text{otherwise} \end{cases} \\ s_a \ominus_{\text{havg}} s_b &:= \text{avg}(\text{havg}(s_a, s_b), \text{avg}(s_a, s_b)) \end{aligned}$$

Addenda

Public D* DiaCollo Instances

Historical Corpora

- ▶ Deutsches Textarchiv (1600–1900)
- ▶ Die Grenzboten (1841–1922)
- ▶ Polytechnisches Journal (1820–1931)

Newspaper Corpora

- ▶ Berliner Zeitung (1994–2005)
- ▶ Tagesspiegel (1996–2005)
- ▶ ZEIT (1946–2016)

Synchronic Corpora

- ▶ DWDS Kernkorpus (1900–1999)
- ▶ Blogs (2003–2014)
- ▶ Film Subtitles (1916–2014)

Aggregated Corpora

- ▶ DTA+DWDS (1600–1999)
- ▶ public (+newspapers, 1600–2016)

Non-German Corpora

- ▶ APWCF (fr, 1644–1647)
- ▶ NHESS (en, 2001–2016)

CLARIN Corpora (non-public)

- ▶ PP Berliner Zeitung (1945–1993)
- ▶ PP Neues Deutschland (1946–1990)
- ▶ PP Neue Zeit (1945–1994)

Fiendishly Awkward Questions: Corpora

Can I use DiaCollo on my own corpus?

- sure – check out the [DiaColloDB](#) and [DiaColloDB::WWW](#) distributions on CPAN
 - ▶ [cpanm](#) is handy for batch installations
- UNIX-like environment is assumed (various flavors of Linux work great)
- KWIC-links and [DDC profiles](#) require a separate [DDC index and server](#)

What languages are supported?

- pretty much any written language ought to work: DiaCollo is *language-agnostic*

What corpus formats are supported?

- input data must be encoded in [UTF-8](#)
- only [pre-tokenized](#) and [pre-annotated](#) formats , e.g.
 - ▶ [DDCTabs](#): text-dump of [DDC search engine](#) index data
 - ▶ [JSON](#): structured [JSON](#) data conforming to [DiaColloDB::Document](#) conventions
 - ▶ [TCF](#): CLARIN-D “[Text Corpus Format](#)” as used by [WebLicht](#)
 - ▶ [TEI](#): basic handling for pre-tokenized [TEI-like XML](#) data (slow!)
- see [DiaColloDB::Document \(3pm\)](#) for an up-to-date list

Fiendishly Awkward Questions: Corpora

Why must I tokenize and annotate my corpus myself?

- one tool \Leftrightarrow one job
- language agnosia \rightsquigarrow flexibility
- DiaCollo is **not** an all-singing+dancing, one-stop-shopping text analysis tool
 - (*and almost certainly never will be*)
- consider [CLARIN-D WebLicht](#) for a generic corpus annotation framework

Can you annotate, index, and/or host my corpus for me?

- maybe . . . we should probably talk later

Can I use DiaCollo to directly compare different corpora?

- . . . on the command-line:
 - ▶ pass a `list://` URL to `dcdb-query.perl` or `dcdb-www-server.perl`
 - ▶ beware the `fudge` and `extend` properties!
- . . . from the `dwds.de/dstar` WWW GUI: only for [pre-aggregated corpora](#)
 - ▶ generic implementation: *work in progress (stage 0: planning)*

What is ‘DDC’, and why might I care?

- “DiaLing/DWDS Concordancer” . . . sometimes “*Diabolically Defective Crust*”
 - search engine used by DWDS and DTA projects at the BBAW
 - required for DiaCollo KWIC-link approximations and DDC relation
 - configuration & usage ↵ **BTSOTD** (“*beyond the scope . . .*”)

Fiendishly Awkward Questions: Corpora

How large does my corpus need to be in order to get reliable results?

- more relevant = ***epoch totals*** $r_N(e)$, rather than global corpus totals
 - ▶ consider increasing slice parameter (E) \rightsquigarrow reducing diachronic granularity
- “good” epoch size depends on ***relative frequency*** of target phenomenon
 - ▶ depends in turn on request parameters query, date, groupby (q, H, G)
 - ▶ see Gabrielatos et al. (2012) for a more detailed discussion
- beware ***compile-time filters*** and ***server-side pruning***
 - ▶ indexing option `-use-all-the-data` disables filters (native, TDF)
 - ▶ `#FMIN 1` query operator disables server-side pruning (DDC)
- ***corpus artifacts*** are always possible
 - ▶ e.g. “Pferdebuckel” (raw), “*Krise↔Tolstoj*” (KWIC)
- completely subjective, non-rigorous, & informal recommendation (**YMMV**):
 - ▶ your chances are pretty good if
 - $$\min \left\{ f(\llbracket q \rrbracket, e), f([g], e) \right\} \geq 100$$
 - ▶ but also interesting results from small corpora ***well below this threshold!***



Fiendishly Awkward Questions: Runtime

Can I download DiaCollo results for offline use?

- static tabular formats (Text, HTML, JSON): yes
 - ▶ use the “Raw URL” link for static tabular formats (Text, HTML)
- static canvas snapshots (bubble, cloud): yes
 - ▶ use the “Download” button in the upper right of the display canvas
- interactive GUI (bubble, cloud): yes
 - ▶ use your browser’s “Save As (Web-Page, complete)” function
- google motion charts (“gmotion”) don’t support offline use

How can I restrict the profile to immediate predecessors?

- use the [DDC relation](#) with a [phrase query](#), e.g. “*=2 Mann” #FMIN 1
- see [Example “What Makes a ‘Man’”](#)

Why does my collocant appear as a collocate for itself?

- self-collocations are never counted for identical tokens ($d = 0$);
 - cf. [“Native Co-ccurrence Relation”](#)
- collocated tokens of a single type are counted twice; cf. [NEAR\(Krise,Krise,4\)](#)
 - ▶ yes, this is a wart, but it’s not the wart you probably think it is

Fiendishly Awkward Questions: Runtime

Why does my collocate item g “disappear” in epoch e ?

- it may have been eliminated by *compile-time filters* or *server-side pruning*
 - ▶ try using the [DDC relation](#) with the #FMIN 1 operator
- it may not be among the k -best collocates in epoch e
 - ▶ k -best pruning occurs *independently* for each epoch
 - ▶ try raising kbest parameter (k) and/or setting the global flag
 - ▶ try using groupby restrictions (H) to select only the collocate(s) of interest

Why does the D3 date-slider (bubble, cloud) “snap” to epoch boundaries?

- DiaCollo result sets are *discrete*, cf. [DiaColloDB::Profile::Multi \(3pm\)](#)
- D3 format size and color are *linearly interpolated* between epochs by the GUI
 - ▶ possible future improvement: unit granularity + moving average smoothing

Fiendishly Awkward Questions: Runtime

**Why does the collocation pair (q, g) appear at epoch e ?
*(even though I know it doesn't really occur until later)***

- epochs are labeled by their minimum possible element, $\tilde{E}(y) = E \lfloor \frac{y}{E} \rfloor$
- epoch label e represents the date interval $[e .. e + E - 1]$
 - ▶ e.g. for slice $E = 10$, epoch “1980” represents the interval 1980–1989

Why don't the corpus KWIC links always return exactly f_{12} hits?

- DiaCollo itself does *not* create or maintain a full-text index (one tool \Leftrightarrow one job)
- retrieval of corpus hits \rightsquigarrow independent DDC server
 - ▶ DDC context query generated on-the-fly for each collocation pair
- compile-time filters \rightsquigarrow *approximate* results only
 - ▶ no equivalent DDC query expression for e.g. wgood, pbad, ...
- to ensure exact results, use the DDC relation with the #FMIN 1 operator

Forensic Analysis Questions: Errors

Error: *DiaColloDB::Document::CLASS: cannot load file ...*

- your input corpus does not appear to be formatted correctly
- did you specify the correct `-dclass=CLASS` option to `dcdb-create.perl`?

Error: *No 'query' parameter specified!*

- your request did not include a `query (q)` parameter
- appears in WWW GUI before any request has been submitted
 - ▶ nothing to see here, move along

Error: *No data to display!*

- no index entries matched your `request`
- usual suspects: `compile-time filters` or `server-side pruning`
 - ▶ check parameters using `dcdb-info.perl` or WWW 'info' link
 - ▶ see `DiaColloDB (3pm)` for details on what the various properties mean
- try using the `DDC relation` with the `#FMIN 1` operator

Forensic Analysis Questions: Errors

Error: *You cannot submit queries from an offline data set!*

- you attempted to submit a new request to an static GUI snapshot
 - ▶ e.g. as produced a browser's "Save As" function
- submit your request to a "live" index-wrapper instead

Error: *Variable 'ddc_url_root' not set: KWIC links disabled!*

- your DiaCollo index is not associated with any running [DDC server](#)
- run a DDC server process for your corpus, and set the `ddcServer` option

Error: *500 Internal Server Error*

- this is just an [HTTP status code](#), *not* an error message (and not very informative)
- keep reading for some (hopefully) more useful diagnostics

Error: *ttk_process(): template error: undef error - [MESSAGE]*

- something went wrong in the WWW GUI (still not very informative)
- actual error message begins with **[MESSAGE]**

Forensic Analysis Questions: Errors

Error: *... called at FILE.pm line XYZ*

- this is a [stack trace](#) of the error
- only the first line or two is likely to be informative

Error: *parseQuery(): ... could not parse query: syntax error: ...*

- your [query](#) (*q*) parameter could not be parsed
- consult the “[Query Syntax](#)” section of the DiaCollo help page

Error: *align(): cannot align non-trivial multi-profiles of unequal size*

- you tried to compare two profiles with incompatible epoch partitions
 - ▶ $\mathcal{E}_{a \bowtie b}$ could not be defined: $1 < |\mathcal{E}_{E_a}| \neq |\mathcal{E}_{E_b}| > 1$
- see “[Comparison Profiles](#)”

Forensic Analysis Questions: Errors

Error: *... abstract method called*

- I probably forgot to implement something; please let me know!

Error: *... timeout elapsed*

- the DDC relation's server request took too long to complete
- your query may be too complex to be handled gracefully – try raising #FMIN

Error: *no 'ddcServer' key defined*

- you tried to use the DDC relation without declaring a DDC server
- **EITHER** edit your index header.json:
`"ddcServer": "HOST:PORT"`
OR use the `-O=ddcServer=HOST:PORT` option to ddc-query.perl
 - ▶ replace *HOST* and *PORT* with values appropriate for your DDC server

I think I found a big nasty stinky ugly creepy crawlly bug!



- it's entirely possible that you have, but before you pick up the bat-phone ...
 - ▶ have you read (and tried to understand) the [documentation](#)? ([RTFM](#))
 - ▶ have you read (and tried to understand) the [error message](#), if any? ([RTFEM](#))
 - ▶ have you thought about what might have gone wrong? ([UYFB](#))
 - ▶ "*Simplify, simplify*" – have you tried a less complex request? ([Thoreau](#))
- ... if none of the above help, [please](#) e-mail me a *precise description* of:
 - ▶ **what you wanted** and/or expected
 - ▶ **what you tried**, including full URL(s) if applicable
 - ▶ **what went wrong** and/or was unexpected

Disclaimer: *neither the author nor the BBAW condones physical violence against users!*

References

- K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.
- M. Davies. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157, 2012. URL http://davies-linguistics.byu.edu/ling450/davies_corpora_2011.pdf.
- J. Didakowski and A. Geyken. From DWDS corpora to a German word profile – methodological problems and solutions. In A. Abel and L. Lemnitzer, editors, *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information*, (OPAL X/2012). IDS, Mannheim, 2013. URL http://www.dwds.de/static/website/publications/pdf/didakowski_geyken_internetlexikograf
- T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
- S. Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2005. URL <http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/>.
- S. Evert. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin, 2008. URL http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_manuscript.pdf.

References

- J. R. Firth. *Papers in Linguistics 1934–1951*. Oxford University Press, London, 1957.
- C. Gabrielatos, T. McEnery, P. J. Diggle, and P. Baker. The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics*, 17(2):151–175, 2012.
doi:10.1075/ijcl.17.2.01gab. URL
<http://www.jbe-platform.com/content/journals/10.1075/ijcl.17.2.01gab>.
- A. Geyken. Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv. In I. Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, volume 4 of *Thesaurus Linguae Aegyptiae*, pages 221–234, Berlin, Germany, 2013. URL <http://nbn-resolving.de/urn:nbn:de:kobv:b4-opus-24424>.
- A. Geyken and T. Hanneforth. TAGH: A complete morphology for German based on weighted finite state automata. In *Finite State Methods and Natural Language Processing, 5th International Workshop, FSMNLP 2005, Revised Papers*, volume 4002 of *Lecture Notes in Computer Science*, pages 55–66. Springer, Berlin, 2006. doi:10.1007/11780885_7.
- B. Hamp and H. Feldweg. GermaNet – a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.
- V. Henrich and E. Hinrichs. GernEdiT – the GermaNet editing tool. In *Proceedings LREC 2010*, pages 2228–2235, 2010. URL
http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf.



References

- G. Heyer, U. Quasthoff, and T. Wittig. *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*. IT lernen. W3L-Verlag, 2006. ISBN 9783937137308. URL <https://books.google.de/books?id=i2JjAAAACAAJ>.
- B. Jurish. A hybrid approach to part-of-speech tagging. Technical report, Project “Kollokationen im Wörterbuch”, Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, 2003. URL <http://kaskade.dwds.de/~jurish/pubs/dwdst-report.pdf>.
- B. Jurish. DiaCollo: On the trail of diachronic collocations. In K. De Smedt, editor, *CLARIN Annual Conference 2015 (Wrocław, Poland, October 14–16 2015)*, pages 28–31, 2015. URL <http://www.clarin.eu/sites/default/files/book%20of%20abstracts%202015.pdf>.
- B. Jurish and A. Gerstenberg. Exploring the internal heterogeneity of a corpus of classical French with DiaCollo. In *Global Philology Open Conference*. Universität Leipzig, February 2017. URL <http://kaskade.dwds.de/~jurish/pubs/jg2017apwcf-diacollo-slides.pdf>.
- B. Jurish, C. Thomas, and F. Wiegand. Querying the deutsches Textarchiv. In U. Kruschwitz, F. Hopfgartner, and C. Gurrin, editors, *Proceedings of the Workshop “Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities” (MindTheGap 2014)*, pages 25–30, Berlin, Germany, March 2014. URL http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf.

References

- B. Jurish, A. Geyken, and T. Werneke. DiaCollo: diachronen Kollokationen auf der Spur. In *Proceedings DHd 2016: Modellierung – Vernetzung – Visualisierung*, pages 172–175, March 2016. URL <http://dhd2016.de/boa.pdf#page=172>.
- A. Kilgarriff and D. Tugwell. Sketching words. In M.-H. Corréard, editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, EURALEX, pages 125–137, 2002. URL <http://www.kilgarriff.co.uk/Publications/2002-KilgTugwell-AtkinsFest.pdf>.
- A. Kilgarriff, P. Rychlý, P. Smrž, and D. Tugwell. The sketch engine. In *Proceedings of Euralex*, pages 105–116, 2004.
- A. Kilgarriff, A. Herman, J. Busta, P. Rychlý, and M. Jakubíček. DIACRAN: a framework for diachronic analysis. In F. Formato and A. Hardie, editors, *Proceedings of Corpus Linguistics 2015*, pages 65–70, UCREL, Lancaster, 2015.
- L. Lemnitzer and C. Kunze. *Computerlexikographie: Eine Einführung*. Gunter Narr Verlag, Tübingen, 2007. URL <http://www.ssg-bildung.ub.uni-erlangen.de/computerlexikographie.pdf>.
- C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.
- F. Moretti. *Distant reading*. Verso Books, 2013.

References

- J. Richling. Referenzkorpus Altdeutsch (Old German reference corpus): Searching in deeply annotated historical corpora, 2011. Talk presented at the conference *New Methods in Historical Corpora*, 29–30 April, 2011. Manchester, UK.
- P. Rychlý. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9, 2008. URL <http://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf>.
- H. D. Thoreau. *Walden*. [1854] 1995. URL <http://www.gutenberg.org/ebooks/205>.
- L. Wittgenstein. *Philosophische Untersuchungen*. Oxford, 1953.