Exploring diachronic collocations with DiaCollo

Bryan Jurish
jurish@bbaw.de

Universität Potsdam, Institut für Linguistik
19th June, 2017

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

Overview

The Situation

- Diachronic Text Corpora
- Collocation Profiling
- Diachronic Collocation Profiling

DiaCollo

- Requests & Parameters
- Profiles, Diffs & Indices

Gory Details

- Corpus Indexing
- Co-occurrence Relations
- Scoring & Comparison Functions

Examples

Summary & Conclusion

The Situation: Diachronic Text Corpora

- heterogeneous text collections, especially with respect to ***date of origin***
  - other partitionings potentially relevant too, e.g. by author, text class, etc.
- increasing number available for linguistic & humanities research, e.g.
  - *Deutsches Textarchiv (DTA)* *(Geyken 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

CLARIN-D

> *"Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache"*
> — 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)*
- distributional semantics *(Schütze 1992; Sahlgren 2006)*
- "text mining" / "distant reading" *(Heyer et al. 2006; Moretti 2013)*

The Situation: Related Work

Conventional (synchronic) Collocation Profiling

- well understood & widely accepted *(e.g. Manning & Schütze 1999; Evert 2005)*

⇝ can't handle (temporal) ***heterogeneity***!

Diachronic Studies: Manual Corpus Partitioning

- Baker et al. (2008): 10 epochs, 1 year each
- Sagi et al. (2009): 5 epochs, ca. 100 years each
- Gulordava & Baroni (2011): 2 epochs, 10 years each
- Scharloth et al. (2013): 3400 epochs, ca. 1 week each (+smoothing)
- Kim et al. (2014): 160 epochs, 1 year each

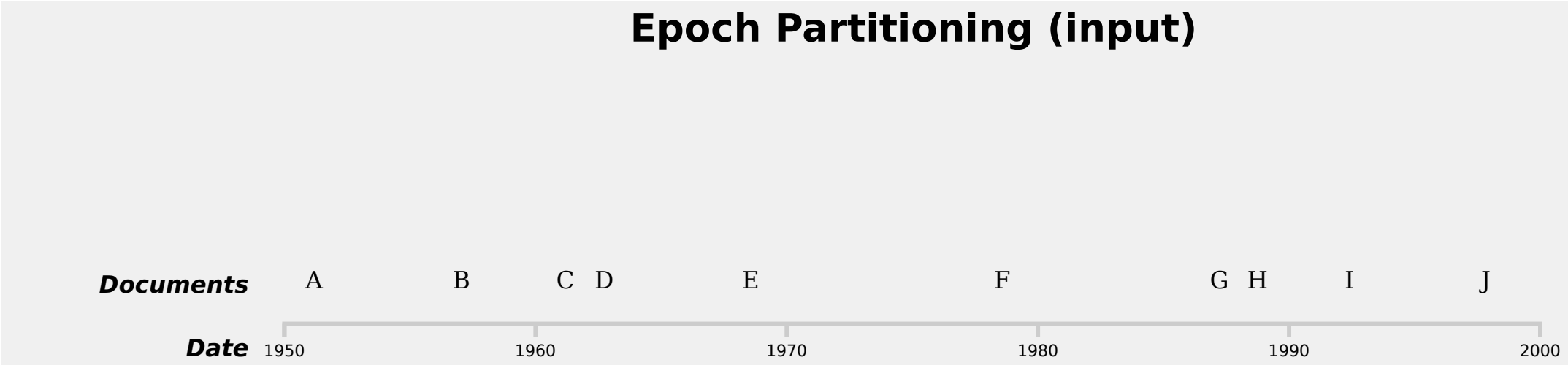
⇝ **Gabrielatos et al. (2012)**: ***epoch granularity*** depends on ***research question***!

"Latent" Distributional Approximations

- Wang & McCallum (2006): "Topics Over Time" (LDA)
- Sagi et al. (2009): LSA model w.r.t. 2000 most frequent content-bearing collocates
- Kim et al. (2014): series of vector space models à la Mikolov et al. (2013)

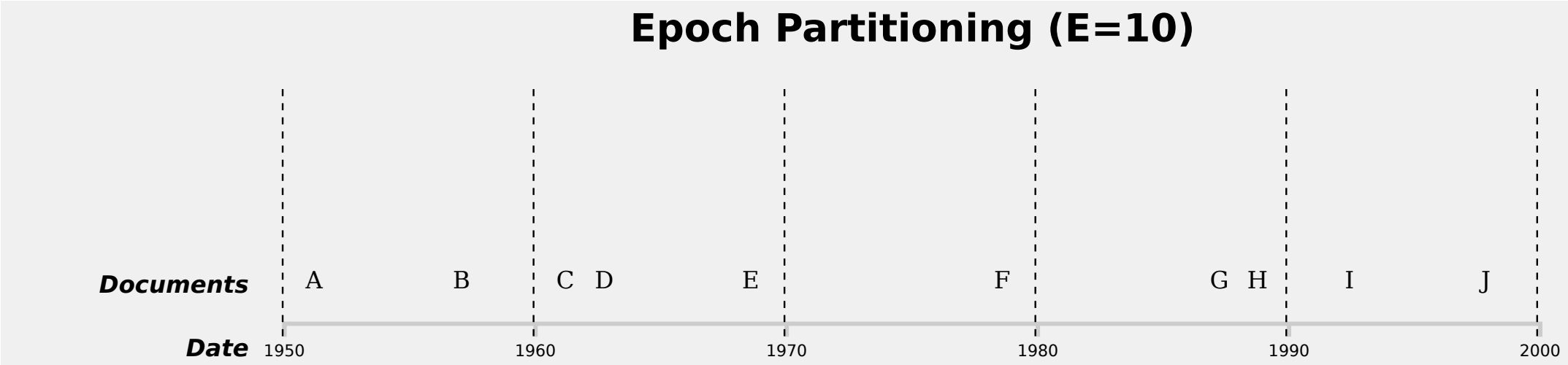
⇝ compile-time parameters, approximate counts ⇒ ***not viable***!

Manual Corpus Partitioning



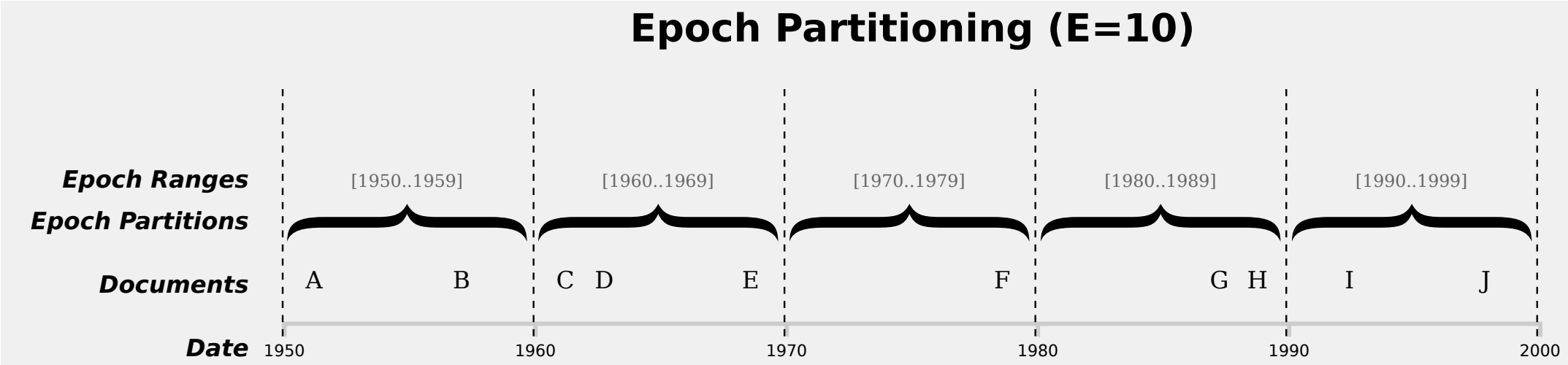
- input corpus with documents $\{A, B, \ldots, J\}$ over date range (1950–1999)

Manual Corpus Partitioning



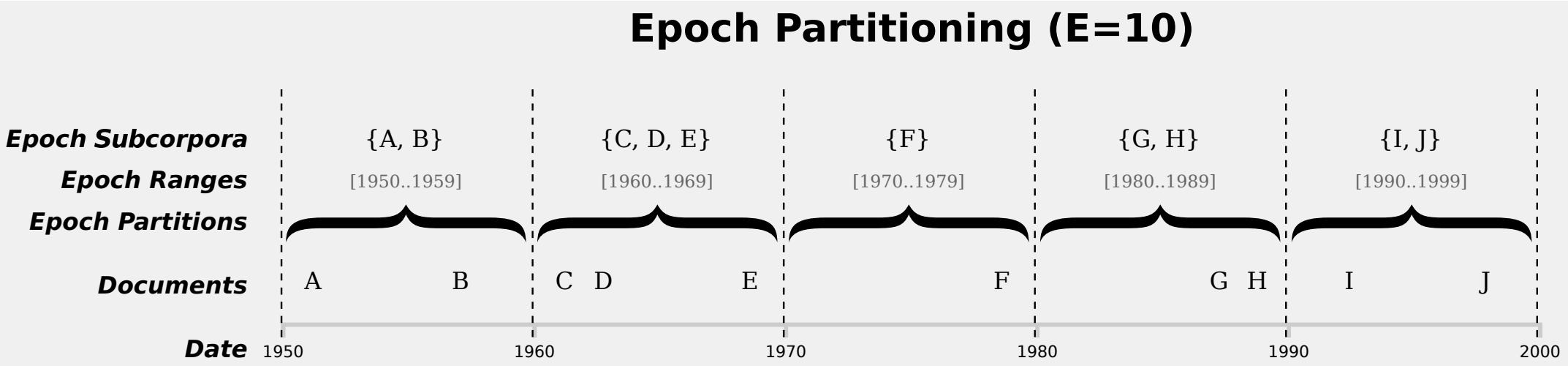
- input corpus with documents $\{A, B, \ldots, J\}$ over date range (1950–1999)
- partition by decade ($E = 10$)

Manual Corpus Partitioning



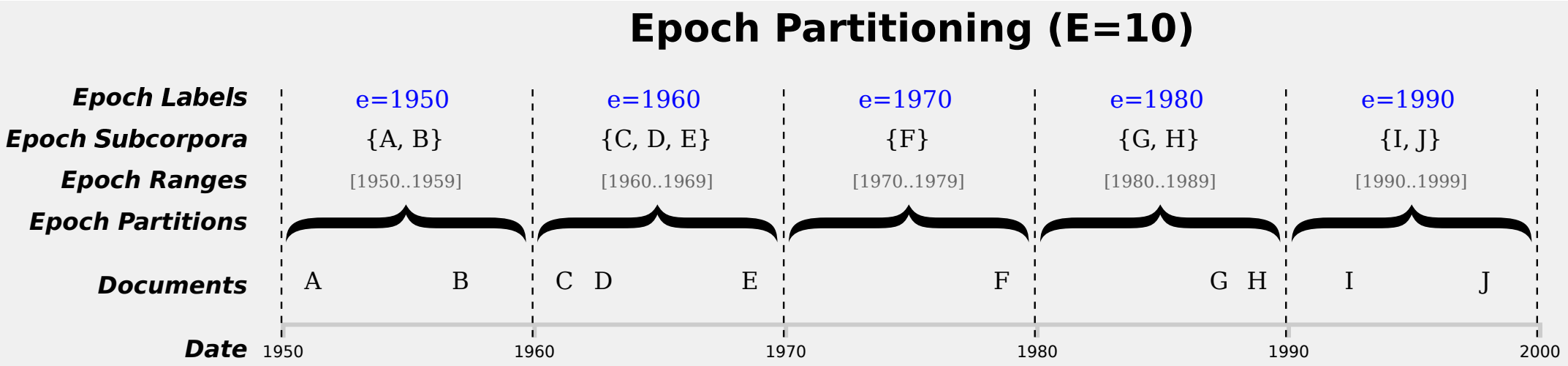
- input corpus with documents $\{A, B, \ldots, J\}$ over date range (1950–1999)
- partition by decade ($E = 10$)

Manual Corpus Partitioning



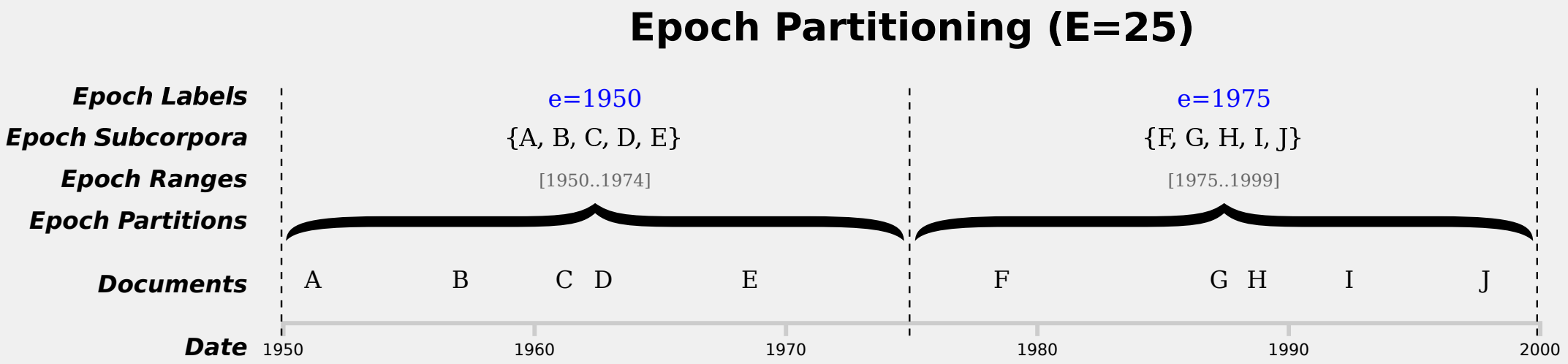
- input corpus with documents $\{A, B, \ldots, J\}$ over date range (1950–1999)
- partition by decade ($E = 10$)
- collect epoch-wise subcorpora

Manual Corpus Partitioning



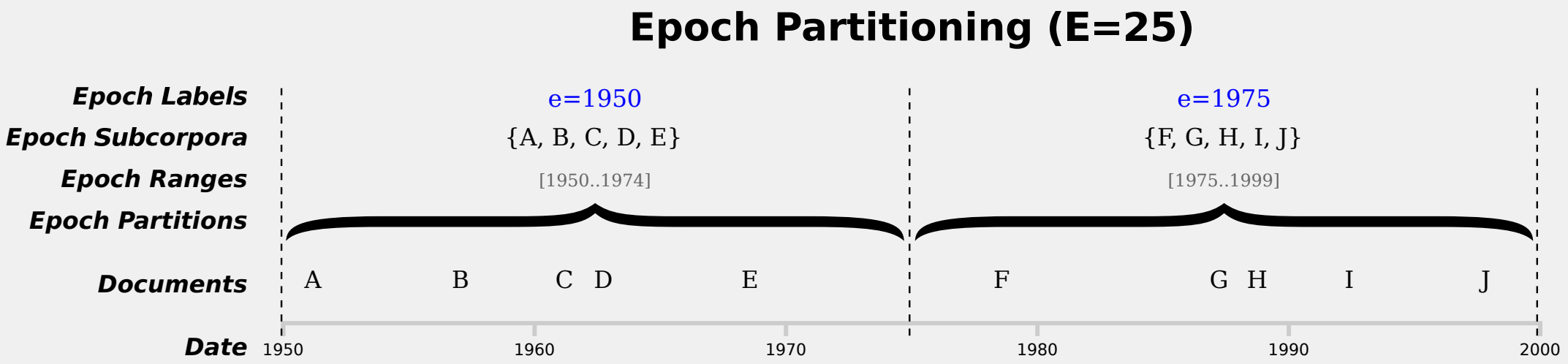
- input corpus with documents $\{A, B, \ldots, J\}$ over date range (1950–1999)
- partition by decade ($E = 10$)
- collect epoch-wise subcorpora
- label sub-corpora (e.g. by minimum date) and analyze independently

Manual Corpus Partitioning



- input corpus with documents $\{A, B, \ldots, J\}$ 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



- input corpus with documents $\{A, B, \ldots, J\}$ 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?*

. . . 

Diachronic Collocation Profiling

CLARIN-D

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)
  - *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: Requests & Parameters

- request-oriented RESTful service *(Fielding 2000)*
- accepts user requests as set of *parameter=value* pairs
- parameter passing via URL query string or HTTP POST request
- common parameters:

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

DiaCollo: Profiles, Diffs & Indices

CLARIN-D

Profiles & Diffs

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

Indices & Attributes

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

Appetizer

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

Gory Details

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.
  - `-O=wbad=`*REGEX* : surface form blacklist regex
  - `-O=pgood=`*REGEX* : PoS whitelist regex
  - `-tfmin=`*FREQ* : minimum global term-tuple frequency
  - `-lfmin=`*FREQ* : minimum global lemma frequency
  - `-cfmin=`*FREQ* : minimum co-occurrence frequency

Basic Definitions

Corpus Data

- a corpus $\mathcal{C}$ is list of N tokens t_i — $\mathcal{C} = t_1 t_2 \ldots t_N$
- each token is an n_A-tuple of attribute values — $t_i \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_{n_A}$
- each token is associated with a unique non-negative integer date (year) — $\mathrm{Y}(t_i) \in \mathbb{N}$

Corpus Domain

- lexical domain (term vocabulary) — $\mathcal{W} = \bigcup_{i=1}^{N} \{t_i\} \subseteq \mathcal{A}_1 \times \cdots \times \mathcal{A}_{n_A}$
- temporal domain (dates) — $\mathcal{Y} = \bigcup_{i=1}^{N} \{\mathrm{Y}(t_i)\} \subset \mathbb{N}$

Common Notation

- attribute projection — $t[j] = a_j$ for $t = \langle a_1, \ldots, a_n \rangle$

  ...for attribute-lists — $t[J] = \langle t_{j_1}, \ldots, t_{j_{n_J}} \rangle$ for $J = \langle j_1, \ldots, j_{n_J} \rangle$
- equivalence classes — $[u]_{T/J} = \{t \in T \mid t[J] = u\} \subseteq T$

Runtime Data: Requests and Profiles

DiaCollo Request

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

runtime user input parameters:

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

Raw Co-occurrence Frequency Profile

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

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

- $r_{\mathrm{N}} : \mathcal{E} \to \mathbb{N}$ the total number of corpus co-occurrences by epoch *(N)*
- $r_1 : \mathcal{E} \to \mathbb{N}$ independent collocant frequency by epoch *(f1)*
- $r_2 : \mathcal{E} \times \mathcal{W}[G] \to \mathbb{N}$ independent collocate frequency by epoch *(f2)*
- $r_{12} : \mathcal{E} \times \mathcal{W}[G] \to \mathbb{N}$ 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 $C = s_1 \ldots s_{n_S}$ of break-units ("sentences") $s_i = t_{i1} \ldots t_{in_{s_i}}$,

$$\begin{aligned} I_{12} \quad &: \quad \mathcal{W} \to \big(\mathcal{W} \to (\mathcal{Y} \to \mathbb{N})\big) \\ &: \quad \langle w, v, y \rangle \mapsto \sum_{i=1}^{n_S} \sum_{j=1}^{n_{s_i}} \sum_{d=-\ell}^{\ell} \mathbb{1}[d \neq 0 \;\&\; t_{ij} = w \;\&\; t_{i(j+d)} = v \;\&\; \mathrm{Y}(t_{ij}) = y] \end{aligned}$$

- ***Beware:*** compile-time filters (`pgood`, `tfmin`, etc.) influence index content!
  - `cfmin` option prunes by co-frequency $\quad \mathrm{f}(w, v, y) < f_{\texttt{cfmin}} \Rightarrow I_{12}(w, v, y) = 0$
- independent "frequencies" $I_1(w, y)$, $I_{\mathrm{N}}(y)$ computed as true marginals:

$$\begin{aligned} I_1 &: \mathcal{W} \times \mathcal{Y} \to \mathbb{N} : \langle w, y \rangle \mapsto \textstyle\sum_{v \in \mathcal{W}} I_{12}(w, v, y) \\ I_{\mathrm{N}} &: \qquad \mathcal{Y} \to \mathbb{N} : \qquad y \mapsto \textstyle\sum_{w \in \mathcal{W}} I_1(w, y) \end{aligned}$$

Native Co-occurrence Relation: Context Window

$\ell = 3$

Input Text	The	fat	cat	sat	on	the	fuzzy	cat	.

Native Co-occurrence Relation: Context Window

$\ell = 3$

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

Native Co-occurrence Relation: Context Window

$\ell = 3$

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	~~.~~

Native Co-occurrence Relation: Context Window

$\ell = 3$

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	

Native Co-occurrence Relation: Context Window

$\ell = 3$

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	t_5	

Native Co-occurrence Relation: Context Window

$\ell = 3$

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		$\boxed{t_1}$	$\langle t_2 \rangle$	t_3			t_4	t_5	

$j{=}1$
$d{=}1$

$\rightsquigarrow$

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

Native Co-occurrence Relation: Context Window

$\ell = 3$

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		$\boxed{t_1}$	t_2	$\langle t_3 \rangle$			t_4	t_5	

$j{=}1$
$d{=}2$

$\rightsquigarrow$

$$I_{12} = \Big\{ \langle \text{fat}, \text{cat} \rangle \mapsto 1, \langle \text{fat}, \text{sit} \rangle \mapsto 1 \Big\}$$

Native Co-occurrence Relation: Context Window

$\ell = 3$

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		$\boxed{t_1}$	t_2	t_3			$\langle t_4 \rangle$	t_5	

$j{=}1$
$d{=}3$

$\rightsquigarrow$

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

Native Co-occurrence Relation: Context Window

$\ell = 3$

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		$\langle t_1 \rangle$	$\boxed{t_2}$	t_3			t_4	t_5	

$j{=}2$

$d{=}{-1}$

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

Native Co-occurrence Relation: Context Window

$\ell = 3$

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	$\boxed{t_2}$	$\langle t_3 \rangle$			t_4	t_5	

$j{=}2$
$d{=}1$

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

Native Co-occurrence Relation: Context Window

$\ell = 3$

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	$\boxed{t_2}$	t_3			$\langle t_4 \rangle$	t_5	

$j{=}2$
$d{=}2$

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

Native Co-occurrence Relation: Context Window

$\ell = 3$

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	$\boxed{t_2}$	t_3			t_4	$\langle t_5 \rangle$	

$j{=}2$

$d{=}3$

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

Native Co-occurrence Relation: Context Window

$\ell = 3$

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		$\langle t_1 \rangle$	$\langle \underline{t_2} \rangle$	$\boxed{t_3}$			$\langle t_4 \rangle$	$\langle \underline{t_5} \rangle$	

$j{=}3$

$d{=}*$

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

Native Co-occurrence Relation: Context Window

$\ell = 3$

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		$\langle t_1 \rangle$	$\langle \underline{t_2} \rangle$	$\langle t_3 \rangle$			$\boxed{t_4}$	$\langle \underline{t_5} \rangle$	

$j{=}4$

$d{=}*$

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

Native Co-occurrence Relation: Context Window

$\ell = 3$

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	$\langle \underline{t_2} \rangle$	$\langle t_3 \rangle$			$\langle t_4 \rangle$	$\boxed{t_5}$	

$j{=}5$

$d{=}*$

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

Native Co-occurrence Relation: Runtime

CLARIN-D

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

- find collocant tuple(s) $[\![q]\!]$, e.g. $[\![\texttt{\$lemma=love}]\!] = [\texttt{love}]_{\mathcal{W}/a_{\text{lemma}}}$ $\quad [\![q]\!] \subseteq \mathcal{W}$
- raw index lookup: $\hat{I}_q : \mathcal{Y} \times \mathcal{W} \to \mathbb{N} : \langle y, v \rangle \mapsto \sum_{w \in [\![q]\!]} I_{12}(w, v, y)$
- group collocates by attributes G: $\hat{I}_{q,G} : \mathcal{Y} \times \mathcal{W}[G] \to \mathbb{N} : \langle y, g \rangle \mapsto \sum_{v \in [g]_{\mathcal{W}/G}} \hat{I}_q(y, v)$
- apply request filter restrictions H: $\hat{I}_{q,G,H} = \hat{I}_{q,G} \restriction \text{ext}(H) : \mathcal{Y} \times \mathcal{W}[G] \to \mathbb{N}$
- aggregate by epoch E: $\hat{I}_{q,G,H,E} : \mathcal{E}_E \times \mathcal{W}[G] \to \mathbb{N} : \langle e, g \rangle \mapsto \sum_{y \in [e]_E} \hat{I}_{q,G,H}(y, g)$
  - where $\widetilde{E} : \mathcal{Y} \to \mathbb{N} : y \mapsto E\lfloor \frac{y}{E} \rfloor \quad ; \quad \mathcal{E}_E = \widetilde{E}(\mathcal{Y}) \quad ; \quad [e]_E = \widetilde{E}^{-1}(e)$
- finalize raw frequency profile $R_Q = \langle r_{\text{N}}, r_1, r_2, r_{12} \rangle$

$$
\begin{aligned}
r_{\text{N}}(e) &= \sum_{y \in [e]_E} I_{\text{N}}(y) \\
r_1(e) &= \sum_{y \in [e]_E} \sum_{w \in [\![q]\!]} I_1(w, y) \\
r_2(e, g) &= \sum_{y \in [e]_E} \sum_{v \in [g]_{\mathcal{W}/G}} I_1(v, y) \\
r_{12}(e, g) &= \hat{I}_{q,G,H,E}(e, g)
\end{aligned}
$$

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

TDF Co-occurrence Relation: Indexing

("term × document matrix" profile type)

- "co-occurrence" $\rightsquigarrow$ anywhere within the selected break unit ("document")
- relatively coarse index granularity (no proximity constraints)
- for corpus partitioned into documents $\mathrm{Doc} = \{d_1, \ldots, d_{n_D}\}$, store:
  - sparse term-document frequency matrix $\boxed{\mathrm{tdf} : \mathcal{W} \times \mathrm{Doc} \to \mathbb{N}}$
  - date counts $\boxed{\mathrm{yf} : \mathcal{Y} \to \mathbb{N} : y \mapsto \sum_{w \in \mathcal{W}} \sum_{d \in \mathrm{dy}^{-1}(y)} \mathrm{tdf}(w, d)}$
  - document dates and bibliographic metadata $\boxed{\mathrm{dy} : \mathrm{Doc} \to \mathcal{Y}}$
- 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 q independently as:
  - set of terms $[\![q]\!]_{\mathcal{W}}$ $\quad [\![q]\!]_{\mathcal{W}} \subseteq \mathcal{W}$
  - set of documents $[\![q]\!]_{\mathrm{Doc}}$ $\quad [\![q]\!]_{\mathrm{Doc}} \subseteq \mathrm{Doc}$
- index lookup with collocate grouping:

$$\begin{aligned}\hat{I}_{\mathrm{tdf}:q,G} &: \mathcal{Y} \times \mathcal{W}[G] \to \mathbb{N} \\ &: \langle y, g\rangle \mapsto \sum\nolimits_{d \in [\![q]\!]_{/y}} \min\left\{\left(\sum\nolimits_{w \in [\![q]\!]_{\mathcal{W}}} \mathrm{tdf}(w,d)\right), \left(\sum\nolimits_{v \in [g]_{\mathcal{W}/G}} \mathrm{tdf}(v,d)\right)\right\}\end{aligned}$$

  - where $[\![q]\!]_{/y} = [\![q]\!]_{\mathrm{Doc}} \cap \mathrm{dy}^{-1}(y)$
- candidate filtering and epoch aggregation as for native index
- final raw frequency profile $R_Q = \langle r_{\mathrm{N}}, r_1, r_2, r_{12}\rangle$

$$\begin{aligned} r_{\mathrm{N}}(e) &= \sum\nolimits_{y \in [e]_E} \mathrm{yf}(y) \\ r_1(e) &= \sum\nolimits_{y \in [e]_E} \sum\nolimits_{w \in [\![q]\!]_{\mathcal{W}}} \sum\nolimits_{d \in [\![q]\!]_{/y}} \mathrm{tdf}(w,d) \\ r_2(e,g) &= \sum\nolimits_{y \in [e]_E} \sum\nolimits_{v \in [g]_{\mathcal{W}/G}} \sum\nolimits_{d \in \mathrm{dy}^{-1}(y)} \mathrm{tdf}(v,d) \\ r_{12}(e,g) &= \hat{I}_{\mathrm{tdf}:q,G,H,E}(e,g) \end{aligned}$$

DDC Co-occurrence Relation

CLARIN-D

("ddc" profile type)

- "co-occurrence" $\rightsquigarrow$ as returned by a ***DDC*** search engine query
  - 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 $R_Q = \langle r_{\mathrm{N}}, r_1, r_2, r_{12} \rangle$

$$
\begin{aligned}
r_{\mathrm{N}} &= \lambda_q \times \texttt{COUNT(* \#SEP) \#BY[date/}E\texttt{]} \\
r_1 &= \lambda_q \times \texttt{COUNT(KEYS(}[\![q \& H]\!]\texttt{ \#SEP \#BY[}G\texttt{=1]) \#SEP) \#BY[date/}E\texttt{]} \\
r_2 &= \lambda_q \times \texttt{COUNT(KEYS(}[\![q \& H]\!]\texttt{ \#SEP \#BY[}G\texttt{=2]) \#SEP) \#BY[date/}E\texttt{,}G\texttt{=2]} \\
r_{12} &= \texttt{COUNT(}[\![q \& H]\!]\texttt{ \#SEP \#BY[date/}E\texttt{,}G\texttt{=2])}
\end{aligned}
$$

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

Scoring & Pruning: Basics

- φ maps raw frequency profiles to scalar association scores $\varphi : \mathbb{R}^4 \to \mathbb{R}$
- score profiles $p_{Q,e}$ computed independently for each epoch $e \in \mathcal{E}_E$:

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

- k-best pruning within each epoch: $\hat{p}_{Q,e} = p_{Q,e} \restriction \mathrm{best}_k(p_{Q,e})$
  - "global" profiles prune by global corpus association score:

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

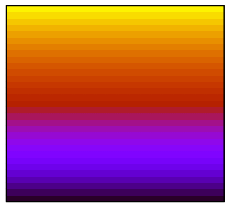
  - alternative: user-specified cutoff threshold
- final diachronic profile maps epoch-labels to epoch-local profiles:

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

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 $r_{\mathrm{N}}(e)$
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 $\frac{1}{2}$

$$\varphi_{\mathrm{f}}(w_1, w_2) = f_{12}$$

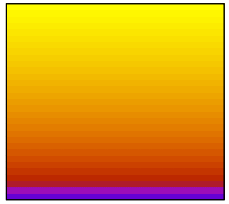


- immediately interpretable, but not very robust
- Zipf distribution leads to "lopsided" visualizations
- values may not be comparable across slices (e.g. for non-balanced corpora)
- many false positives with high-frequency collocates
- **not** generally a good measure of collocate affinity

Score Functions: lf (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 $r_{\mathrm{N}}(e)$
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 $\frac{1}{2}$

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

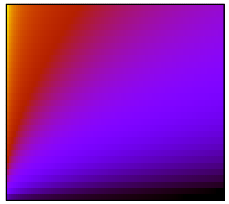


- better visual scaling than raw frequency
- otherwise shares raw frequency's shortcomings

Score Functions: mi (pointwise MI × 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 $r_{\mathrm{N}}(e)$
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 $\frac{1}{2}$

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



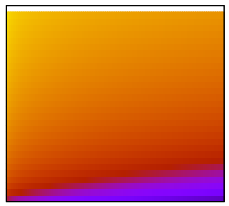
- used by first version of Sketch Engine *(Kilgarriff et al. 2004)*
- PMI gives code-length change for (optimal) joint vs. independent encodings
- PMI alone is very sensitive to low-frequency items ($\rightsquigarrow$ longer codes)
  - *post-hoc* workaround: include log-frequency coefficient
- some preference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

Score Functions: ll (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 $r_{\mathrm{N}}(e)$
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 $\frac{1}{2}$

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

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

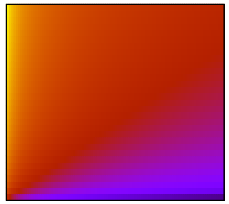


- 1-sided variant of the binomial log likelihood ratio *(Dunning 1993; Evert 2008)*
  - only "attracting" collocate pairs are assigned positive values
- null hypothesis H_0 filters out "uninteresting" high-frequency collocates
- very sensitive to fixed & formulaic expressions $\rightsquigarrow$ ***poor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

Score Functions: ld (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 $r_{\mathrm{N}}(e)$
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 $\frac{1}{2}$

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



(Rychlý 2008)

- "lexicographer-friendly" association score
- less susceptible to low-frequency outliers than PMI $\times$ log-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`) $\ominus : \mathbb{R}^2 \to \mathbb{R}$
- 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 $p_{Q_a \ominus Q_b, e_{ab}} : \mathrm{Dom}_{Q_a \ominus Q_b / e_{ab}} \to \mathbb{R} : g \mapsto p_{Q_a, e_a}(g) \ominus p_{Q_b, e_b}(g)$
  - $e_{ab} = \langle e_a, e_b \rangle \in \mathcal{E}_{a \bowtie b}$ an aligned epoch pair
  - $\mathrm{Dom}_{Q_a \ominus Q_b / e_{ab}} \subseteq \mathrm{dom}(p_{Q_a, e_a}) \cup \mathrm{dom}(p_{Q_b, e_b})$ characteristic for $\ominus$ at e_{ab}:
    - "pre-trimmed" operations $= \mathrm{dom}(\hat{p}_{Q_a, e_a}) \cup \mathrm{dom}(\hat{p}_{Q_b, e_b})$
    - "restricted" operations $= \mathrm{dom}(p_{Q_a, e_a}) \cap \mathrm{dom}(p_{Q_b, e_b})$
- prune and collect

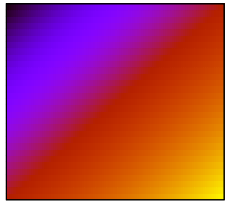
$$\hat{p}_{Q_a \ominus Q_b, e_{ab}} = p_{Q_a \boxminus Q_b, e_{ab}} \restriction \mathrm{best}_k(p_{Q_a \ominus Q_b, e_{ab}})$$
$$\hat{P}_{Q_a \ominus Q_b} : \mathcal{E}_{a \bowtie b} \to \mathbb{R}^{\mathcal{W}[G]} : e_{ab} \mapsto \hat{p}_{Q_a \ominus Q_b, e_{ab}}$$

  - 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: $s_\mathrm{a} = p_{Q_a,e_a}(g)$
s_b	2nd score value operand given collocate g: $s_\mathrm{b} = p_{Q_b,e_b}(g)$

$$s_\mathrm{a} \ominus_\mathrm{diff} s_\mathrm{b} := s_\mathrm{a} - s_\mathrm{b}$$

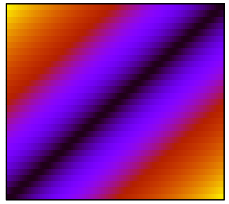


- pre-trimmed
- asymmetric
- selects collocates strongly associated only with Q_a

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: $s_\mathrm{a} = p_{Q_a,e_a}(g)$
s_b	2nd score value operand given collocate g: $s_\mathrm{b} = p_{Q_b,e_b}(g)$

$$s_\mathrm{a} \ominus_\mathrm{adiff} s_\mathrm{b} := |s_\mathrm{a} - s_\mathrm{b}| \quad ; \quad \boxminus_\mathrm{adiff} := \ominus_\mathrm{diff}$$

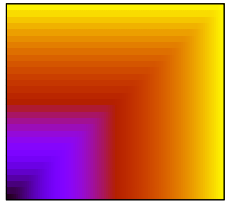


- pre-trimmed
- symmetric
- selects based on $|s_\mathrm{a} - s_\mathrm{b}|$, but reports raw difference $s_\mathrm{a} - s_\mathrm{b}$
- returns most extreme differences among strong collocates of Q_a and Q_b
- sign of returned score indicates association preference for Q_a $(+)$ or Q_b $(-)$

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: $s_\mathrm{a} = p_{Q_a,e_a}(g)$
s_b	2nd score value operand given collocate g: $s_\mathrm{b} = p_{Q_b,e_b}(g)$

$$s_\mathrm{a} \ominus_\mathrm{max} s_\mathrm{b} := \max\{s_\mathrm{a}, s_\mathrm{b}\}$$

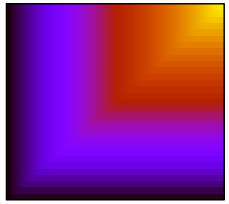


- pre-trimmed
- symmetric
- selects only stronger of the operand association scores
- potentially useful for discovering collocates deserving further investigation

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: $s_\mathrm{a} = p_{Q_a,e_a}(g)$
s_b	2nd score value operand given collocate g: $s_\mathrm{b} = p_{Q_b,e_b}(g)$

$$s_\mathrm{a} \ominus_\mathrm{min} s_\mathrm{b} := \min\{s_\mathrm{a}, s_\mathrm{b}\}$$

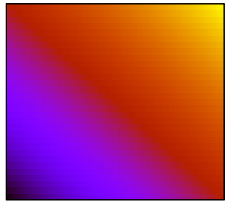


- restricted
- symmetric
- selects only weaker of the operand association scores
- high scores indicate similar strong association preferences
- very sensitive to sparse data problems (missing data $\rightsquigarrow$ zeroes)

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: $s_\mathrm{a} = p_{Q_a,e_a}(g)$
s_b	2nd score value operand given collocate g: $s_\mathrm{b} = p_{Q_b,e_b}(g)$

$$s_\mathrm{a} \ominus_\mathrm{avg} s_\mathrm{b} := \frac{s_\mathrm{a}+s_\mathrm{b}}{2}$$

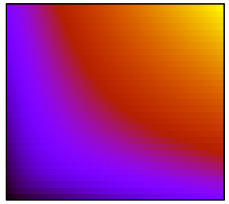


- restricted
- symmetric
- selects strong associations for either Q_a or Q_b, preferring shared associations
- ***not*** very sensitive to non-uniform operand values
  - high scores do not necessarily indicate similar collocation behavior

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: $s_\mathrm{a} = p_{Q_a,e_a}(g)$
s_b	2nd score value operand given collocate g: $s_\mathrm{b} = p_{Q_b,e_b}(g)$

$$s_\mathrm{a} \ominus_\mathrm{havg} s_\mathrm{b} :\approx \frac{2 s_\mathrm{a} s_\mathrm{b}}{s_\mathrm{a} + s_\mathrm{b}}$$



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

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

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=l,p%3DNE

1950–1959
- Berlin blockade aftermath

1960–1969
- anti-government protests & strikes in France

1970–1979
- Nixon & Brandt resignations; Iranian revolution

1980–1989
- *Solidarność* in Poland; Soviet war in Afghanistan; Schmidt coalition collapses

1990–1999
- wars in ex-Yugoslavia, Kosovo, & Chechnya; financial crises in Asia & Mexico

2000–2009
- global financial crisis

2010–2016
- 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)

[source: T. Werneke]

Example 1: Selected Lemma-Clouds

1980–1989:

2010–2016:

Example 2: Lexicography

CLARIN-D

'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

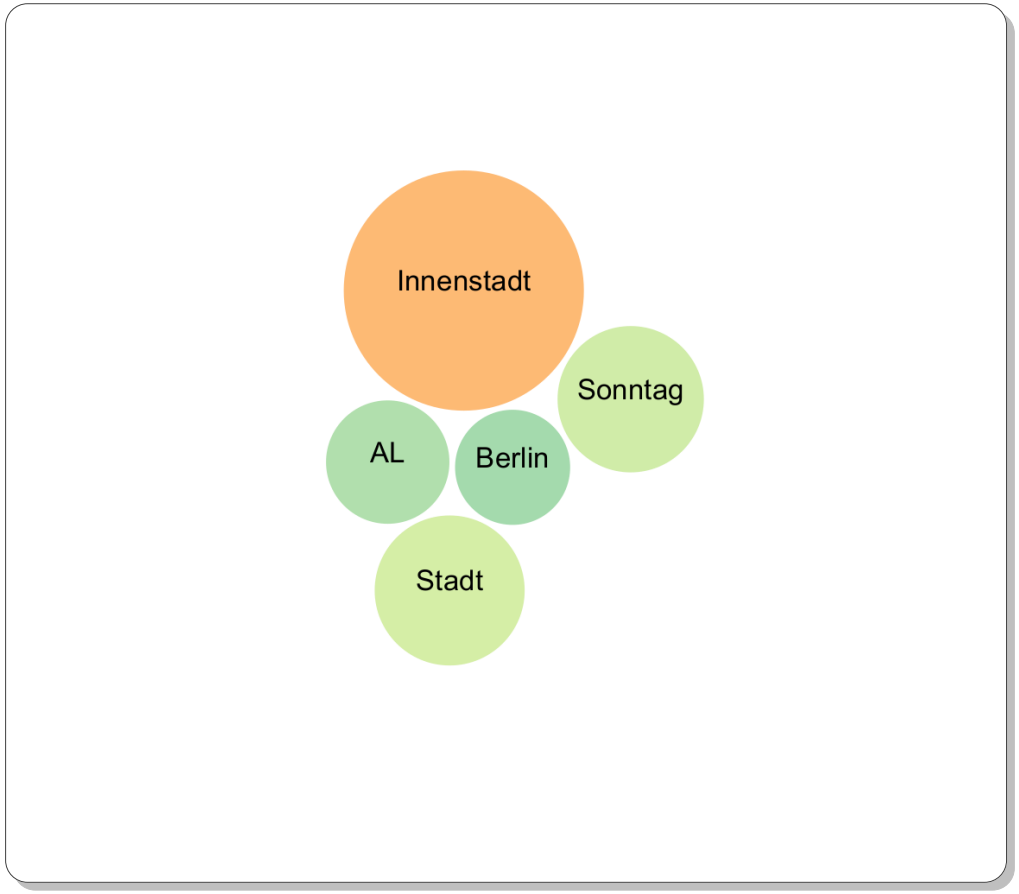
The case of *autofrei* ("automobile-free")

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

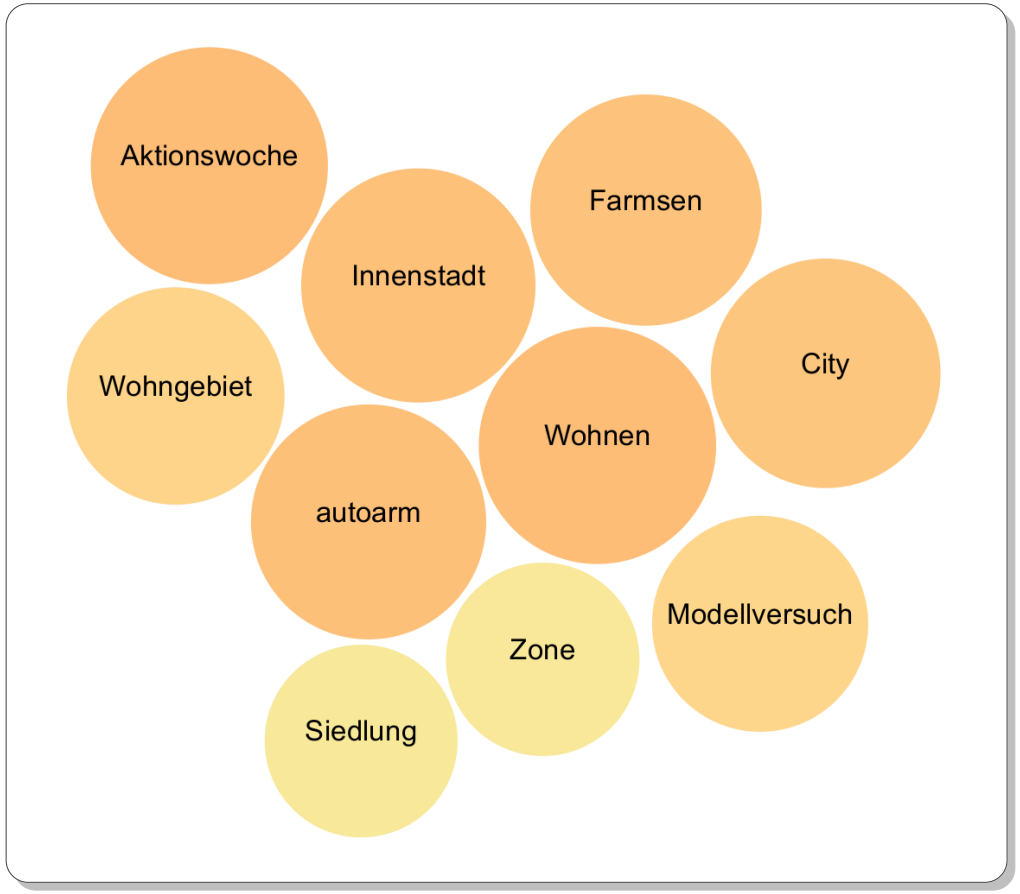
[source: A. Geyken]

Example 2: Selected Bubble-Charts

1985–1989



1990–1994



Example 3: Revolution

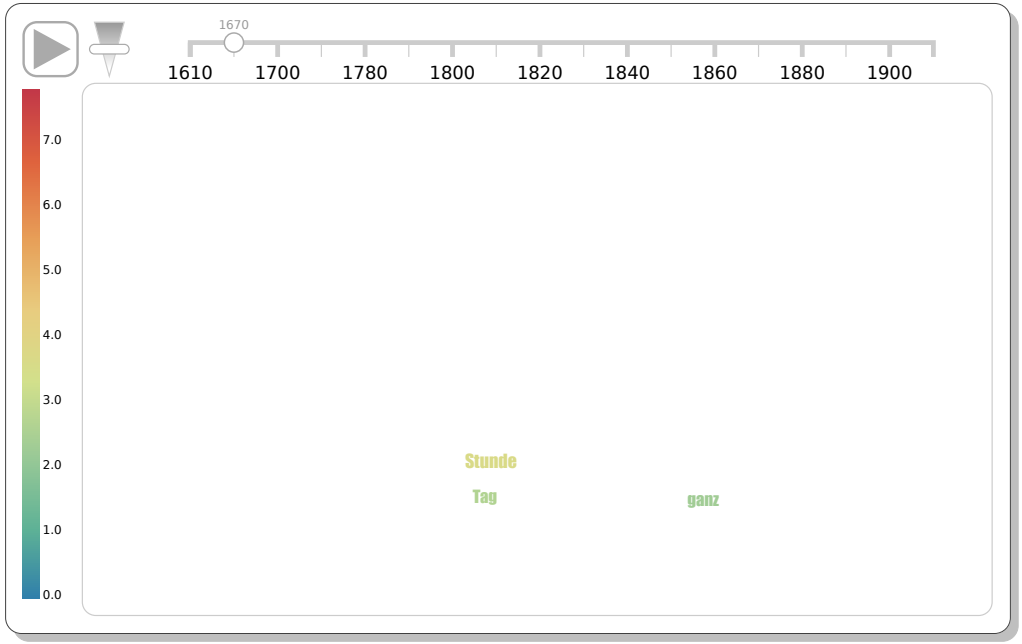
well, you know . . .

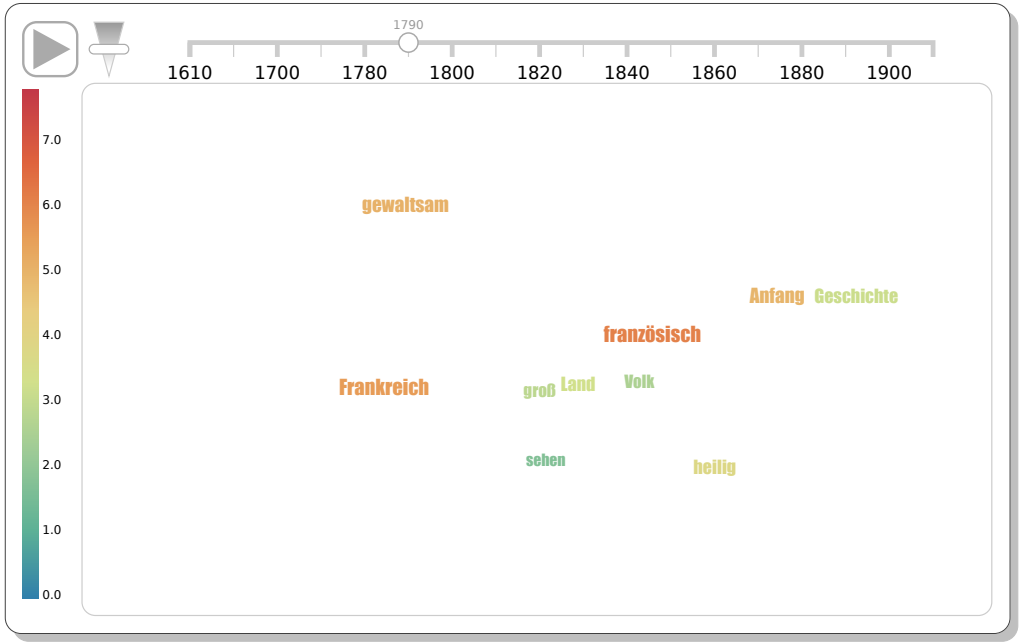
http://kaskade.dwds.de/dstar/dta/diacollo/?q=Revolution&ds=10&f=cloud

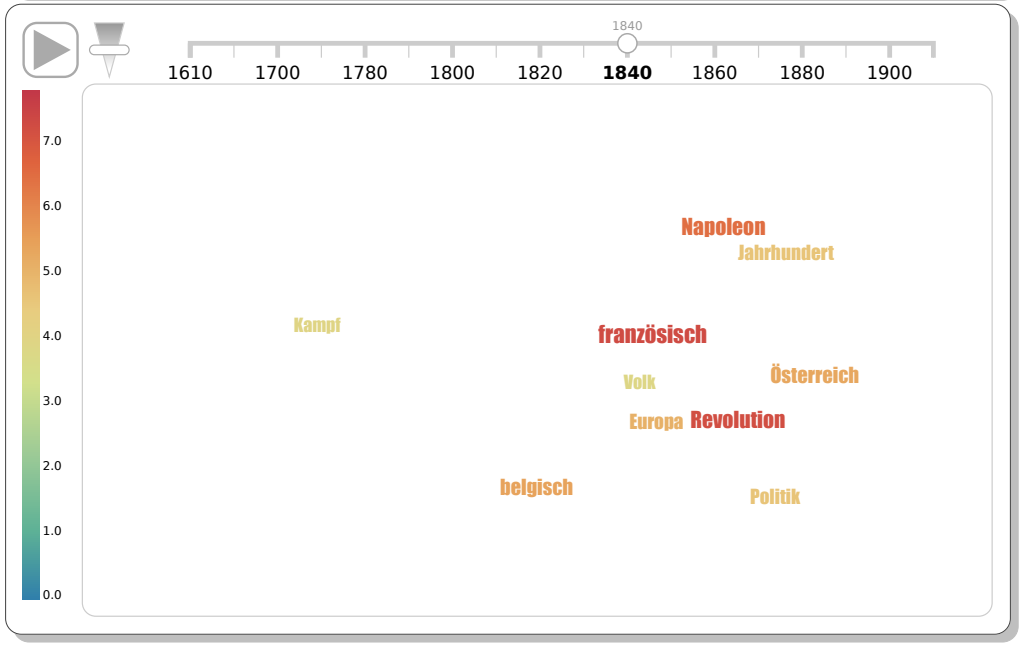
- $<$ **1770**: only 'rotation' sense
  - *ganz, Stunde, Tag* ("entire, hour, day")
- $\geq$ **1770**: 'dramatic change'
  - *menschlich* ("human")
- $\geq$ **1790**: French Revolution
  - *französisch, Frankreich* ("French, France")
- $\geq$ **1840**: violent political upheaval (Napoleonic era)
  - *Napoleon*
- $\geq$ **1860**: industrial revolution
  - *Industrie, industriell* ("industry, industrial")

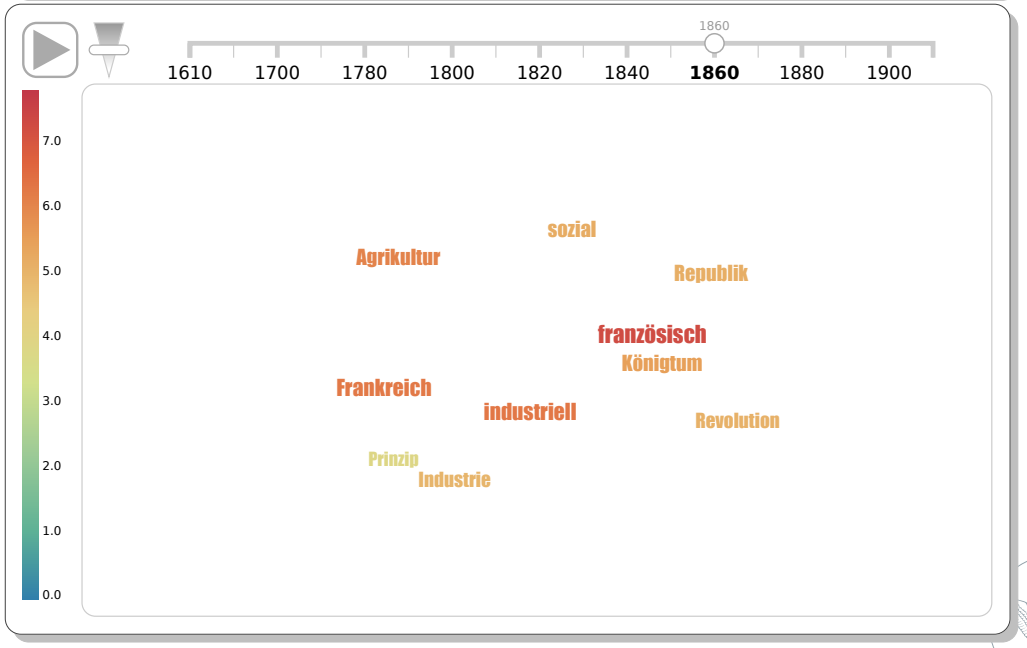
[source: L. Lemnitzer, J. Lennon, P. McCartney]

Example 3: Selected Lemma-Clouds









Example 4: Gender & Cultural Bias

CLARIN-D

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

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

Disclaimer

- historical corpus data can reveal persistent cultural biases
- linked collocation data does not reflect the opinions of the 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*** $\rightsquigarrow$ *berühmt, ehrlich, gelehrt, tapfer, weise, . . .*
  - ***Frau*** $\rightsquigarrow$ *betrübt, lieb, schön, tugendreich, verwitwet, . . .*
- differences grow less pronounced in late 18th & 19th centuries

Example 4: Selected Lemma-Clouds

1725–1749:

1825–1849:

Example 5: What Makes a 'Man'?

CLARIN-D

'[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 5: Selected Lemma-Clouds

1700–1799

(`diff=adiff`)

1800–1899

(`diff=adiff`)

Example 6: Genealogy of Terminology

CLARIN-D

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 6: Lemma-Clouds

differences

(`diff=adiff`)

similarities

(`diff=havg`)

Example 7: Pronominal Adverbs by Genre

CLARIN-D

'[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 ⇝ *demnach, 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 7: Selected Lemma-Clouds

1650–1699:

1950–1999:

Example 8: 400 Years of Potables

CLARIN-D

'[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=l&G=1

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

Remarks

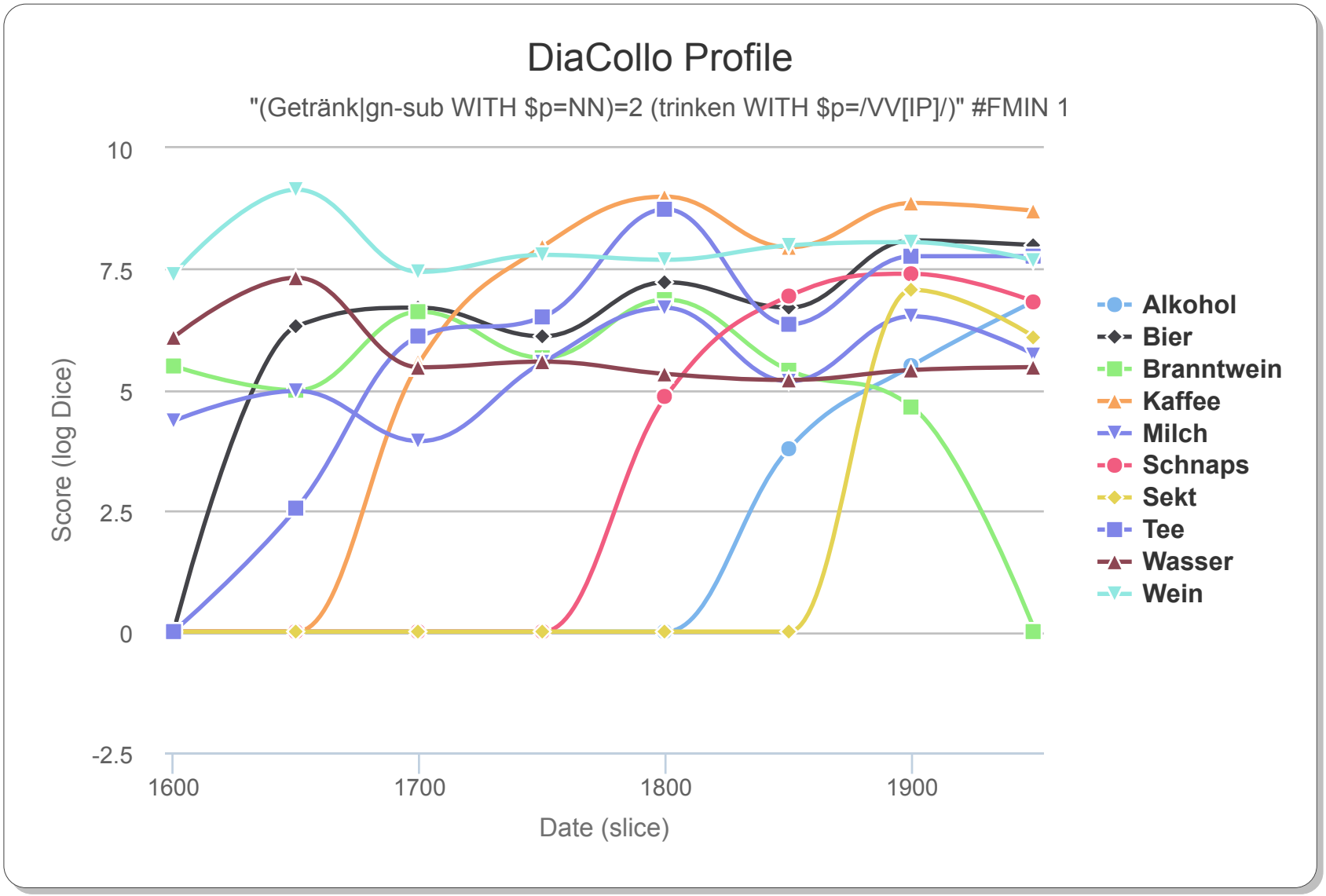
- uses DDC back-end for fine-grained data acquisition
- uses GermaNet thesaurus-based lexical expansion for ***Getränk*** ("beverage")
  (Hamp & Feldweg 1997; Lemnitzer & Kunze 2007; Henrich & Hinrichs 2010)
- considers only those target terms immediately preceding verb ***trinken*** ("to drink")
- "global" 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 8: Time Series $(k = 10)$



Summary & Conclusion

CLARIN-D

Diachronic Collocation Profiling

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

DiaCollo

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

Applications

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

CLARIN-D

— The End —

Thank you for listening!

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

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

http://metacpan.org/release/DiaColloDB

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Addenda

Public D* DiaCollo Instances

CLARIN-D

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

CLARIN-D

Can I use DiaCollo on my own corpus?

- sure – check out the DiaColloDB and DiaColloDB::WWW distributions on CPAN
  - `cpanm` is handly 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

CLARIN-D

Why must I tokenize and annotate my corpus myself?

- one tool ⇔ one job
- language agnosia ⇝ 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)*

Fiendishly Awkward Questions: Corpora

CLARIN-D

What is 'DDC', and why might I care?

- "DiaLing/DWDS Concordancer" *. . . sometimes "Diabolically Defective Cruft"*
- 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 . . . ")*

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Fiendishly Awkward Questions: Corpora

CLARIN-D

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

- more relevant = ***epoch totals*** $r_{\mathrm{N}}(e)$, 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 artefacts*** are always possible
  - e.g. "Pferdebuckel" (raw), "*Krise*$\leftrightarrow$*Tolstoj*" (KWIC)
- completely subjective, non-rigorous, & informal recommendation (YMMV):
  - your chances are pretty good if $\min\left\{\mathrm{f}\left(\llbracket q \rrbracket, e\right), \mathrm{f}\left([g], e\right)\right\} \geq 100$
  - but also interesting results from small corpora ***well below this threshhold***!

Fiendishly Awkward Questions: Runtime

CLARIN-D

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-cccurrence 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

CLARIN-D

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

CLARIN-D

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, $\widetilde{E}(y) = E\lfloor \frac{y}{E} \rfloor$
- 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

CLARIN-D

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

CLARIN-D

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

CLARIN-D

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

CLARIN-D

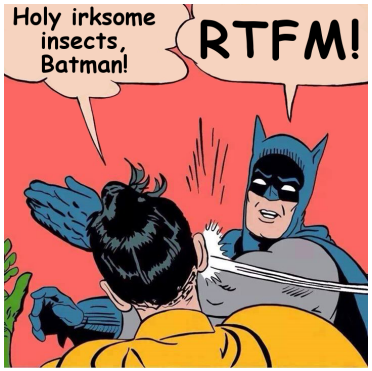
Error: *. . . abstract method called*

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

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

Forensic Analysis Questions: Bugs



I think I found a big nasty stinky ugly creepy crawly 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

P. Baker, C. Gabrielatos, M. Khosravinik, M. Krzyżanowski, T. McEnery, and R. Wodak. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3):273–306, 2008.

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 Auto-*
matic 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.

References

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.

R. T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000. URL https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

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.

References

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.

K. Gulordava and M. Baroni. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July 2011. ACL. URL http://www.aclweb.org/anthology/W11-2508.

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.

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.

References

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, 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.

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.

References

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.

Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65. ACL, June 2014. URL http://www.aclweb.org/anthology/W14-2517.

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.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. URL https://arxiv.org/abs/1301.3781.

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.

E. Sagi, S. Kaufmann, and B. Clark. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on Geometrical Models of Natural Language Semantics*. ACL, March 2009. URL http://www.aclweb.org/anthology/W09-0214.

M. Sahlgren. *The Word-Space Model*. PhD thesis, Gothenburg University, 2006.

J. Scharloth, D. Eugster, and N. Bubenhofer. Das Wuchern der Rhizome. linguistische Diskursanalyse und Data-driven Turn. In D. Busse and W. Teubert, editors, *Linguistische Diskursanalyse. Neue Perspektiven*, pages 345–380. VS Verlag, Wiesbaden, 2013. URL http://www.scharloth.com/files/Rhizom_Zeit.pdf.

H. Schütze. Word space. In *Advances in Neural Information Processing Systems 5, [NIPS Conference, Denver, Colorado, USA, November 30 - December 3, 1992]*, pages 895–902, 1992. URL http://papers.nips.cc/paper/603-word-space.

References

H. D. Thoreau. *Walden*. [1854] 1995. URL http://www.gutenberg.org/ebooks/205.

X. Wang and A. McCallum. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433, New York, 2006. ACM. doi:10.1145/1150402.1150450.

L. Wittgenstein. *Philosophische Untersuchungen*. Oxford, 1953.