Using DiaCollo for Historical Research

Bryan Jurish
Berlin-Brandenburg Academy of Sciences and Humanities, Berlin
jurish@bbaw.de

Maret Nieländer
Georg Eckert Institute for International Textbook Research, Braunschweig
nielaender@leibniz-gei.de

CLARIN Annual Conference 2019
Leipzig, Germany

1st October, 2019

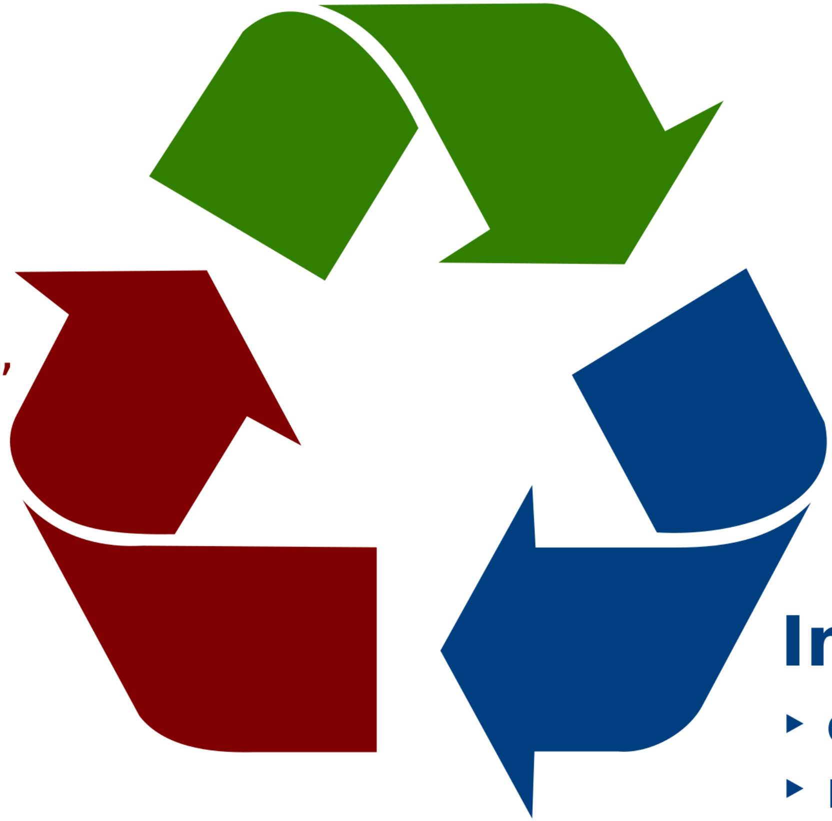
Overview

- Collaborative software development
- Corpora & collocations
- DiaCollo: diachronic collocation profiling
- Use case: Education policy in *Die Grenzboten*
- Summary & conclusion

Software Development Cycle

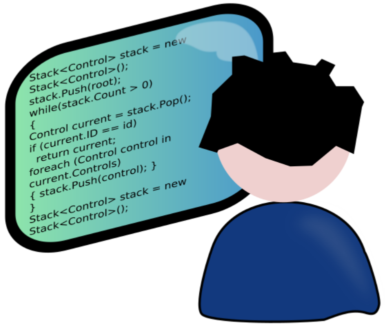
Planning

- identify desiderata & bugs
- sketch next steps



Evaluation

- testing “in the wild”
- user feedback



```
Stack<Control> stack = new
Stack<Control>();
stack.Push(root);
while(stack.Count > 0)
{
Control current = stack.Pop();
if (current.ID == id)
  return current;
foreach (Control control in
current.Controls)
{ stack.Push(control); }
}
Stack<Control> stack = new
Stack<Control>();
```

Implementation

- coding & documentation
- release & deployment

Corpora & Collocations

Diachronic Text Corpora

- heterogeneous with respect to to ***date of origin***
- should expose temporal effects of e.g. ***semantic shift***, ***discourse trends***
- problematic for conventional NLP tools (which assume **homogeneity**)

Collocation Profiling

(Church & Hanks 1990; Manning & Schütze 1999; Evert 2005)

> *"You shall know a word by the company it keeps"* — J. R. Firth

- **prompt** user for target **collocant** term(s) of interest (w_1)
- **lookup** all candidate **collocates** (w_2) co-occurring with w_1
- **rank** candidates by association score
  - score function $\varphi(f_1, f_2, f_{12}, N)$ approximates **relevance** of w_2 to w_1
  - "chance" co-occurrences with high-frequency w_2 should be **filtered out**!
  - statistical method $\rightsquigarrow$ requires **large data sample**

Diachronic Collocation Profiling

The Problem: (temporal) heterogeneity

- conventional collocation extractors assume **corpus homogeneity**
- co-occurrence frequencies are computed only for **word-pairs** (w_1, w_2)
- influence of **occurrence date** (and other document properties) is irrevocably lost

A Solution (sketch)

- represent terms as n-tuples of independent attributes, **including occurrence date**
- partition corpus **on-the-fly** into **user-specified intervals** (“date slices”, “epochs”)
- collect independent epoch-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: Development

Planning & Evaluation

- in collaboration with DWDS lexicographers & CLARIN-D historians

Implementation

- Perl+PDL API, CLI, client/server
  - RESTful D* **web-service** + GUI
- various output & visualization formats, e.g.
  - TSV, JSON , HTML, Highcharts, d3-cloud, . . .
- **batteries not included**
  - tokenization, annotation, full-text search, . . .
- **garbage in ⇝ garbage out**
  - "messy" corpora ⇝ unsatisfying results

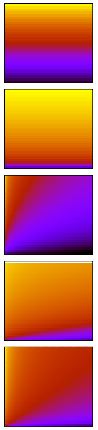
Deployment

- successfully applied to 70 distinct curated corpora at the BBAW, including:
  - Royal Society *Philosophical Transactions* (1665–1869, 9.8K documents, 35M tokens)
  - *Deutsches Textarchiv* (1600–1900, 3.6K documents, 205M tokens)
  - *DWDS Zeitungen* (1946–2019, 16M documents, 6.3G tokens)

DiaCollo: Scoring & Comparison Functions

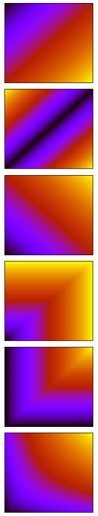
Selected Score Functions

f	raw collocation frequency	$= f_{12}$
lf	collocation log-frequency	$= \log_2(f_{12} + \varepsilon)$
mi	pointwise MI × 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)	$\approx \mathrm{sgn}(f_{12} \vert f_1, f_2) \times \log \frac{\mathrm{L}(H_0)}{\mathrm{L}(H_1)}$
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	$= s_\mathrm{a} - s_\mathrm{b}$
adiff	absolute score difference	$= \lvert s_\mathrm{a} - s_\mathrm{b} \rvert$
avg	arithmetic average	$= \frac{s_\mathrm{a} + s_\mathrm{b}}{2}$
max	maximum	$= \max\{s_\mathrm{a}, s_\mathrm{b}\}$
min	minimum	$= \min\{s_\mathrm{a}, s_\mathrm{b}\}$
havg	harmonic average	$\approx \frac{2 s_\mathrm{a} s_\mathrm{b}}{s_\mathrm{a} + s_\mathrm{b}}$



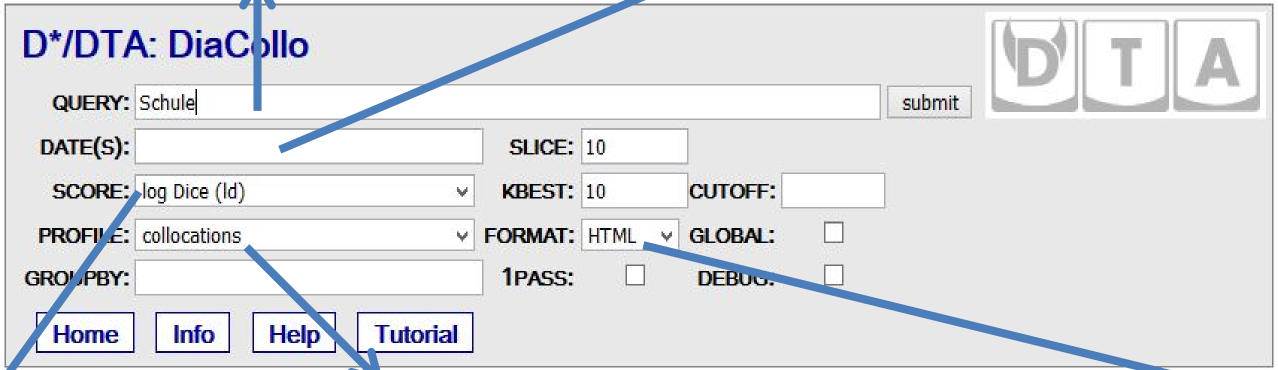
CLARIN-D

Use Case: Education Policy in *Die Grenzboten*

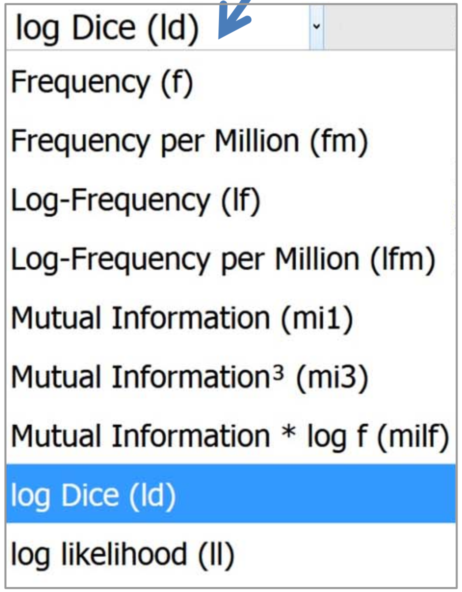
‘*Schule*’: DiaCollo Query (DTA)

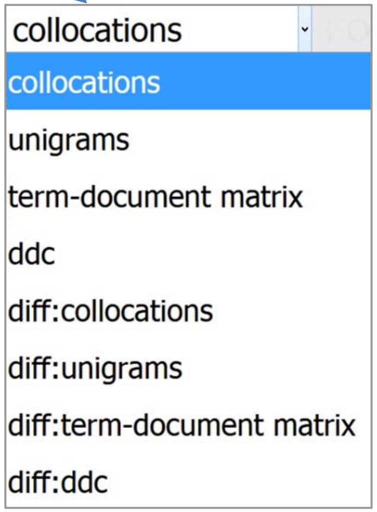
Target as:
LEMMA(s), e.g. Maske
/REGEX/ , e.g. /^Masken.*$/
DDC QUERY

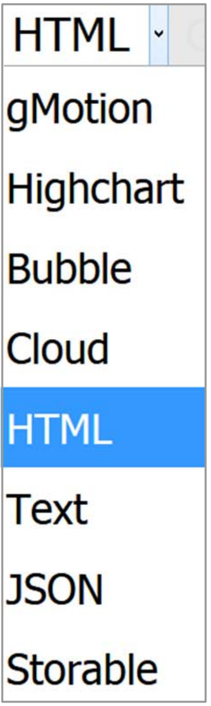
Target date(s):
DATE(s), e.g. 1900:1999 or *:* or 1900:*
/REGEX/, e.g. /^18[345]/


D*/DTA: DiaCollo

QUERY: Schule
DATE(S):
SLICE: 10
SCORE: log Dice (ld)
KBEST: 10
CUTOFF:
PROFILE: collocations
FORMAT: HTML
GLOBAL: □
GROUPBY:
1PASS: □
DEBUG: □

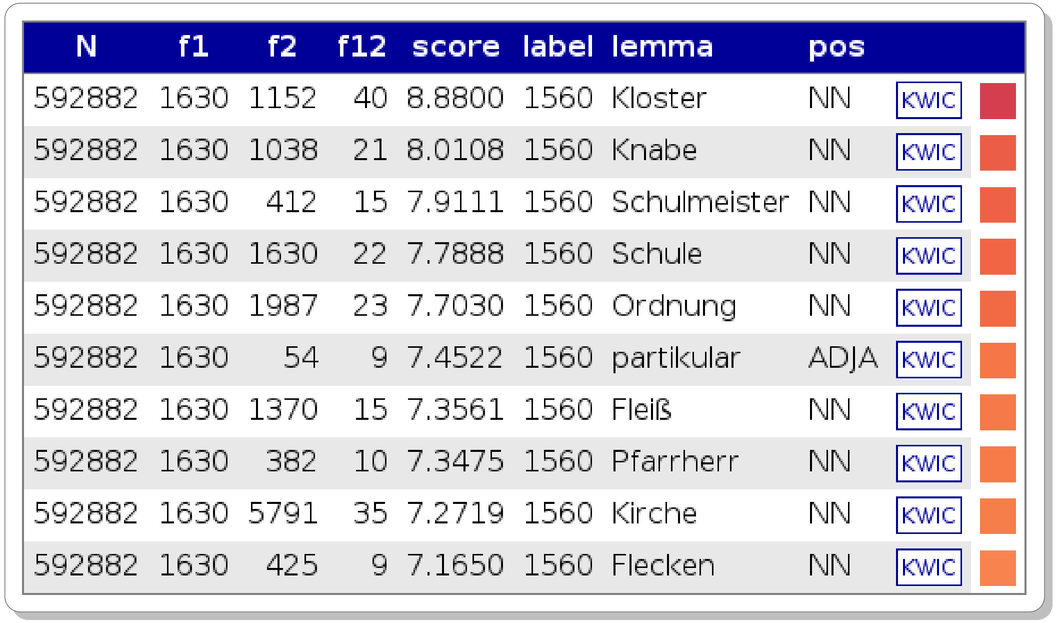

log Dice (ld)
Frequency (f)
Frequency per Million (fm)
Log-Frequency (lf)
Log-Frequency per Million (lfm)
Mutual Information (mi1)
Mutual Information³ (mi3)
Mutual Information * log f (milf)
log Dice (ld)
log likelihood (ll)





'*Schule*': DiaCollo Collocates (DTA: HTML)

1560–1569

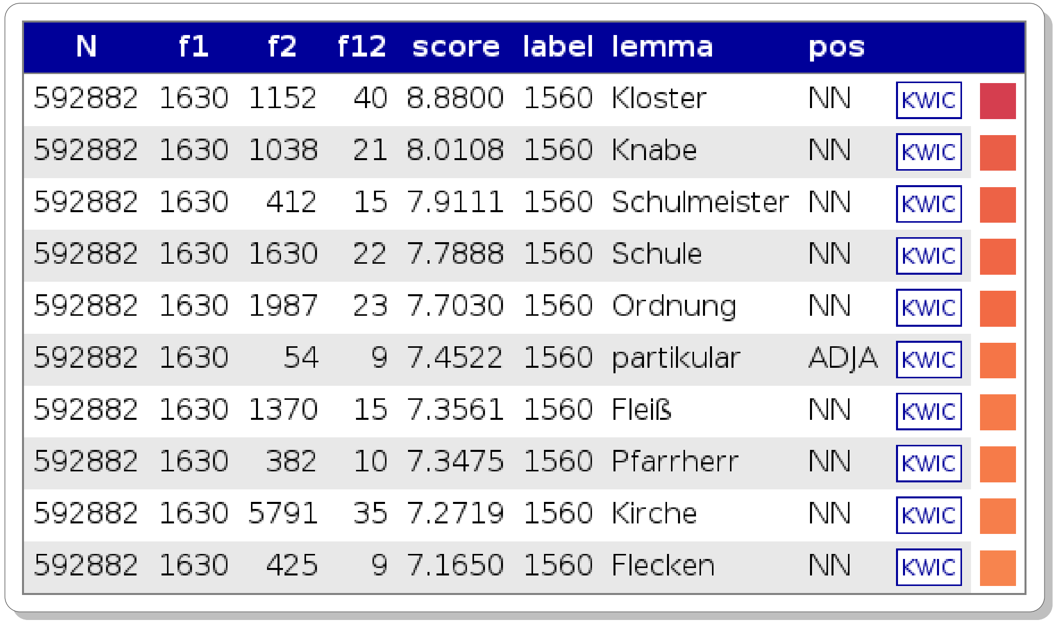


N	f1	f2	f12	score	label	lemma	pos		
592882	1630	1152	40	8.8800	1560	Kloster	NN	KWIC	🟥
592882	1630	1038	21	8.0108	1560	Knabe	NN	KWIC	🟧
592882	1630	412	15	7.9111	1560	Schulmeister	NN	KWIC	🟧
592882	1630	1630	22	7.7888	1560	Schule	NN	KWIC	🟧
592882	1630	1987	23	7.7030	1560	Ordnung	NN	KWIC	🟧
592882	1630	54	9	7.4522	1560	partikular	ADJA	KWIC	🟧
592882	1630	1370	15	7.3561	1560	Fleiß	NN	KWIC	🟧
592882	1630	382	10	7.3475	1560	Pfarrherr	NN	KWIC	🟧
592882	1630	5791	35	7.2719	1560	Kirche	NN	KWIC	🟧
592882	1630	425	9	7.1650	1560	Flecken	NN	KWIC	🟧

- association with religious institutions
  - *Kloster* ("cloister")
  - *Pfarrherr* ("pastor")
  - *Kirche* ("church")

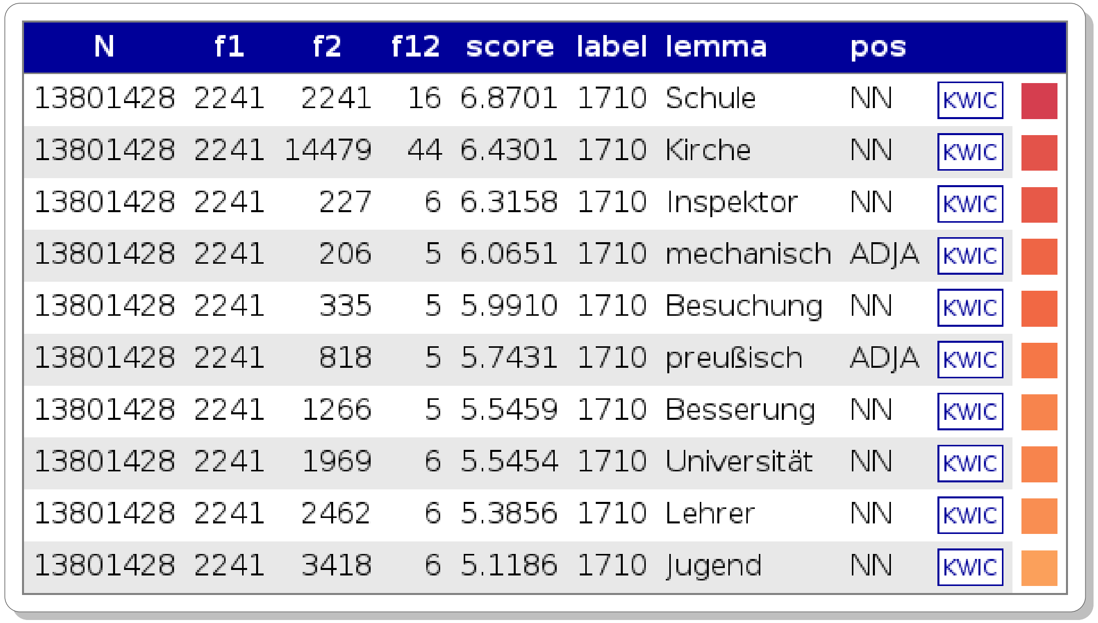
‘*Schule*’: DiaCollo Collocates (DTA: HTML)

1560–1569



N	f1	f2	f12	score	label	lemma	pos		
592882	1630	1152	40	8.8800	1560	Kloster	NN	KWIC	🟥
592882	1630	1038	21	8.0108	1560	Knabe	NN	KWIC	🟥
592882	1630	412	15	7.9111	1560	Schulmeister	NN	KWIC	🟥
592882	1630	1630	22	7.7888	1560	Schule	NN	KWIC	🟧
592882	1630	1987	23	7.7030	1560	Ordnung	NN	KWIC	🟧
592882	1630	54	9	7.4522	1560	partikular	ADJA	KWIC	🟧
592882	1630	1370	15	7.3561	1560	Fleiß	NN	KWIC	🟧
592882	1630	382	10	7.3475	1560	Pfarrherr	NN	KWIC	🟧
592882	1630	5791	35	7.2719	1560	Kirche	NN	KWIC	🟧
592882	1630	425	9	7.1650	1560	Flecken	NN	KWIC	🟧

1710–1719



N	f1	f2	f12	score	label	lemma	pos		
13801428	2241	2241	16	6.8701	1710	Schule	NN	KWIC	🟥
13801428	2241	14479	44	6.4301	1710	Kirche	NN	KWIC	🟥
13801428	2241	227	6	6.3158	1710	Inspektor	NN	KWIC	🟥
13801428	2241	206	5	6.0651	1710	mechanisch	ADJA	KWIC	🟧
13801428	2241	335	5	5.9910	1710	Besuchung	NN	KWIC	🟧
13801428	2241	818	5	5.7431	1710	preußisch	ADJA	KWIC	🟧
13801428	2241	1266	5	5.5459	1710	Besserung	NN	KWIC	🟧
13801428	2241	1969	6	5.5454	1710	Universität	NN	KWIC	🟧
13801428	2241	2462	6	5.3856	1710	Lehrer	NN	KWIC	🟧
13801428	2241	3418	6	5.1186	1710	Jugend	NN	KWIC	🟧

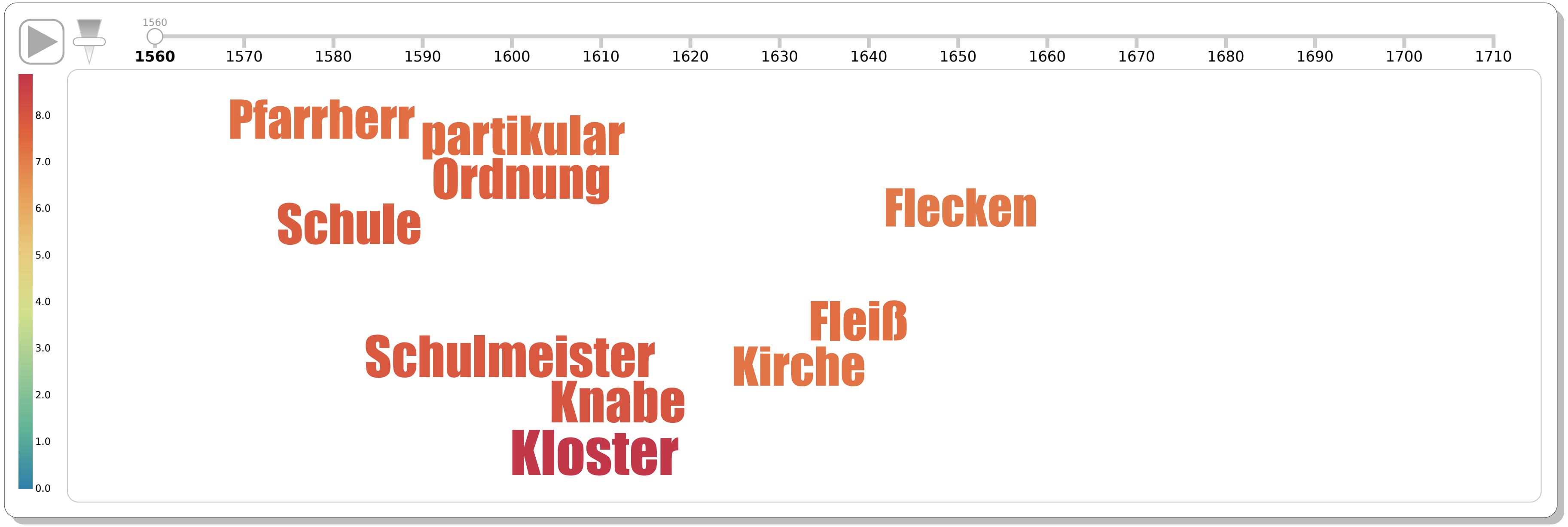
- association with religious institutions
  - *Kloster* (“cloister”)
  - *Pfarrherr* (“pastor”)
  - *Kirche* (“church”)

- stronger secular associations
  - *Inspektor* (“inspector”)
  - *preußisch* (“prussian”)
  - *Universität* (“university”)
- trend continues as time progresses

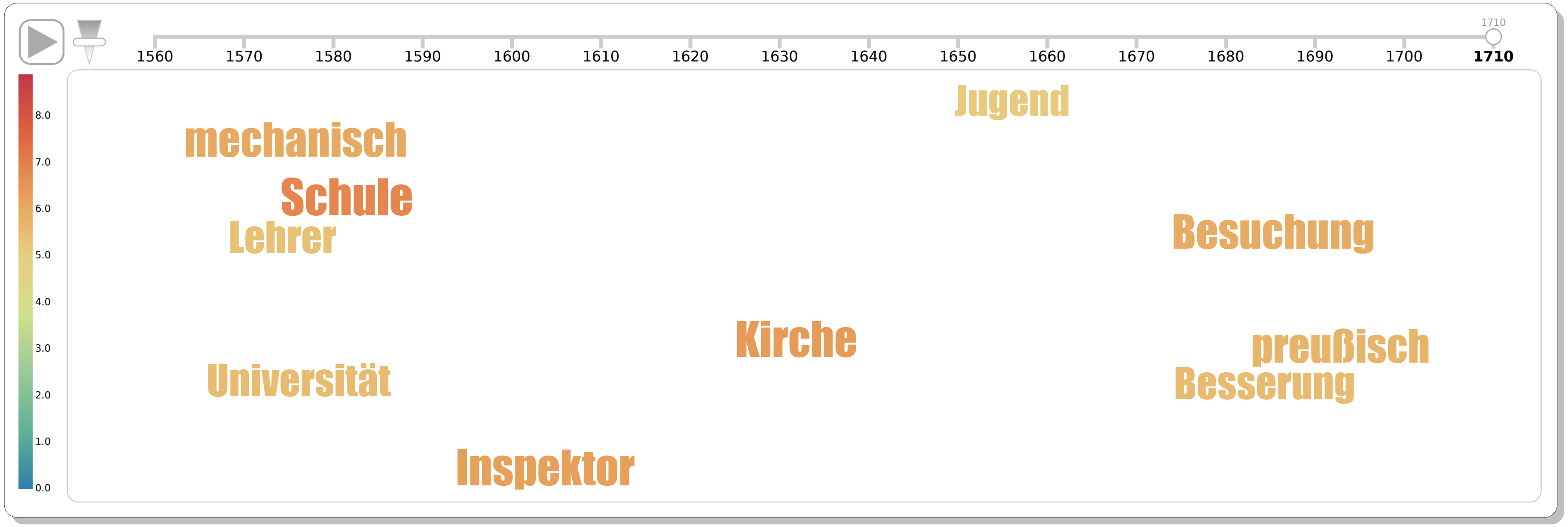
‘*Schule*’: DiaCollo Collocates (DTA: lemma-cloud)

CLARIN-D

1560s:



1710s:



GEORG ECKERT INSTITUT Leibniz-Institut für internationale Schulbuchforschung

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Die Grenzboten Corpus

Die Grenzboten 67. 2. 1908
Die Grenzboten 63. 4. 1904
Grenzboten 52. 3. 1893
Grenzboten 52. 4. 1893
Die Grenzboten 65. 2. 1906

Image: SuUB Bremen

http://brema.suub.uni-bremen.de/grenzboten
http://www.deutschestextarchiv.de/doku/textquellen#grenzboten

- *Die Grenzboten* (“the messengers from the border(s)”) was a bi-weekly national-liberal German language periodical published 1841–1922
- covered a wide range of politics, literature, and the arts throughout the ‘long’ nineteenth Century
- 270 volumes (ca. 187,000 pages) digitized, OCR’ed, and structured by the SuUB Bremen in the context of a DFG-Project
  - integrated into the corpus research infrastructure of the *Deutsches Textarchiv* at the BBAW CLARIN Service Center

Are *Die Grenzboten* concerned with education?

Step 1: query corpus vocabulary database (LexDB)

- identify relevant terms in the corpus, e.g. *Schule* (“school”), 1840–1899
  - . . . in the *Deutsches Textarchiv*: 101.52 per million tokens
  - . . . in *Die Grenzboten* : **237.29** per million tokens

Step 2: query DiaCollo

- identify strong collocates for *Schule* (“school”)
- identify possible debates in the corpus via query results
- close reading in the texts via “keyword-in-context” (KWIC) hyperlinks

Education Policy & Religion

Collocate ‘*Kirche*’ (“church”)

- persistently prominent throughout the entire *Grenzboten* corpus
- 1850s–1880s: *konfessionell* (“confessional”)
- 1890s–1910s: *Religionsunterricht* (“religious education”)

Refining the Search

- restrict to attributive adjective collocates (`GROUPBY: l,p=ADJA`)

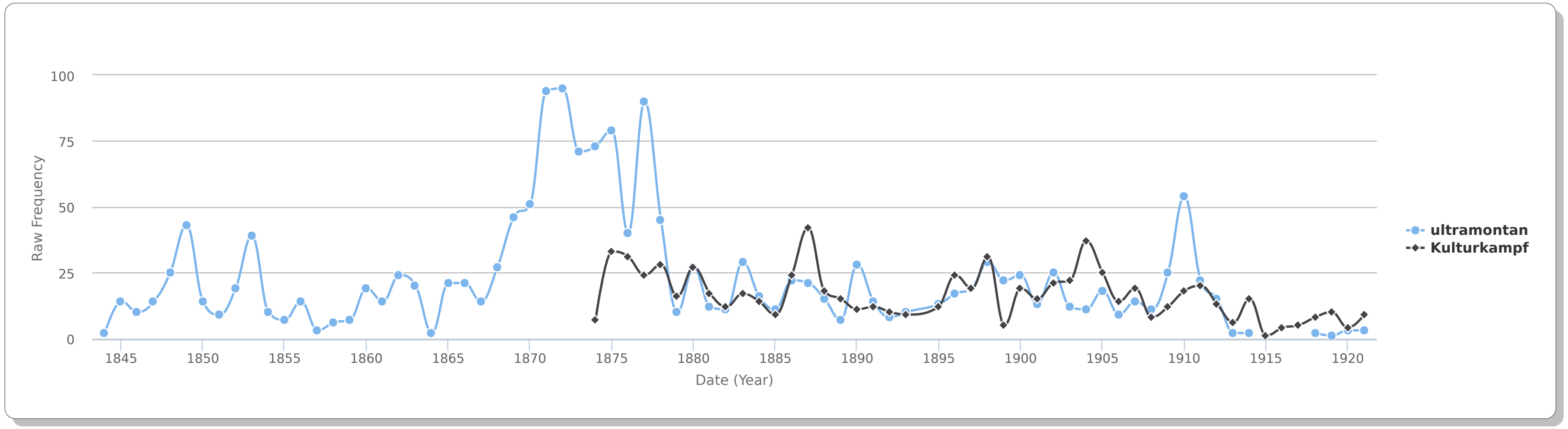
protestantisch	(“protestant”)	1860s
katholisch	(“Catholic”)	1860s-1870s
evangelisch	(“Protestant, Evangelical”)	1860s-1870s
konfessionell	(“confessional”)	1860s-1880s
kirchlich	(“churchly”)	1870s

- collocates related to church & religious confession peak in the 1860s–1870s
- also prominent: *öffentlich* (“public”; 1840s, 1870s–1900s)
  - KWIC ⇝ stance of publicly funded schools w.r.t. church influence in education

Education Policy: *Kulturkampf*

Kulturkampf ("cultural struggle")

- rights & influences of state (Prussia) vs. church (Pope Pius IX)
- *ultramontan* ("ultramontane") ⇝ staunch supporters of the Catholic Church



Refining the Search: GermaNet thesaurus + paragraph search window

(Hamp & Feldweg 1997; Henrich & Hinrichs 2010)

- corpus hits show evidence for anti-Catholic opinions in debates on education
  - who should be in charge of education and curricula?
  - how to deal with different religious denominations in schools?

Upshot

- some important aspects of debate are **not** apparent from initial naïve DiaCollo queries
- informed curiosity & focused investigation leads to very satisfying results

Summary & Conclusion

Collaborative Development

- cyclic process ⇝ *feedback loop*
- elusive common ground ⇝ *terminology, research methodology*

DiaCollo

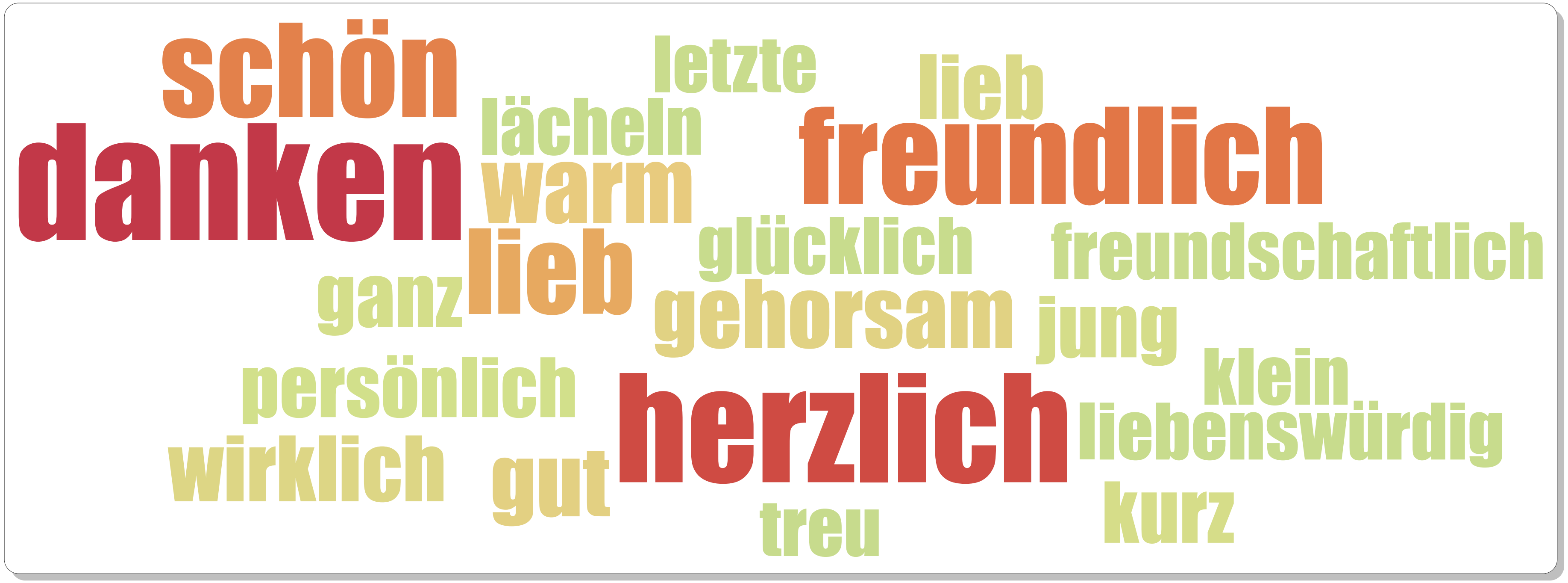
- diachronic text corpora ⇝ *semantic shift, discourse trends*
- conventional tools ⇝ *implicit assumptions of homogeneity*
- diachronic profiling ⇝ *date-dependent lexemes*

. . . as a tool for historical research

- fluent “blended”/“scalable” reading ⇝ *distant ↔ close reading*
- digital corpora (sources) ⇝ *quantity, quality, legal issues*

CLARIN-D

— The End —



Thank you for listening!

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

http://metacpan.org/release/DiaColloDB

GEORG ECKERT INSTITUT Leibniz-Institut für internationale Schulbuchforschung

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

References

References

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

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/.

J. R. Firth. *Papers in Linguistics 1934–1951*. Oxford University Press, London, 1957.

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.

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, 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.

H. Kermes, S. Degaetano, A. Khamis, J. Knappen, and E. Teich. The Royal Society corpus: From uncharted data to corpus. In *Proceedings of LREC 2016*, Portoroz, Slovenia, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/summaries/792.html.

References

C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

A. Stulpe and M. Lemke. Blended reading. In M. Lemke and G. Wiedemann, editors, *Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*, pages 17–61. Springer, 2016. doi:10.1007/978-3-658-07224-7_2.