Visualizing Semantic Change with DiaCollo

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

Thomas Werneke
Centre for Contemporary History, Potsdam
werneke@zzf-potsdam.de

Genealogies of Knowledge:
Translating Political and Scientific Thought across Time and Space
University of Manchester

8th December, 2017

Overview

The Situation

- Diachronic Text Corpora
- Collocation Profiling
- Diachronic Collocation Profiling

DiaCollo

- Requests & Parameters
- Profiles, Diffs & Indices

Use Cases

- *Die Grenzboten* corpus
- *Die Grenzboten* & anti-semitism
- *Die Grenzboten* & education policy

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)*
  - Royal Society Corpus (RSC) *(Kermes et al. 2016)*
  - 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

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:
  - Royal Society *Philosophical Transactions* (1665–1869, 9.8K documents, 35M tokens)
  - 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: 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)*

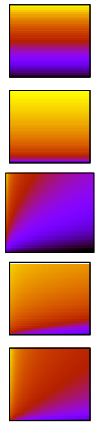
Indices & Attributes

- compile-time filtering of native indices: frequency 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

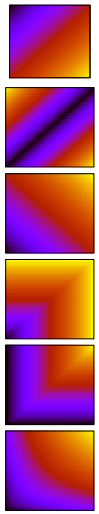
Selected Score Functions

- **f** raw collocation frequency $= f_{12}$
- **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) $\approx \mathrm{sgn}(f_{12} | f_1, f_2) \times \log \frac{\mathrm{L}(H_0)}{\mathrm{L}(H_1)}$
- **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 $= |s_\mathrm{a} - s_\mathrm{b}|$
- **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 $\approx \frac{2 s_\mathrm{a} s_\mathrm{b}}{s_\mathrm{a} + s_\mathrm{b}}$



Use Cases

Die Grenzboten Corpus

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
  - coverage of civic life, opinions, and debates surrounding the revolution of 1848, the restoration period, industrialization, the German Empire (*Kaiserreich*), and the First World War ⇝ valuable source for a broad range of disciplines
- 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

Use Cases

Basic Idea

- explore the corpus of national-liberal cultural history in order to...
  - test and demonstrate the utility of DiaCollo itself,
  - and that of other NLP software tools provided by CLARIN-D

  ...for the field of historical semantics.
- method: "blended reading"

Use Case 1

- *Die Grenzboten*'s stance towards Jewish people (Thomas Werneke)

Use Case 2

- *Die Grenzboten*'s content with respect to education policy (Maret Nieländer)

Blended Reading

CLARIN-D

≈ ***Close Reading + Distant Reading***

"Unter dem Begriff des 'blended reading' schlagen wir eine Strategie im Sinne einer Best Practice vor, die semiautomatische Analyseverfahren mit klassischer Textlektüre so integriert, dass sozialwissenschaftliche Erkenntnispotenziale, die sich auf die Auswertung großer Textdatenmengen stützen, optimal ausgeschöpft werden."

"With the concept of 'blended reading' we propose a best-practice strategy which integrates semi-automatic analysis techniques with traditional (hermeneutic, 'close') reading such that the potential for socioscientific insights supported by the evaluation of large text data-sets is optimally exploited."

— Stulpe & Lemke (2016)

Blended Reading: anti-semitism

CLARIN-D

Searching for anti-semitic ressentiments within the corpus

Step 1

- isolate a promising time slice
  - choosing the 1880s, covering the '*Antisemitismusstreit*'

Step 2

- identify collocates of the term *Jude* ("jew")
  - select unusual collocates typically relating to ressentiments

Step 3

- analyze text passages via key-word-in-context (KWIC)
  - conventional source exegesis

Blended Reading: anti-semitism

Collocates related to anti-semitic ressentiments / stereotypization

Query: 50 strongest common noun collocates of the term '*Jude*', 1880–1889

Four collocates evoked particular curiosity:

- *Zahl* ("number")
- *Vermehrung* ("reproduction, proliferation")
- *Wucher* ("usury")
- *Handel* ("trade")

Close Reading Example: '*Vermehrung*'∼"proliferation"

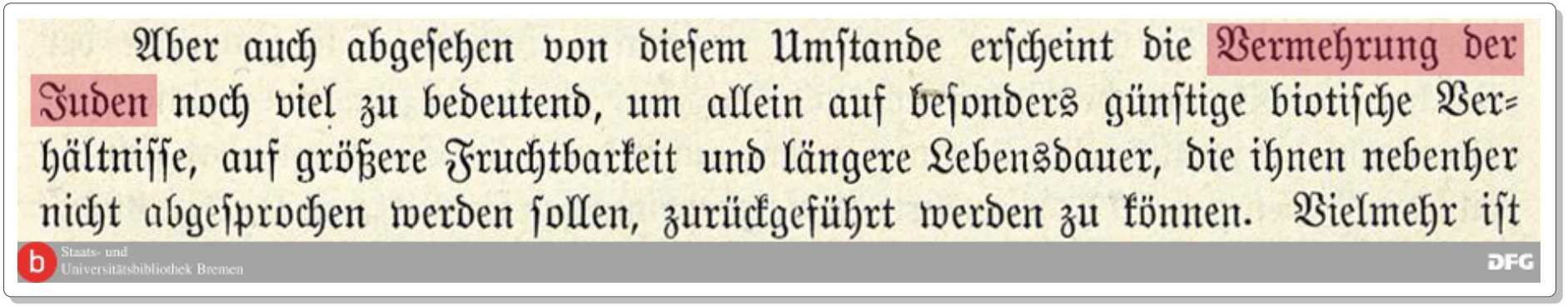
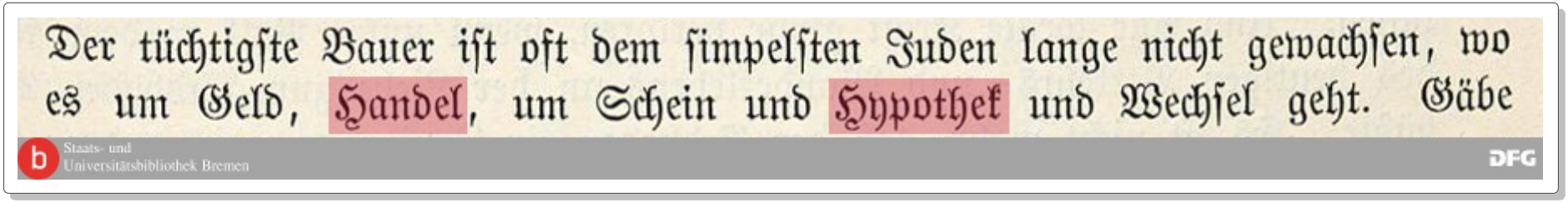

Aber auch abgesehen von diesem Umstande erscheint die Vermehrung der Juden noch viel zu bedeutend, um allein auf besonders günstige biotische Verhältnisse, auf größere Fruchtbarkeit und längere Lebensdauer, die ihnen nebenher nicht abgesprochen werden sollen, zurückgeführt werden zu können. Vielmehr ist

Image: SuUB Bremen

"But, apart from these circumstances, the ***proliferation*** [Vermehrung] ***of the Jews*** *still appears too signicant to be attributed only to particularly favorable biotic conditions, to greater fertility and longer lifespan, which by the way should not be denied them."*

— In: "Die Juden in Österreich." *Die Grenzboten*, Jg. 41 (1882), p. 629.

Close Reading Example: '*Handel*'∼"trade"


Der tüchtigste Bauer ist oft dem simpelsten Juden lange nicht gewachsen, wo es um Geld, Handel, um Schein und Hypothek und Wechsel geht. Gäbe

Staats- und Universitätsbibliothek Bremen DFG

Image: SuUB Bremen

> *"The most capable farmer often cannot measure up to the simplest Jew when it comes to money and **trade**, to certificate and **mortgage** and exchange."*

— In: "Erdboden." *Die Grenzboten*, Jg. 59 (1900), p. 271.

Use Case 1: Conclusions

- We have found evidence of stereotypizations, pointing to a cultural racism.
- The authors in part convey their arguments together with typical cultural anti-semitic ressentiments.

Blended Reading: *Schule* ("school")

CLARIN-D

Step 1: query corpus vocabulary database (LexDB)

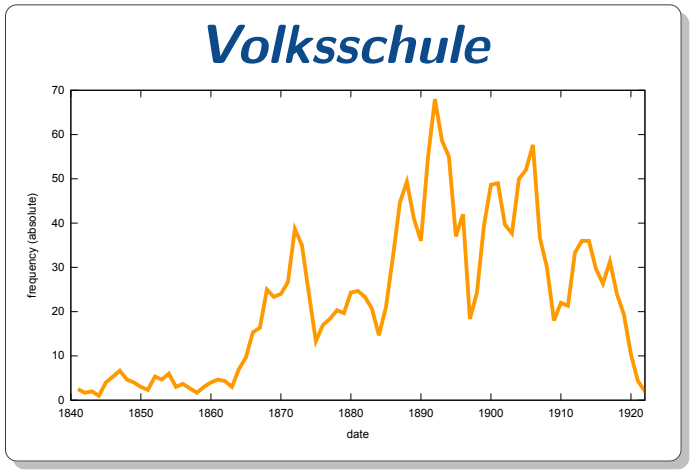
- identify relevant terms within the corpus (e.g. *Schule*, 1840–1899)
  - ...in the *Deutsches Textarchiv*: 98.79 per million tokens
  - ...in *Die Grenzboten* : 237.86 per million tokens
- select interesting terms
  - identify high-frequency compounds containing *Schule*

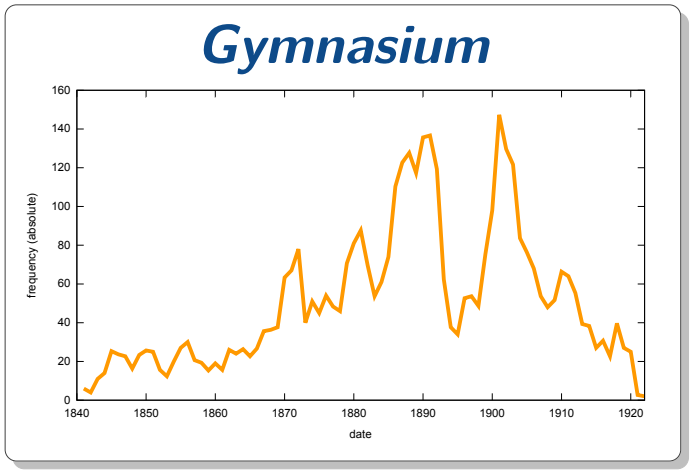
Step 2: query DiaCollo

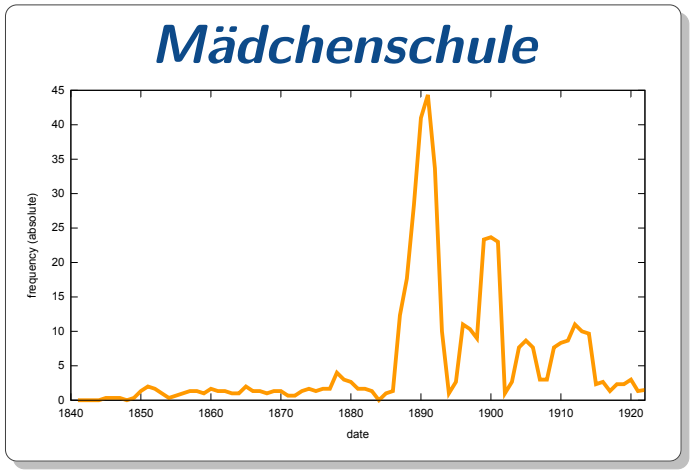
- identify strong adjective, common noun, & proper name collocates
- identify possible debates in the corpus via query results
- close reading in the texts via "key-word-in-context" (KWIC) hyperlinks

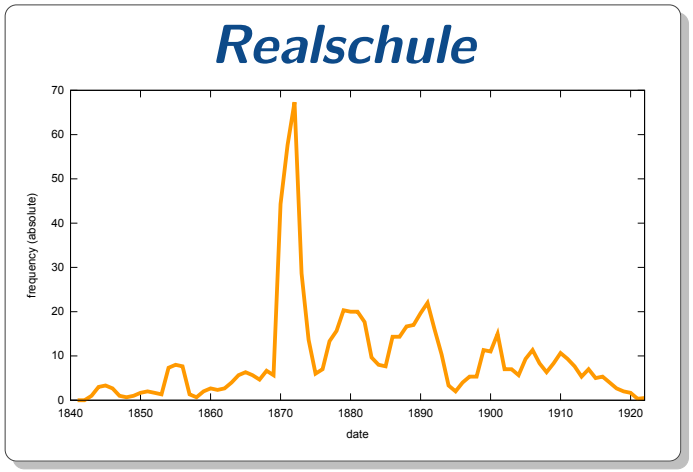
Blended Reading: Frequencies & Time Series

CLARIN-D





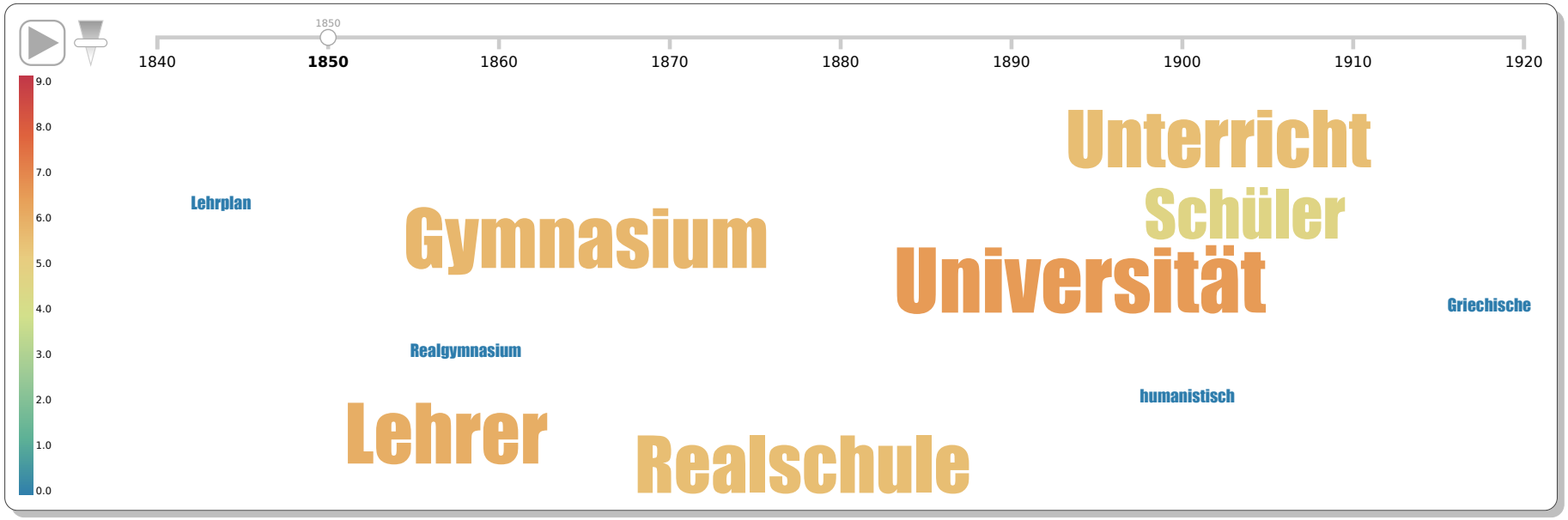




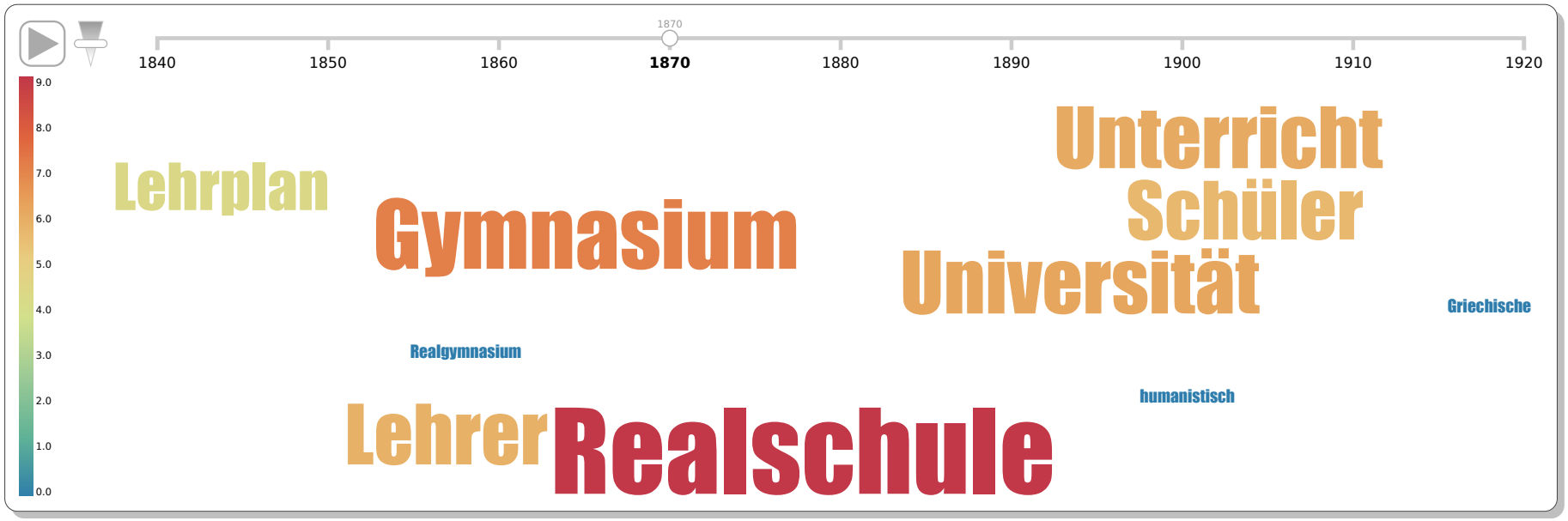
Lemma	Frequency
Schule	20960
Hochschule	2156
Volksschule	1860
Realschule	762
Mädchenschule	450
Mittelschule	416
Bürgerschule	280
Fortbildungsschule	269
Fachschule	248
Vorschule	237
Elementarschule	234
Malerschule	186
Privatschule	184
Einheitsschule	183
Tochterschule	171
Kunstschule	168
Gelehrtenschule	159
Klosterschule	141
Dorfschule	126
Gewerbeschule	124
Kriegsschule	116
Lateinschule	110
Militärschule	95
Knabenschule	92
Oberrealschule	88
Staatsschule	85
Musikschule	83

‘*Gymnasium*’: DiaCollo Collocates

1850s:



1870s:



'*Erziehung*': DiaCollo Collocates

CLARIN-D

("education, upbringing")

Attributive Adjectives (ADJA)

- 1850s: *häuslich, sittlich* ("domestic, moral")
- 1870s: *weiblich, ästhetisch* ("feminine, aesthetic")
- 1910s: *national*, later ***staatsbürgerlich*** ("national, civic")

Common Nouns (NN)

- *Menschengeschlecht* ("humankind" $\rightsquigarrow$ Lessing's *opus magnus* of 1780)
- 1860s: *Jugend*, *Kind* ("youth, child")
- 1870s: *Unterricht*, *Schule* & *Ausbildung* appear ("instruction, school, training")

Proper Names (NE)

- 1840s: *Frankreich* ("France")
- 1860–1880s: Hegel, Lessing, Fichte, Schiller, Rousseau
- 1890–1910s: Paul Güßfeldt, Georg Kerschensteiner

Education Policy & Religion

Collocate '*Kirche*' ("church")

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

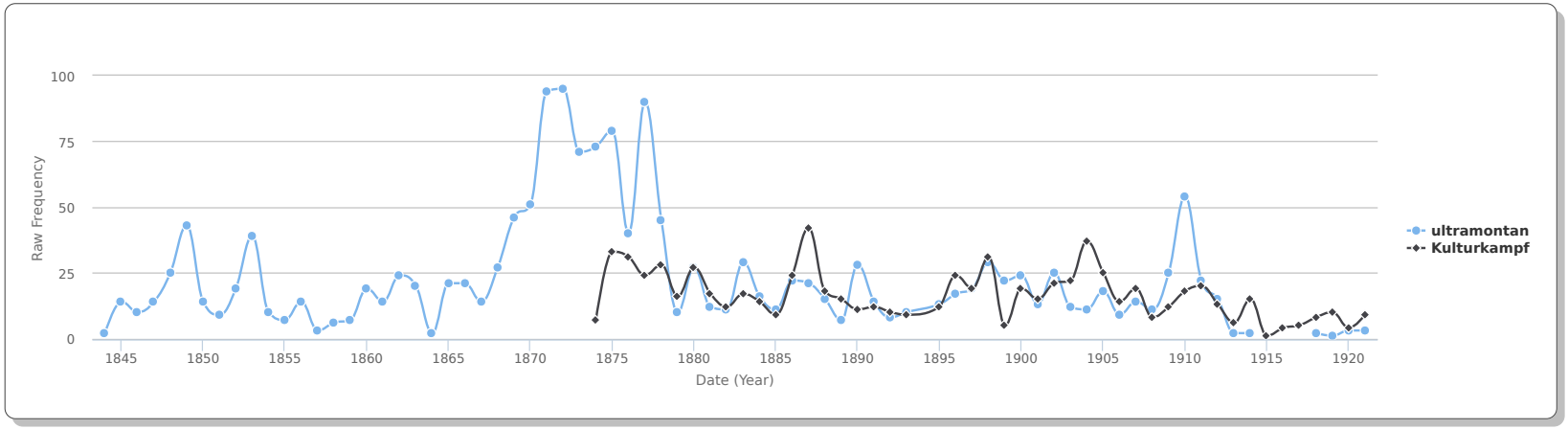
Refining the Search

- 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?

Use Case 2: Conclusions

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

Summary & Conclusion I

Diachronic Collocation Profiling

- "meaning is use" ⇝ *. . . in temporal context*
- 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-level granularity*
- DDC/D* integration ⇝ *fine-grained queries, corpus KWIC links*
- RESTful web service ⇝ *external API, online visualization*

Summary & Conclusion II

DiaCollo's Application in the field of Historical Semantics

Advantages

- support for semiasological methods
- inductive and explorative methods
- fluent "blending" from distant to close reading (and back)

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

GEORG ECKERT INSTITUT Leibniz-Institut für internationale Schulbuchforschung

ZZF POTSDAM

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.

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

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.

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.

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.

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.

References

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.

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.

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, 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.

References

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

F. Moretti. *Distant reading*. Verso Books, 2013.

N.N. Die Juden in Österreich. *Die Grenzboten*, Jg. 41:627–634, 1882. URL http://brema.suub.uni-bremen.de/grenzboten/periodical/titleinfo/168690.

N.N. Erdboden. *Die Grenzboten*, Jg. 59:265–272, 1900. URL http://brema.suub.uni-bremen.de/grenzboten/periodical/titleinfo/292689.

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.

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