Auto-IA Workshop HMTM Hannover

TEXT MINING FOR SOCIAL SCIENTISTS

Lexicometrics

Gregor Wiedemann | g.wiedemann@leibniz-hbi.de
Media Research Methods Lab
Leibniz-Institute for Media Research | Hans-Bredow-Institut

Andreas Niekler | aniekler@informatik.uni-leipzig.de
Abteilung Automatische Sprachverarbeitung, Institut für Informatik
Universität Leipzig

This lecture

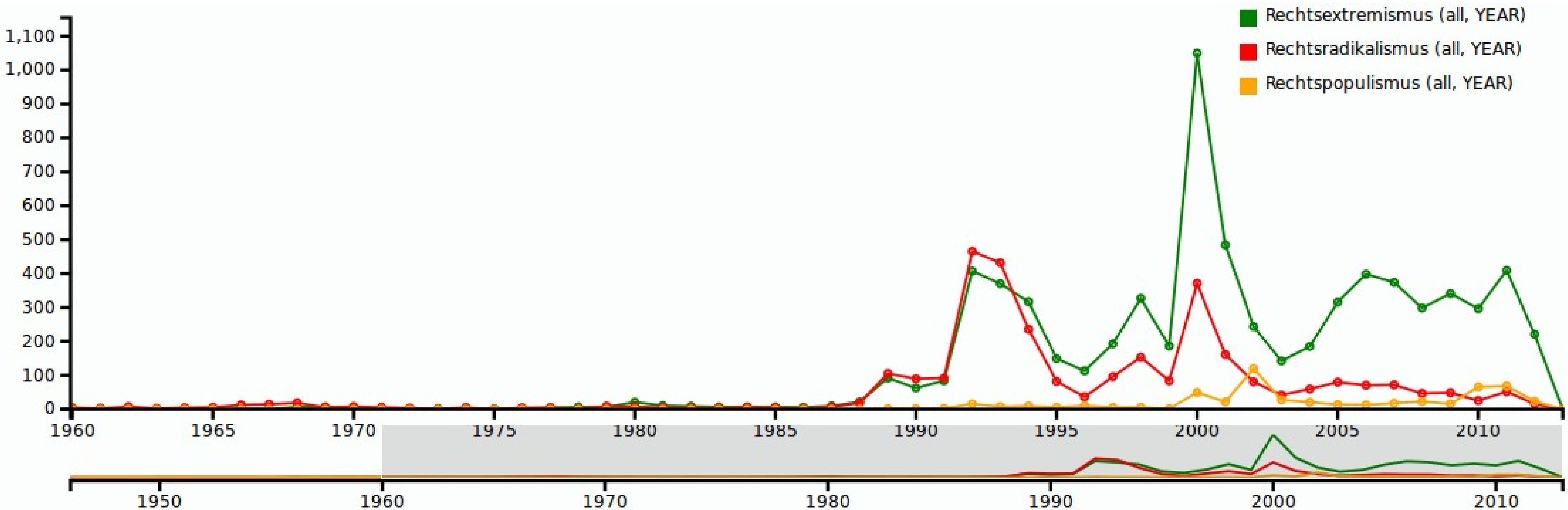
1. Frequency analysis
2. Key term extraction
3. Cooccurrence analysis

Frequency analysis

Frequency analysis

- Motivation: Analysis: comparing frequencies of units of analysis per context
  - 1) between different UoA
  - 2) in different collections
  - 3) over time
- Possible Units of Analysis (UoA):
  - **terms** → in CA we often will concentrate on those
  - concepts (set of terms), ...
  - documents, paragraphs, ...
  - linguistic units (sentences, punctuation marks, vowels, ...)
- Context Units
  - term frequency: frequency of a term within a document / entire collection
  - document frequency: frequency of documents containing a term

Frequency analysis



- **Problems of „term as events":**
  - distribution of language data → keep Zipf's law in mind
  - „burstiness of terms"
    - → probability of a word occuring again after seen once increases drastically
    - → use log(tf(w)) or df(w) ?
  - varying collection sizes → normalize frequencies by collection size!

Zipf's law

ASV
Automatische Sprachverarbeitung

- George K. Zipf 1935: observation on distribution of terms in a corpus
  - List types of a corpus by frequency (n) and assign a rank (r) such that the most frequent type has rank 1
  - rank of a word multiplied by its frequency is roughly constant (k):

$$r \times n \approx k$$

type	frequency n	rank r	r * n
sich	1.680.106	10	16.801.060
immer	197.502	100	19.750.200
Mio	66.116	500	18.059.500
Medien	19.041	1000	19.041.000
Miete	3.755	5000	18.775.000
vorläufige	1.664	10000	16.640.000

[Data: Projekt „Deutscher Wortschatz“]

<u>Implication:</u> from Zipf's Law can be derived that roughly 50% of the vocabulary occurs only once in every document/collection!

(see Heyer/Quasthoff/Wittig 2006)

Aggregation / Normalization

- analytic insights in frequency analysis comes from relating / comparing frequency information (→ e.g. time series)
- frequencies are easily summable to create information on higher / more abstract levels

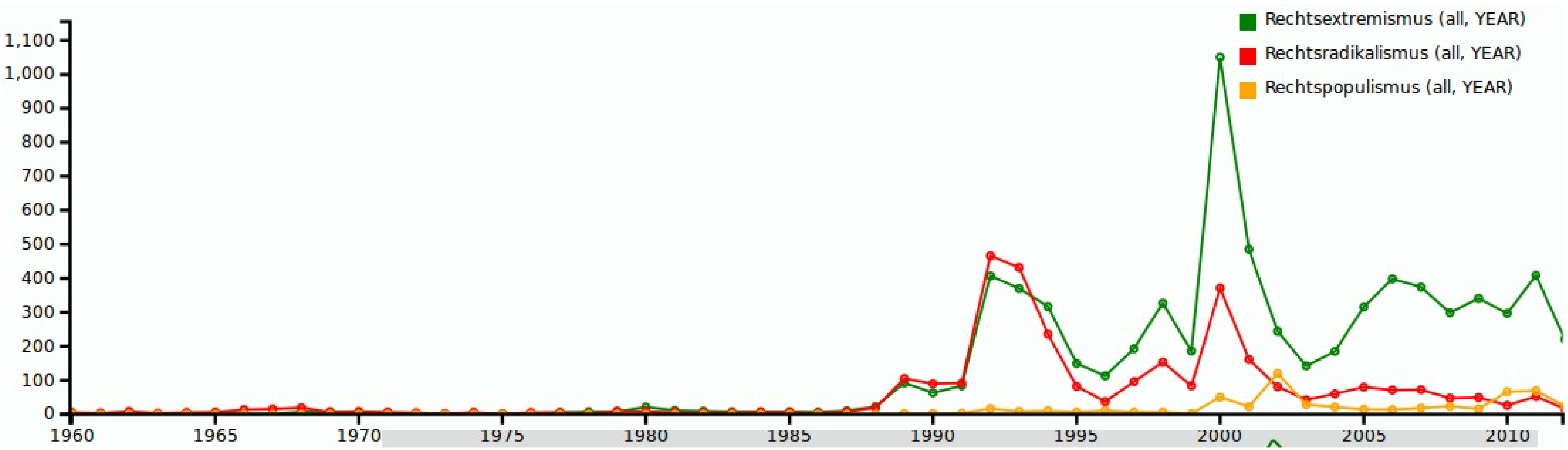
```
– tf(w, d) – frequency of term w in document d
– tf(w, m) – sum of tf(w, d) for all d in month m
– tf(w, y) – sum of tf(w, y) for all d in year y, or
                          tf(w, m) for all m in y
```

- in many cases absolute frequencies may be hard to interprete
- normalization: relative frequencies by all terms/documents in base population

$$tfnorm_w = \frac{tf(w,y)}{\sum_{d \in D_y} \sum_{t \in V} tf(t,d)} \qquad dfnorm_w = \frac{df(w,y)}{|D_y|}$$

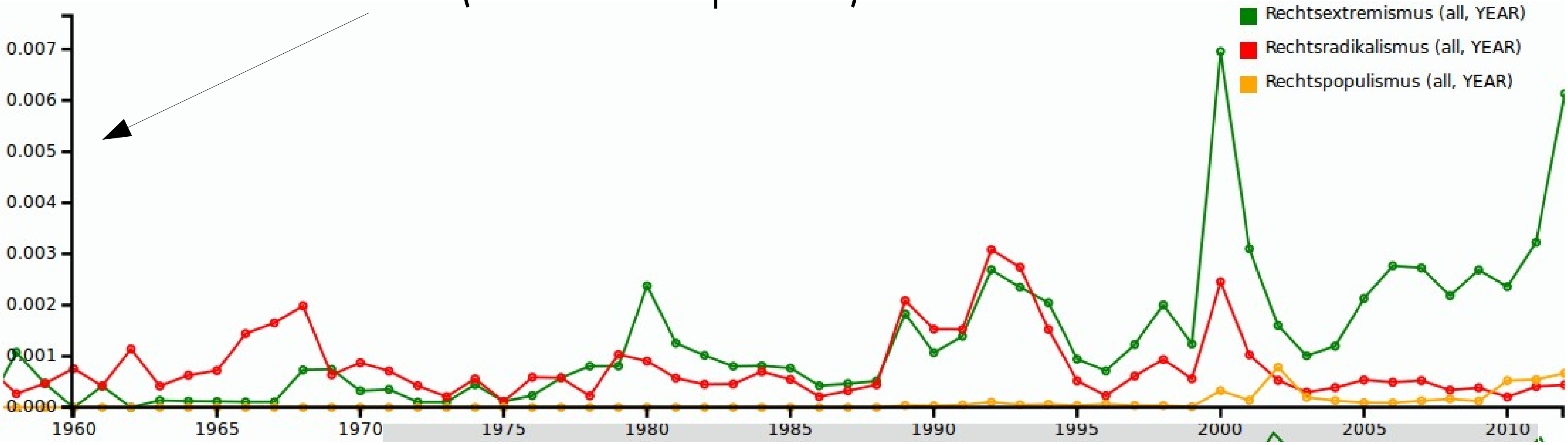
D_y – set of documents in year y
V – vocabulary

Frequency analysis



Frequency analysis

relative scale (document frequencies)



Visualization as word clouds

Concepts / dictionaries

- dictionaries (list of words) may be compiled to count conceptual events
  - e.g. basic approach of *sentiment analysis* → identification of subjective mood in source materials
    - positive terms: {good, awesome, brilliant, gorgious, ...}
    - negative term: {bad, aweful, horrifying, devastating, ...}

→ Tutorial 3

  - intersection of discoursive fields:
    - war terminology: {blitz, bomb, formation, neutral zone, red zone, kamikaze, ...} measured in articles about soccer, american football or quidditch
  - operationalization of theoretical hypthothesis
    - TINA rethorics: {no alternative, no other possibility, impossible, indispensable, ...} with respect to different policy fields
- **caution:**
  - Should all events count equally? (e.g. sentiments)
  - does occurrence match appropriate context? (feature-/aspect based sentiments)

Applying frequency analysis

- Context matters!
  - counting simple occurrence usually neglects contexts
  - but, right contexts can be assured by previous selection strategies
    - e.g. counting „no alternative" in documents on European politics compared to a general corpus
  - ← Applying filter beforehand increases chances to generate informative data
- Utilization of frequency data for description / identification of
  - content shares → e.g. pie chart
  - trends / time series → e.g. line chart
- consider normalization strategies

Key term extraction

Key term extraction

- One task, many names:
  - „Terminology mining, term extraction, term recognition, or glossary extraction, is a subtask of information extraction. The goal of terminology extraction is to *automatically* extract relevant terms from a given corpus." [Wikipedia]
- Extended task: „Ontology learning"
  - key terms and their (hierarchical) relationships (e.g. is-a, part-of, hypernym/hyperonym, synonym/antonym relations)
- Evaluation:
  - judgements on relevancy done by human experts

- approaches based on:
  - Frequency
    - Frequency
    - TF-IDF
  - Difference corpus
    - Log likelihood
    - Characteristic elements diagnostics

Frequency

- Assumption
  - the more frequent, the more important
  - removing stop words helps to identify more relevant terms
- Evaluation
  - language is Zipf distributed
  - raw frequency does not cover relevancy well

- example:
  - protest data TAZ (2000-2009)
- approach to get n most relevant terms
  1) create DTM from corpus
  2) compute vector v of column sums
  3) order v in decreasing order
  4) output item 1 to n of v

type/frequency		type/frequency (sw removed)	
die	22807	polizei	2072
der	19938	menschen	1492
und	11426	demonstration	1105
den	7180	berlin	982
das	5560	uhr	968
von	5145	demonstranten	961
auf	5013	kundgebung	700
mit	4957	samstag	666
sich	4668	neonazis	651
dem	4205	worden	632
ein	4188	straße	625
nicht	3909	npd	572
für	3858	berliner	558
eine	3625	jahr	537
ist	3486	jahren	534
des	3308	teilnehmer	532
sie	3299	rechten	484
auch	3115	seien	477
gegen	3070	demo	472
als	2521	motto	459

TF-IDF

- remember TF-IDF from Information Retrieval:
  - relevancy is correlated with term frequency and inversed document frequency

$$N = |D|$$

$$idf_w = \log\left(\frac{N}{n_w}\right)$$

$$weight_w = tf_{wd} \cdot idf_w$$

polizei	7.089975
rund	6.731556
neonazis	6.309970
uhr	6.270761
samstag	6.236580
kundgebung	6.021211
menschen	5.990043
npd	5.892383
sie	5.598942
gestern	5.563909
ist	5.556133
etwa	5.476461
berlin	5.457175
aufmarsch	5.453003
demonstranten	5.437748
hatten	5.436246
teilnehmer	5.336355
rechten	5.246831
nicht	5.204938
unter	4.991625

Corpus comparison

- Difference based Term Extraction methods follow a different approach:
  - comparing frequencies in a target corpus T with frequencies in a general comparison corpus C
  - significant deviation in T from expected term distribution measured in C is considered as relevancy criterion
- Tests used in CA
  - Log Likelihood (Dunning 1993; Rayson/Garside 2000)
  - Characteristic elements diagnostics (Lebart/Salem 1994)

Log Likelihood

- Contingency Table

	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

- Log Likelihood
  - E1 = c*(a+b) / (c+d)
  - E2 = d*(a+b) / (c+d)
  - LL = 2*((a*log (a/E1)) + (b*log (b/E2)))

	LL	Frq
NPD	7867,59	1157
Demonstration	7789,70	1295
Demo	7098,27	829
Demonstranten	5463,47	1042
Kundgebung	5306,27	790
Neonazis	5224,08	751
Polizei	5165,54	2262
Aufmarsch	3704,68	468
Gegendemonstranten	2811,12	320
Neonazi	2565,88	305
taz	2438,61	474
Anti	2380,52	232
Antifa	2237,23	243
Demonstrationen	1841,17	400
Teilnehmer	1722,93	582
Teilnehmern	1464,79	335
Bündnis	1390,89	396
Nazis	1377,34	359
Motto	1370,38	468
Protest	1324,34	412

Summary

- lot's of approaches to extract terms...
- corpus comparison methods are usually better than frequency based methods
- disadvantage?: terms are observed independently of each other
- LL and „characteristic elements diagnostic" well established in corpus linguistic literature

Cooccurrence Analysis

Cooccurrence Analysis

- Structuralist semantics (F. de Saussure):
  - syntagmatic relation: signifiers which occur conjointly complement w.r.t function and content
  - paradigmatic relation: signifiers which occur in similar contexts have similar function w.r.t. grammar and content → cp. distributional hypothesis
- Computing cooccurrences
  - local context C(w): set of words that occur in the same 'window' as *w*
  - global context G(w): set of words which occur conjointly with *w* in a *statistically significant* manner
  - windows: sentences, paragraphs, documents, headlines, k left/right neighbour words

Cooccurrence Analysis

The sun is shining.	$C_{sentence}$(sun) = {The, is, shining}
The sun is burning.	$C_{sentence}$(sun) = {The, is, burning}
The light is shining.	$C_{sentence}$(light) = {The, is, shining}

G(sun) = {The, is, shining, burning}

G(sun) ~ G(light)

Cooccurrence Analysis

- Counting co-occurrence
  - => focus on high frequent events in text data (Zipf's law!)
  - maximal frequency pair: „the – of“
- Determine significance of co-occurrence
  - statistical test measuring „surprise“
  - => better captures semantic characteristics of a text
  - there is not *the* single measure

Cooccurrence Analysis

- statistical significancy
  - measure of deviation from random conjoint occurrence
- measurements
  - n_a – windows containing A
  - n_b – windows containing B
  - n_{ab} – windows containing A and B
  - n – number of all windows

(bag of words within windows)

- significance measures
  - Frequency (baseline*)
  - Dice
  - Mutual Information
  - Log Likelihood

* remember Zipf!

Examples

Eingabe	logl	dice	baseline	MI
Abfall	radioaktiv	radioaktiv	d-	Abklingzeit
Abfall	Tonne	entsorgen	und	Bodenwurzel
Abfall	entsorgen	Endlager	in	Chemie-Praktikum
Abfall	Endlager	Entsorgung	werden	Dosenbier-Trinker
Abfall	werden	hochradioaktiv	ein	STAWA
Zink	Kupfer	Blei	d-	Verzinken
Zink	Blei	Kupfer	und	Eisengegenstand
Zink	und	Cadmium	%N%	Hartlot
Zink	Cadmium	Zinn	ein	Bismut
Zink	Silber	Nickel	in	stolberger
Montag	am	am	d-	VHS-Öffnungszeit
Montag	%N%	abend	am	Focus-Tag
Montag	Uhr	Uhr	%N%	Einzelhandlesverband
Montag	abend	Freitag	in	FIS-Sicherheitsexperte
Montag	in	kommend	ein	Freischützstras

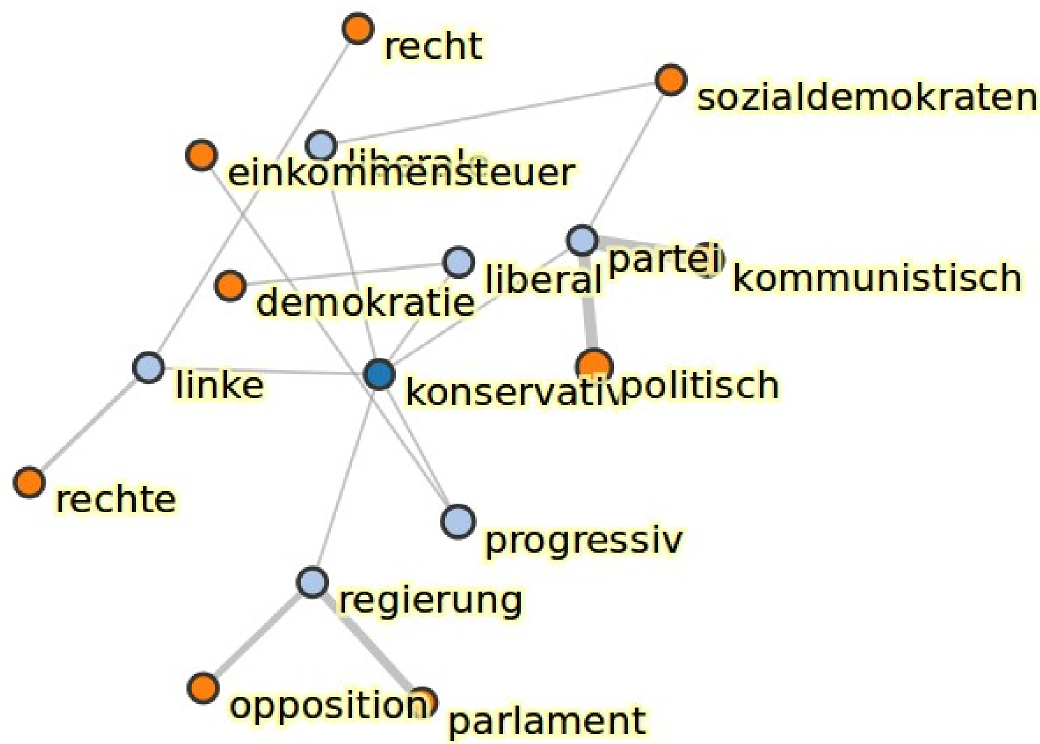
Application in Social Science

- (change of) meaning may be inferred from cooccurrence results
- cooccurrence analysis → comparison of different result sets
  - change of context units (neighbours, sentence, document, ...)
  - filter terms by POS-/NE-types
  - tracking change of global contexts by comparing time ranges

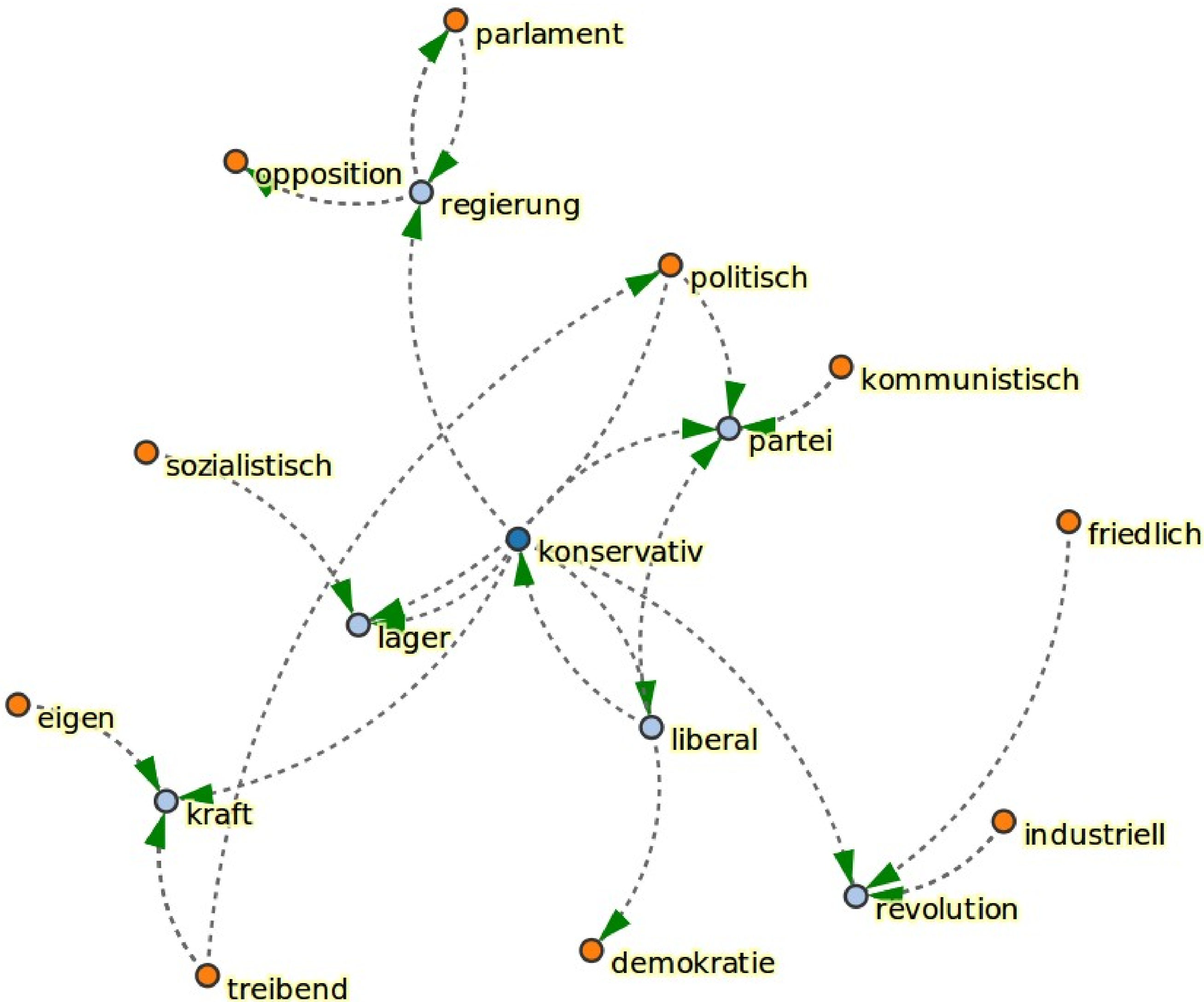
- Visual analytics:
  - tables
  - graphs
  - KWIC-Lists

Visualization

- cooccurrences = network structure
  → visualization as graph
  - nodes : terms
  - edges : cooccurrence relation
- e.g. additional information:
  - edge width: significancy value
  - node color: order of cooccurrence
- Caution:
  - algorithms for graph drawing produce outcomes which are not necessarily semantically interpretable!



Visualization



Visualization

- KWIC-lists: „Keyword in context“ (H.P. Luhn, 1966)
  - selection of snippets by single keyword
  - centering display around key word

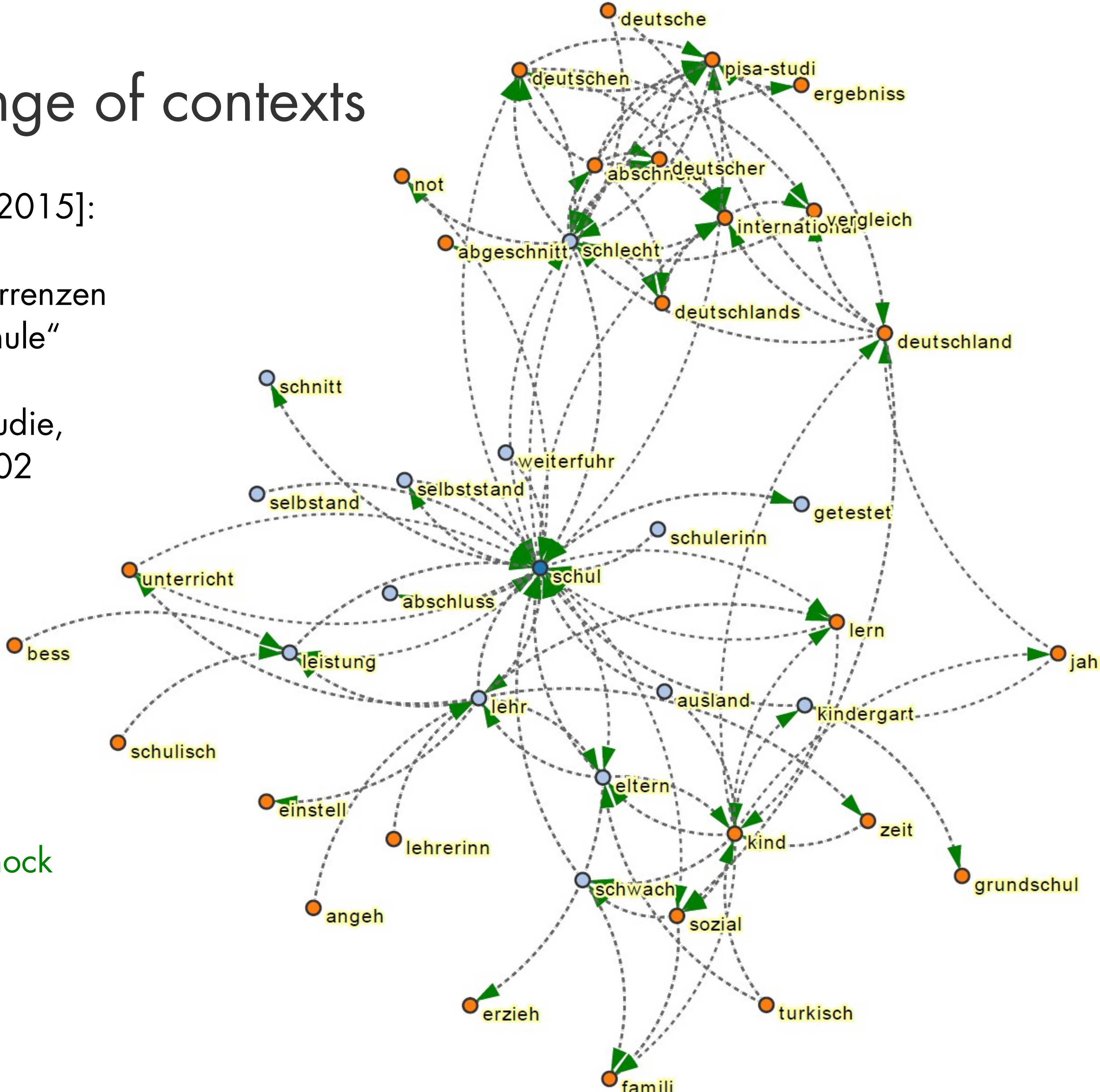
… Antitrustpolitik unterstützten Festhalten an	konservativen	Idealen, die der modernen …
… Die FDP steht für liberal	konservativen	Egoismus. Wofür stehen AL …
… und Sozialpolitiker sind ratlos.	Konservative	Politiker plädieren für härtere Strafen …
… auf den massiven Widerstand der	konservativen	Mehrheit im Bundesrat. Ihr …
… Sozialisten), manche Engländer (vornehmlich	Konservative	) und Italiener (vor allem …
… an anderen geschliffen. Der	konservativen	Regierung Margaret Thatchers fällt …
… Radikalen unter ihnen haben die "	konservative	Revolution" auf ihr Panier geschrieben. …

change of contexts

[Maas 2015]:

Kookkurrenzen
zu „Schule“

PISA-Studie,
2001/02



PISA shock

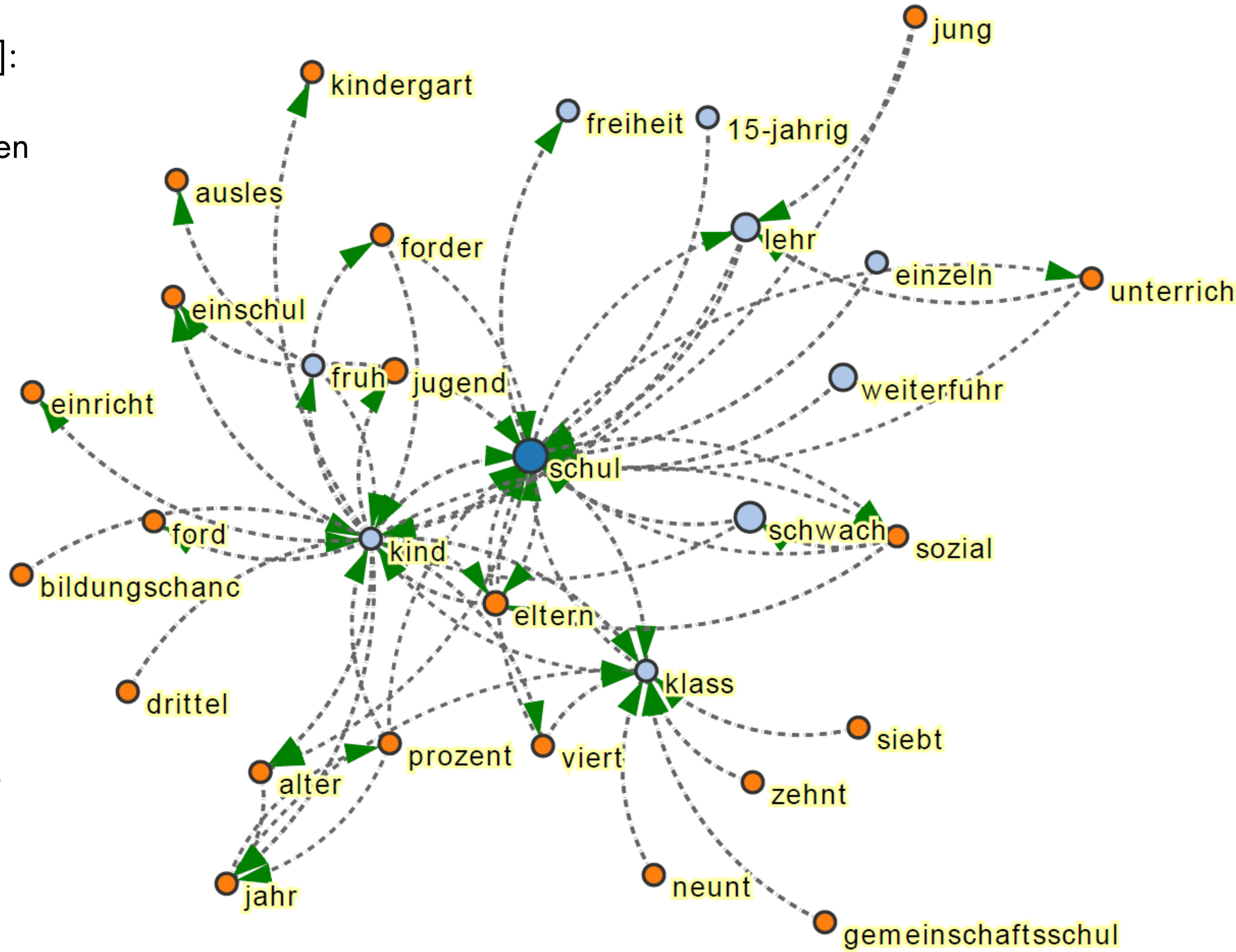
change of contexts

[Maas 2015]:

Kookkurrenzen
zu „Schule“

PISA-Studie,
2004/05

School types

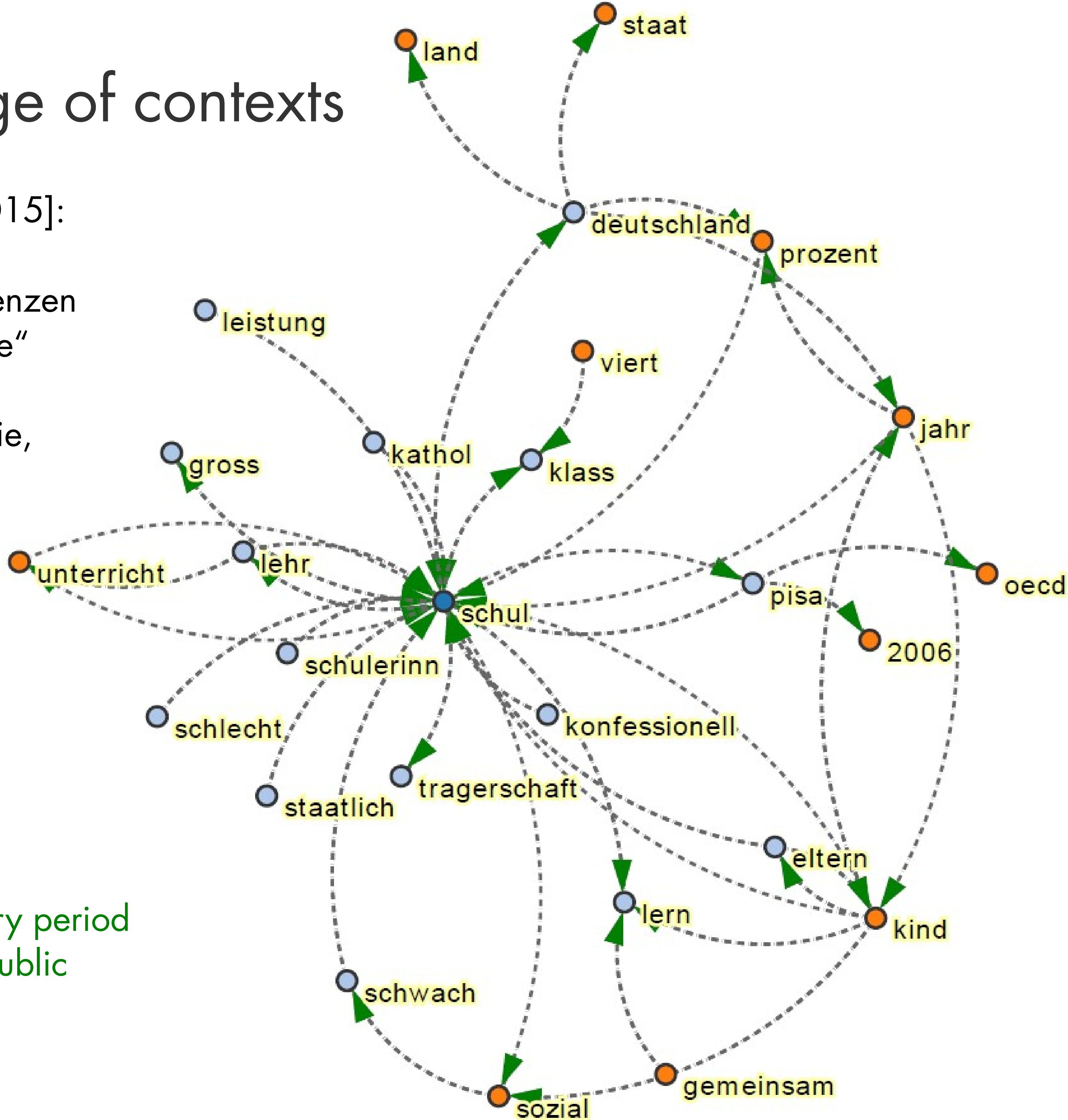


change of contexts

[Maas 2015]:

Kookkurrenzen
zu „Schule“

PISA-Studie,
2007/08



Elementary period
Private/public

Summary

- Cooccurrence analysis:
  - global contexts → meaning of terms („discourse level“)
  - significancy of cooccurrence relation is crucial
- „visual hermeneutics“ / distant reading of collections through graphical representations
- informational enrichment by creative filtering:
  - different sub collections
  - time ranges
  - person names / NE
  - certain POS-types
  - …

What differences in results do you expect from different windows?

- sentences
- paragraphs
- documents
- headlines
- k left/right neighbour words

Summary Lexicometrics

- Applications
  - 1. Frequency analysis
  - 2. Key terms / characteristic elements
  - 3. Concordance (local context) → KWIC lists
  - 4. Cooccurrence / collocation (global context)
- 5. application: Dimension reduction by multivariate statistics (not covered in this lecture)
  - Multidimendional scaling
  - Correspondence analysis
  - Prinicipal component analysis
- Context selection:
  - interpretation of contrasting results of different subselections of the base population