Universität Zürich[UZH]

Department of Computational Linguistics

Automated Text Analysis in the 'Text Crunching Center' (TCC)

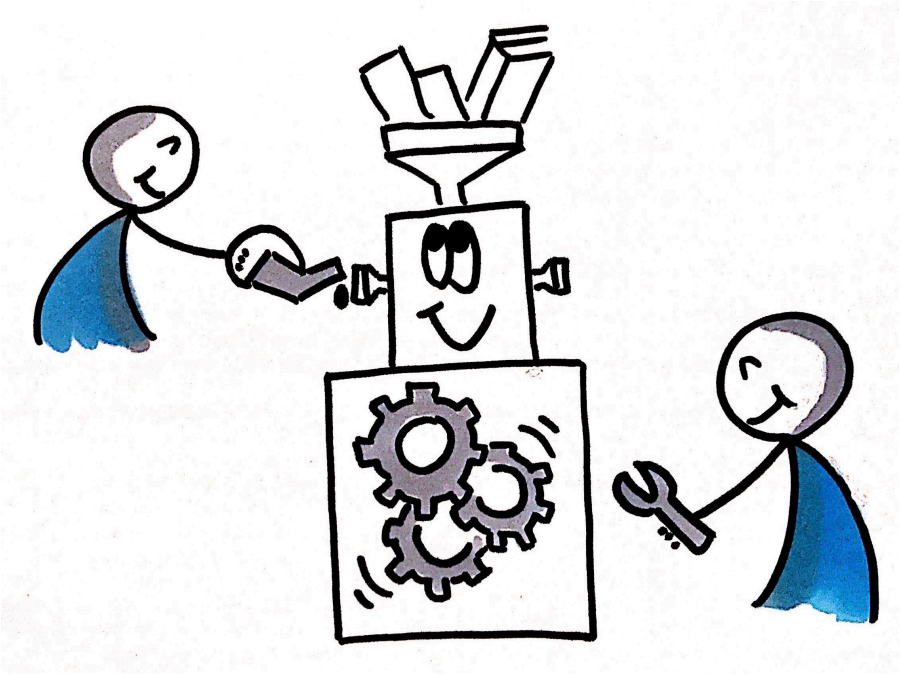
Or: Exploring the world of semantics from our home office

PD Dr. Gerold Schneider

Department of Computational Linguistics

gschneid@cl.uzh.ch

https://www.cl.uzh.ch/de/TCC.html



Abstract

This talk introduces the Text Crunching Centre (TCC) which is a Computational Linguistics and Digital Humanities service hosted at the University of Zurich.

We present teasers from text analytics involving social, political and historical studies. And a bit of Covid-19, too.

1. Offers of the Text Crunching Center (TCC)

We offer our expertise in the following areas:

- Text Analytics
- Text Mining
- Digital Humanities
- Machine Translation
- ...

We offer consulting and support in

- Digitalisation
- Processing of text, including multilingual and historical texts
- Advice on tools, software and best practices
- Help with project applications and joint projects
- Ready-made solutions
- Training, coaching, workshops

Contents / Some of Our Methods (black: case study)

Case Studies:

- 2. Stylometry and Stylistics (Donald Trump)
- 3. Document classification (US politics)
- 4. Topic Modelling (Democratisation)
- 5. Sentiment Detection (Migration)
- 6. Conceptual Maps & Distributional Semantics (Lifestyles, History, Covid, Medicine)

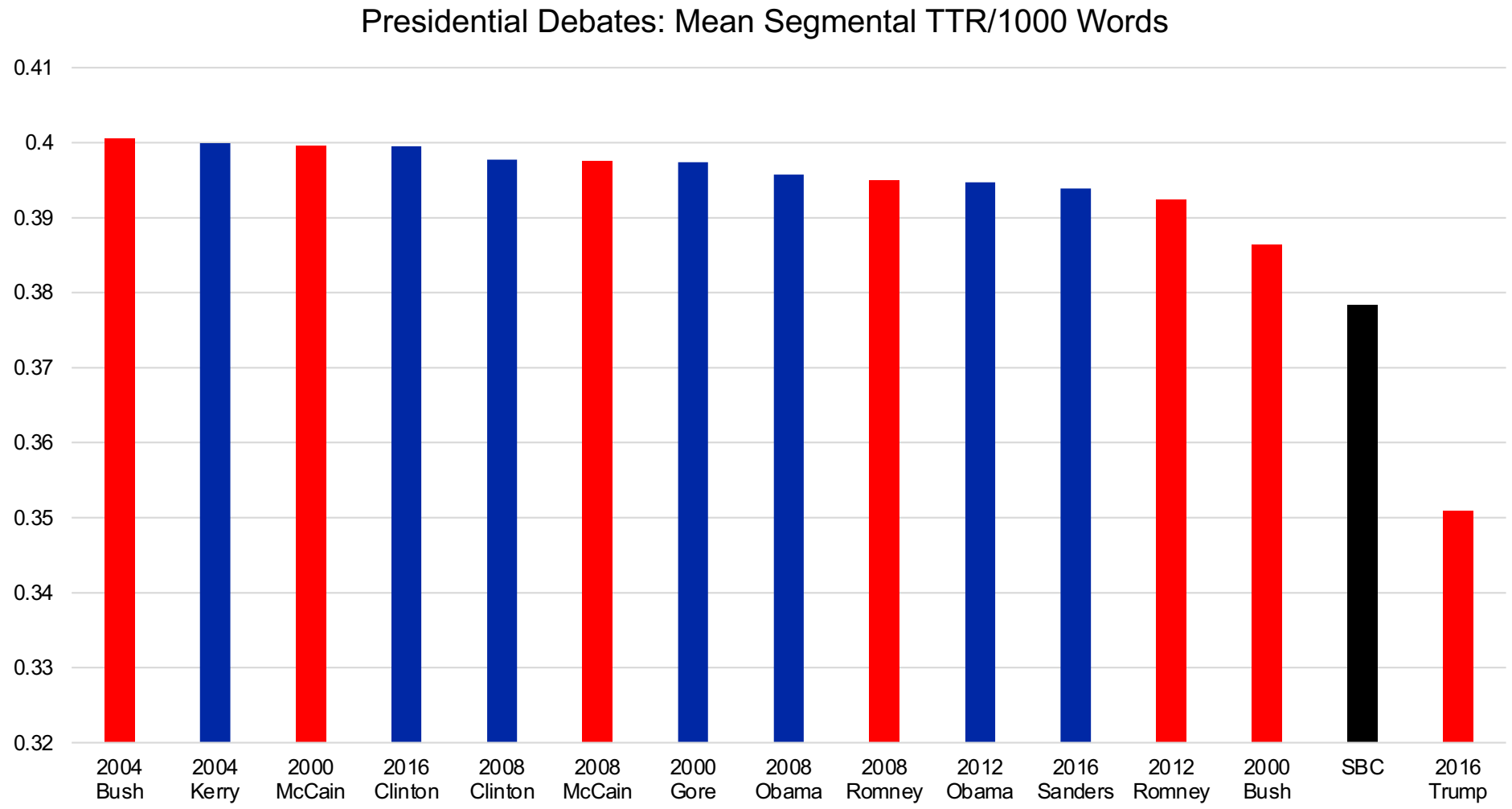
Other methods that we use:

- Cognitive Language Models
- Clustering
- Keyword Detection
- Collocation
- Named Entity Recognition
- Network Analysis
- Chatbots
- Relation Mining
- Machine Translation
- Neural Networks
- Sequence Learning

Universität Zürich[UZH]

2. Stylistics: Vocabulary richness of Donald Trump

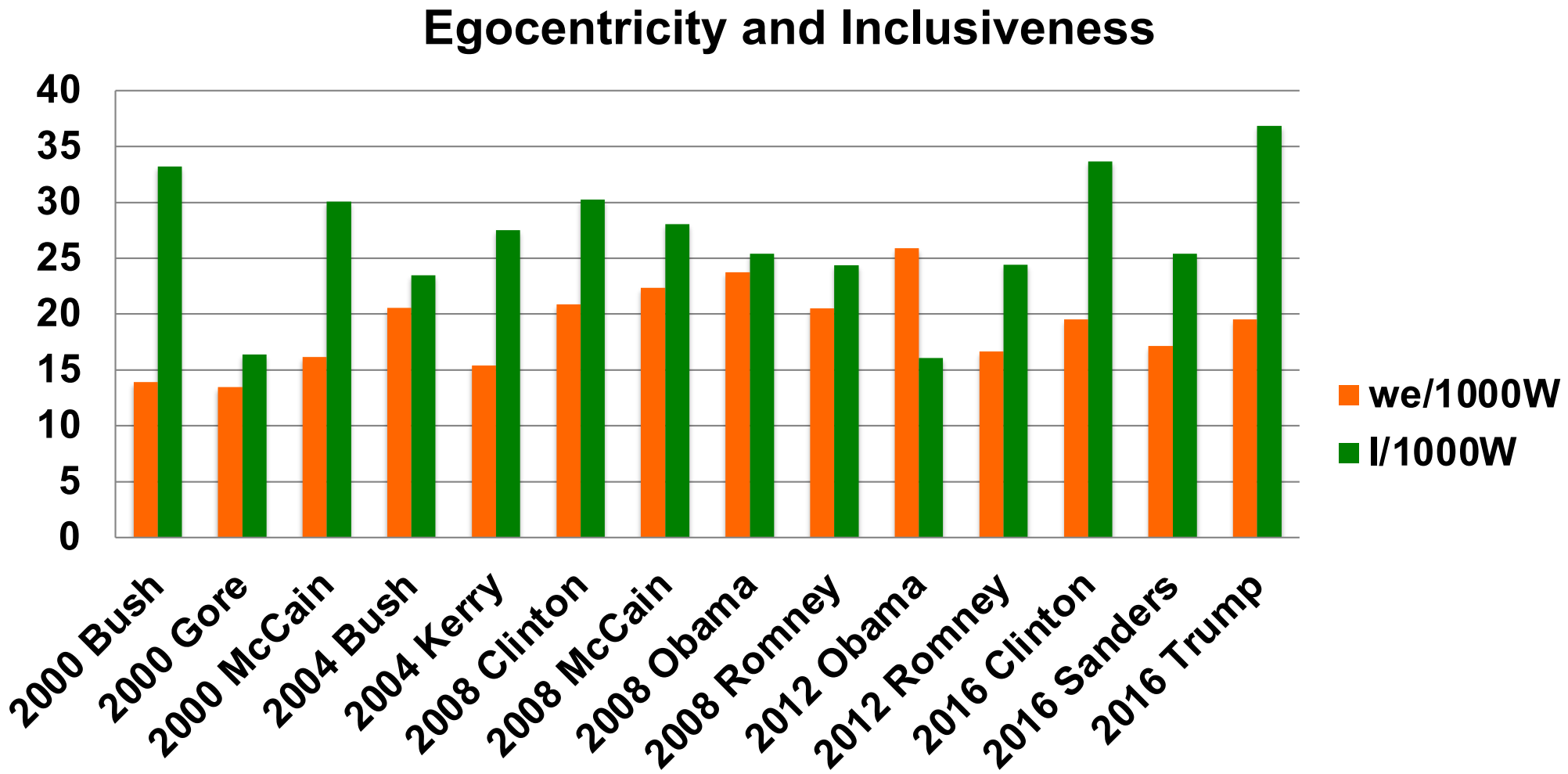
Department of Computational Linguistics



Vocabulary Richness

Ronan, Patricia, and Gerold Schneider (2020). "A Man Who Was Just an Incredible Man, an Incredible man: Age Factors and Coherence in Donald Trump's Spontaneous Speech". In Ulrike Schneider and Matthias Eitelmann (eds.), *Linguistic Enquiries into Donald Trump's Language. From 'Fake News' to 'Tremendous Success'*. London: Bloomsbury. 62-84.

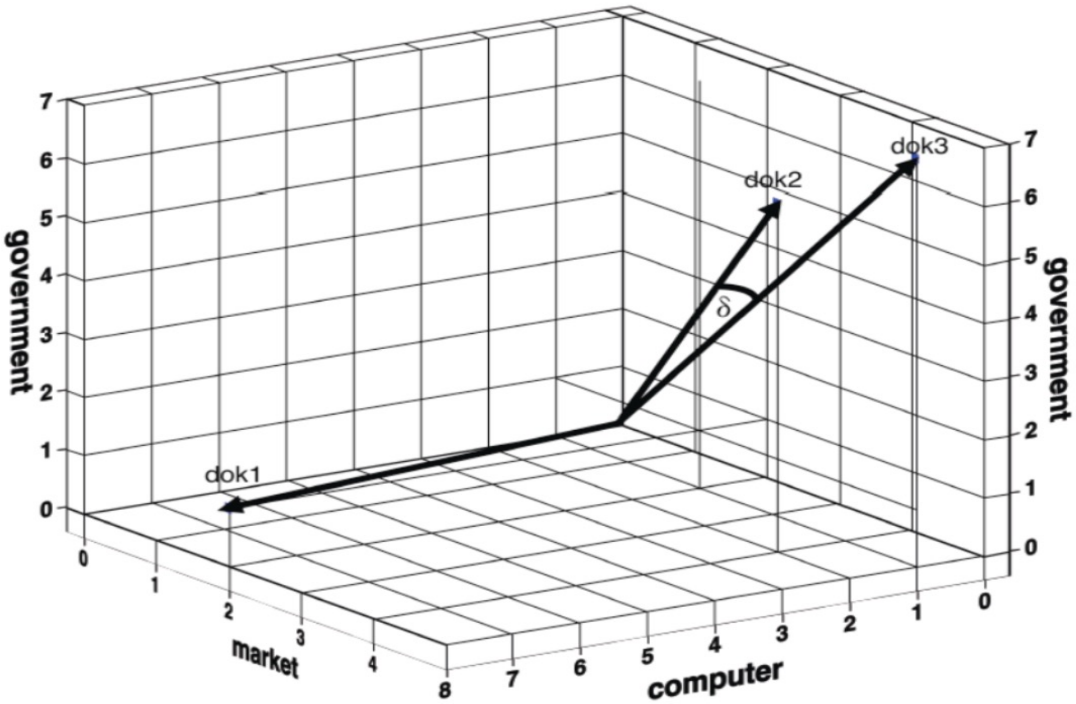
2. Stylistics: Donald Trump



Egocentricity (*I*) and Inclusiveness (*we*)

3. Document Classification: Method

One of the most versatile methods (e.g. Grimmer & Stewart 2013)



Words as Vectors

- Document 2 and 3 are similar as the angle δ between their vectors is small → probably in the same class
- If two documents are very closely related according to the vector space model, this means that they are similar, that they use similar words
- Even if some terms may differ (e.g. *cosmonaut* vs. *astronaut*), many words will be the same in documents of the same topic (e.g. *orbit, space, rocket, launch, defense, satellite*)

Frequency for	market	computer	government
Doc 1	2	8	1
Doc 2	3	2	6
Doc 3	4	1	7

3. Document Classification

Detect pre-annotated, often binary classes.
Even subtle semantic differences can often be detected; but depends heavily on data.

Teaser: US Speeches (< 2013)

Prediction of Party Affiliation (Republican / Democrat) based CORPS II Corpus: 8 mio words, 3618 Speeches (Guerini et al. 2013).
We use speakers that have at least 10 speeches (and only American ones).

Logistic regression achieves 95-98% accuracy

Actual vs. Predicted	actual dem	actual rep
predicted dem	**1787**	36
predicted rep	131	**1294**

# Reden	Name
889	Bill Clinton
427	George W. Bush
388	Ronald Reagan
356	Dick Cheney
347	Barack Obama
316	John F. Kennedy
107	Michelle Obama
102	Margaret Thatcher
93	Laura Bush
61	Richard M. Nixon
53	Al Gore
51	Alan Keyes
43	Joe Biden
36	Condoleezza Rice
26	John Kerry
18	Hillary Rodham Clinton
13	Lynne Anne Vincent Cheney
13	Howard Dean
10	John Edwards

Die typisch republikanischsten Merkmale sind:

Merkmal	F-score
've	0.6455
're	0.6443
nation	0.6443
it_'s	0.6336
men	0.6333
–	0.6312
i_'m	0.6286
'm	0.6286
you_all	0.6273
freedom	0.6261
we_'re	0.6254
well	0.6224
<PERIOD>_he	0.6219
<PERIOD>_and	0.6203
great	0.6192
's	0.6177
one	0.6159
government	0.6158
america	0.6153
military	0.6147

Typisch **un**republikanischste Merkmale (Auswahl):

Merkmal	F-score
nra	0.0014
equal_pay	0.0014
of_climate	0.0014
racial_<COMMA>	0.0014
insurance_program	0.0014
high-wage	0.0014
our_steel	0.0014
without_health	0.0014
in_clean	0.0013
together_across	0.0013
campaign_finance	0.0013
to_hillary	0.0013
service_program	0.0013
fugitives	0.0013
stalkers	0.0013
our_planet	0.0013
financial_system	0.0013
after_high	0.0013
student_loans	0.0013
toxic_waste	0.0013
<PERIOD>_hillary	0.0013
from_welfare	0.0013
national_service	0.0013
more_police	0.0013

4. Topic Modelling: Method

Feature lists from document classification are useful but related features are not together, the method cannot abstract from words to concepts.

Topic models: generative probability model describes how likely it is that

– the given documents belong to certain topics (like in document classification)

– these topics generate the given words

Documents and words are given, topics are fitted

In terms of conditional probabilities: the probability of

$$p(topic \mid document) \cdot p(word \mid topic)$$

is maximized

Good at revealing semantically related subject areas & associations

Delivers topics, key words, weights. Advantages:

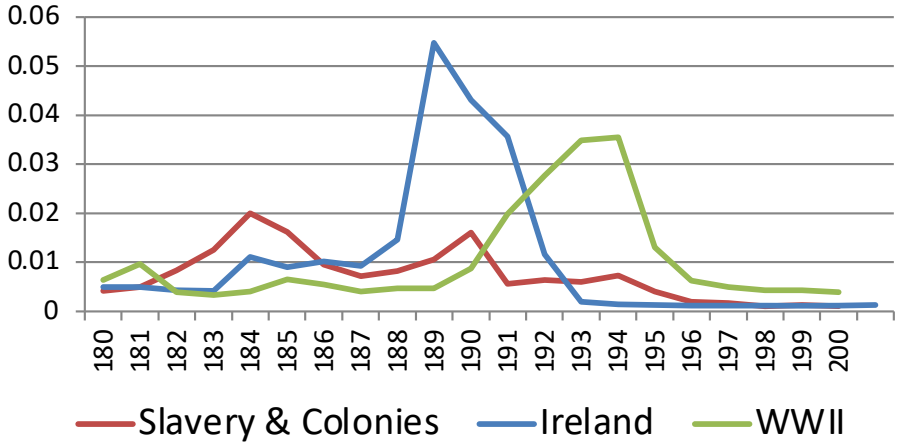
– completely **data-driven,** using **distributional semantics**

– **domain-independent, language-independent**

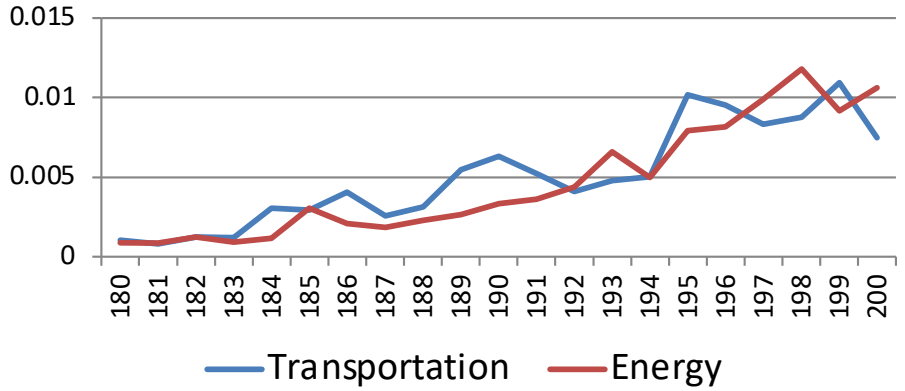
4. Topic Modelling: British Parliamentary Debates (Hansard Corpus)

Department of Computational Linguistics

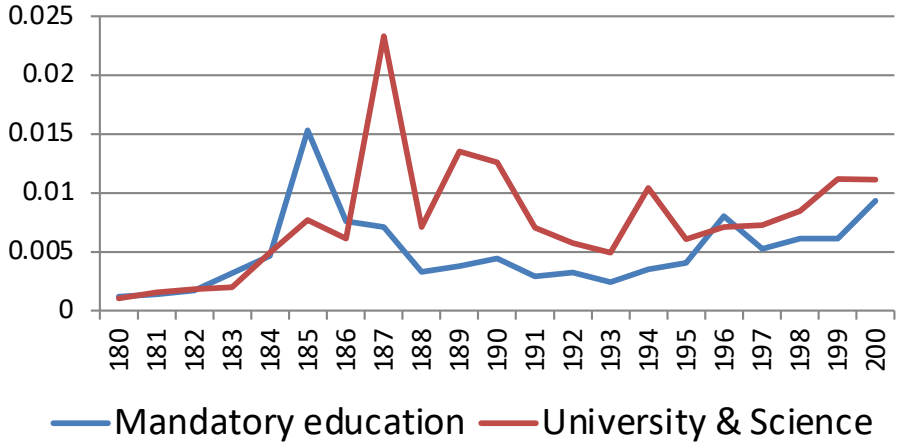
Historical facts



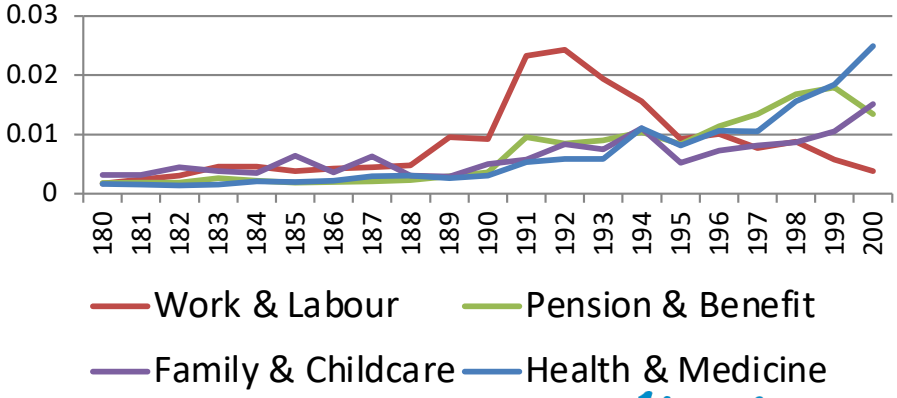
Democratisation through technical innovations



Democratisation through education



Democratisation through social protections



4. Topic Modelling: British Parliamentary Debates

Topic Change and democratisation (indeces)

Pearson correlations between topic change and democratisation and economic indeces

	topic 0	topic 1	topic 6	topic 12	topic 15	topic 33	topic 50	topic 64	topic 78
Pearson with DEMOC	0.22	-0.13	-0.41	-0.87	-0.85	-0.72	0.37	0.34	-0.49
Pearson with POLITY	0.21	0.05	-0.36	-0.88	-0.87	-0.77	0.36	0.41	-0.40
Pearson with GDP	0.32	-0.31	-0.19	-0.54	-0.58	-0.55	-0.11	0.27	-0.62
Pearson with EDUC	0.03	-0.42	-0.36	-0.66	-0.70	-0.73	-0.07	-0.02	-0.78
Key-words	university	ireland	election	court	government	bank	war	school	colony
	student	secretary	vote	case	treaty	debt	Germany	education	slave
	education	chief	constituency	judge	France	money	World	child	governor
	research	county	candidate	offense	Spain	interest	peace	teacher	island
	science	dublin	representation	justice	Russia	exchequer	nation	parent	canada

Pearson correlations between topic change and democratisation and economic indeces

	topic 20	topic 30	topic 38	topic 39	topic 44	topic 46	topic 51	topic 59	topic 72	topic 74	topic 92
Pearson with DEMOC	0.60	0.74	0.85	0.88	0.76	0.32	0.88	0.72	0.73	0.71	0.91
Pearson with POLITY	0.52	0.67	0.77	0.85	0.67	0.38	0.80	0.67	0.65	0.63	0.87
Pearson with GDP	**0.96**	0.32	0.85	0.73	**0.98**	0.60	0.88	0.56	0.99	0.85	0.65
Pearson with EDUC	**0.94**	0.36	0.93	0.80	**0.96**	0.52	0.93	0.64	0.98	0.81	0.76
Key-words	European	defence	pension	railway	health	police	water	trade	information	child	company
	community	air	benefit	transport	service	crime	oil	union	report	woman	business
	European	aircraft	people	service	hospital	officer	energy	worker	document	family	private
	union	force	insurance	line	medical	force	gas	strike	survey	young	profit
	policy	equipment	allowance	rail	patient	home	electricity	industrial	record	marriage	firm

Schneider, Gerold and Maud Reveilhac (accepted for publication). "Colloquialisation, Compression and Democratisation in Parliamentary Debates". In *Exploring Language and Society with Big Data: Parliamentary discourse across time and space,* ed. Minna Korhonen, Haidee Kotze and Jukka Tyrkkö. *Studies in Corpus Linguistics* Series. Amsterdam: Benjamins.

5. Sentiment on Migration

RQs:

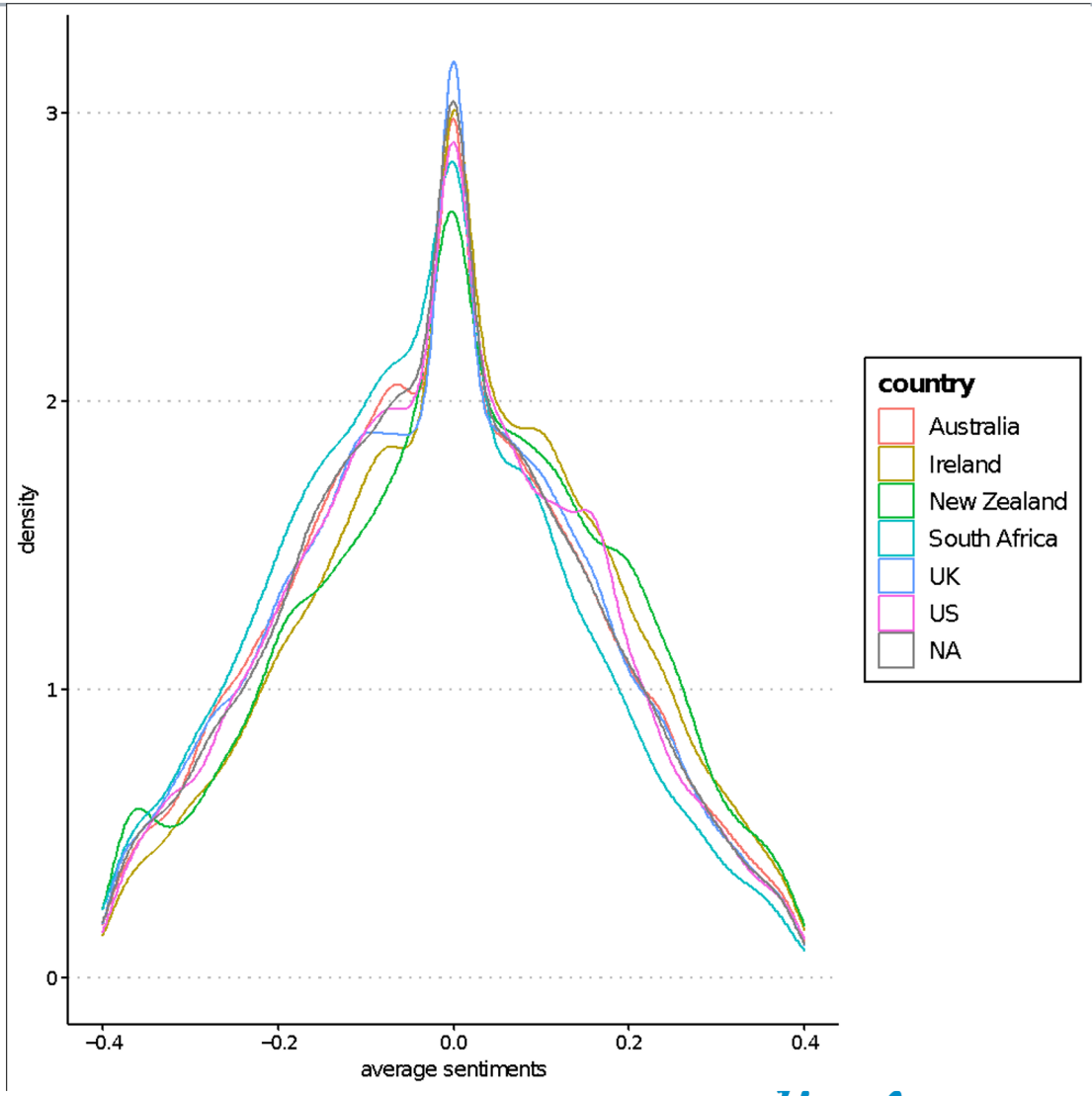
1. Which ideologies surround migration in social media?

2. What differences exist internationally?

3. Are the findings related to population surveys?

Data:

We used the Twitter API and identify **opinion makers** from the US, Britain, Ireland, South Africa and Australia.
We then retrieved their followers & randomly sampled **100'000 active followers**.



5. Sentiment: Correlations with survey data

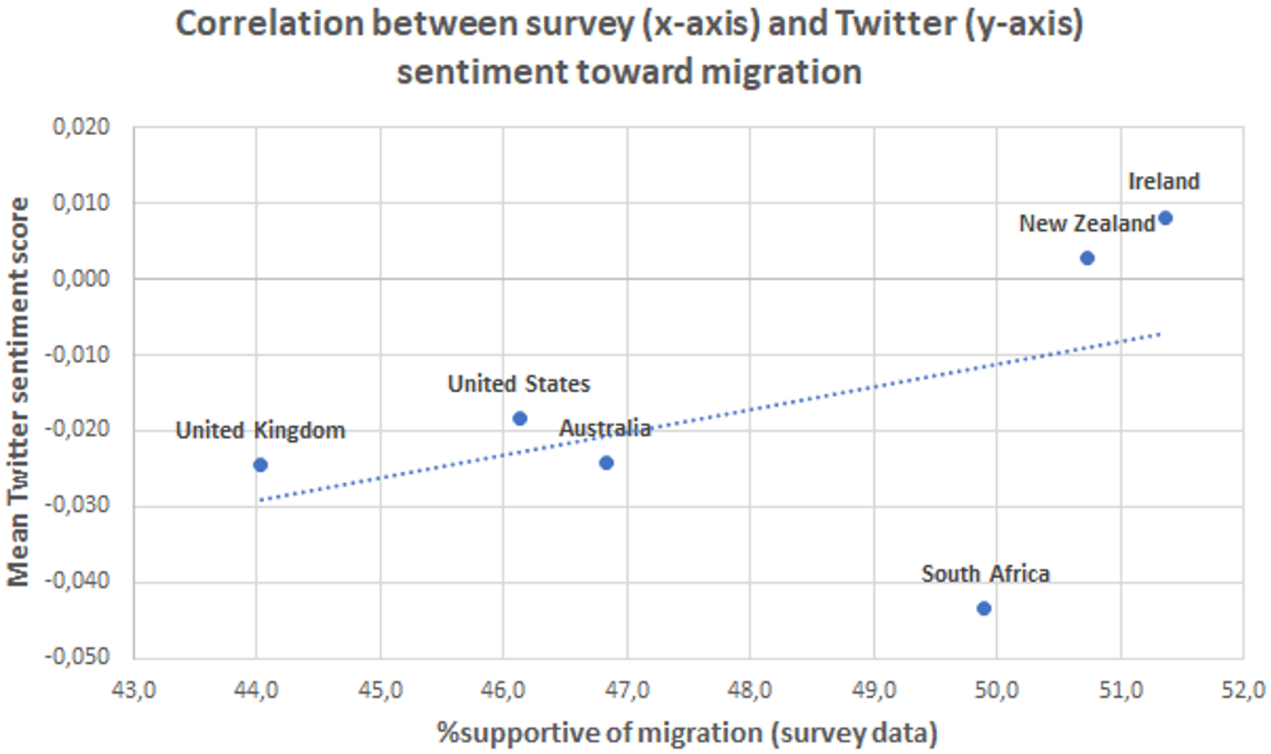


Figure: Correlations

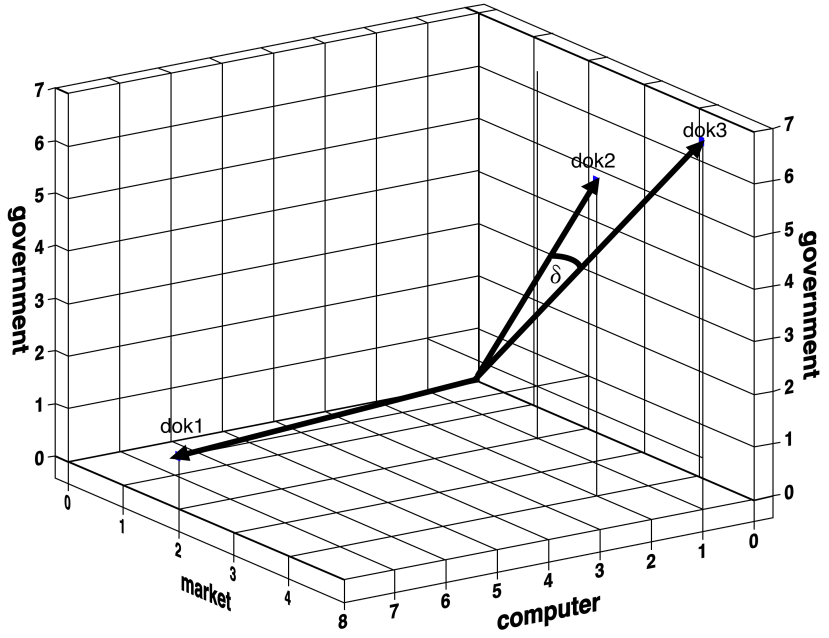
There is a (Spearman) correlation of 0.6 between survey measured attitudes and Twitter sentiment toward migration.

South Africa is an outlier. W/o SA: 0.9

6. Distributional Semantics: Method

Document Classification uses Document-Term Matrices

Frequency	*market*	*computer*	*government*
dok1	2	8	1
dok2	4	2	6
dok3	5	1	7



- Context Windows e.g.[-10 w_0 +10] give us Term-Term Matrices

	dog	hyena	cat
runs	1	1	4
barks	5	2	0

TABLE 1 Distributional vectors representing the words *dog*, *hyena* and *cat*.

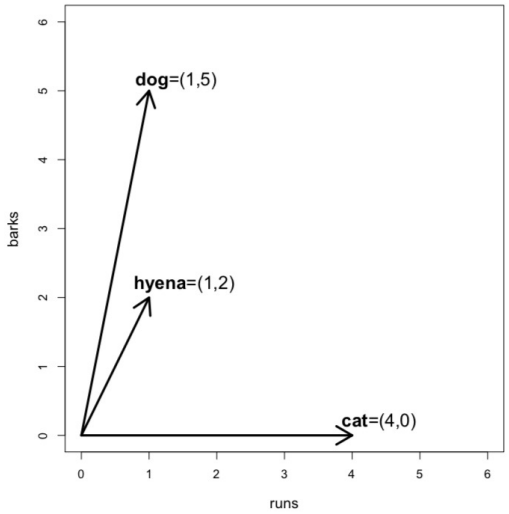


FIGURE 1 Geometric representation of the vectors in Table 1.

Source: Baroni et al. 2014:13

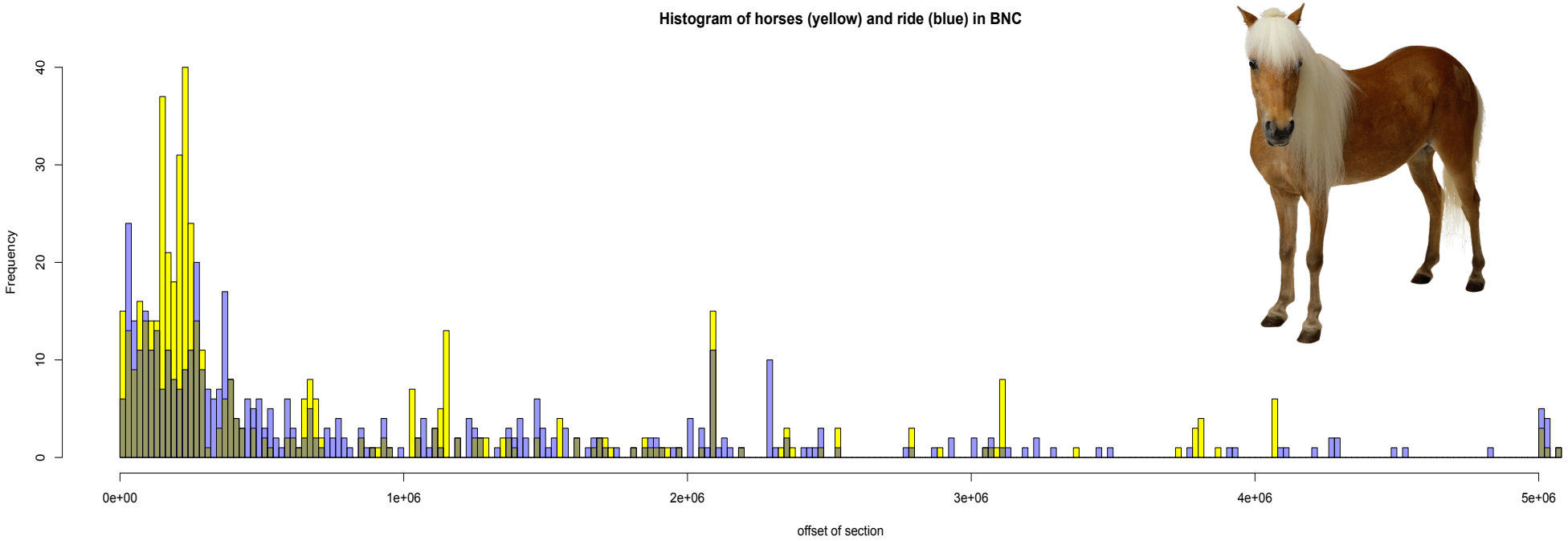
6. Method: Conceptual Maps

Kernel Density Estimation calculates semantic distances between words

Also a distributional semantic method, with extra large observation windows

Idea: semantically similar words often co-occur in the discourse

E.g. *horse* and *ride* in the BNC



Similar words are plotted close to each other on the arising concept map

We use textplot (http://dclure.org/tutorials/textplot-refresh/).) & gephi

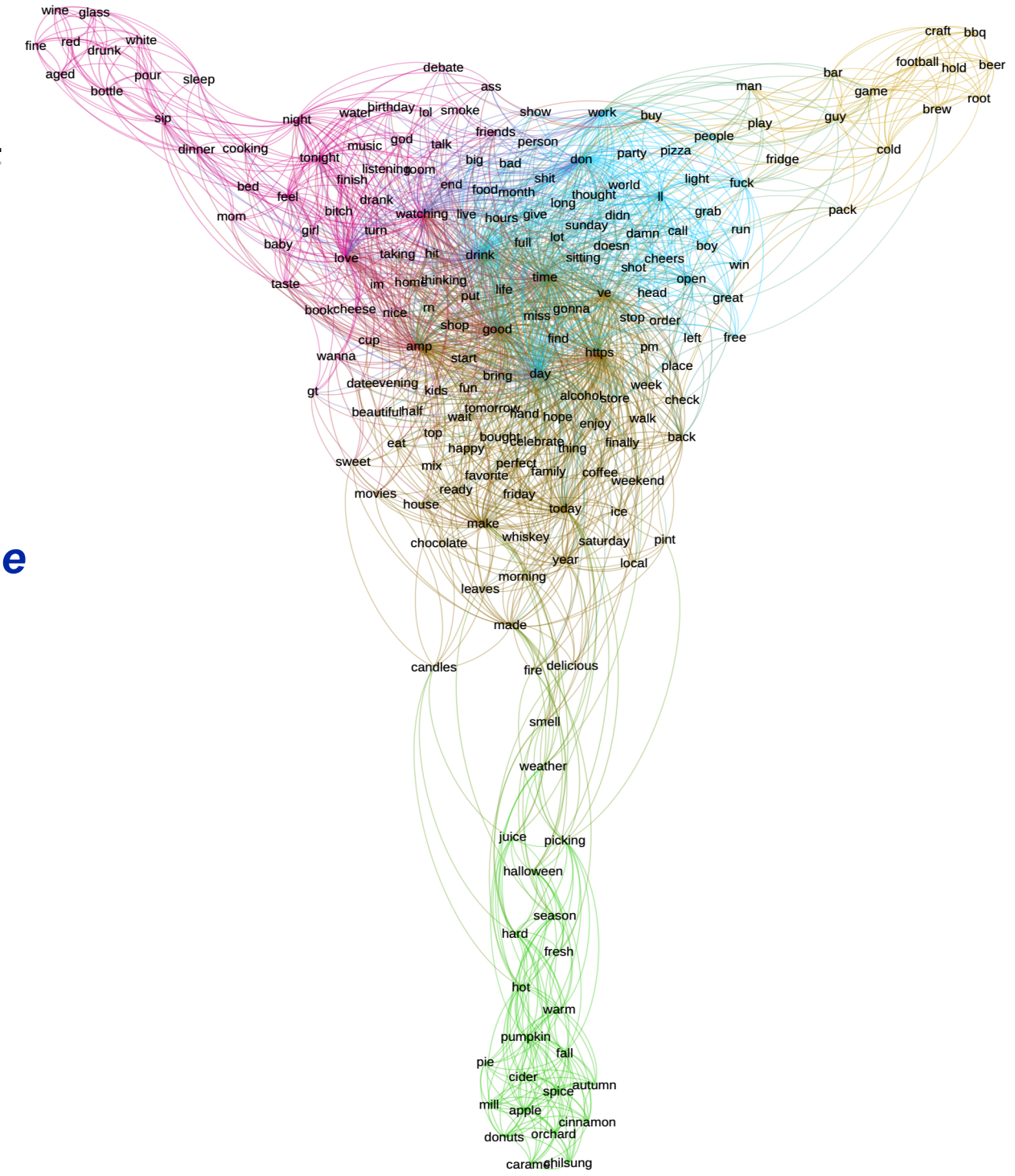
Universität Zürich[UZH]

Department of Computational l

6. Conceptual Maps

A light-hearted topic, for a client in food industry:

Tweets on *beer, cider, wine*



6. Distributional Semantics: Details

```
closest_to(training_COHA,"war")
        word similarity to "war"
1        war           1.0000000
2         ii           0.6759675
3       wars           0.6532365
4      world           0.6085764
5    vietnam           0.5716265
6   conflict           0.5645766
7     korean           0.5624021
8  indochina           0.5484778
9   outbreak           0.5452774
10 bloodiest           0.5346753
```

```
closest_to(training_COHA,"computer")
        word similarity to "computer"
1    computer              1.0000000
2   computers              0.8736527
3    software              0.8694296
4          pc              0.7722187
5        risc              0.7653417
6     desktop              0.7466422
7         ibm              0.7336524
8  technology              0.7303213
9     digital              0.7264610
10        pcs              0.7253547
```

```
closest_to(training_COHA,"churchill")
      word similarity to "churchill"
1    churchill           1.0000000
2       attlee           0.7811835
3         eden           0.7306978
4    macmillan           0.6983202
5      winston           0.6946622
6  chamberlain           0.6871389
7        bevin           0.6769213
8  beaverbrook           0.6766671
9      baldwin           0.6438595
10      stalin           0.6347145
```

```
closest_to(training_COHA,"strike")
        word similarity to "strike"
1     strike              1.0000000
2    walkout              0.8627721
3   stoppage              0.7623656
4   shutdown              0.6856940
5    strikes              0.6843846
6   walkouts              0.6842624
7    wildcat              0.6830367
8    lockout              0.6750034
9   strikers              0.6574749
10 picketing              0.6517628
```

Evaluation (**precision**):

Go through list

Recall is more difficult.

Useful for showing

- Changes in word meaning
- Associations of different actors

Digging into History

here COHA from 1940s, with 2000 terms

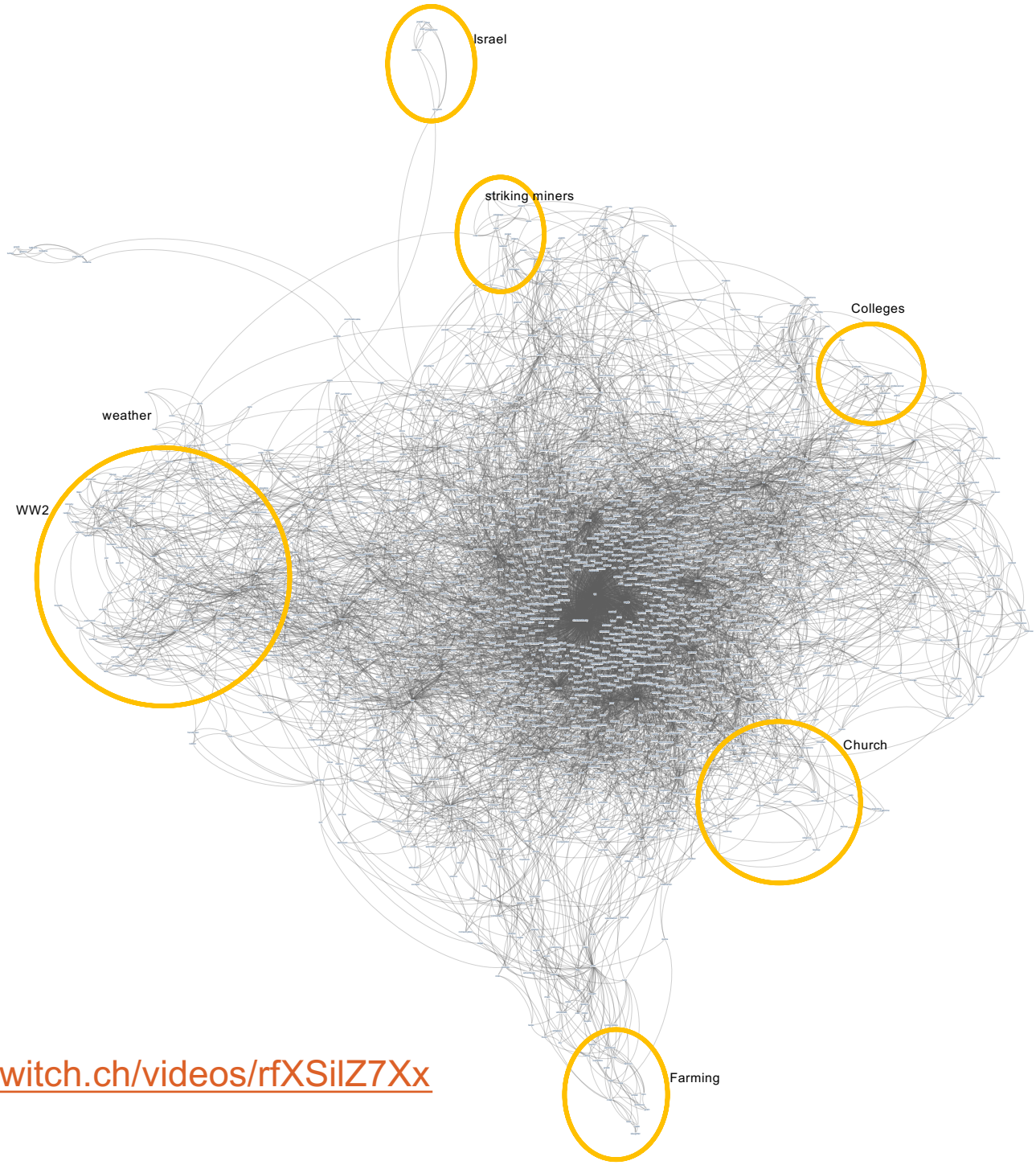
Evaluation suggestions

1. Similar terms together, e.g. pearl & harbor reich & nazi
← Distr. Semantics

2. Separation of independent topics, e.g. Japanese attacks
← TM Predictive Validity

3. Arrangement of Metadata, e.g. decades
← Recall, completeness

Full presentation: https://tube.switch.ch/videos/rfXSilZ7Xx



Does this work with any text domain?

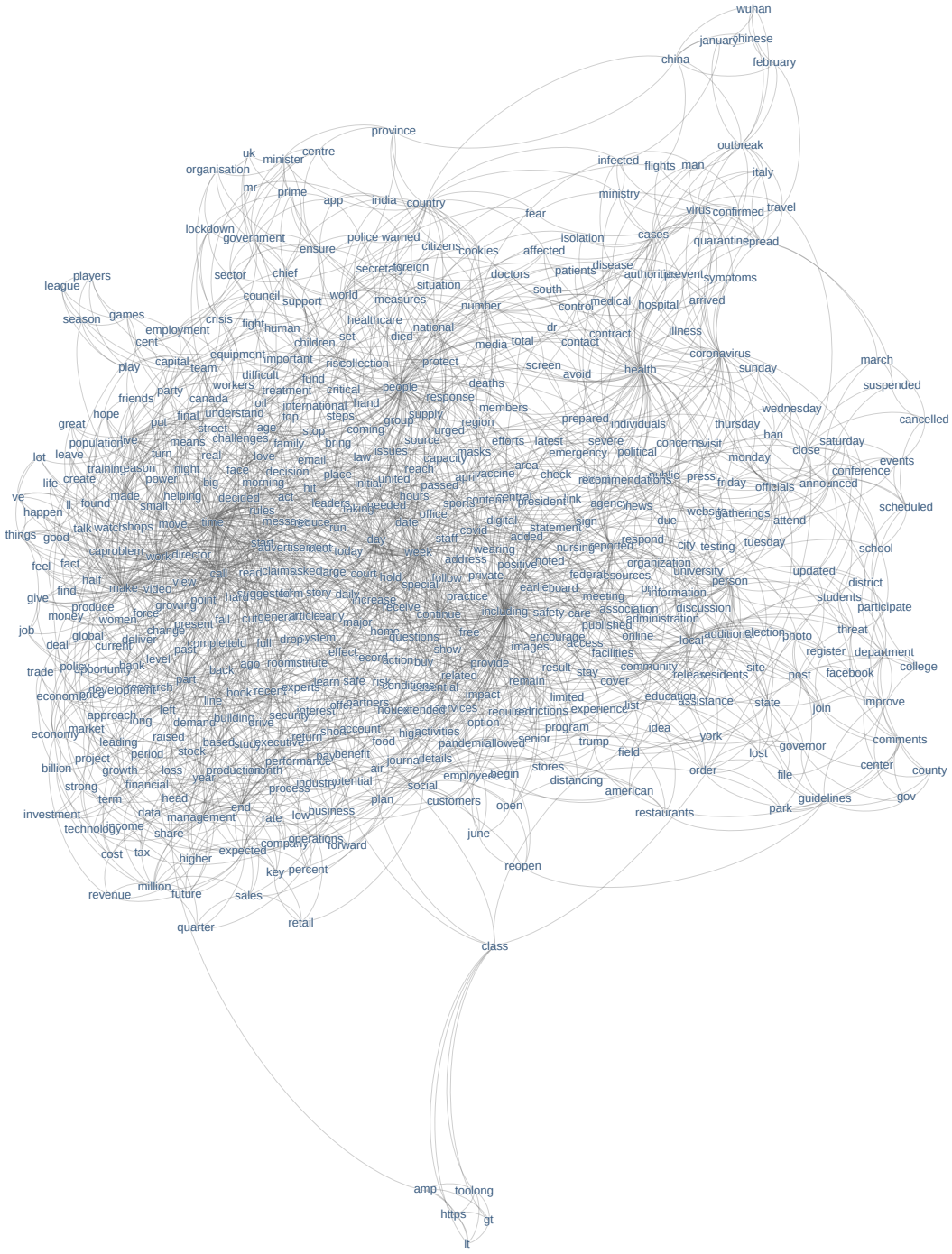
Let's try a small Covid-19 corpus,

from January to May 2020,

3.3 million words

We can discern

- Start in Jan/Feb in China
- Spread to Italy and UK
- Cancellation of events, fast pace
- Hospitals, contact rules
- Schools closing, restaurants closing
- Hope for re-opening
- Financial support



Conclusions

Presented TCC and several of our methods, with

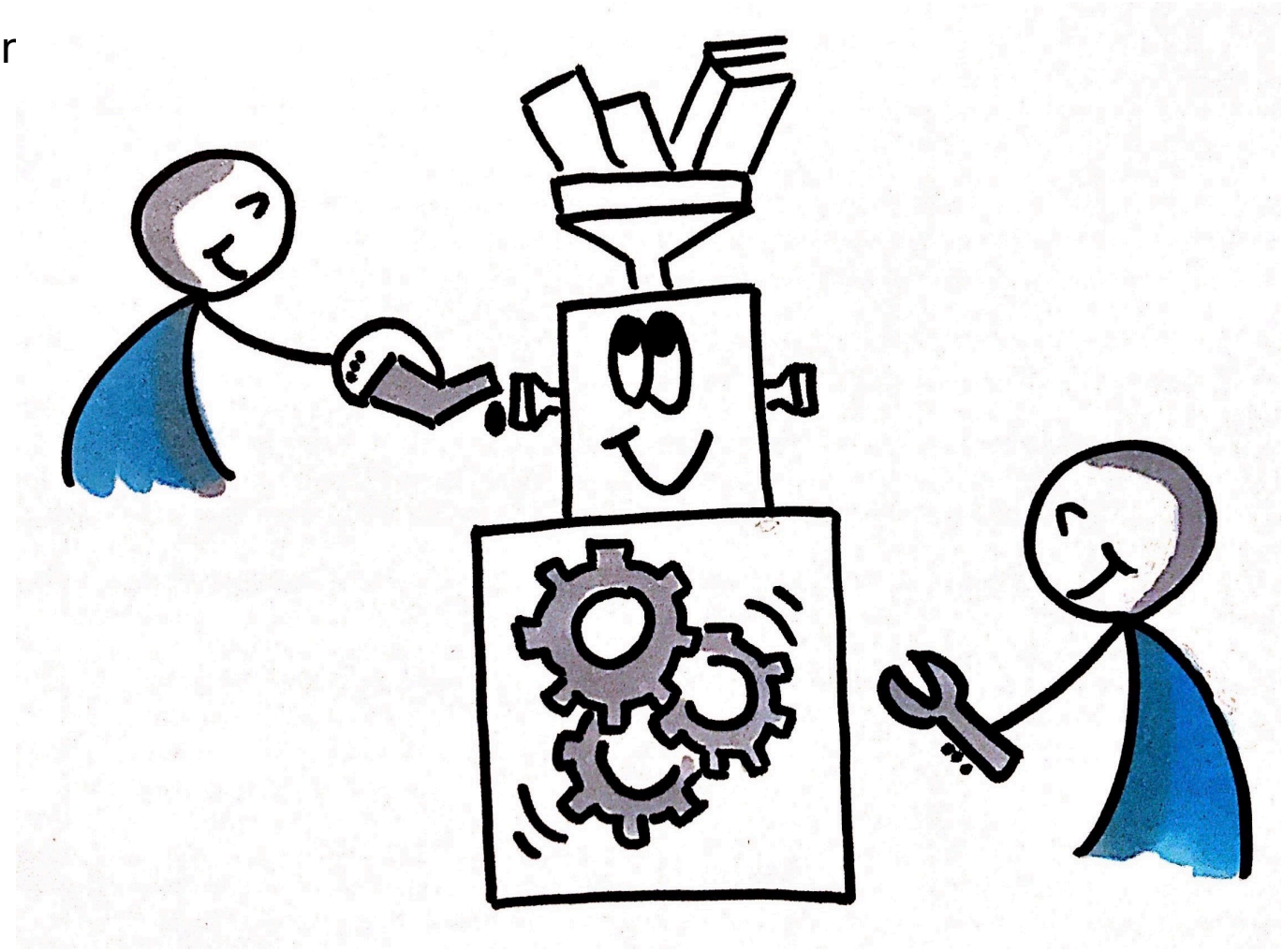
Case Studies:

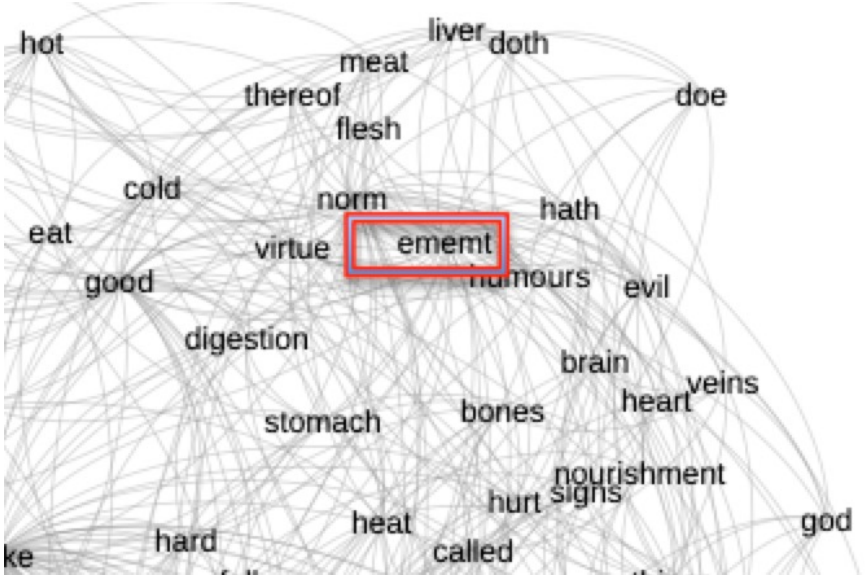
- 2. Stylometry and Stylistics (Donald Trump): simple approaches can get you far
- 3. Document classification (US politics): feature weights as keyword metric
- 4. Topic Modelling (Democratisation): from words to concepts, data-driven correlation to democratic indeces, history and society
- 5. Sentiment Detection (Migration): international differences, correlation to surveys
- 6. Conceptual Maps (Lifestyles, History, Covid, Medicine): overview, exploration, need of interpretation
- If you are interested, contact us or come to a workshop, for example tomorrow: https://www.ibme.uzh.ch/en/Biomedical-Ethics/Agenda/Conferences-and-Workshops/Excellence-in-Patient-Care-Symposium-2021/Program.html

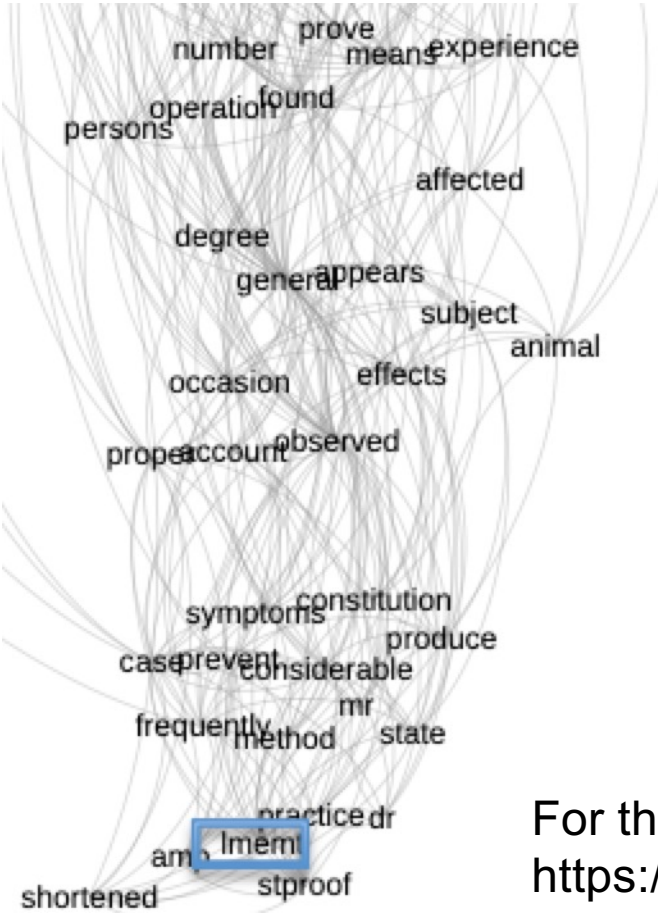
Thank you / Q & A

Thank you for your atter

Questions & Answers





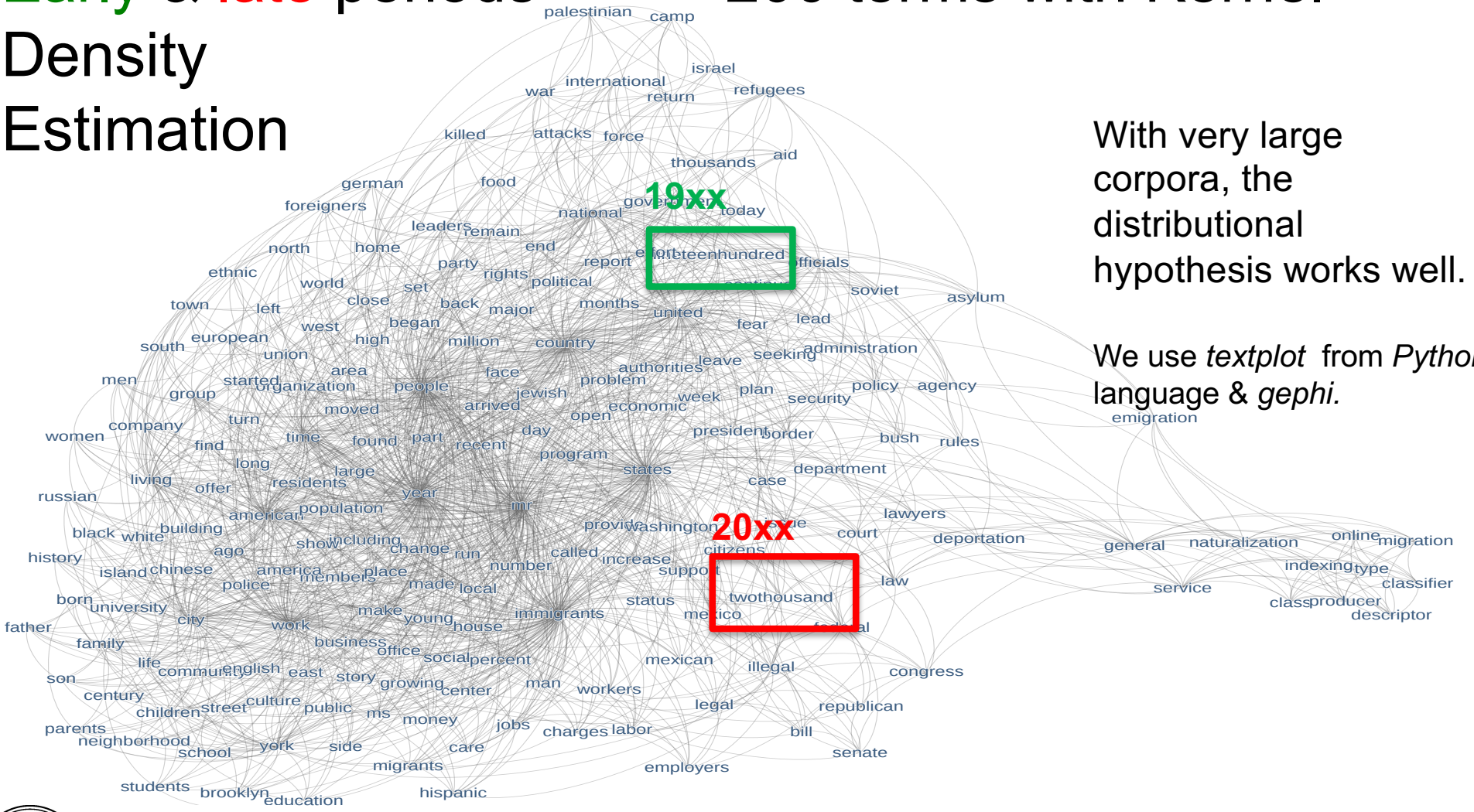


Conceptual map of the 200 most frequent content words in EMEMT (1500-1700) vs. LMEMT (1700-1800)

For the full story:
https://tube.switch.ch/videos/951fe35c

Association with terms from the migration domain: NYT 19xx vs 20xx (1987 to 2011)

Early & late periods 200 terms with Kernel Density Estimation



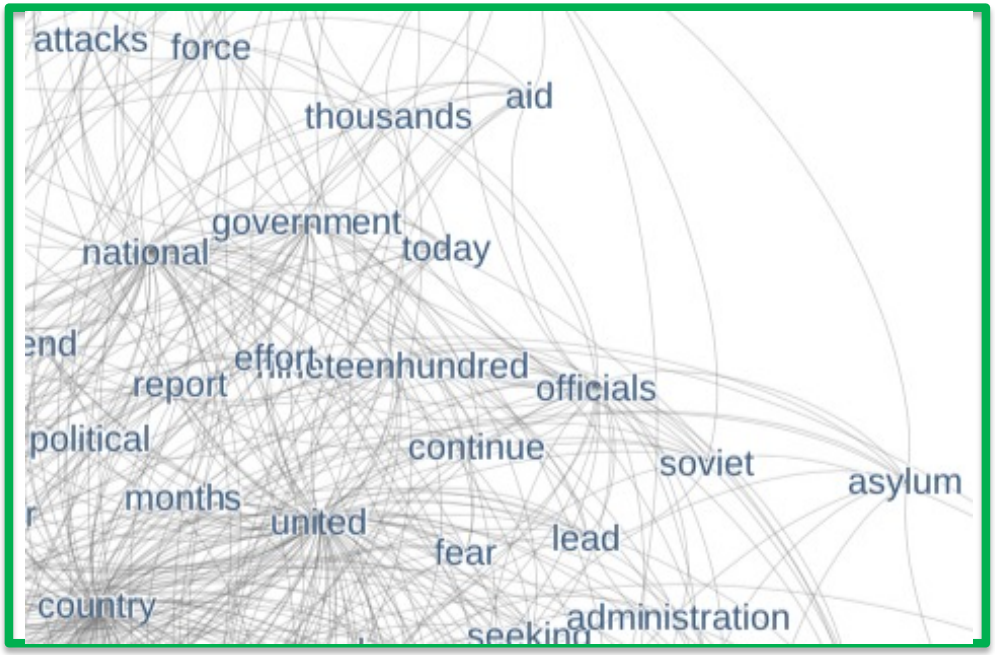
With very large corpora, the distributional hypothesis works well.

We use *textplot* from *Python* language & *gephi.*

Association with terms from the migration domain

Early & late periods using Kernel Density Estimation

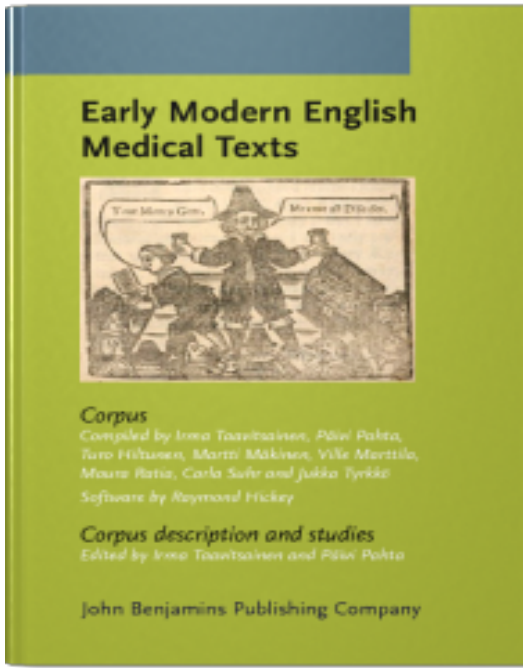
19xx: aid, united nations, attacks



20xx: (il)legal, law, mexico

Data: Medical History

EMEMT (1500-1700) & LMEMT (1700-1800). ~2 million words each, VARDed & manually checked



Early Modern English Medical Texts **(2010)**

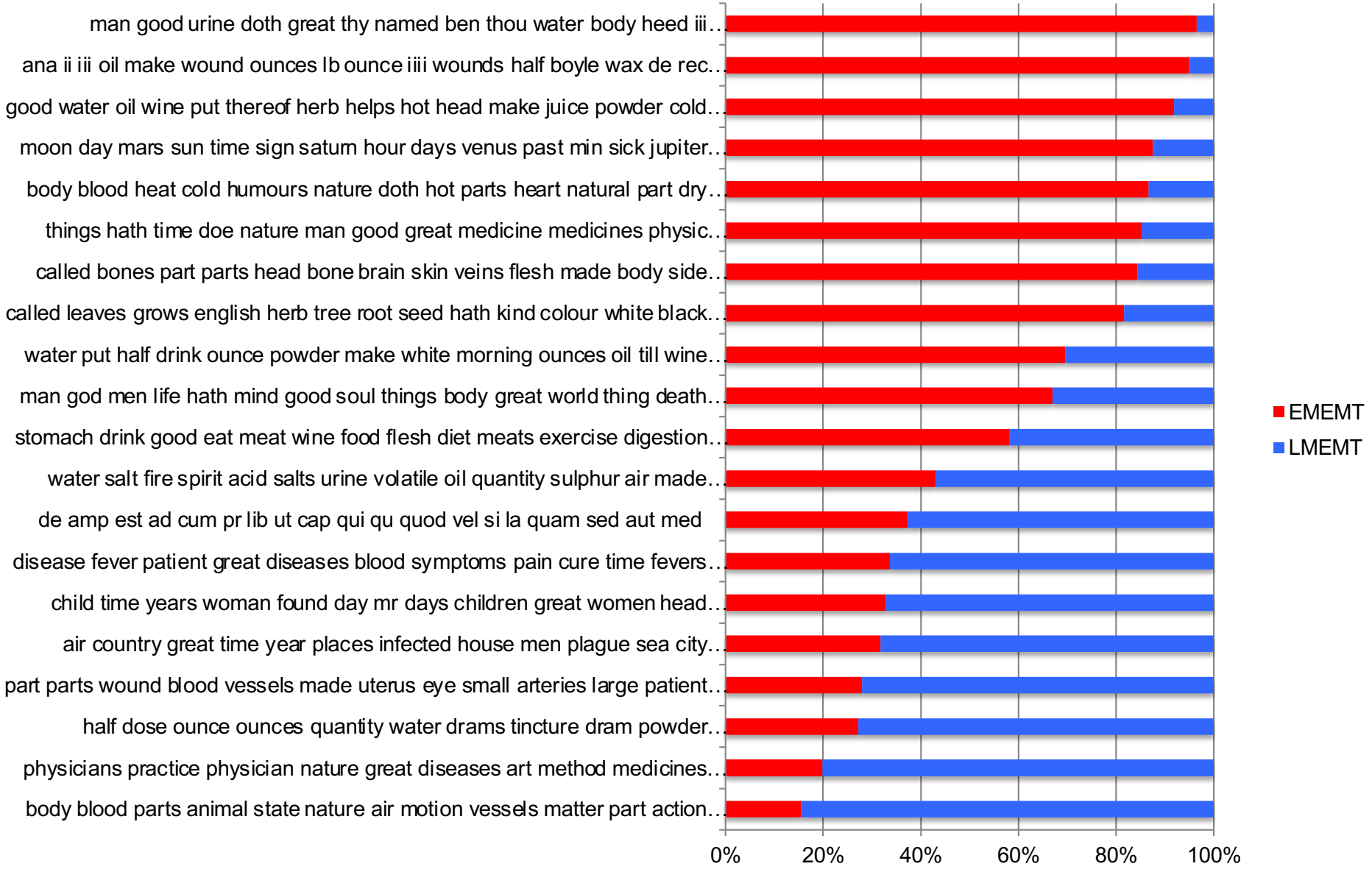
Late Modern English *Medical Texts* **(2019)**

20 topics: more prominent in EMEMT

EMEMT/ LMEMT	TOPIC	EMEMT	LMEMT	KEYWORDS	COMMENT
27.6746	14	4.39%	0.16%	man **good** urine doth great thy named ben thou water body heed iii betokens ye ii **evil** stone blood	good vs. evil
18.6624	1	3.51%	0.19%	ana ii iii oil make wound **ounces lb ounce** iiii wounds half boyle wax de rec add **roses powder**	recipes
11.2918	15	6.48%	0.57%	good **water oil wine** put thereof **herb** helps hot head make juice powder cold stomach therewith eyes ye made	recipes
6.9937	13	2.08%	0.30%	**moon day mars sun time sign saturn hour days venus** past min sick jupiter planet noon planets mercury signs	astrology, superstition
6.4759	9	**10.63%**	1.64%	body blood heat cold **humours** nature doth hot parts heart natural part dry hath moist **spirits humour** air reason	humours
5.7593	0	**12.00%**	2.08%	**things** hath **time** doe nature **man good great** medicine medicines physic thing doth book physicians made physician cure part	ragbag, unspecific
5.4022	3	5.63%	1.04%	**called** bones part parts head bone brain skin veins flesh made body side neck sinews hath mouth members substance	Logocentric, scholastic
4.4537	11	3.44%	0.77%	**called** leaves grows english **herb tree root seed** hath kind colour white black flowers **pepper** small trees call sea	herbals, recipes,
2.2873	2	8.99%	3.93%	water put half drink ounce powder make white morning ounces oil till wine boil warm good leaves pint day	recipes,
2.0241	4	7.57%	3.74%	man **god** men life hath mind good **soul** things body great world thing death make thou fear health love	religion
1.3865	5	4.74%	3.42%	stomach drink good eat meat wine food flesh diet meats exercise digestion water bread milk strong hot health made	regimen
0.7550	7	3.74%	4.95%	water salt fire spirit acid salts urine volatile oil quantity sulphur air made waters earth parts colour particles liquor	
0.5945	12	0.92%	1.55%	de amp est ad cum pr lib ut cap qui qu quod vel si la quam sed aut med	Latin terms in all periods
0.5084	10	6.07%	11.94%	disease fever **patient** great diseases blood **symptoms** pain cure time fevers body cold stomach day urine medicines case violent	

20 topics: more prominent in LMEMT

EMEMT/ LMEMT	TOPIC	EMEMT	LMEMT	KEYWORDS	COMMENT
0.1842	17	2.51%	**13.64%**	**body blood parts animal** state nature air motion **vessels matter part action effects** natural fluids life quantity animals degree	Science, empirical
0.2474	8	5.74%	**23.22%**	**physicians practice physician** nature great diseases art **method medicines knowledge** time make proper **dr** present medicine reason	professional
0.3726	6	1.85%	4.96%	half **dose ounce ounces quantity** water **drams tincture dram powder** medicine spirit wine drops grains syrup make mix ingredients	recipes
0.3880	18	3.42%	8.82%	**part parts wound blood vessels** made uterus eye small arteries large patient found bladder **operation artery** wounds great side	invasive medicine
0.4650	16	2.08%	4.47%	air country great time year places infected house men **plague sea city** number persons years **hospital** infection town **london**	hospitals, society
0.4876	19	4.20%	8.61%	**child time years woman** found day mr days children great women head made months water dead hours case told	childbirth
0.5084	10	6.07%	11.94%	disease fever **patient** great diseases blood **symptoms** pain cure time fevers body cold stomach day urine medicines case violent	
0.5945	12	0.92%	1.55%	de amp est ad cum pr lib ut cap qui qu quod vel si la quam sed aut med	Latin terms in all periods
0.7550	7	3.74%	4.95%	water salt fire spirit acid salts urine volatile oil quantity sulphur air made waters earth parts colour particles liquor	
1.3865	5	4.74%	3.42%	stomach drink good eat meat wine food flesh diet meats exercise digestion water bread milk strong hot health made	
2.0241	4	7.57%	3.74%	man **god** men life hath mind good **soul** things body great world thing death make thou fear health love	religion
2.2873	2	8.99%	3.93%	water put half drink ounce powder make white morning ounces oil till wine boil warm good leaves pint day	recipes
4.4537	11	3.44%	0.77%	**called** leaves grows english **herb tree root seed** hath kind colour white black flowers **pepper** small trees call sea	herbals, recipes
5.4022	3	5.63%	1.04%	**called** bones part parts head bone brain skin veins flesh made body side neck sinews hath mouth members substance	called: scholastic


man good urine doth great thy named ben thou water body heed iii…
ana ii iii oil make wound ounces lb ounce iiii wounds half boyle wax de rec…
good water oil wine put thereof herb helps hot head make juice powder cold…
moon day mars sun time sign saturn hour days venus past min sick jupiter…
body blood heat cold humours nature doth hot parts heart natural part dry…
things hath time doe nature man good great medicine medicines physic…
called bones part parts head bone brain skin veins flesh made body side…
called leaves grows english herb tree root seed hath kind colour white black…
water put half drink ounce powder make white morning ounces oil till wine…
man god men life hath mind good soul things body great world thing death…
stomach drink good eat meat wine food flesh diet meats exercise digestion…
water salt fire spirit acid salts urine volatile oil quantity sulphur air made…
de amp est ad cum pr lib ut cap qui qu quod vel si la quam sed aut med
disease fever patient great diseases blood symptoms pain cure time fevers…
child time years woman found day mr days children great women head…
air country great time year places infected house men plague sea city…
part parts wound blood vessels made uterus eye small arteries large patient…
half dose ounce ounces quantity water drams tincture dram powder…
physicians practice physician nature great diseases art method medicines…
body blood parts animal state nature air motion vessels matter part action…
0%
20%
40%
60%
80%
100%
EMEMT
LMEMT

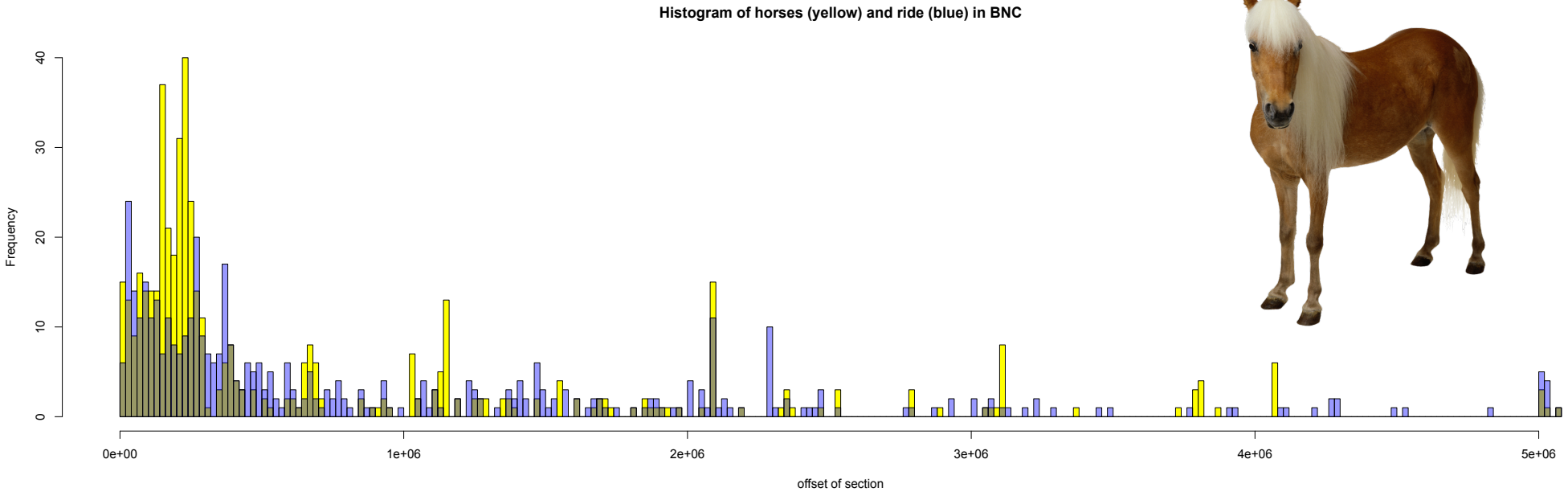
Conceptual Maps: Method

Kernel Density Estimation calculates semantic distances between words

Also a distributional semantic method

Idea: semantically similar words often co-occur in the discourse

E.g. *horse* and *ride* in the BNC

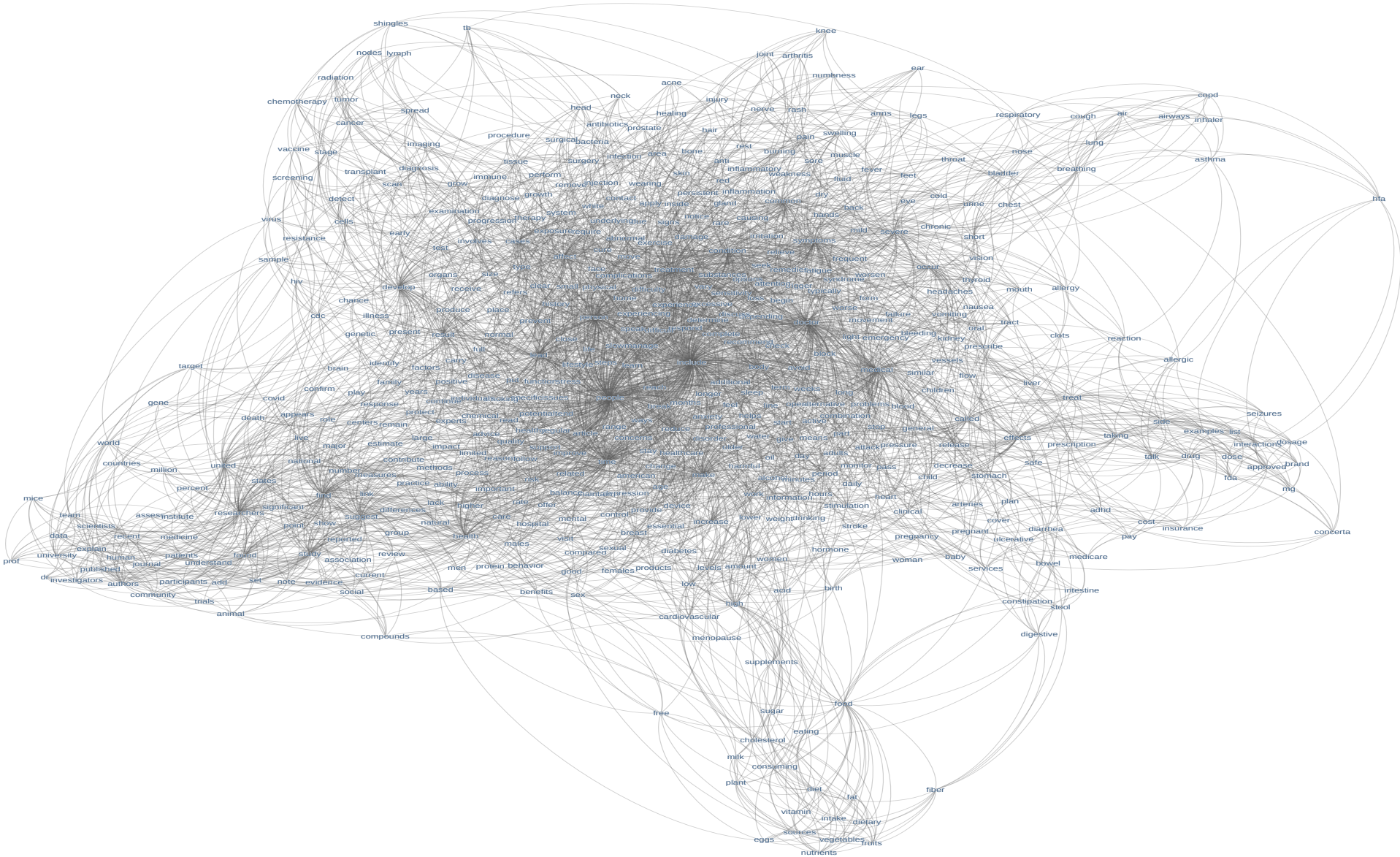


Universität Zürich[UZH]

Conceptual maps: 2000 medical articles

Department of Computational Linguistics

https://www.kaggle.com/trikialaaa/2k-clean-medical-articles-medicalnewstoday



Distributional Semantics: Beipackzettel Patients ←→ Experts

Distributional semantics for synonymy:

	word	similarity to "krank"
1	krank	1
2	traurig	0.488198287057208
3	fühlen	0.487254304485187
4	verwirrt	0.482285979365288
5	matt	0.471478138888479
6	kurzatmig	0.468269510091778
7	unwohl	0.464858218737638
8	zittrig	0.460731892217573
9	reizbar	0.451883650417177
10	fühlt	0.439395384324387
11	ausgetrocknet	0.437100053211833
12	hyperaktiv	0.434033504322587
13	reden	0.428721364102472
14	erbricht	0.427210229099717
15	wassermangel	0.425165456638084
16	insulinbedarf	0.424917730602283
17	dinge	0.423592010815547
18	unpassenden	0.423240004431008
19	aufgeregt	0.419300024940486
20	gelegenheiten	0.414182580492214

← Patients

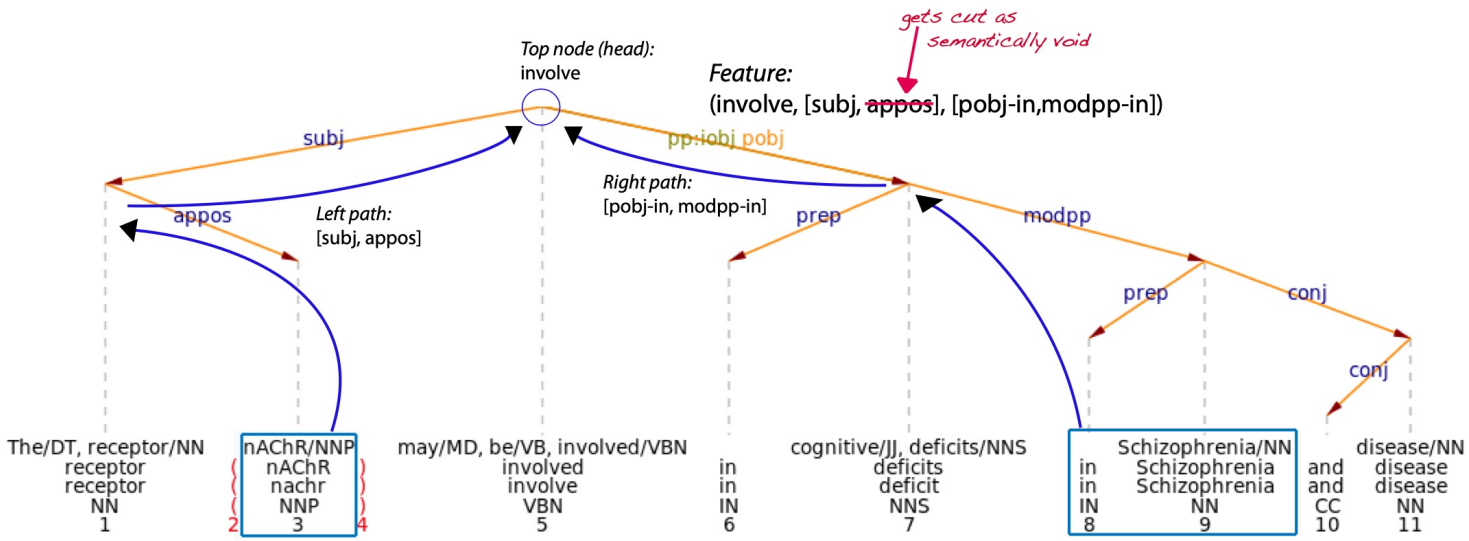
VS

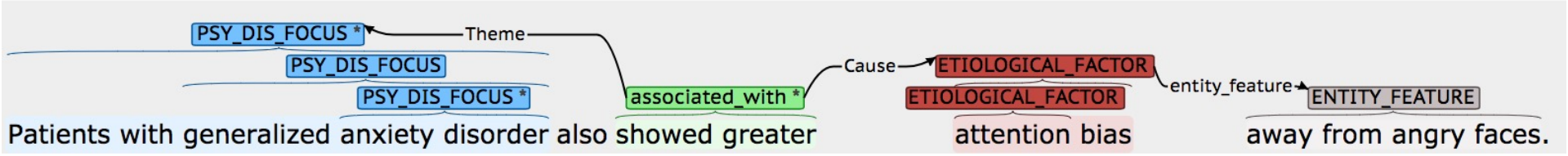
Experts →

	word	similarity to "krank"
1	krank	0.999999999999999
2	21642	0.466164886394565
3	knochenmarktransplantierte	0.451129683506542
4	abschlossen	0.433968247070756
5	augenproblemen	0.42722477575409
6	lamivudinrefraktären	0.425924326423023
7	crystal	0.423800775919397
8	augenkontrollen	0.421896277200166
9	virusstämmen	0.419803058456844
10	eingetroffen	0.416453027612025
11	doppelblinder	0.41502762233347
12	studienbeginn	0.412907260189627
13	elektrolytbalance	0.411730665489956
14	rfhe2	0.411239974203691
15	479	0.410315336369673
16	kernstudie	0.409860401156079
17	vorwarnung	0.408712765795287
18	phosphene	0.40614890128617
19	vertiefter	0.405817730683872
20	oft	0.400641094732256

Relation Mining: Who does what to whom?

We use automatic syntactic and thematic structures to improve relation detection. We have obtained high results in several shared tasks





Text Crunching Center (TCC)

We offer our expertise in the following areas:

- Text Analytics
- Text Mining
- Sentiment Detection
- Digital Humanities
- Machine Translation
- ...

We offer consulting and support in

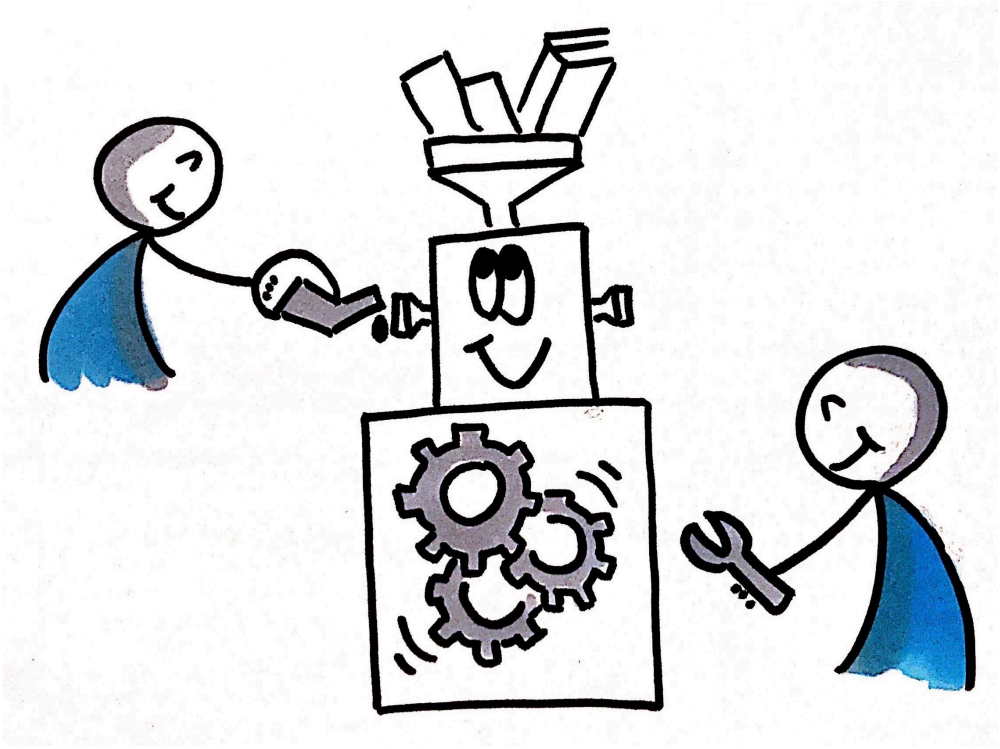
- Digitalisation
- Processing of text, including multilingual and historical texts
- Advice on tools, software and best practices
- Help with project applications and common projects
- Ready-made solutions

Q & A

Thank you for your attention!

Discussion / Question & Answers

(Reserve slides in the following)



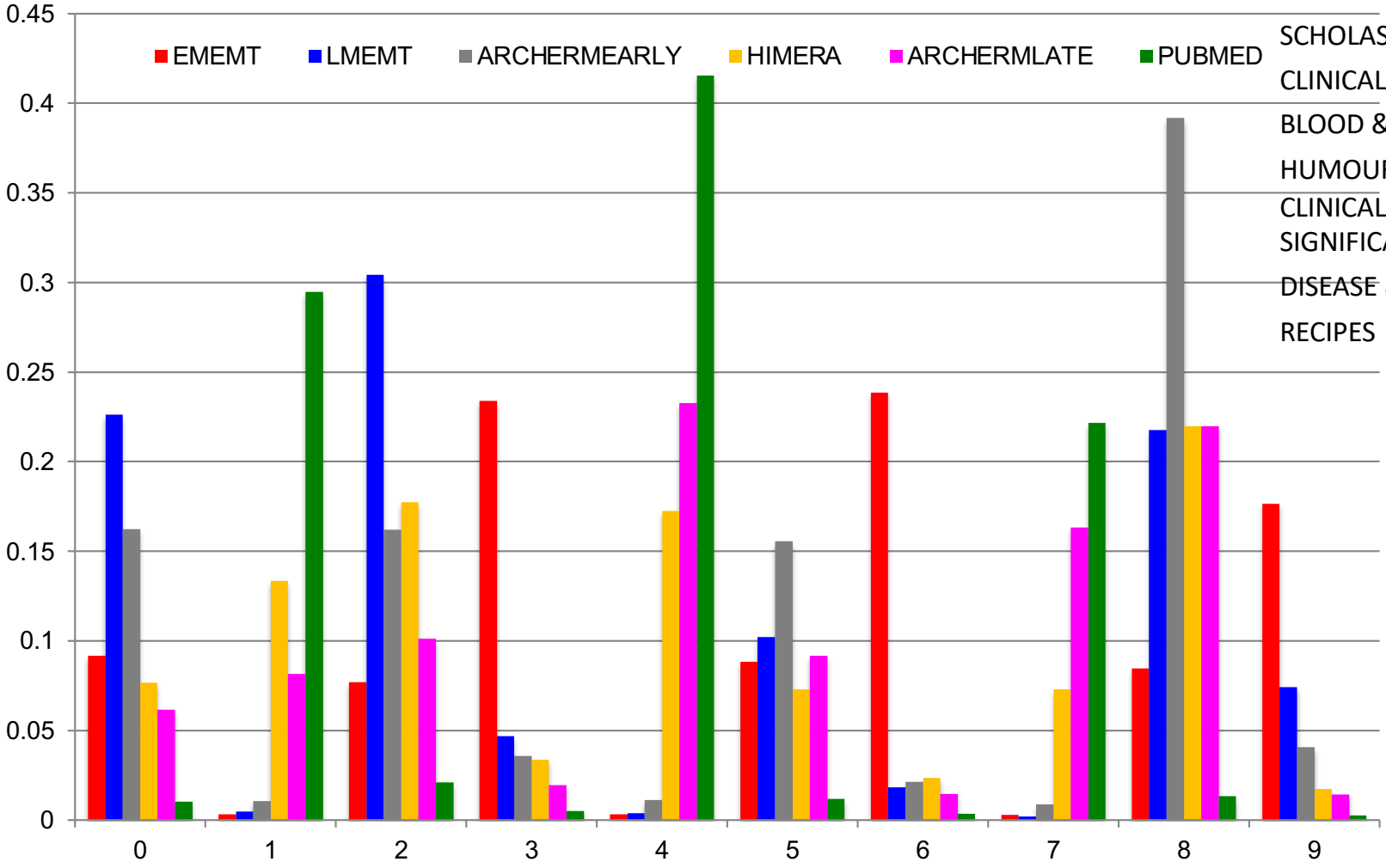
Universität Zürich[UZH]

Department of Computational Linguistics

Topic Models: Medical Topics & Style from 1500-2019

Topic Model:10 topics: Influence per corpus:

EMEMT (1500-1700) LMEMT (1700-1800) ARCHERMEARLY (1650-1900)
HIMERA (1850-1960) ARCHERMLATE (1900-2000) PUBMED (1960-2000)



Manual Topic Label	
BODY PARTS	0
CLINICAL STUDY: RISK	1
PROFESSIONALING PRACTICE	2
SCHOLASTIC	3
CLINICAL STUDY: DIAGNOSIS	4
BLOOD & WOUNDS	5
HUMOURS	6
CLINICAL STUDY: SIGNIFICANCE	7
DISEASE & SYMPTOMS	8
RECIPES	9

Syntactic Complexity

Left branching

Left branching is cognitively more demanding to process, as the head comes later.

Yngve Total: most significant predictor in Cheung & Kemper (1992)

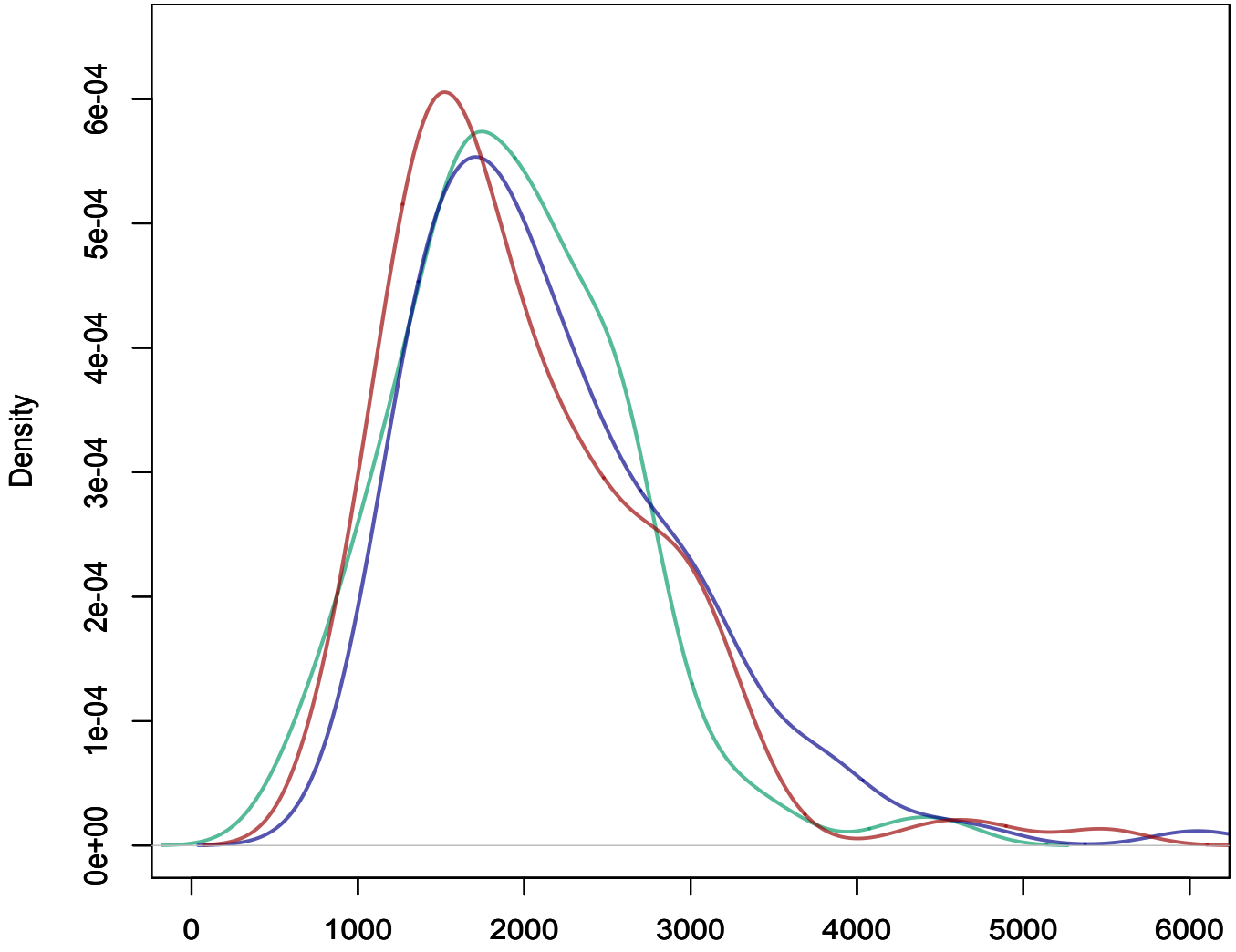
Mean ≠ Median, significant differences

Longer sentences → more scope for left branching

Group	Age	Ø of Yngve Total*	Median	Ø Yngve Total / Sentence length
1	20-35	1933.57	1867	185.52
2	40-55	**2174.83**	1965	131.58
3	65-80	1985.43	**1794**	**124.24**

Syntactic Complexity

Density of left-branching: considerably lower in group 3



Distributional Semantics: Evaluation

Testing the distributional hypothesis: Karlgren & Sahlgren (2001) used a distributional model on the TOEFL word similarity test.
https://www.sics.se/%7emange/papers/KarlgrenSahlgren2001.pdf

Above 50% is typically considered as PASS, random baseline is 25%.

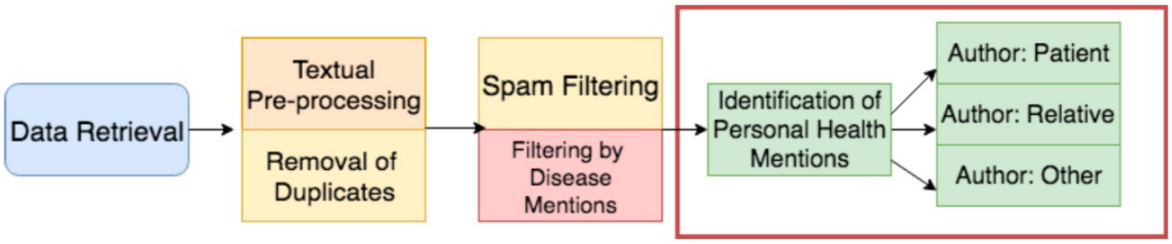
Results of Karlgren & Sahlgren (2001)

TABLE 26.1

Average Results (±1.5) in Percent of Correct Answers to the TOEFL-test
Tr. means truncation length, **WS** means 'word stems', and **PoS+WS** means 'part-of-speech tagged word stems'.

Linguistic analysis	Context window 1+1	2+2	3+3	4+4	Average (±0.73)
None	64.5	67.0	65.3	65.5	65.6
Tr. 6	55.0	57.5	57.3	55.3	56.3
Tr. 8	61.5	64.3	62.0	63.3	62.8
Tr. 10	66.0	68.5	66.3	66.3	66.8
Tr. 12	64.8	65.3	63.8	64.8	64.6
WS	**63.5**	**70.8**	**72.0**	**66.0**	**68.1**
PoS+WS	**66.0**	**64.5**	**65.0**	**65.5**	**65.3**
Average (±0.56)	63.0	65.4	64.5	63.8	

Document Classification: Personal Health-Mention Ident.



- **Identification of Personal Health Mentions (PHM)**
  - Is the micro-post about a specific patient?
  - Is the micro-post authored by the patient itself or by a relative?
- **Participation in Social Media Mining for Health Shared Task (SMM4H 2019)**
  - Task 4: Generalizable identification of personal health experience mentions
  - Focus of shared task: generalize from two given health contexts/health conditions (influenza vaccination and infection) to three unknown health contexts in test set

Document Classification: Personal Health-Mention Ident

Results of our Deep Learning Classification (BERT)

Team	Acc	F1	P	R
Health concerns in all contexts				
UZH	**0.8772**	**0.8727**	0.8392	**0.9091**
ASU1	0.8456	0.8036	**0.9783**	0.6818
UChicagoCompLx	0.8316	0.7913	0.9286	0.6894
MIDAS@IIITD	0.8211	0.783	0.8932	0.697
TMRLeiden	0.793	0.7256	0.9398	0.5909
CLaC	0.6386	0.4607	0.7458	0.3333
Health concerns in Context 1: Flu virus (infection/vaccination)				
UZH	0.9438	0.9474	0.9101	0.9878
UChicagoCompLx	0.925	0.9231	0.973	0.878
ASU1	0.925	0.9221	0.9861	0.8659
MIDAS@IIITD	0.8875	0.88	0.9706	0.8049
TMRLeiden	0.8625	0.8493	0.9688	0.7561
CLaC	0.6625	0.5645	0.8333	0.4268
Health concerns in Context 2: Zika virus, travel plans changes				
UZH	0.7536	0.7385	0.7059	0.7742
MIDAS@IIITD	0.6667	0.5818	0.6667	0.5161
ASU1	0.6957	0.5116	0.9167	0.3548
UChicagoCompLx	0.6377	0.4681	0.6875	0.3548
TMRLeiden	0.6377	0.4186	0.75	0.2903
CLaC	0.5362	0.2	0.4444	0.129
Health concerns in Context 3: Zika virus, reducing mosquito exposure				
UZH	0.8393	0.7692	0.75	0.7895
MIDAS@IIITD	0.8214	0.6667	0.9091	0.5263
ASU1	0.8036	0.5926	1.0	0.4211
UChicagoCompLx	0.8036	0.5926	1.0	0.4211
TMRLeiden	0.7857	0.5385	1.0	0.3684
CLaC	0.6964	0.3704	0.625	0.2632

Results Task 4 (PHM Identification)

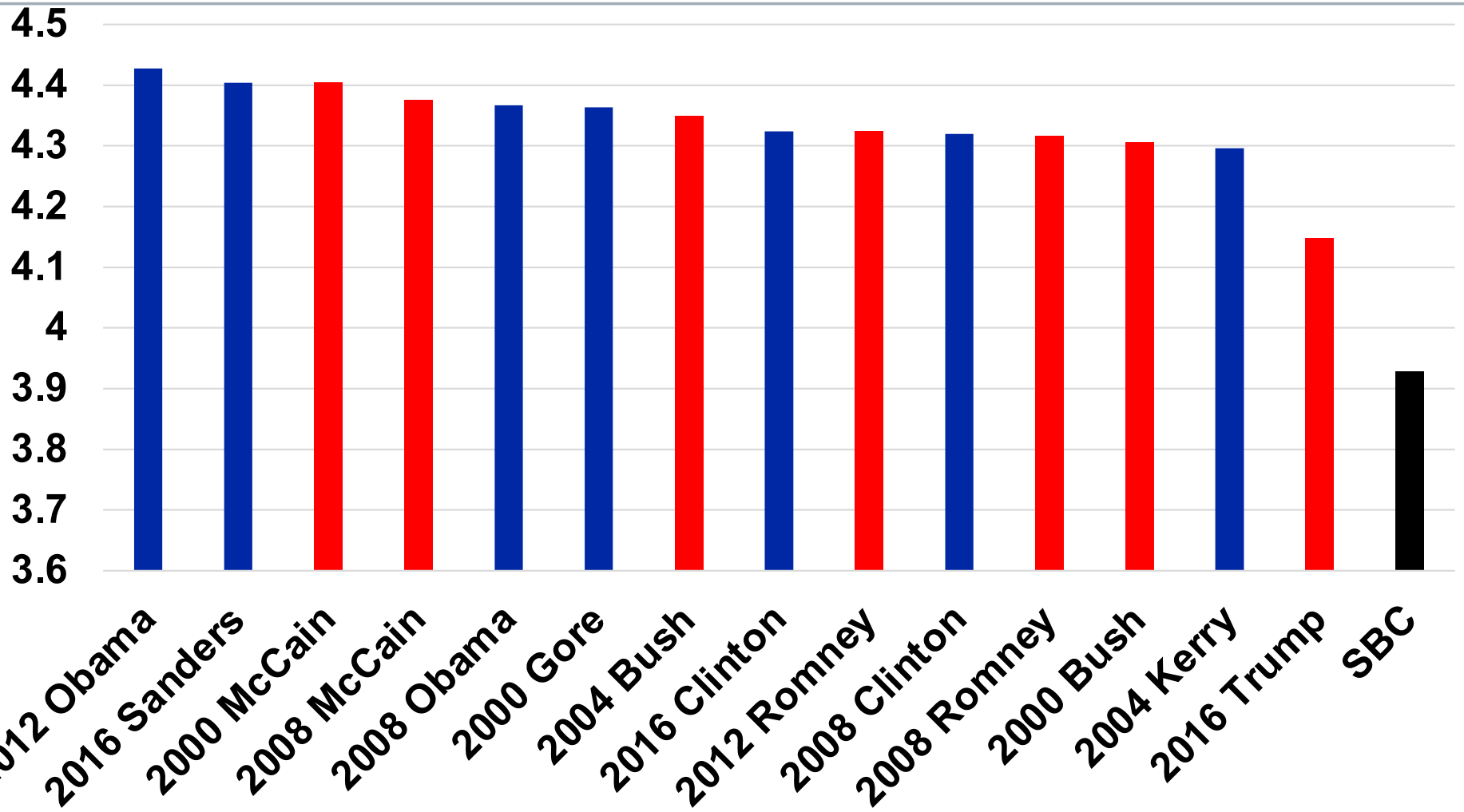
Distributional Semantics: patterns of social & historical change

Poverty in Dickens and contemporary writers

Distributional semantics for synonymy, compared to *poverty* in CLMET

Distributional semantics		green = empathy & social criticism			red=disgust & misery			
SYNONYMS/ASSOC OF POVERTY		WORDS		1780-1850	WORDS		1850-1920	WORDS
	Dickens	1619929		CLEMT p2	1640497		CLMET p3	1535183
	word	sim to "poverty"		word	sim to "poverty"		word	sim to "poverty"
1	poverty	1	1	poverty	1	1	poverty	1
2	debauchery	0.5651157	2	debasing	0.5461153	2	degradation	0.5695389
3	wrongs	0.5636513	3	misery	0.5338886	3	destitution	0.5530797
4	cupidity	0.5542695	4	cravings	0.5214956	4	miseries	0.5468322
5	breasts	0.5442365	5	violating	0.5152474	5	dregs	0.5398224
6	wealth	0.5413223	6	indigence	0.5092006	6	alleviate	0.5265905
7	oppression	0.5365471	7	punishments	0.5033317	7	compensations	0.5220415
8	sickness	0.5335124	8	debase	0.4981087	8	squalid	0.5176405
9	riches	0.5302013	9	hardens	0.4974971	9	misery	0.5104598
10	unrelenting	0.5268691	10	untaught	0.4946344	10	penury	0.5019935
11	joys	0.5214431	11	degradation	0.4936348	11	squalor	0.4984847
12	griefs	0.5176553	12	immoderate	0.4760417	12	commiseration	0.4903199
13	hardship	0.51688	13	unassisted	0.4756063	13	privations	0.4876888
14	baseness	0.5152118	14	automaton	0.4745053	14	brotherhood	0.4855808
15	privation	0.5132133	15	luxury	0.4723517	15	sufferings	0.485163
16	barbarous	0.5130689	16	extravagance	0.4713179	16	lice	0.4811208
17	destitute	0.5102119	17	tutors	0.4689418	17	toil	0.4800149
18	heartless	0.5081498	18	profligacy	0.4685716	18	intoxication	0.4794735
19	sordid	0.5050025	19	wretchedness	0.4675212	19	thriftless	0.4794436
20	purest	0.5030571	20	destitution	0.4667001	20	hovels	0.4791035

Presidential Debates: Word Length



Average word length

Poverty in Dickens: Synonyms/Associations

Frequency of ‘poor’ and ‘poverty’ in Dickens, compared to CLMET

per 1 million words, with heatmap (red = high)		
poor	*poverty*	**Book or Corpus**
928.045536	61.86970241	Dickens Christmas Carol
522.568241	19.56137802	Dickens David Copperfield
410.830999	5.33546752	Dickens Great Expectations
545.872078	9.411587547	DickensHard Times
805.321268	55.32741535	Dickens Nicholas Nickleby
313.292506	36.27597442	Dickens Pickwick Papers
626.42656	57.60244234	Dickens Tale Of Two Cities
614.872461	31.05416467	Dickens OliverTwist
393.178336	45.52422226	CLMET_3.1_1 (1710-1780)
471.119489	36.28543825	CLMET_3.1_2 (1780-1850)
414.520845	32.10004943	CLMET_3.1_3 (1850-1920)

Poverty in Dickens: Synonyms/Associations

Distributional semantics for synonymy, compared to *poverty* in CLMET

Distributional semantics		green = empathy & social criticism			red=disgust & misery			
SYNONYMS/ASSOC OF POVERTY		WORDS		1780-1850	WORDS		1850-1920	WORDS
	Dickens	1619929		CLEMT p2	1640497		CLMET p3	1535183
	word	sim to "poverty"		word	sim to "poverty"		word	sim to "poverty"
1	poverty	1	1	poverty	1	1	poverty	1
2	debauchery	0.5651157	2	debasing	0.5461153	2	degradation	0.5695389
3	wrongs	0.5636513	3	misery	0.5338886	3	destitution	0.5530797
4	cupidity	0.5542695	4	cravings	0.5214956	4	miseries	0.5468322
5	breasts	0.5442365	5	violating	0.5152474	5	dregs	0.5398224
6	wealth	0.5413223	6	indigence	0.5092006	6	alleviate	0.5265905
7	oppression	0.5365471	7	punishments	0.5033317	7	compensations	0.5220415
8	sickness	0.5335124	8	debase	0.4981087	8	squalid	0.5176405
9	riches	0.5302013	9	hardens	0.4974971	9	misery	0.5104598
10	unrelenting	0.5268691	10	untaught	0.4946344	10	penury	0.5019935
11	joys	0.5214431	11	degradation	0.4936348	11	squalor	0.4984847
12	griefs	0.5176553	12	immoderate	0.4760417	12	commiseration	0.4903199
13	hardship	0.51688	13	unassisted	0.4756063	13	privations	0.4876888
14	baseness	0.5152118	14	automaton	0.4745053	14	brotherhood	0.4855808
15	privation	0.5132133	15	luxury	0.4723517	15	sufferings	0.485163
16	barbarous	0.5130689	16	extravagance	0.4713179	16	lice	0.4811208
17	destitute	0.5102119	17	tutors	0.4689418	17	toil	0.4800149
18	heartless	0.5081498	18	profligacy	0.4685716	18	intoxication	0.4794735
19	sordid	0.5050025	19	wretchedness	0.4675212	19	thriftless	0.4794436
20	purest	0.5030571	20	destitution	0.4667001	20	hovels	0.4791035

Tonalität und Sentiment

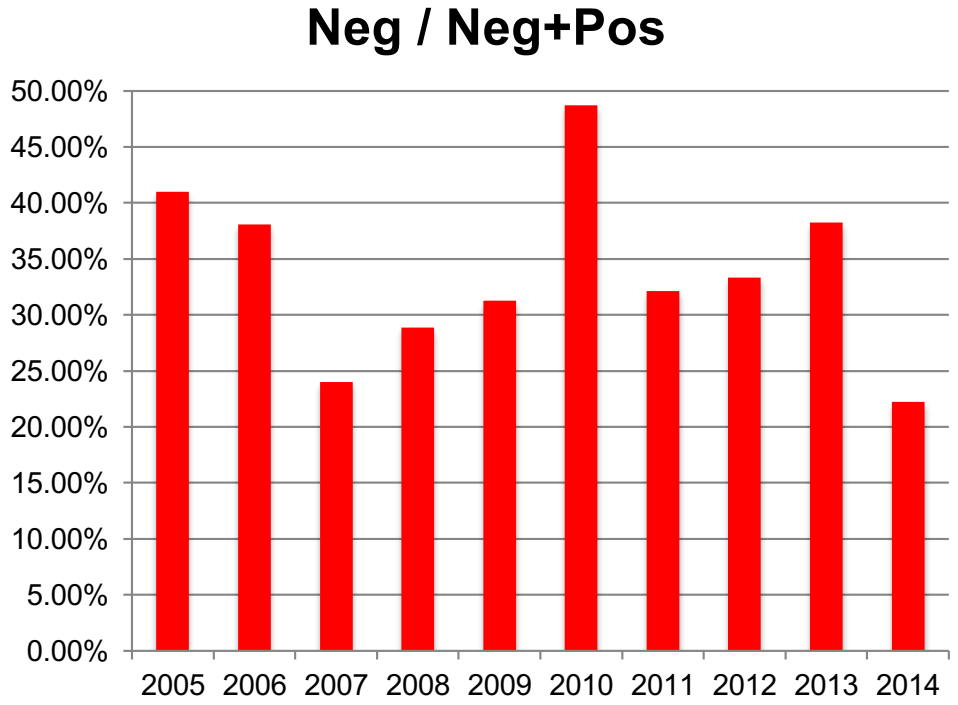
Basiert auf Wörterbüchern, lernt aber auch aus dem Kontext zur Domänenanpassung (z.B. Klenner et al. 2014).

Teaser zu Stuttgart 21:

“Stuttgart21” & “öffentlicher Verkehr”

in deutschen Zeitungen: 2005 – 2014.

- 1309 Artikel, 689’000 Wörter
- Allgemein viele negative Artikel, Peak in 2010



3.2. Tonalität & Sentiment: was war da in Stuttgart im 2010?

SPIEGEL ONLINE POLITIK

Politik | Wirtschaft | Panorama | Sport | Kultur | Netzwelt | Wissenschaft | Gesundheit | einestages | Karriere | Uni | Reise | Auto | Stil

Nachrichten > Politik > Deutschland > Stuttgart 21 > "Stuttgart 21"-Räumung: Bürgerkrieg im Schlossgarten

"Stuttgart 21"-Räumung: Bürgerkrieg im Schlossgarten

Von Josef-Otto Freudenreich, Stuttgart

Bislang war der Bürgerprotest gegen "Stuttgart 21" friedlich - jetzt ist die Lage eskaliert: Bei der Räumung des Baugeländes hat die Polizei Tränengas und Wasserwerfer eingesetzt, viele Demonstranten wurden verletzt. Beobachter machen Regierungschef Mappus für den Gewaltausbruch verantwortlich.

Donnerstag, 30.09.2010 – 19:07 Uhr

Drucken | Senden | Merken

Nutzungsrechte | Feedback

Kommentieren | 1516 Kommentare

Teilen Empfehlen 58 Twittern 8 +1

Es regnet Tränengas. Kinder, Schüler, alte Frauen und Männer fallen übereinander, werden hochgehoben und dorthin geschleppt, wo der scharfe Strahl der Wasserwerfer nicht mehr hinreicht. Manche Gesichter sind blutüberströmt, die Augen brennen höllisch, der Atem wird knapp. Szenen wie in Wackersdorf, sagen ältere Semester, die über einschlägige Demo-Erfahrungen verfügen. Aber es ist nicht Wackersdorf, es ist der Stuttgarter Schlossgarten, die gute Stube der Schwaben, in der sie sonntags mit ihren Kindern spielen.

Interaktive Grafik