University of Zurich[UZH]

Institute of Computational Linguistics

Linguistically Motivated Trend Identification

Gerold Schneider & Michael Amsler

University of Konstanz &

University of Zürich

Institute of Computational Linguistics

SwissText 2016, June 8

ZHAW Winterthur

Overview

1) Motivation
2) Neologism Detection for German
3) Coupling Neologisms with Sentiment Analysis
4) Neologism Detection for English
5) Tracing the Neologism
6) Topic modeling
7) Conclusions

1. Motivation

Correlation: Attention to political issues ~ salience of topics

But words and language themselves change over time

1) We track the introduction of **new terms** (MWT)

   We detect their invention/ﬁ xation, and sources

   New terms often indicate a new perspective.

   We detect **shift of topics or tonality**:

2) Changing associations according to **sentiment detection**

3) Changing associations according to **co-occurrence statistics & topic-modelling**

For this study, we conduct this analysis for German and English

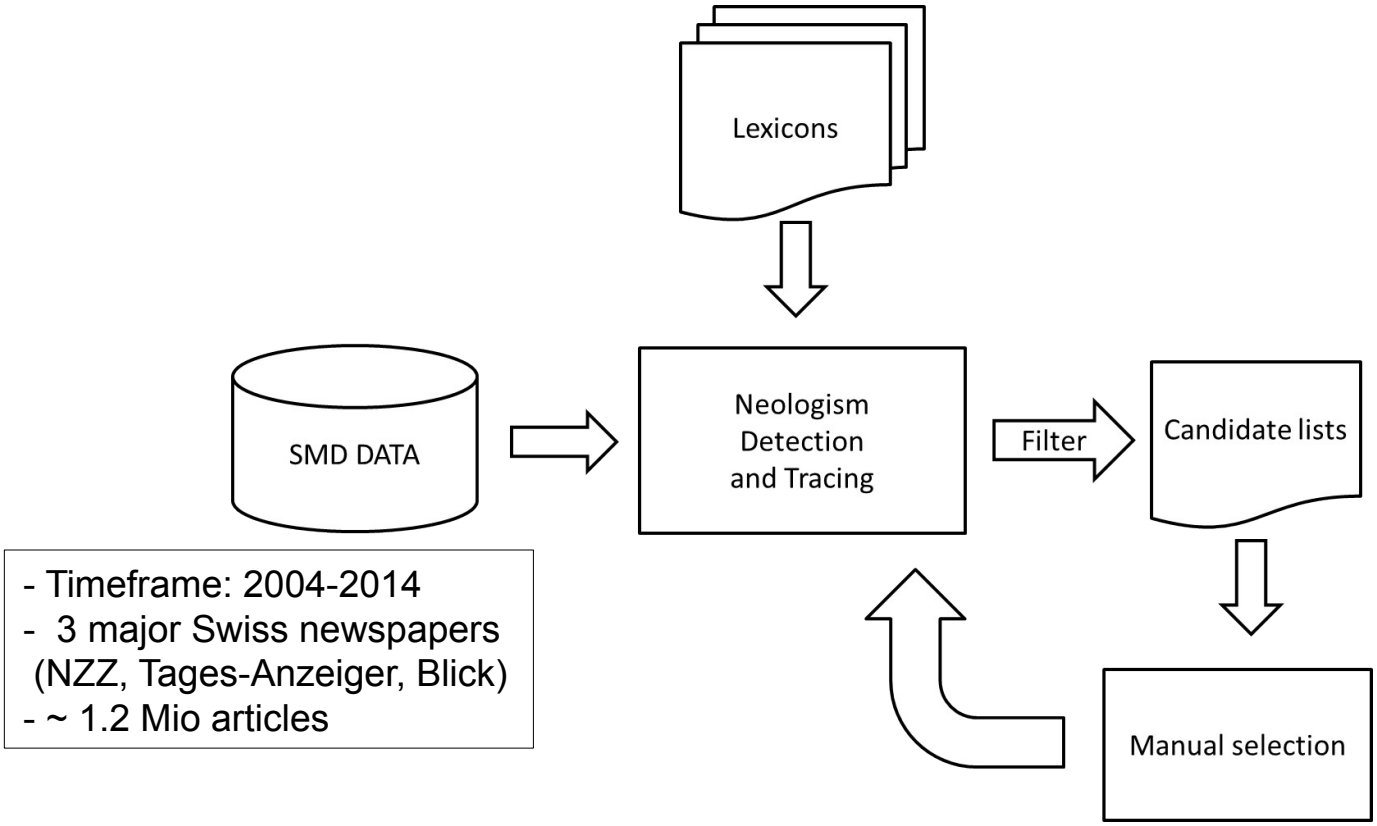
   using news corpora (Swiss media and New York Times/CNN)

2. Neologism Detection for German

- Definition elements of neologism:
  - New term, but not ad-hoc created occasionalism
- We define the following criteria:
  - Neologisms should be new (i.e. not occur in an assembled lexicon of “known words”)
  - Neologisms should occur repeatedly and in multiple sources
- German very productive, especially noun compounds

  Fussball + Europameisterschaft = Fussballeuropameisterschaft
- Method:
- First idea: Single-unit words are relatively easy to detect if occurring the first time (relative to what?)
- But: Abundance requires different forms of filtering

Neologism Detection for German: Schema



Schema for the detection of German neologisms within the SMD corpus

Neologisms: filtered candidates

Corpus	German	English (direct translation)
All_party_articles	Heiratsstrafe	marriage punishment
	Scheininvalide	people pretending disability
	Superreich	super rich
	Weissgeldstrategie	white (clean) money strategy
All_campaign_articles	Politgeograf	political geographer
	Ärztestopp	stop of (admission of) physicians
	Fumoir	smoking room
	vorbeipolitisieren	to politicise off target (to beat about the bush)
All_klimawandel_articles	Klimaabkommen	climate agreement
	Klimaschutzmassnahme	climate protection measure
	Klimasünde	climate sin
	Klimakrieg	climate war
All_IV_articles	Scheininvalide	people pretending disability
	Integrations-massnahme	(refugee) integration measure
	Frauenrentenalter	retirement age of women
	Abbauvorlage	(cost) reduction bill

Neologisms: Party preferences

Party	Term				
	Scheininvalide	Heiratsstrafe	Weissgeld-strategie	Superreich	Rentenklau
SVP	**0.88**	0.56	0.52	0.47	0.43
FDP	0.29	0.54	**0.66**	0.42	0.43
CVP	0.24	**0.81**	0.3	0.25	0.43
SP	0.45	0.45	0.56	**0.67**	**0.86**

Conditional probability for parties, given the neologism, i.e $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Neologism Detection for German: interim results

- Simple but effective approach
- Definition criteria seem to potentially scrape out good candidates
- Proves to be applicable also to smaller subsamples of the corpus

  → produces even better candidate lists

- Considering the found neologisms in the All_party_subcorpus we also point to the fact that they represent strong stance and conjure up derogatory associations

Neologisms: Results II

Party	Term				
	Scheininvalide	Heiratsstrafe	Weissgeld-strategie	Superreich	Rentenklau
SVP	**0.88**	0.56	0.52	0.47	0.43
FDP	0.29	0.54	**0.66**	0.42	0.43
CVP	0.24	**0.81**	0.3	0.25	0.43
SP	0.45	0.45	0.56	**0.67**	**0.86**

Conditional probability for parties, given the neologism, i.e $P(A|B) = \frac{P(A \cap B)}{P(B)}$

3. Coupling Neologisms and Sentiment Analysis

FREITAG, 13. JUNI 2003

Tages-Anzeiger

SCHWEIZERISCHE TAGESZEITUNG WWW.TAGESANZEIGER.CH AUFLAGE 234 518 111. JAHRGANG, NR. 134 FR. 2.50 (inkl

der Politik
de. Doch was
r Begriff? 11

Die Zürcher Mode-Königin sagt, wie man sich bei Hitze nicht zeigen sollte. 15

Nationaltrainer Köbi Kuhn über sein Erfolgsreze und seinen Führungsstil. 43

Volk
hat sich ges-
nen bezahl-
chieden. Er-
rend 14 Wo-
80 Prozent
sersatzord-
en von jähr-
ken würden
ahlt. In fünf
nabzüge um
2012 um ein

Christoph Blocher will gegen «Scheininvalide» vorgehen

Die Diskussion über ein höheres Rentenalter und die maroden Pensionskassen hält Blocher für sekundär. Das grösste Problem sei die Invalidität.

bedarf sieht der Präsident der Zürcher SVP bei der Invalidenversicherung (IV). Hier gebe es gigantische Missbräuche. Ein Grossteil der psychisch Kranken seien nämlich bloss «Scheininvalide». Manche wollten gar nicht mehr gesund werden, weil sie die IV-Rente einem Lohn vorzö-

am meisten Ärzte und Psychologen tätig sind. «Eigentlich sollte es ja umgekehrt sein», findet er. «Je mehr Ärzte es gibt, desto gesünder sollten die Leute sein.»

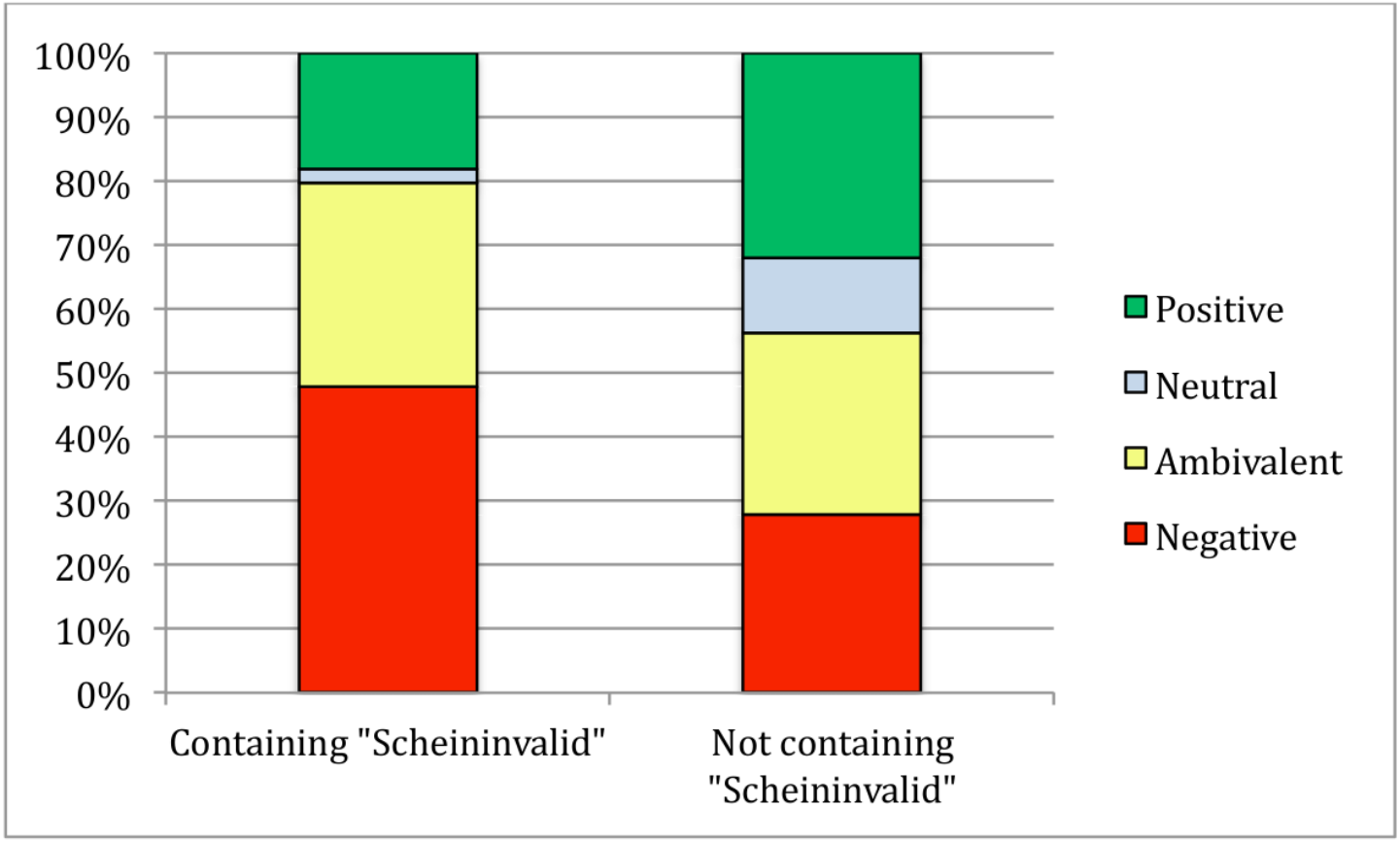
Der Zürcher Nationalrat hat seine Forderungen gestern der Fraktion präsentiert und sich auf der ganzen Linie durchge-

KOMMEN
Nach
bitte
Von Andre
D

Coupling Neologisms and Sentiment Analysis

- We focus on the influence of the neologism “Scheininvalide” (people pretendig disability)
- Setting:
  - from all the articles with mentions of a major party of the government we filter the ones with the mention of the disability insurance (“IV”, “Invalidenversicherung”)
  - We compare the overall tonality in the coverage containing or lacking the neologism
  - We apply a sentiment analysis based on a system which has already has been tested and on other similar cases (see Klenner et al. 2014, Wueest et al. 2014)

Results of overall tonality analysis



Comparison: What leads to the observed negativity?

Not containing „Scheininvalid“				Containing „Scheininvalid“			
Articles 2154				Articles 95			
Term		**Abs. count**	**Occur-rence per article**	**Term**		**Abs. count**	**Occur-rence per article**
German	English			German	English		
Problem	problem	802	0.37	Missbrauch	misuse	76	0.80
Kosten	costs	568	0.26	Problem	problem	48	0.51
Defizit	deficit	336	0.16	Defizit	deficit	27	0.28
Schuld	blame / debt	326	0.15	Krankheit	illness	23	0.24
Kritik	criticism	265	0.12	Beschwerde	complaint	22	0.23

Neologism and Sentiment Analysis

- Clearly more negative overall tonality of the articles with “Scheininvalide”:
  - Negative percentage increases from 28% to 48%
  - Also decrease of positive coverage: from 32% to 18%

- Inspection in negative articles reveals the shift in the topic and the aspects made salient respectively

4. Neologism Detection for English

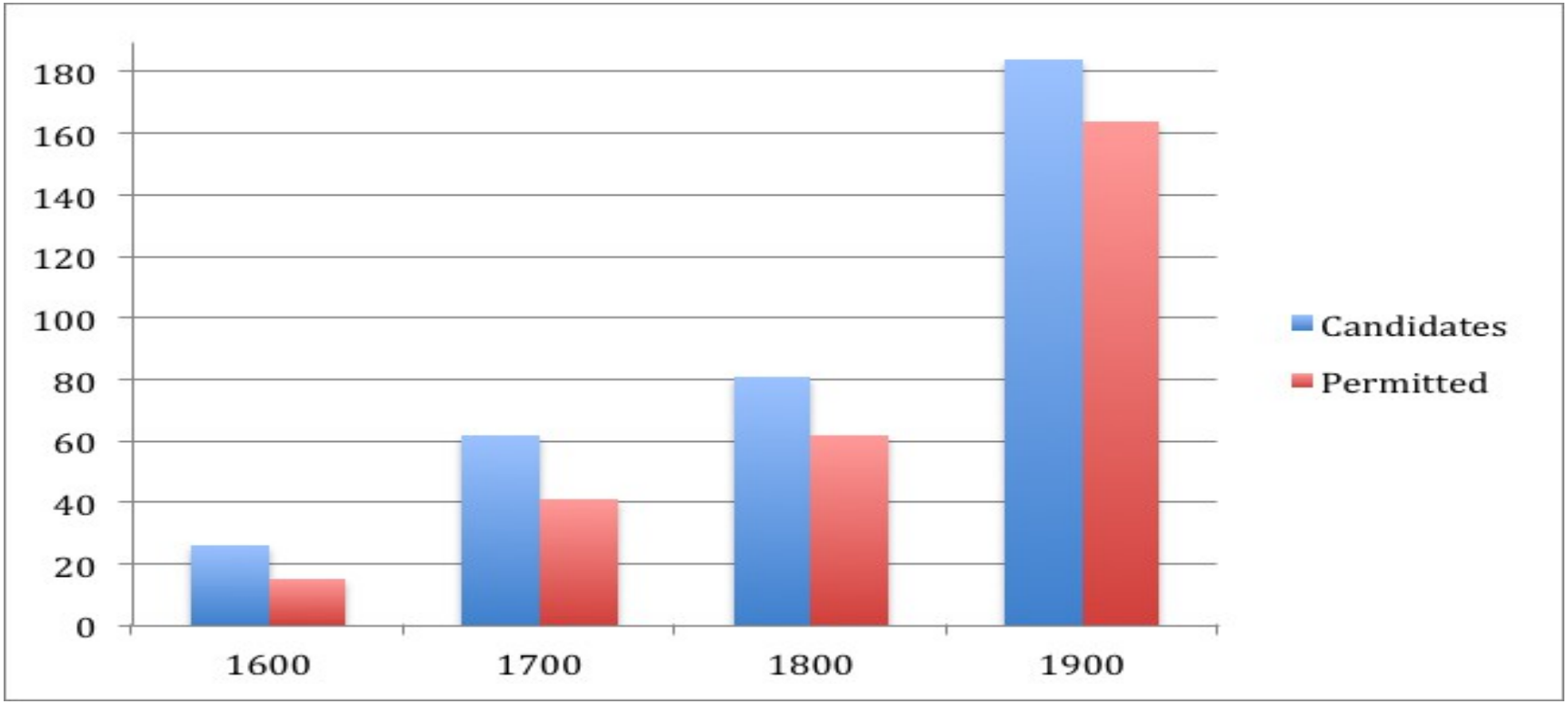
4. Neologism Detection for English

- English compound nouns are written as 2 words, unlike in German
- Compound nouns are a major method for creating neologisms
- Noun-noun sequences have tremendously increased
- They typically derive from a more complex paraphrase, often involving a preposition

4. Neologism Detection for English

Neologism Detection in ARCHER corpus (1600-2000)

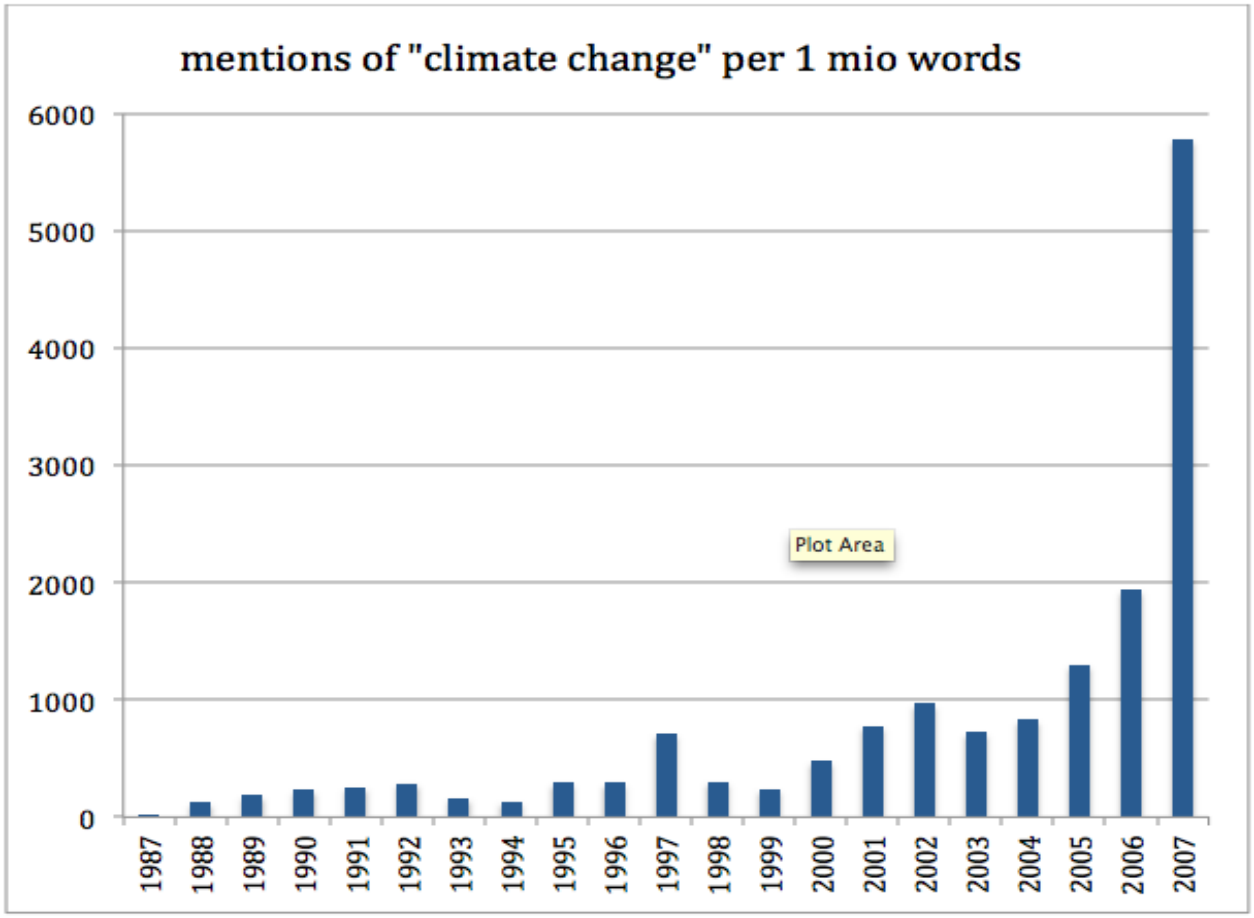
- New noun compound *candidate* = noun-noun sequence licensed by paraphrase
- *Permitted*: those which also pass manual evaluation



5. Tracing a neologism

Tracing the example of 'climate change'

- We now look at the career, life cycle, salience, associations of one neologism
- Salience: frequency in NYT corpus (1987-2007, per 1 mio words)



Associations per period

O/E and T-score as overuse measure

- We split the NYT climate change data into 2 periods: 19=1987-1999 & 20:2000-07
- E(xpected) is homogenous distributions
- Foreach noun compound: how over- or underused is it?
- Overuse 19, and overuse 20, sorted by T-score*F
- Semantic classification
  - Core, e.g. carbon dioxide, climate change
  - Solution frames, e.g. energy efficiency, fuel economy
  - Culprits= problem frames e.g. energy use, power plant
  - Political action, e.g. climate treaty, energy policy
  - Scientific voice, e.g. research group, climate research
  - Other environmental issue, e.g. ozone layer, drinking water
  - Criticism, e.g. greenhouse theory, energy bill

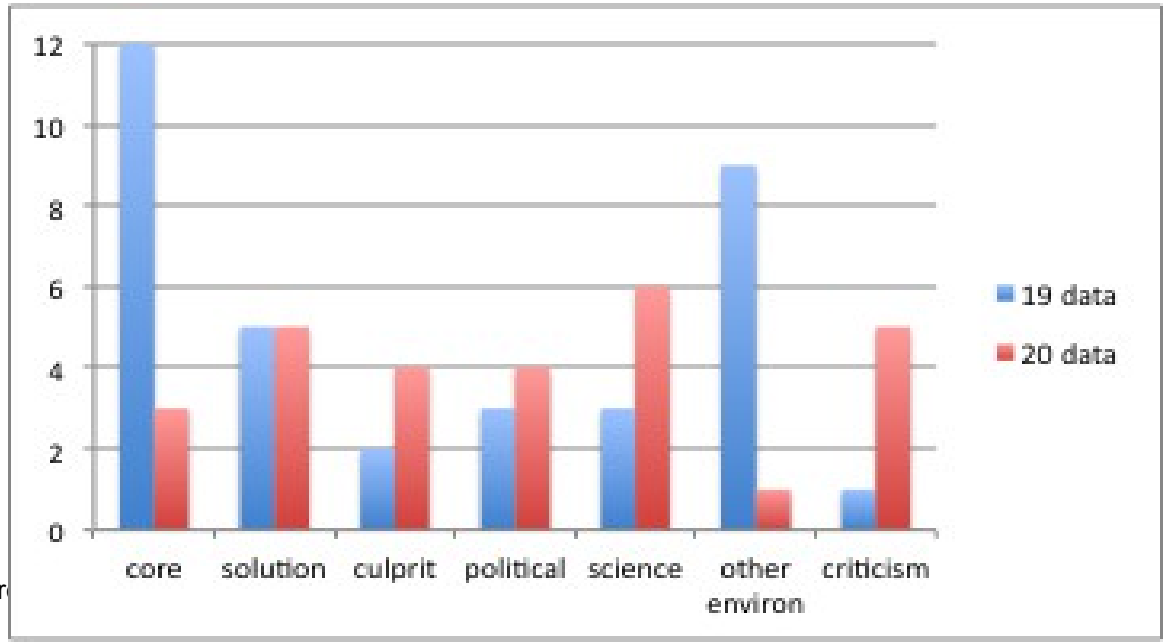
Rank	noun-noun	Pref20OE	Pref20T	Pref20TF
1	carbon-dioxide	0.85	-8.31	-30111.11
2	greenhouse-effect	0.17	-39.12	-21826.52
3	ozone-layer	0.39	-15.89	-6133.80
4	greenhouse-warming	0.07	-30.35	-3186.32
5	surface-temperature	0.27	-15.56	-2692.42
6	ice-age	0.57	-7.98	-2304.88
7	warming-trend	0.67	-6.06	-2025.66
8	trap-heat	0.32	-12.61	-1993.14
9	climate-system	0.50	-8.66	-1853.00
10	air-pollution	0.78	-4.28	-1743.71
11	population-growth	0.45	-8.76	-1454.43
12	ozone-depletion	0.23	-13.15	-1248.94
13	acid-rain	0.52	-7.04	-1148.04
14	energy-use	0.73	-4.21	-1095.39
15	ozone-shield	0.08	-19.90	-1074.62
16	sea-level	0.80	-3.37	-1070.23

Rank	noun-noun	Pref20OE	Pref20T	Pref20TF
1	climate-change	1.09	4.35	16604.00
2	vice-president	1.16	3.45	2780.45
3	fuel-economy	1.23	3.68	1665.05
4	tomorrow-night	1.42	4.38	967.78
5	sea-ice	1.21	2.72	807.88
6	health-care	1.21	2.68	730.40
7	energy-bill	1.35	3.47	663.44
8	energy-policy	1.19	2.33	618.91
9	wind-power	1.26	2.72	541.92
10	task-force	1.22	2.27	417.34
11	energy-plan	1.38	2.85	321.78
12	power-plant	1.11	1.33	313.41
13	attorney-general	1.40	2.89	303.00
14	model-year	1.26	2.21	285.08
15	energy-independence	1.41	2.87	284.55
16	climate-policy	1.34	2.54	279.64
17	et-al	1.38	2.72	271.63

Associations per period

Concepts per semantic group among top 50

– **Semantic classification**	**top from 19 data**	**top two from 20 data**
– Core, e.g.	carbon dioxide,	climate change, sea ice
– Solution frames, e.g.	energy efficiency,	fuel economy, wind power
– Culprits= problem frames e.g.	energy use,	power plant, auto industry
– Political action, e.g.	climate treaty,	energy policy, task force
– Scientific voice, e.g.	research group,	climate research, climate science
– Other environmental issue, e.g.	ozone layer,	drinking water, --
– Criticism, e.g.	greenhouse theory,	energy bill, energy crisis



Associations per period

Yesterday's neologisms are today's keywords

Multiword expression (Sag et al. 2002) often offer a better unit of analysis than tokens. … it is usually easier and more reliable to automatically group together pairs or triples of words that occur together more often than one would expect by chance (Schwartz & Ungar 2015: 84)

There are some documents for which the most relevant keywords are monograms. For example a document on “Tiger”, “Sun”, “football”, etc. … On the other hand, for documents on “Static web page” or “cricket world cup”, there is no mono-gram keyword which can give proper clue about the document. For this type of documents, bi-gram or tri-gram keywords best serve the purpose, thus mono-grams must not be included. (Das et al. 2013: 240)

Associations per period (extra)

Overuse by features of binary document classification

- Logistic regression, using all words: > 90% correct prediction of period
- Most word features are trivial, and not useful. The few interesting ones *are* **noun compounds**

19 Feature	Feature Weight	20 Feature	Feature Weight
economic	0.547	m	1.010
soviet	0.521	--	0.842
science_times	0.481	2005	0.785
page	0.481	2006	0.715
pollution	0.459	letter	0.610
1990	0.455	2004	0.516
nations	0.448	op-ed	0.478
greenhouse_effect	0.438	editorial	0.450
18	0.423	while	0.427
earth	0.419	emissions	0.417
1989	0.402	2007	0.397
atmosphere	0.402	group	0.348
rain	0.398	james	0.348
washington	0.389	kyoto	0.347
expected	0.389	iraq	0.346
trees	0.388	bush_administration	0.342
dr.	0.387	dec.	0.339
such	0.378	editor_re	0.339
effect	0.363	editorial_EOL	0.335
president_clinton	0.356	photo	0.331

Associations per period (extra)

Overuse by features of binary document classification

- Logistic regression, only noun-compund: 83 % correct prediction of period
- Most noun compound features are useful and interesting → keywords

19 Feature	Feature Weight	20 Feature	Feature Weight
auto-maker	2.02	climate-science	1.81
greenhouse-effect	1.98	environment-minister	1.70
greenhouse-warming	1.76	cell-research	1.69
ozone-depletion	1.75	attorney-general	1.49
ozone-shield	1.64	warming-gas	1.48
waste-heat	1.63	administration-official	1.45
summit-conference	1.59	energy-bill	1.40
surface-temperature	1.50	air-pollutant	1.36
day-conference	1.48	carbon-sequestration	1.34
industry-coalition	1.46	energy-plan	1.33
oil-spill	1.41	mass-destruction	1.33
biodiversity-treaty	1.40	energy-independence	1.30
energy-tax	1.39	wind-farm	1.29
research-organization	1.38	assistant-secretary	1.25
policy-center	1.34	campaign-pledge	1.25
missile-attack	1.27	missile-defense	1.24
ozone-layer	1.27	carbon-footprint	1.21
delaying-action	1.26	memory-keeper	1.20
year-term	1.25	energy-secretary	1.19
dengue-fever	1.25	estate-tax	1.18

LDA (Latent Dirichlet Allocation) with 5 topics

Each of the 102271 paragraphs from our NYT data (all 6080 articles containing *climate change* or *global warming*) as separate doc for LDA (Blei et al 2003, Stevens et al. 2012). With 5 topics, weight & keywords:

0 0.19585 climate warming global change emissions greenhouse gases united carbon countries scientists world report dioxide states nations dr heat treaty

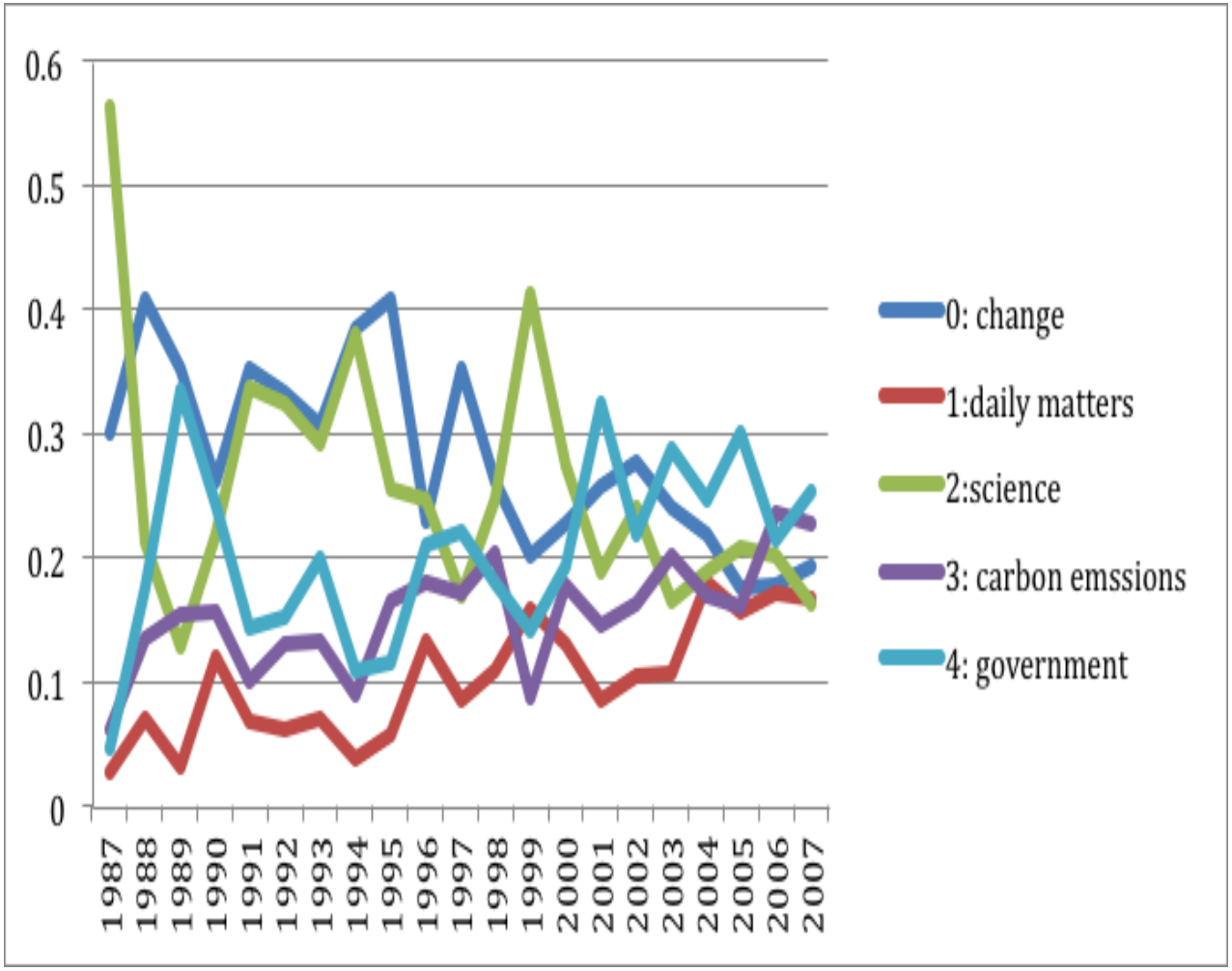
1 0.10515 mr people york book time life city good university home ms day science back don dr years editor school

2 0.15109 years water ice dr sea species scientists year arctic north people long university climate million time ago ocean land

3 0.14514 energy carbon emissions power percent companies oil gas company dioxide coal industry million year environmental fuel billion plants reduce

4 0.18824 mr bush president administration environmental house united states white global american policy change state climate clinton issue political government

LDA Topics change over time



Interpretation:

0 – change is looming: [-]

1 – daily matters: [+] climate change now part of it

2 – science

3 – carbon emissions [+]

We know we really should

4 – politics: no change

100 topics, selection

- Many specific topics are easy to interpret

- Not all are easily interpretable

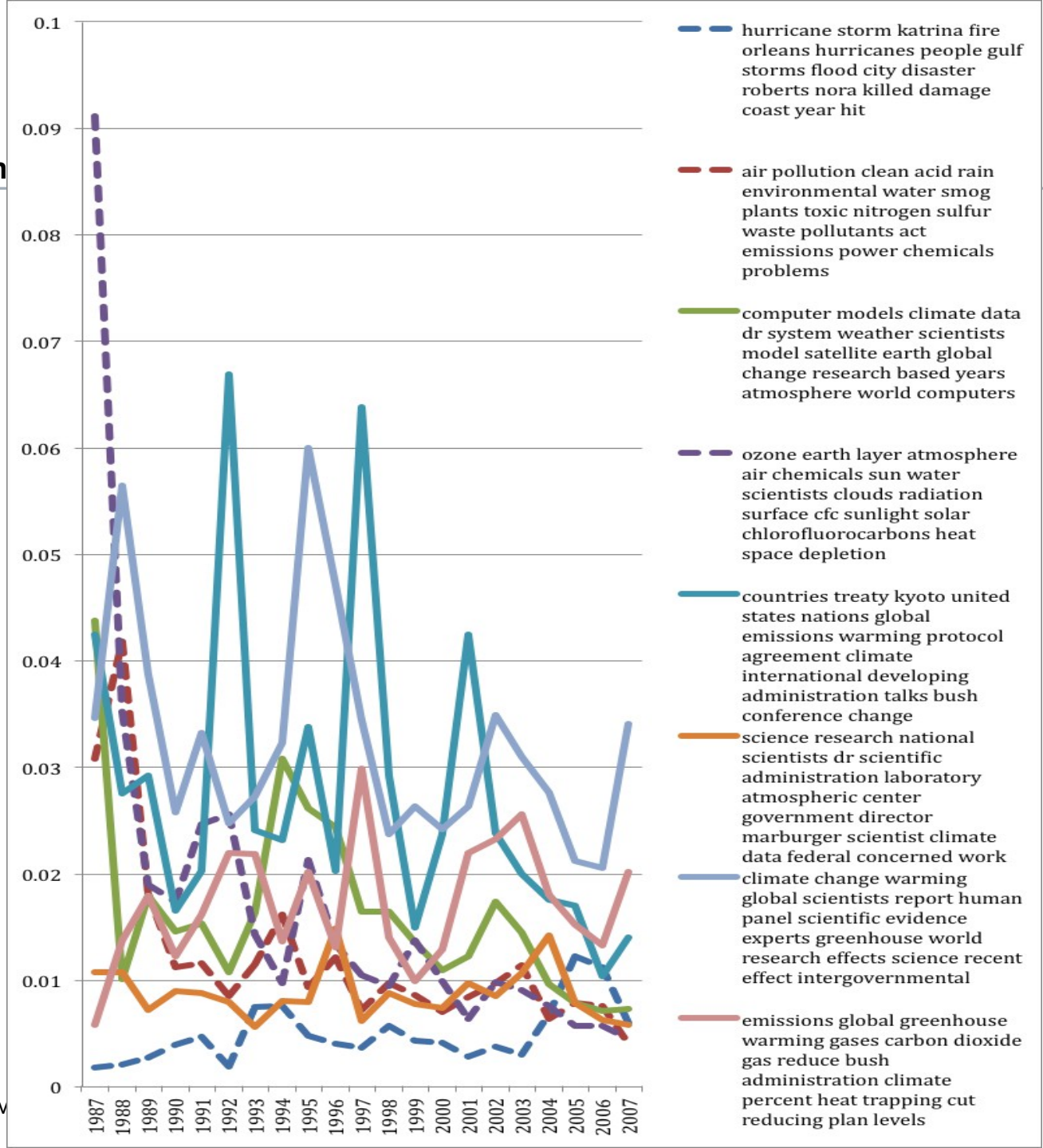
- some topics are closely related

Discussion:

Investigating the development of individual terms as we have done provides an important addition

- hierarchical models

- interactive models



Topic Model using noun compounds only, 5 topics

0 0.03366 climatechange seaice sealevel iceage vicepresident icesheet warmingtrend climatesystem icecap carbondioxide jetstream tippingpoint surfacetemperature etal journalnature whitecity levelrise inconvenienttruth journalscience

1 0.04047 fueleconomy climatechange carbondioxide vicepresident autoindustry fuelefficiency energypolicy energybill windpower powerplant energyefficiency modelyear airpollution energyuse gasmileage energyplan windfarm energyindependence oilconsumption

2 0.04048 healthcare vicepresident summitmeeting climatechange newsconference taxcut massdestruction missiledefense healthinsurance securityadviser cellresearch majorityleader attorneygeneral stockmarket deathpenalty business administrationofficial childcare debtrelief

3 0.01339 tomorrownight org music drinkminimum titlerole sciencefiction soloshow artworld tenorsaxophonist groupshow art titlecharacter climatechange dance icecream settonight rockstar picturecaption museumadmission

4 0.07709 carbondioxide climatechange greenhouseeffect ozonelayer airpollution greenhousegas energyefficiency warmingtrend climatetreaty sulfurdioxide energyuse populationgrowth rainforest trapheat acidrain newsconference vicepresident heatwave scienceadviser

Conclusion

- We have developed several methods to **detect neologisms** in **German** and **English**, especially suited for media news text
- We used therefore **news text corpora (and general corpora)** to automatically create candidates at a manageable scale
- We further measured the **influence of the neologism in the public discourse** using sentiment analysis, overuse stats and topic modeling
- **Noun compounds** as particularly important keywords (fixed, stance)
- We will pursue a **similar strategy to detect and trace syntactical n-grams** aiming at the detection of arguments in forms of fixed phrases
- We need furthermore a coupling with a **citation extraction and attribution** system for a more fine-grained analysis of the news text to keep genuine journalistic content and citations apart.

Thank you for your attention!

Questions?