WZB

Berlin Social Science Center

Introduction to *automated text analyses* in the political sciences

Atelier documents, archives, discours 2018-2019
Université de Lausanne
20.11.2018
1 pm – 5 pm

DR. CHRISTIAN RAUH
www.christian-rauh.eu

Our plan for today

- Promises and pitfalls of automated text analysis
  (or: automated content analysis, text mining, corpus analytics, ...)
  - ➢ Your expectations?
  - ➢ Teaching goal: Enable informed decisions on whether and which automated content analysis methods are suitable for your research
- Discuss the *intuition* and *pragmatic challenges* of the most common political science text analysis methods,
  - Corpus construction and discovery
  - Dictionary-based analyses
  - Text scaling procedures
  - Briefly: topic models, natural language processing, machine learning
- Running example and *tutorials* implemented in R
  Climate change in United Nations General Assembly speeches
- Please bear *your* research questions in mind, apply the discussed ideas to them, and interrupt me whenever something is unclear!

WZB

Part 1
Qualitative, quantitative, and automated

120 most frequent words in Krippendorf 2004: chps. 1-2
Size ~ Relative term frequency

WZB

Content analysis is …

– the analysis of (text) *documents*

- Politics usually happens through written or spoken text
- Which documents matter for your research question?
- Do you cover all of them or only a sample?

– the analysis of *messages*

- Sender → Text → Recipient
- What is your object of inference?

– always *context-dependent*!

- All texts are produced for a purpose
- How does this purpose relate to your inferences?
- Which assumptions do you apply when interpreting the texts?

WZB

Content analysis in between …

– *Positivist* & *interpretative* approaches to scientific inquiry

– *Qualitative* & *quantitative* approaches to social science measurement

WZB

Content analysis as a methodology

– Content analysts vs. newspaper readers

– Reliability, replicability, validity

- ➢ Specify assumptions and benchmarks you apply to texts
- ➢ Detail interpretation / coding / categorization schemes
- ➢ If possible: Validate with external data / information

– Unobstrusive / non-reactive measurement

WZB

Content analysis: A working definition

"Content analysis is a research **technique** for making **replicable** and **valid** **inferences** from **text** (or other meaningful matter) to the **contexts** of their use."

Source: Krippendorf 2004: p. 18

WZB

Why automate?

- Political texts increasingly available in *digitized formats*

- The challenge: *Volume!*
  - Risk of sampling bias
  - Human coding is time- and resource intensive

- The promise: Automated analyses retrieve theoretically relevant concepts from complete text corpora at *comparatively* low cost

- Automated text analyses...
  ... rely on quantitative representations of source texts
  ... often apply statistical models based on *assumptions* about text generation
  ... are extremely reliable, but have to be validated
  ... cannot replace careful and close reading of source texts!

WZB

Four principles of automated text analysis (Grimmer and Stewart 2013)

1. All quantitative models of language are wrong – but some are useful (sometimes)
2. Automated text analyses augment and amplify human interpretation but do not replace it
3. There is no globally best method for automated text analysis
4. Validate, validate, validate!

⇒ Applicability of an automated content analysis can only be judged against *your particular research question* and *theory*!

WZB

Part 2
Corpus construction and discovery

WZB

Acquiring documents

- Which kinds of texts are suitable for automated content analysis?
  - Unit of analysis is usually on *document level* (other units possible)
  - *Focussed* documents preferable (depending on your theoretical concepts)
  - Sufficient *number of words* required (depending on the applied method)
- Typical text sources
  - *Existing corpora*: Other social science projects or linguistic resources
  - *Online databases*: e.g. LexisNexis, Factiva, Gale Cengage (newspapers and press agencies); governments, parliaments, international orgs; etc...
  - *Web scraping*: Press releases, news sites, blogs, Twitter, etc. (see Munzert et al., 2015, Wiley)
  - *Scan / OCR* of printed matter
- Store documents with consistent formats and document names
  - Plain txt files work best (converters freely available)
  - UTF-8 encoding standard for Latin alphabet

WZB

Pre-processing: Turning text into data
(Typical, but not universally applicable steps!)

- Remove...

  ... document "boilerplate" (info not part of the analysed message)

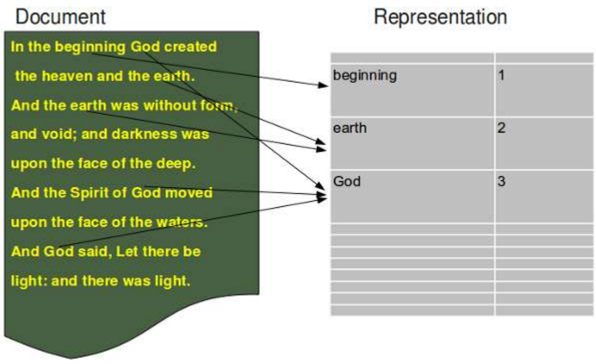
  ... punctuation, capitalization, numbers

  ... very common and very uncommon terms ("stop words", <1% of docs)
- Lemmatization / Stemming
  - Words referring to the same concept mapped to a single root
  - {economy, economic , economically} → economi
- Turn documents into "*bags of words*"
  - Discards the order in which words occur!
  - Unigrams, bigrams ... n-grams

⇒ Document frequency matrix

WZB

“Bags of words” (illustration w/out stemming)



Source: python-course.eu

WZB

Document frequency matrix (illustration)

Label	Titles
c1	Human machine interface for Lab ABC computer applications
c2	A survey of user opinion of computer system response time
c3	The EPS user interface management system
c4	System and human system engineering testing of EPS
c5	Relation of user-perceived response time to error measurement
m1	The generation of random, binary, unordered trees
m2	The intersection graph of paths in trees
m3	Graph minors IV: Widths of trees and well-quasi-ordering
m4	Graph minors: A survey

Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
computer	1	1	0	0	0	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
time	0	1	0	0	1	0	0	0	0
user	0	1	1	0	1	0	0	0	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0

Source: http://web.eecs.utk.edu/~mberry/order/node4.html

WZB

The potential of discovery

- Even if you do not want to apply statistical analyses to your corpus, a look at the aggregated term frequencies may:
  - ... show unknown temporal patterns
  - ... provide contextual information for specific concepts (co-locations, keyword-in-context, synonyms, ...)
  - ... guide selection of individual texts for further human interpretation and coding
  - ... give you an aggregated perspective on the discourse helping to contextualize individual documents therein

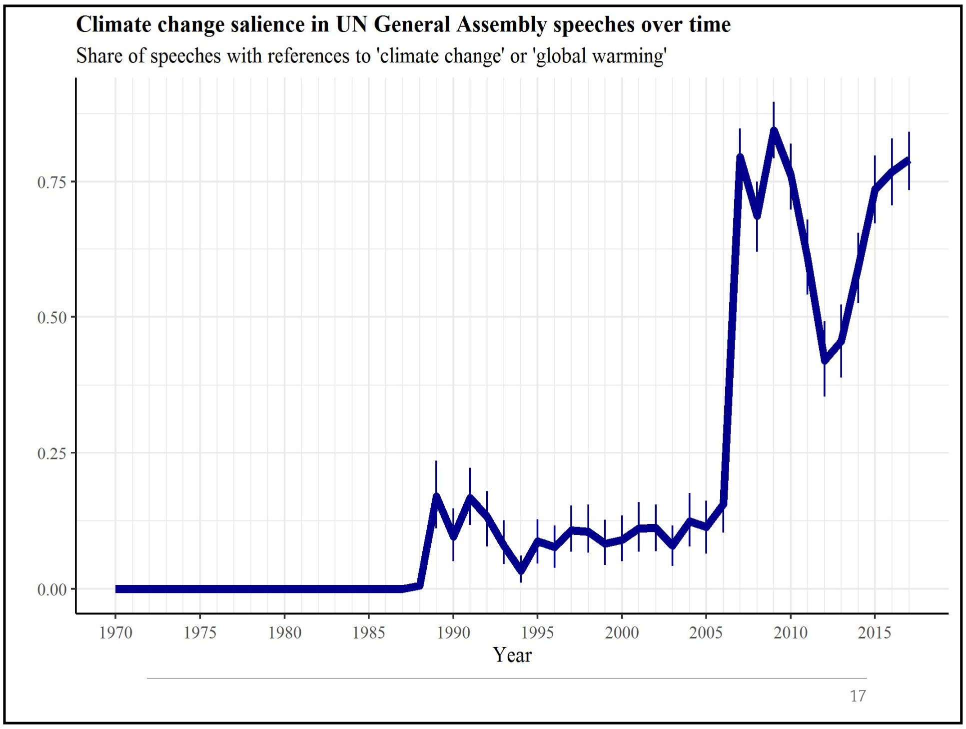
WZB

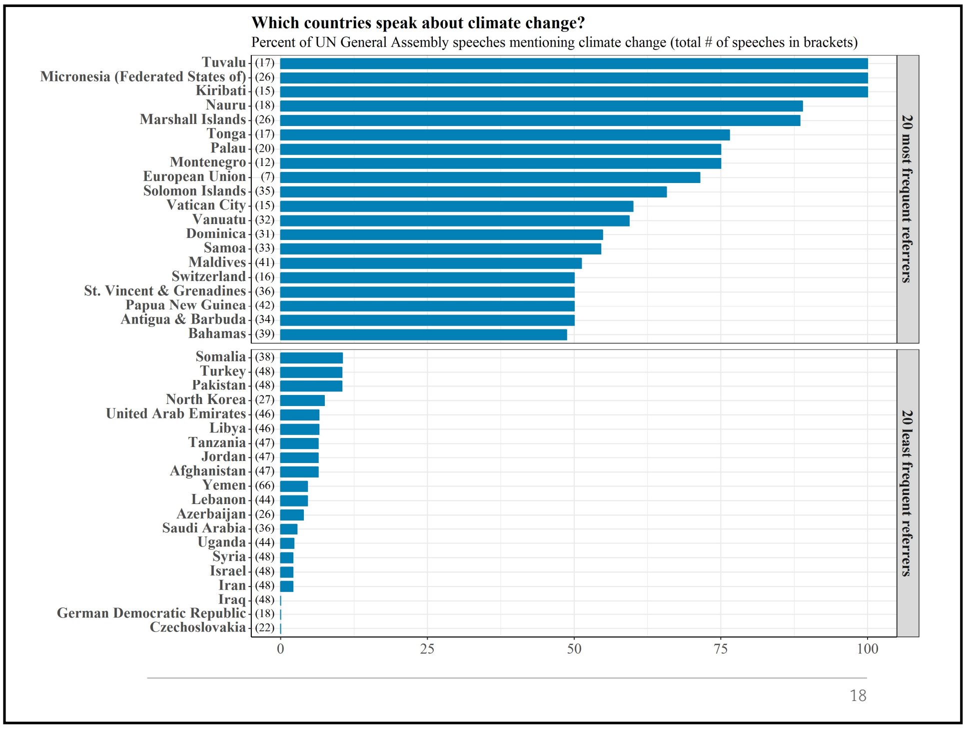
Introducing our running example

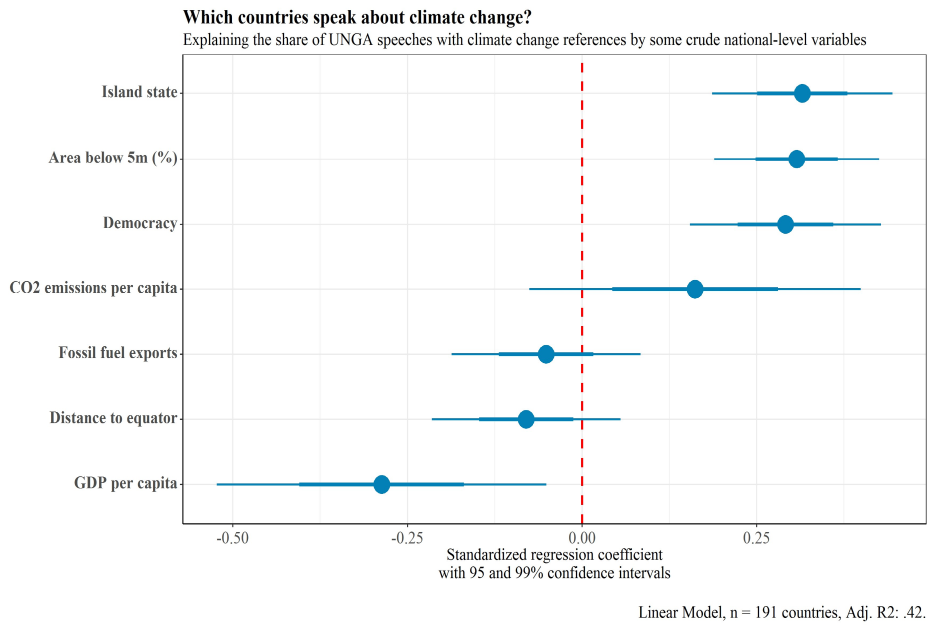
- Speeches in the United Nations General Assembly
  Based on the UNGD corpus assembled by *Baturo, Dasandi, Mikhaylov (2017, R&P)*

- What is the context of these documents we need to have in mind along the sender-message-recipient framework?
  - Who speaks when? With what purpose?
  - The examples will (try to) make inferences about the 'senders', assuming that speeches reflect state positions along the words used

- Climate change as the political issue of interest
  - Identified by speeches referring literally to '*climate(−|)change*' or '*global(−|)warming*'
  - 100-term window around these references to see how and what national delegates say about or associate with climate change
  - For a 'real' analysis, more fine-tuning will most likely be needed, bear with me...


Climate change salience in UN General Assembly speeches over time
Share of speeches with references to 'climate change' or 'global warming'
0.00
0.25
0.50
0.75
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
Year
17


Which countries speak about climate change?
Percent of UN General Assembly speeches mentioning climate change (total # of speeches in brackets)
Tuvalu (17)
Micronesia (Federated States of) (26)
Kiribati (15)
Nauru (18)
Marshall Islands (26)
Tonga (17)
Palau (20)
Montenegro (12)
European Union (7)
Solomon Islands (35)
Vatican City (15)
Vanuatu (32)
Dominica (31)
Samoa (33)
Maldives (41)
Switzerland (16)
St. Vincent & Grenadines (36)
Papua New Guinea (42)
Antigua & Barbuda (34)
Bahamas (39)
20 most frequent referrers
Somalia (38)
Turkey (48)
Pakistan (48)
North Korea (27)
United Arab Emirates (46)
Libya (46)
Tanzania (47)
Jordan (47)
Afghanistan (47)
Yemen (66)
Lebanon (44)
Azerbaijan (26)
Saudi Arabia (36)
Uganda (44)
Syria (48)
Israel (48)
Iran (48)
Iraq (48)
German Democratic Republic (18)
Czechoslovakia (22)
20 least frequent referrers
0
25
50
75
100
18


Which countries speak about climate change?
Explaining the share of UNGA speeches with climate change references by some crude national-level variables
Island state
Area below 5m (%)
Democracy
CO2 emissions per capita
Fossil fuel exports
Distance to equator
GDP per capita
-0.50
-0.25
0.00
0.25
Standardized regression coefficient
with 95 and 99% confidence intervals
Linear Model, n = 191 countries, Adj. R2: .42.

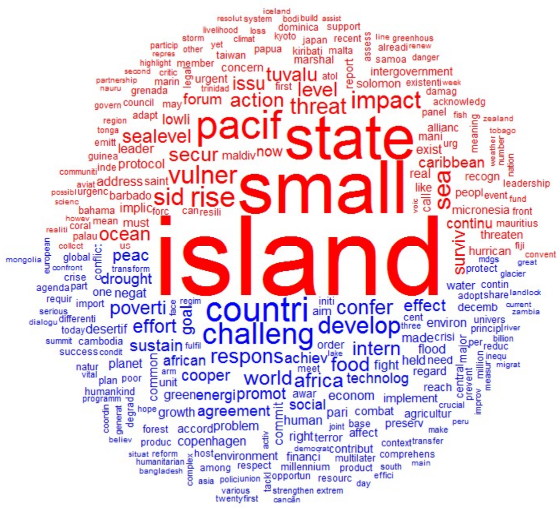

develop
nation
countri
unit
intern
state
global
world
challeng
island
sustain
secur
agreement
framework
action
convent
need
small
effort
impact
must
effect
commit
also
address
confer
issu
communiti
year
respons
peopl
threat
support
human
econom

After 2006

Until 2006

21

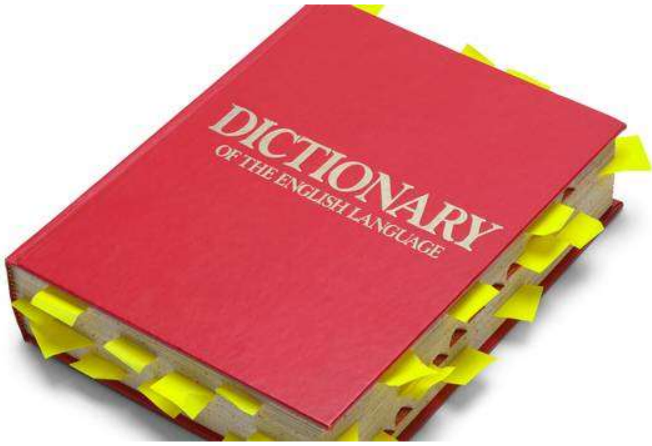
Island states



All others

WZB

Part 3
Dictionary-based text analysis



23

WZB

Basic idea of dictionary-based text analyses

- Presence of or rate at which a *set of predefined key words* occurs in a document is used to classify or scale the document into/along theoretically relevant categories

+ Intuitive and easy to apply
+ Replicable and expandable to various theoretical concepts

24

WZB

Example I: Debating the EU in national parliaments (Rauh & De Wilde 2018, EJPR)

- Question: Do national parliamentary debates enhance the public accountability of EU decision-making?

  (If so, they should mirror EU authority, decision-making, public demand, and feature a balance of government and opposition…but party strategies…)

- Text analysis approach
  - Build a full-text corpus of parliamentary debates in various EU member states (get it here: www.bit.ly/ParlSpeech)
  - Find typical ways of referencing the EU polity, politics, and policies on n-gram level (= reading lots of speeches!)
  - Generalize these examples by regular expressions and build a respective *dictionary*
  - Count, normalize and aggregate EU references to party-month level
  - Relate this to relevant external data

WZB

Rauh & De Wilde (2018): Text data

	Period available	N speeches	Ø speeches per month	Ø terms per speech	Unique terms
DE: *Bundestag*	1991-03 / 2013-09	149,553	607.94	550.82	600,925
ES: *Congreso*	1989-11 / 2015-10	131,986	515.57	526.92	360,012
NL: *Tweede Kamer*	1994-12 / 2015-11	787,879	3396.03	165.57	401,471
UK: *House of Commons*	1988-11 / 2015-01	1,463,637	5361.31	202.07	1,037,450

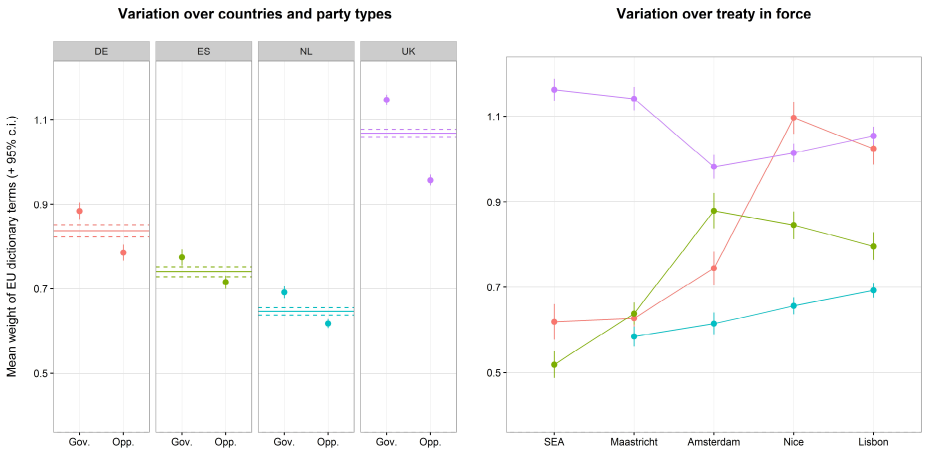
Table 1: Domestic plenary debate corpora

Rauh & De Wilde (2018): Dictionary

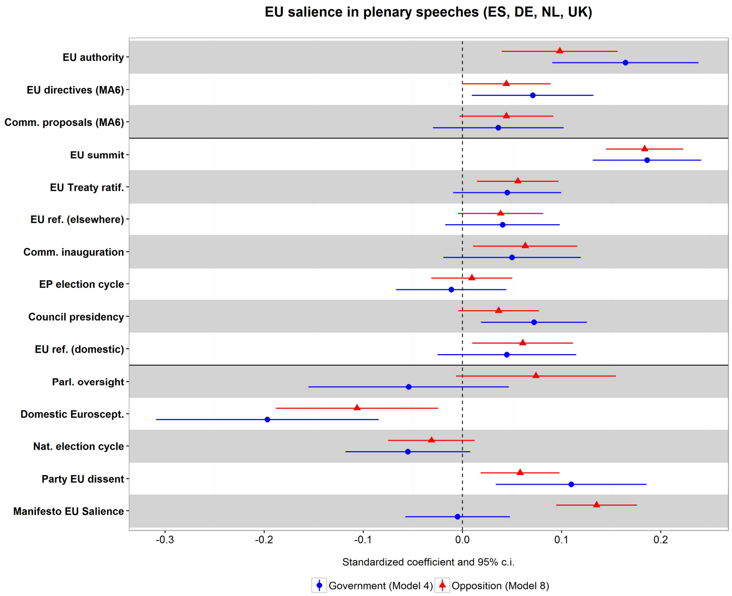
(English version, other languages more complex)

<table>
<tr><th>EU polity</th><th>EU politics</th><th>EU policy</th></tr>
<tr><td>(eu|ec)[1-9]{1,2}</td><td>ecb</td><td>(c|e)sdp</td></tr>
<tr><td>(european|europe's|eu's) constitutional treaty</td><td>ecj</td><td>(common|european) foreign and security polic(y|ies)</td></tr>
<tr><td>(rome|maastricht|amsterdam|nice|lisbon) treat(y|ies)</td><td>ep</td><td>(common|european) security and defen(s|c)e polic(y|ies)</td></tr>
<tr><td>ec('s){0,1}</td><td>european (official(s){0,1}|civil servant(s){0,1})</td><td>eurozone|euro zone|euro area</td></tr>
<tr><td>economic and monetary union</td><td>european (politics|policy)</td><td>cfsp</td></tr>
<tr><td>eec('s){0,1}</td><td>european central bank</td><td>european ([a-z]*){0,1}polic(y|ies)</td></tr>
<tr><td>emu</td><td>european commission(er|ers){0,1}</td><td>european ([a-z]*){0,1}(act(s){0,1}|bill(s){0,1}|law(s){0,1}|legislation(s){0,1}|statute(s){0,1}</td></tr>
<tr><td>eu('s){0,1}</td><td>european competenc(e|es|ies)</td><td>european ([a-z]*){0,1}(aim(s){0,1}|goal(s){0,1}|target(s){0,1}</td></tr>
<tr><td>euratom('s){0,1}</td><td>european council</td><td>european ([a-z]*){0,1}decision(s){0,1}</td></tr>
<tr><td>european ([a-z]*){0,1}(integration|unification|cooperation)</td><td>european court of justice</td><td>european ([a-z]*){0,1}directive(s){0,1}</td></tr>
<tr><td>european_communit(y|ies)</td><td>european election(s){0,1}</td><td>european ([a-z]*){0,1}engagement(s){0,1}</td></tr>
<tr><td>european (economic |atomic energy)communit(y|ies)</td><td>european executive</td><td>european ([a-z]*){0,1}guideline(s){0,1}</td></tr>
<tr><td>european institutions</td><td>european level(s){0,1}</td><td>european ([a-z]*){0,1}(measure(s){0,1}|action(s){0,1})</td></tr>
<tr><td>european project(s){0,1}</td><td>european member state(s){0,1}</td><td>european ([a-z]*){0,1}(provision(s){0,1}|prescription(s){0,1}</td></tr>
<tr><td>european treat(y|ies)</td><td>european parliament</td><td>european ([a-z]*){0,1}(requirement(s){0,1}|allowance(s){0,1}</td></tr>
<tr><td>european_union('s){0,1}</td><td>european procedure(s){0,1}</td><td>european ([a-z]*){0,1}(standard(s){0,1}|norm(s){0,1}</td></tr>
<tr><td>single european act</td><td>european summit(s){0,1}</td><td>european ([a-z]*){0,1}agenda(s){0,1}</td></tr>
<tr><td>treat(y|ies) of (rome|maastricht|amsterdam|nice|lisbon)</td><td>mep(s){0,1}</td><td>european ([a-z]*){0,1}budget(s){0,1}</td></tr>
<tr><td>treaty establishing a constitution for europe</td><td>policy on europe</td><td>european ([a-z]*){0,1}f(u|o)nd(s){0,1}</td></tr>
<tr><td>treaty on (the functioning of the){0,1}european union</td><td></td><td>european ([a-z]*){0,1}programme(s){0,1}</td></tr>
<tr><td></td><td></td><td>european ([a-z]*){0,1}regulation(s){0,1}</td></tr>
<tr><td></td><td></td><td>european ([a-z]*){0,1}strateg(y|ies)</td></tr>
<tr><td></td><td></td><td>european (case-law|jurisprudence|legal)</td></tr>
<tr><td></td><td></td><td>european (single |internal)market{0,1}</td></tr>
<tr><td></td><td></td><td>european [a-z]* union</td></tr>
<tr><td></td><td></td><td>european currenc(y|ies)</td></tr>
<tr><td></td><td></td><td>european mandate(s){0,1}</td></tr>
<tr><td></td><td></td><td>police and judicial cooperation in criminal matters</td></tr>
<tr><td></td><td></td><td>single currency</td></tr>
<tr><td></td><td></td><td>stability and growth pact</td></tr>
</table>

Rauh & De Wilde (2018): Descriptive results



Rauh & De Wilde (2018): Multivariate results



WZB

Example II: NGOs in the public discourse on the WTO (Rauh & Bödeker 2014 mimeo)

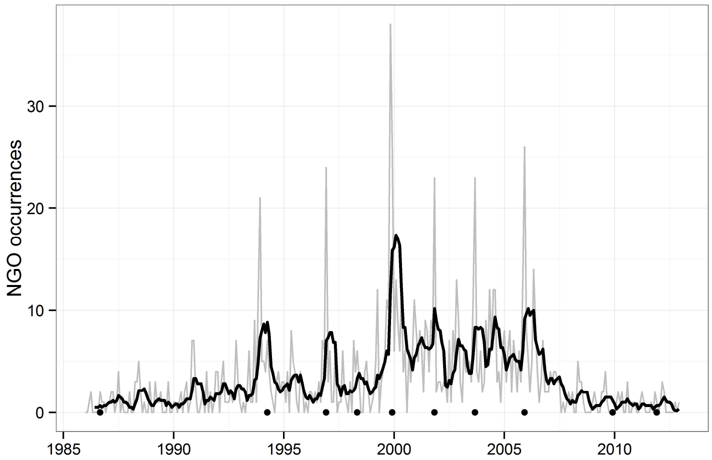
- Question: Which (type of) non-governmental actors participate in the public discourse on the World Trade Organization?

- Approach
  - Financial Times (UK), Straits Times (S'pore), New York Times (US)
  - Download, parse and clean all 11.388 articles from LexisNexis that mention the WTO in headline or lead (1985-2012)
  - Generate encompassing list of all transnational NGOs (Sources: WTO stakeholder directory and UN ECOSOC database)
  - Tag and count the occurrence of each of these NGOs in the articles
  - Manually classify tagged NGOs as 'business' or 'public interest'
  - Aggregate and analyse the data

30

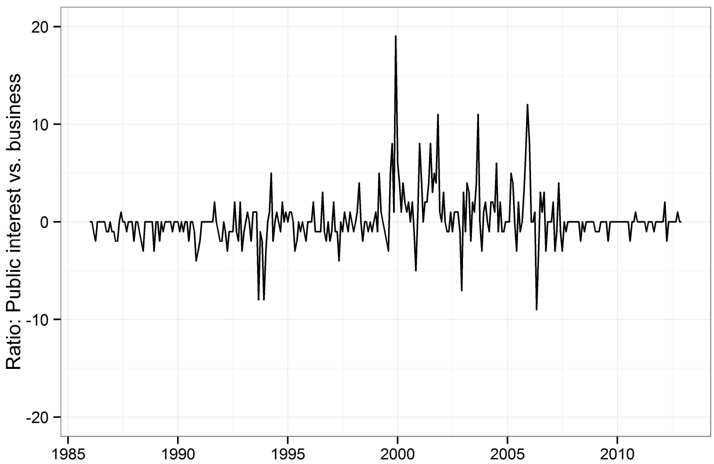
WZB

Example II (cont.)



WZB

Example II (cont.)



WZB

Example II (cont.)

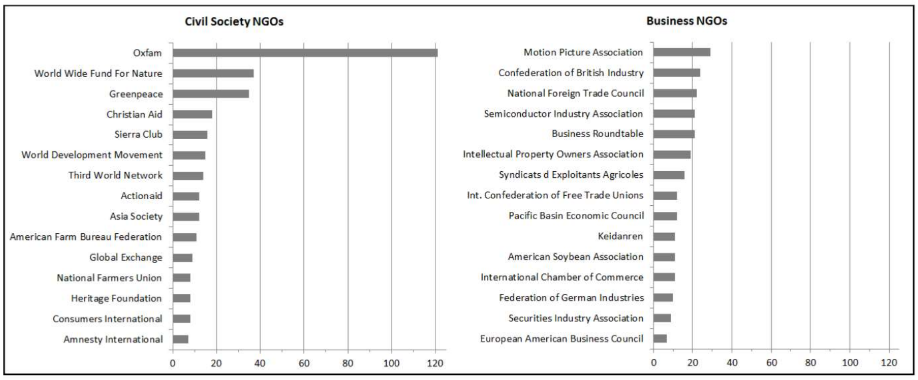


Figure 5: The 15 most present NGOs from the civil society and the business sector

33

WZB

Typical dictionary application: Sentiment analysis

- Typical application: Sentiment analyses
  - Do the analysed message convey information positively or negatively (tone)?
  - Sentiment dictionary: List of terms with individual tone scores; usually ranging between -1 (negative) and 1 (positive)
  - Sentiment at document level: rate at which positively or negatively connoted words occur (often: relative to the overall number of terms in document)

34

WZB

Exemplary sentiment dictionary (Young and Soroka 2011)

Exemplary terms (stemmed) from the *Lexicoder Sentiment dictionary*

Positive connotation	Negative connotation
ALLEVIAT*	ABSURD*
BENEFIT*	BELLIGEREN*
COOPERAT*	CONFRONT*
DESERV*	CONTAGIOUS*
EXCITE*	FALTER*
FAIR*	HELPLESS*
OUTSTAND*	IDEOLOGUE*
PERFECT*	LOSE*
RESOLV*	NEGLECT*
USEFUL*	SCANDAL*
All in all, the LSD has 4,567 unique entries	

35

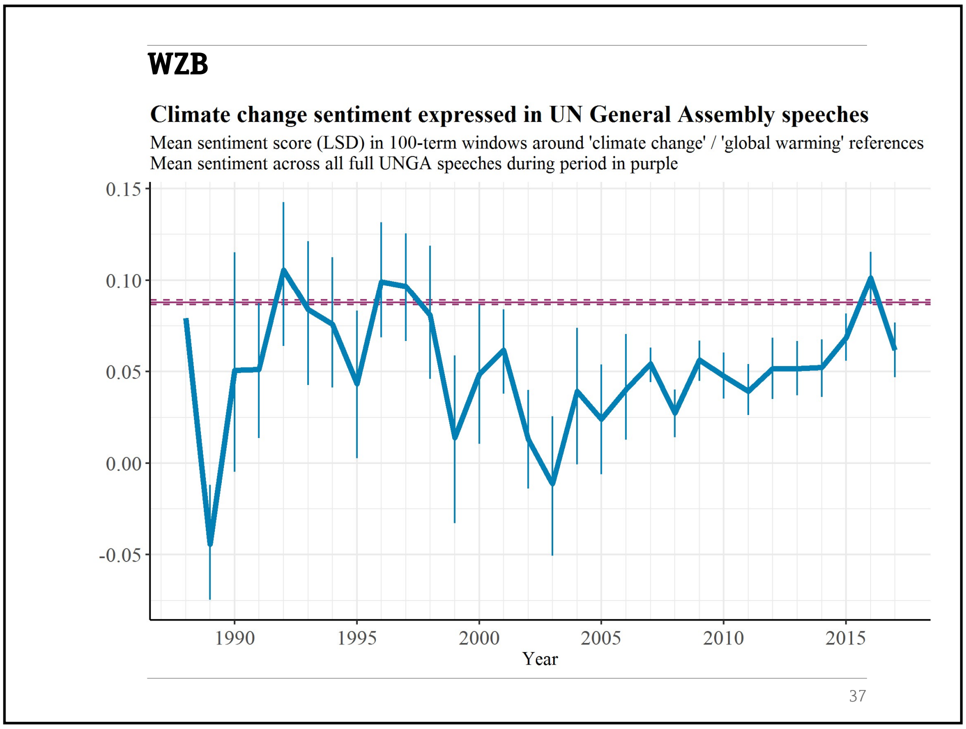
WZB

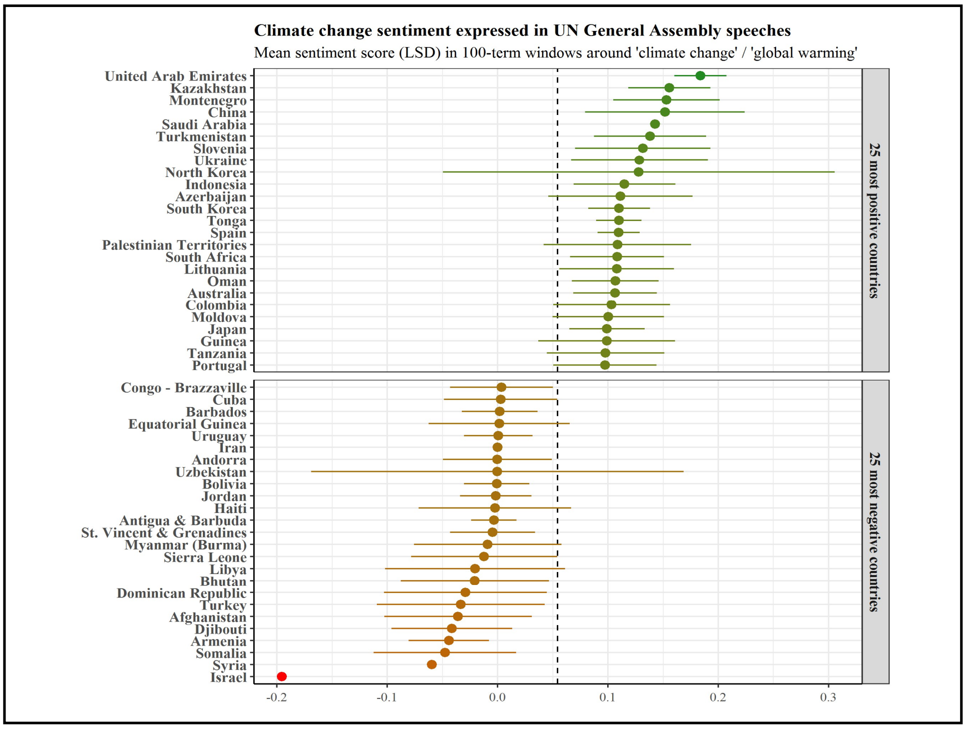
A sentiment analysis applied to our running example

- How positively or negatively do national delegates in the United Nations General assembly speak about climate change?
- ➢ And: Does this meaningfully capture expressed political *positions* on climate change issues?

- Approach
- ➢ Apply the Lexicoder sentiment dictionary with the respective functions in the *quanteda* R package to the corpus created above
- ➢ Normalized sentiment score in 100-term window around climate change references

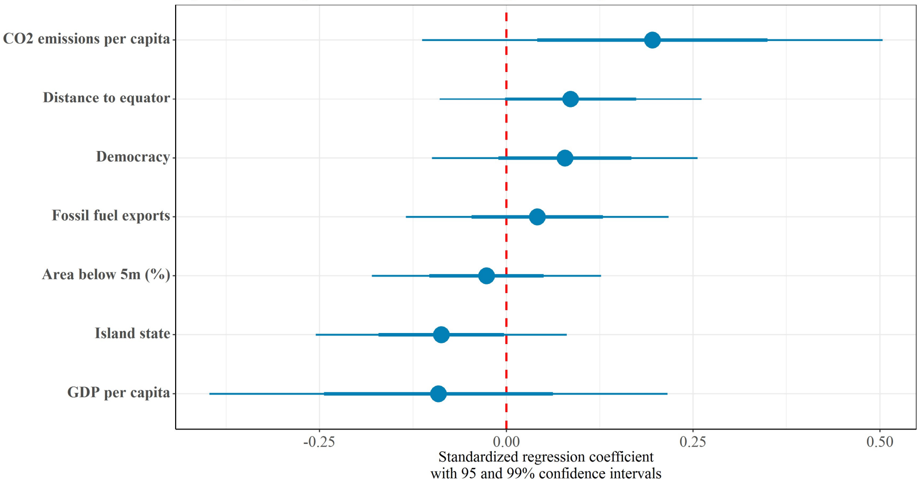
36


WZB
Climate change sentiment expressed in UN General Assembly speeches
Mean sentiment score (LSD) in 100-term windows around 'climate change' / 'global warming' references
Mean sentiment across all full UNGA speeches during period in purple
0.15
0.10
0.05
0.00
-0.05
1990
1995
2000
2005
2010
2015
Year
37


Climate change sentiment expressed in UN General Assembly speeches
Mean sentiment score (LSD) in 100-term windows around 'climate change' / 'global warming'
United Arab Emirates
Kazakhstan
Montenegro
China
Saudi Arabia
Turkmenistan
Slovenia
Ukraine
North Korea
Indonesia
Azerbaijan
South Korea
Tonga
Spain
Palestinian Territories
South Africa
Lithuania
Oman
Australia
Colombia
Moldova
Japan
Guinea
Tanzania
Portugal
Congo - Brazzaville
Cuba
Barbados
Equatorial Guinea
Uruguay
Iran
Andorra
Uzbekistan
Bolivia
Jordan
Haiti
Antigua & Barbuda
St. Vincent & Grenadines
Myanmar (Burma)
Sierra Leone
Libya
Bhutan
Dominican Republic
Turkey
Afghanistan
Djibouti
Armenia
Somalia
Syria
Israel
-0.2
-0.1
0.0
0.1
0.2
0.3
25 most positive countries
25 most negative countries

WZB

What sentiment do states attach to climate change?
Explaining the average sentiment around climate change references by some crude national-level variables



Linear Model, n = 191 countries, Adj. R2: .02.

WZB

Pitfalls of dictionary approaches

- Validity of the derived measures not granted
  - Can the theoretical concepts/objects of interest be captured at term/document level?
  - Do term level scores closely align with the typical word usage in the analysed context?

⇒ Use ‘off-the-shelf‘ term lists developed in other contexts only with extreme caution

⇒ Ideally: Develop your own dictionaries tailored to your research question

⇒ Apply/calculate context-specific baselines

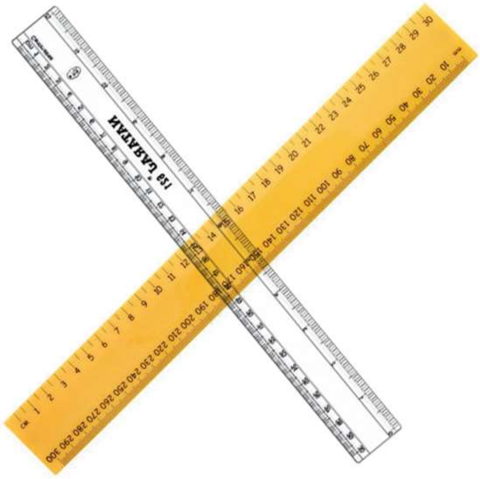
⇒ In any case: Validate your results!
(e.g. against human coders or external data related to your concepts)

⇒ Validated sentiment dictionary for English political language: Young and Soroka (2001, *PC*)

⇒ Validated sentiment dictionary for German political language: Rauh (2018, *JTIP*)

WZB

Part 4
Text scaling



WZB

Automated scaling of texts

- Scaling techniques …
  - … automatically distribute documents across a latent (underlying) scale (dimension)
  - … are used to infer the position of a document's author
  - … were mainly developed in studying the ideological positions that drive party manifestos or political speeches (left-right dimension)
  - … are increasingly applied to other questions such as lobbying success
- Basic idea

  Estimate text positions by focussing on language that discriminates most strongly among the texts (i.e. give strong weight to terms that occur very frequently in some texts but only very infrequently in others)

WZB

Prominent PolSci scaling approaches

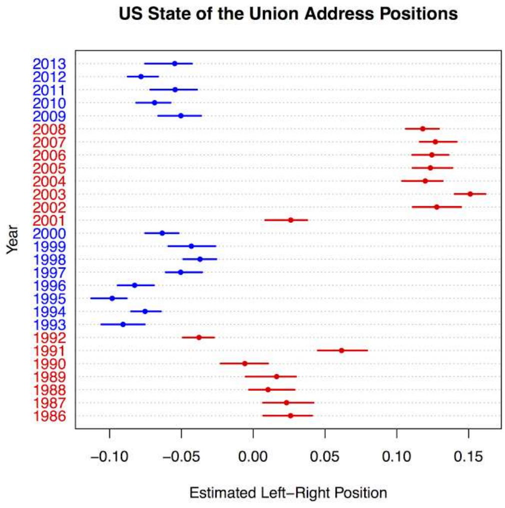
- Unsupervised scaling: *Wordfish* (Slapin and Proksch 2008)
  - Assumes that there is only exactly one dimension structuring the text corpus!
  - Algorithm weights term frequencies so that that there is a maximum distance between the texts in the corpus
  - Rare terms influence the results strongly
  - Resulting positions can only be interpreted relative to each other
  - Content of the scale has to be interpreted ex-post

- Supervised scaling: *Wordscores* (Laver, Benoit and Garry 2003)
  - Researcher supplies reference texts with ‘known’ values across the latent scale
  - Algorithm retrieves and weights the relative term frequencies in these texts
  - Virgin texts are then positioned on the latent dimension along the weights of the terms they contain

43

WZB

Wordfish example



Source: The Monkey Cage / Benjamin Lauderdale

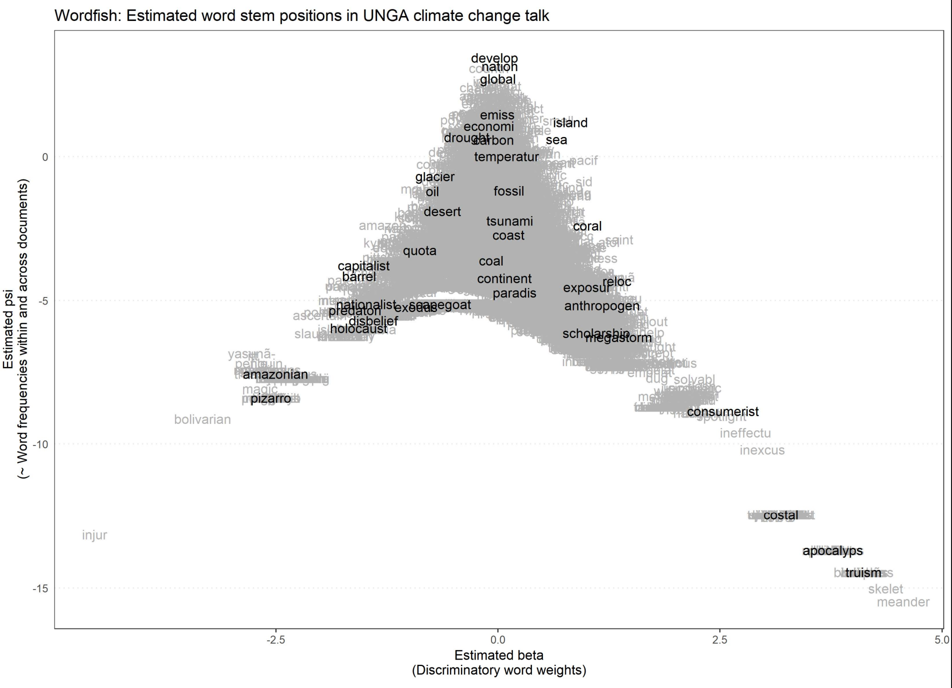
44

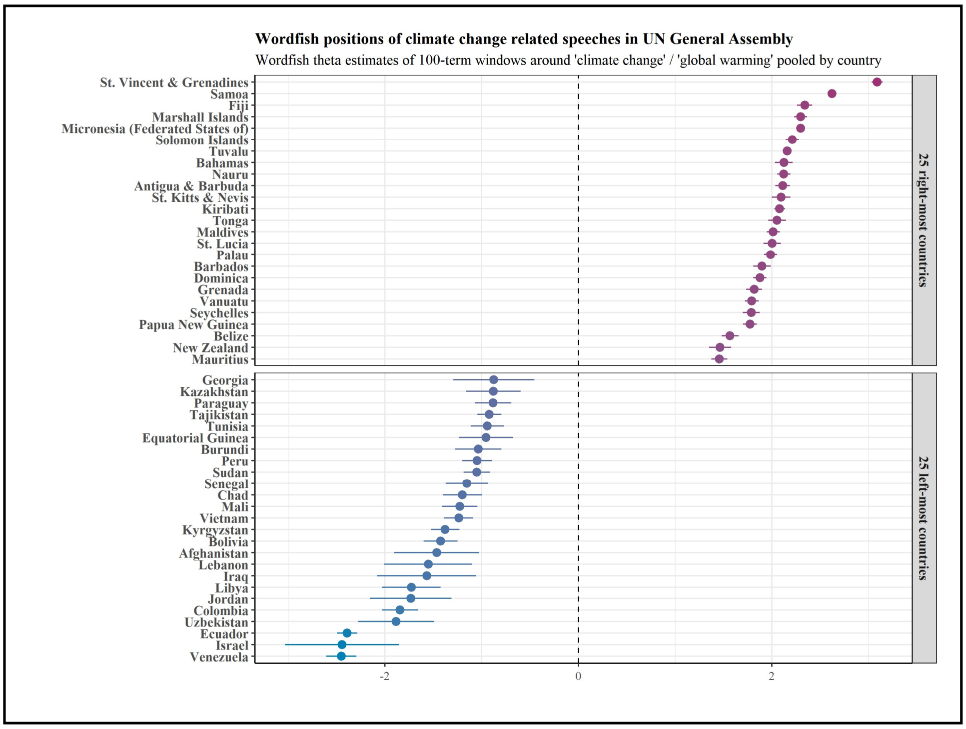
WZB

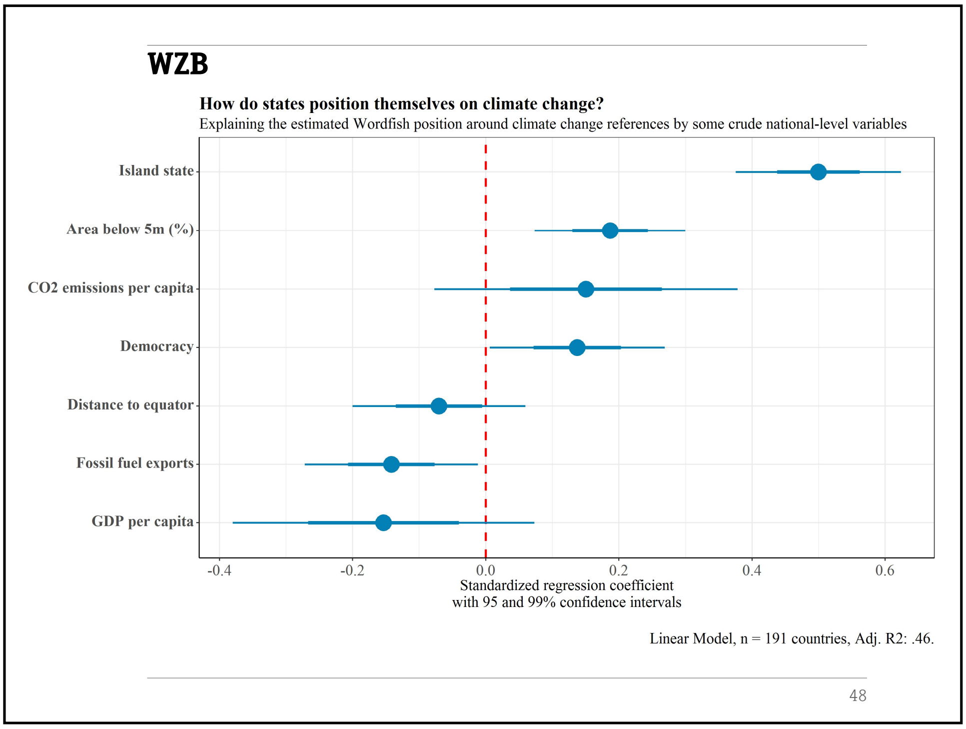
Applying Wordfish to our running example

- What differentiates national delegates in the United Nations General Assembly according to the relative frequency of words they use when speaking about climate change?
- ➢ And: Does this meaningfully capture expressed political *positions* on climate change issues?

- Approach
- ➢ Apply the Wordfish algorithm (as implemented in *quanteda*) to the corpus of 100-term window around climate change references aggregated to country (! pre-processing!)
- ➢ Scrutinize term weights ('betas') and document positions ('thetas')




Wordfish positions of climate change related speeches in UN General Assembly
Wordfish theta estimates of 100-term windows around 'climate change' / 'global warming' pooled by country
St. Vincent & Grenadines
Samoa
Fiji
Marshall Islands
Micronesia (Federated States of)
Solomon Islands
Tuvalu
Bahamas
Nauru
Antigua & Barbuda
St. Kitts & Nevis
Kiribati
Tonga
Maldives
St. Lucia
Palau
Barbados
Dominica
Grenada
Vanuatu
Seychelles
Papua New Guinea
Belize
New Zealand
Mauritius
25 right-most countries
Georgia
Kazakhstan
Paraguay
Tajikistan
Tunisia
Equatorial Guinea
Burundi
Peru
Sudan
Senegal
Chad
Mali
Vietnam
Kyrgyzstan
Bolivia
Afghanistan
Lebanon
Iraq
Libya
Jordan
Colombia
Uzbekistan
Ecuador
Israel
Venezuela
25 left-most countries
-2
0
2


WZB
How do states position themselves on climate change?
Explaining the estimated Wordfish position around climate change references by some crude national-level variables
Island state
Area below 5m (%)
CO2 emissions per capita
Democracy
Distance to equator
Fossil fuel exports
GDP per capita
-0.4
-0.2
0.0
0.2
0.4
0.6
Standardized regression coefficient
with 95 and 99% confidence intervals
Linear Model, n = 191 countries, Adj. R2: .46.
48

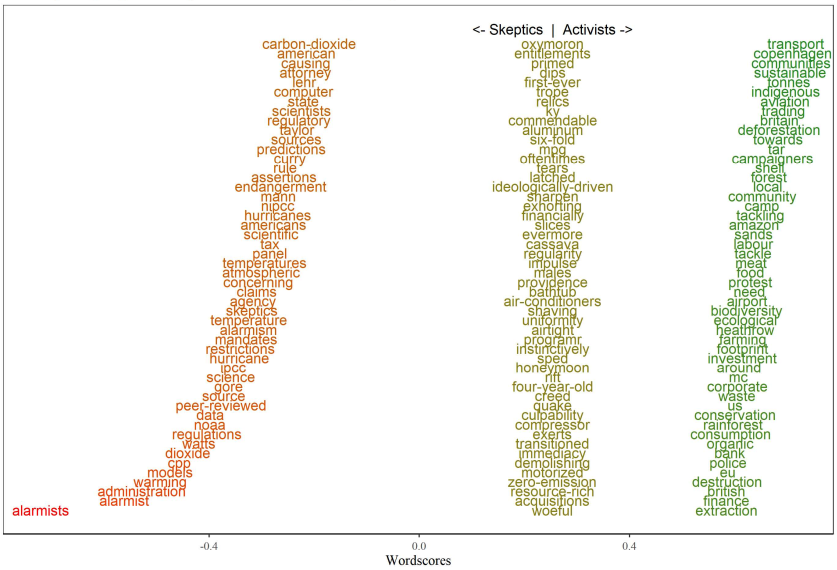
WZB

Applying Wordscores to our running example

- In how far do speeches of national delegates in the UNGA use language of climate sceptics or climate activists?
- ➢ And: Does this meaningfully capture expressed political *positions* on climate change issues?

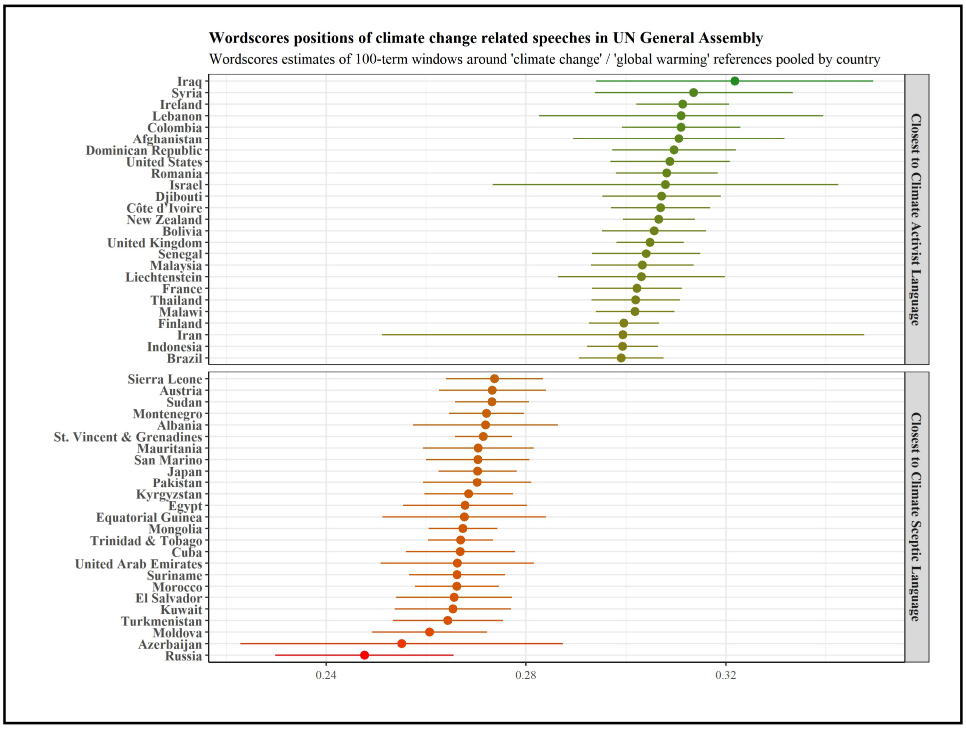
- Approach
  - Corpus of 3000+ reference texts: scrape *climate–change related news* (!) from websites of *The Heartland Institute* (climate change sceptics or deniers; reference score: -1) and *The Ecologist* (climate activists; +1)
  - Train a Wordscores model via *quanteda* on this corpus and analyze the resulting term weights
  - Scale UNGA speeches (pooled by country) along this model and see whether we find something meaningful

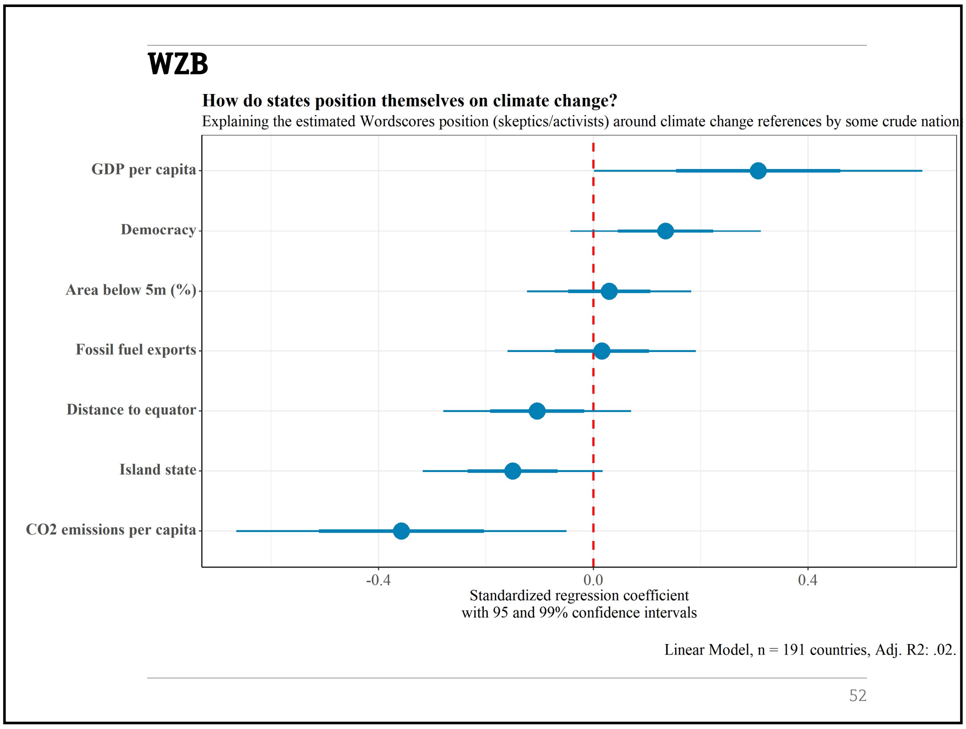
Example of a Wordscores model
Identify language that separates climate change skeptics and activists



Reference texts to train the WS algorithm are climate-change related news from two outlets:
1) 'Heartland Institute' (Skeptics, reference score = -1, n = 1,322)
2) 'The Ecologist' (Activists, reference score = 1, n= 1,851)

Plot shows the terms that are closest to the minimum, mean, and maximum of the estimated score distribution
(61,675 terms scored in total)


Wordscores positions of climate change related speeches in UN General Assembly
Wordscores estimates of 100-term windows around 'climate change' / 'global warming' references pooled by country
Iraq
Syria
Ireland
Lebanon
Colombia
Afghanistan
Dominican Republic
United States
Romania
Israel
Djibouti
Côte d'Ivoire
New Zealand
Bolivia
United Kingdom
Senegal
Malaysia
Liechtenstein
France
Thailand
Malawi
Finland
Iran
Indonesia
Brazil
Closest to Climate Activist Language
Sierra Leone
Austria
Sudan
Montenegro
Albania
St. Vincent & Grenadines
Mauritania
San Marino
Japan
Pakistan
Kyrgyzstan
Egypt
Equatorial Guinea
Mongolia
Trinidad & Tobago
Cuba
United Arab Emirates
Suriname
Morocco
El Salvador
Kuwait
Turkmenistan
Moldova
Azerbaijan
Russia
Closest to Climate Sceptic Language
0.24
0.28
0.32


WZB
How do states position themselves on climate change?
Explaining the estimated Wordscores position (skeptics/activists) around climate change references by some crude nation
GDP per capita
Democracy
Area below 5m (%)
Fossil fuel exports
Distance to equator
Island state
CO2 emissions per capita
-0.4
0.0
0.4
Standardized regression coefficient
with 95 and 99% confidence intervals
Linear Model, n = 191 countries, Adj. R2: .02.
52

WZB

Pitfalls of automated scaling

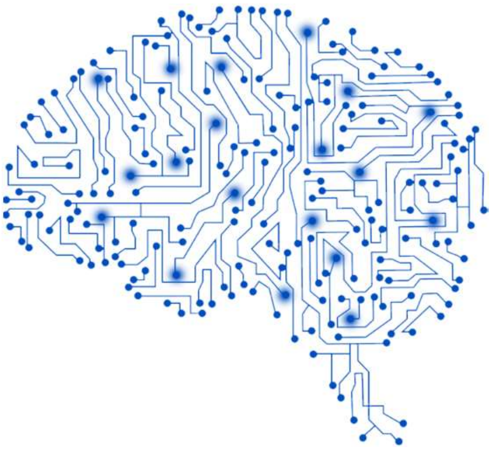
- Scaling works only:
  - with documents that are very focussed on the theorized dimension (cf. party manifestos vs. newspaper articles)
  - if documents come from the same context in which the language is used identically (political speeches vs. news outlets?)

⇒ Scaling procedures make strong assumptions!

⇒ Scaling procedures require particularly careful validation!

WZB

Part 5
Machine learning and topic models (briefly)



WZB

Supervised machine learning – intuition

- Basic idea of supervised classification
  Algorithm 'learns' from (a few) human-coded documents before it automatically classifies (many) 'virgin' texts

- Achieved along four (iterative) steps:

1. Construct a training and a test set from your documents
   - Human coders apply a coding scheme to two subsets of docs (-> session 2)
   - Size depends on doc length, unique language, number of categories etc. but usually a small fraction of the overall corpus is enough
2. 'Learn' classifier function from the training set
   - Training documents used to find a statistical function that best predicts the human-coded categories along the document-term frequencies
   - Different algorithms come with different assumptions
   - The *RTextTools* package implements different algorithms, e.g.

WZB

Supervised machine learning – intuition

3. Validate the classifier in the test set
   - Use the classifier function from the training set to predict the categories of documents in the test set
   - Does you classifier live up to the '*gold standard*' of human coding?
   - If precision is insufficient, go back to step 1: Either your coding scheme has to be re-worked, or the training set has to be expanded
4. Classify the 'virgin' texts
   - If precision is satisfying, you can in a reliable and valid manner classify all remaining documents of so-far unknown categories

⇒ Validation part of the method!
⇒ Required size of human-coded sets decreases with lesser categories, more discriminatory language and longer documents
⇒ 'Representative' samples of training and test documents needed

WZB

Unsupervised learning

- General idea ....

  Algorithm 'learns' both categories and categorization from the distribution of characteristics in the supplied data

- ... applied to text analysis
  - Which words tend to co-occur? Which clusters can be optimized? How can documents be distributed over clusters in a statistically optimal way?
  - Researcher does not supply any theoretical categories *a priori* (only abortion criteria, e.g. number of clusters, in some approaches)
  - Results can only be interpreted *ex post*

- Validation
  - Assessing semantic validity requires much contextual knowledge!
  - Models not generalizable beyond the data and parameters supplied!

57

WZB

A prominent unsupervised approach: *Topic models* (e.g. Blei 2012)

- Typical application
  Identification and distribution of abstract 'topics in large amounts of 'documents' without prior knowledge/assumptions on these topics

- Assumptions
- Topics are defined by frequency distributions of co-occurring terms
- Text were generated by firstly choosing topic composition (possibly several per text) and only secondly by a respective choice of words

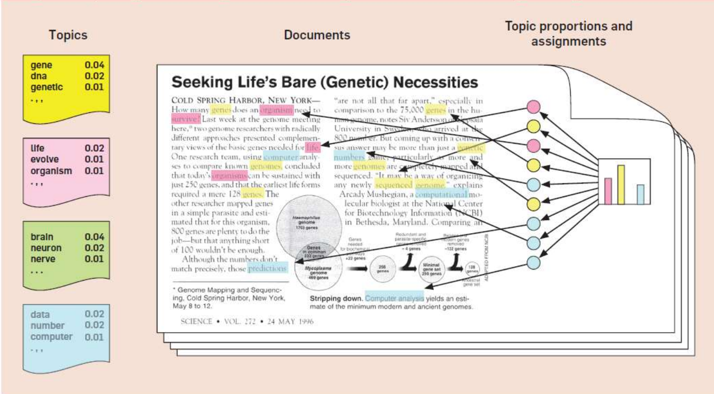
- Estimation
- Algorithm reverse-engineers this assumed text creation process by asking: Which latent topic distribution would explain the observed word frequency distribution best?
- Cluster-analysis on term level, probabilistic distribution of documents

58

WZB

Th logic of topic models presented by their inventor (Blei 2012)

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

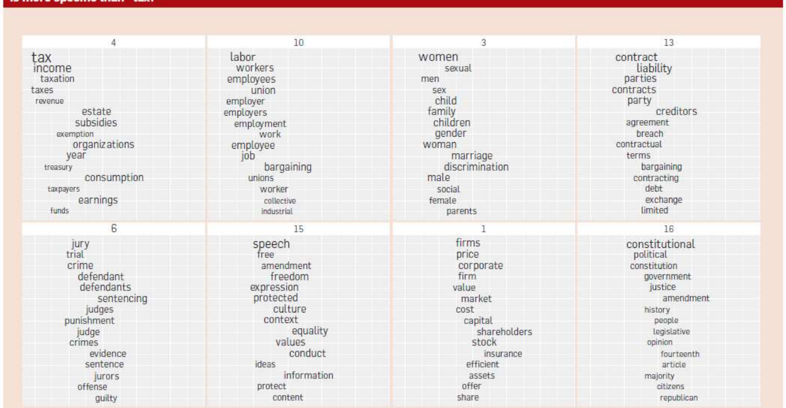


59

WZB

Typical output of topic models (Blei 2012)

Figure 3. A topic model fit to the *Yale Law Journal*. Here, there are 20 topics (the top eight are plotted). Each topic is illustrated with its top-most frequent words. Each word's position along the x-axis denotes its specificity to the documents. For example "estate" in the first topic is more specific than "tax."



60

WZB

Pitfalls of topic models

- Caution!
  - High interpretation demand after the analysis!
  - Results are not very robust and strongly depend on the specific data set and model parameters (seed and number of topics)
  - What exactly is a topic (cf. “frame”, “issue”, “narrative”; “event”)?

⇒ Useful tool to explore very large collections and to narrow down more targeted samples

⇒ Systematic analysis and especially comparisons across topics only with greatest caution (robustness and model fit of topic models is a current research frontier).

WZB

Outlook and conclusions

WZB

What we could not speak about…

- Wort vector models
  Representation of words in high-dimensional spaces allows analysing proximity of concepts and evolution of narratives over time …

- Text similarity / plagiarism measures …
  Analysing changes of word order e.g. highly useful to study consecutive drafts of policies, treaties, etc (e.g. Rauh, 2018) …

- Part-of-speech tagging and grammatical parsing
  Retaining grammatical structure (contrast to bags of words) allows more targeted study of subject-object relations (e.g. predicting conflict intensity from newswires, Schrodt 2011)…

WZB

Promises and pitfalls of automated content analyses

\+ A more complete and reliable analysis of social phenomena
  - Analysis of very large document sets achievable at low cost
  - Reduced / removed sampling bias

+/- Human resources remain significant
  - Dictionary development, coding of reference texts and especially validation requires intense human engagement

\- Context dependency more pronounced
  - Quantitative representations of language cannot abstract from varying contexts (human coders can)

\- Reliability is partially traded against validity
  - Power of automated analyses declines quickly with the complexity of theoretical concepts

WZB

Conclusions

- Automated text analyses are a powerful, yet not a definitive tool for content analysis in the Political and Social Sciences
- The computer allows us to digest larger amounts of information, uncovers patterns on much more aggregated levels, but interpretation, contextualisation, and validation remain key responsibility of the researcher!

Thank you for your attention!

Slides and tutorials available at www.christian-rauh.eu/teaching