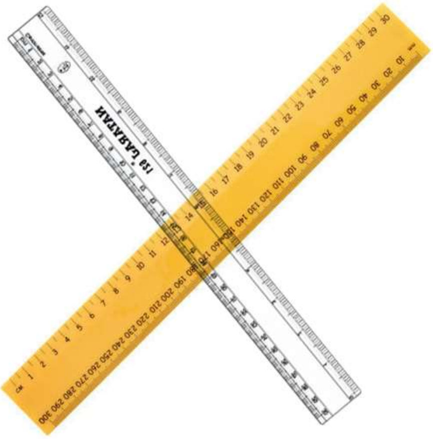
WZB

Text-mining international politics

Session 5
Text scaling approaches



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, e.g.

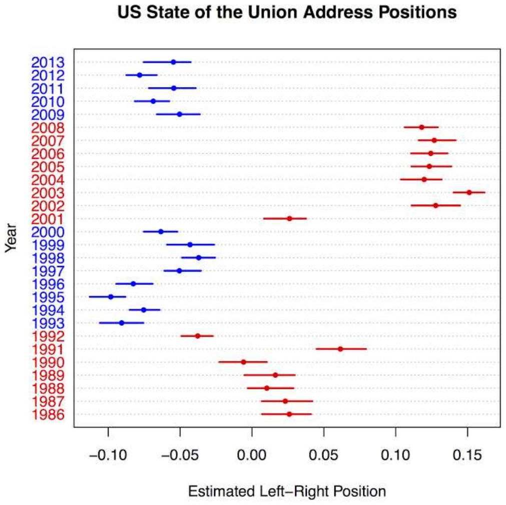
- Basic idea

  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)

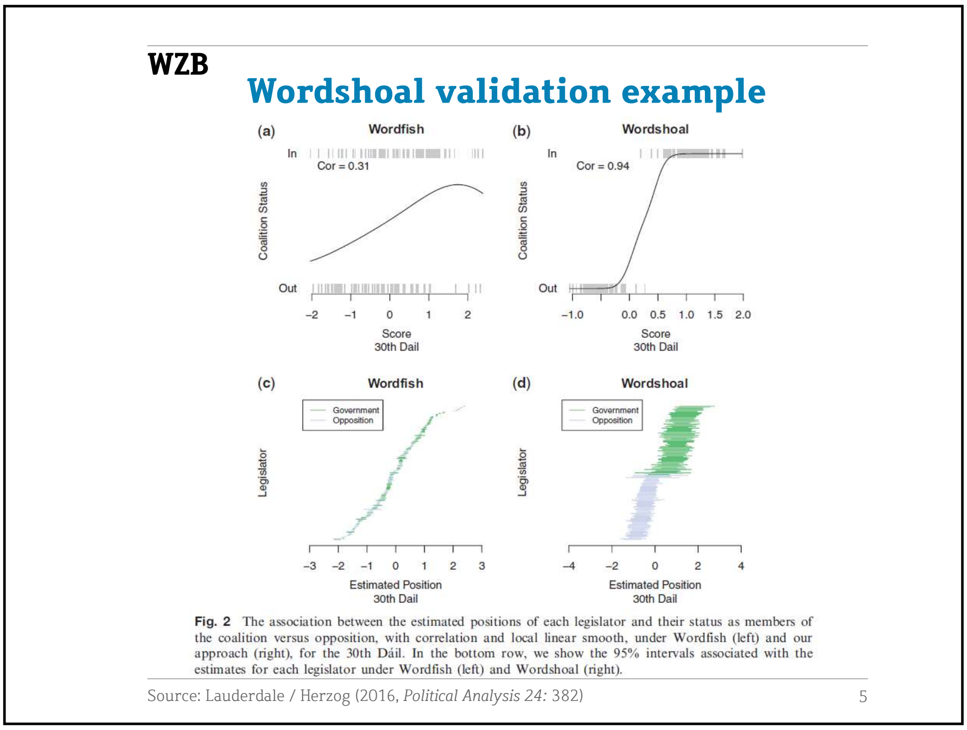
Unsupervised scaling

- *Wordfish* (Slapin and Proksch 2008)
  - Assumes that there is only exactly one latent 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
- *Wordshoal* (Lauderdale and Herzog 2016)
  - Two-stage approach:
    1. Scale variation in word usage with Wordfish for specific 'debates'
    2. Use factor analysis to construct a common scale across debates
  - Relaxes assumption of uni-dimensionality
    Discriminating power of individual words may vary across debates
  - Geared more towards scaling latent *actor* rather than *text* positions
  - Still unsupervised, interpretation only *ex–post*

Wordfish example



Source: The Monkey Cage / Benjamin Lauderdale

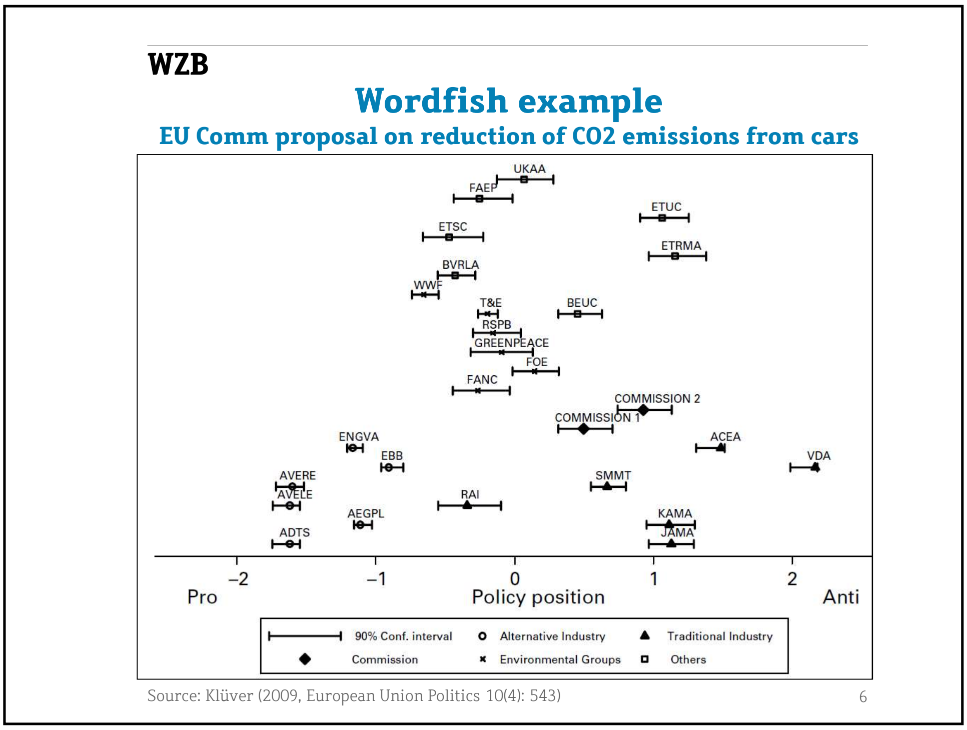


WZB

Wordshoal validation example

Fig. 2 The association between the estimated positions of each legislator and their status as members of the coalition versus opposition, with correlation and local linear smooth, under Wordfish (left) and our approach (right), for the 30th Dáil. In the bottom row, we show the 95% intervals associated with the estimates for each legislator under Wordfish (left) and Wordshoal (right).

Source: Lauderdale / Herzog (2016, *Political Analysis 24:* 382)

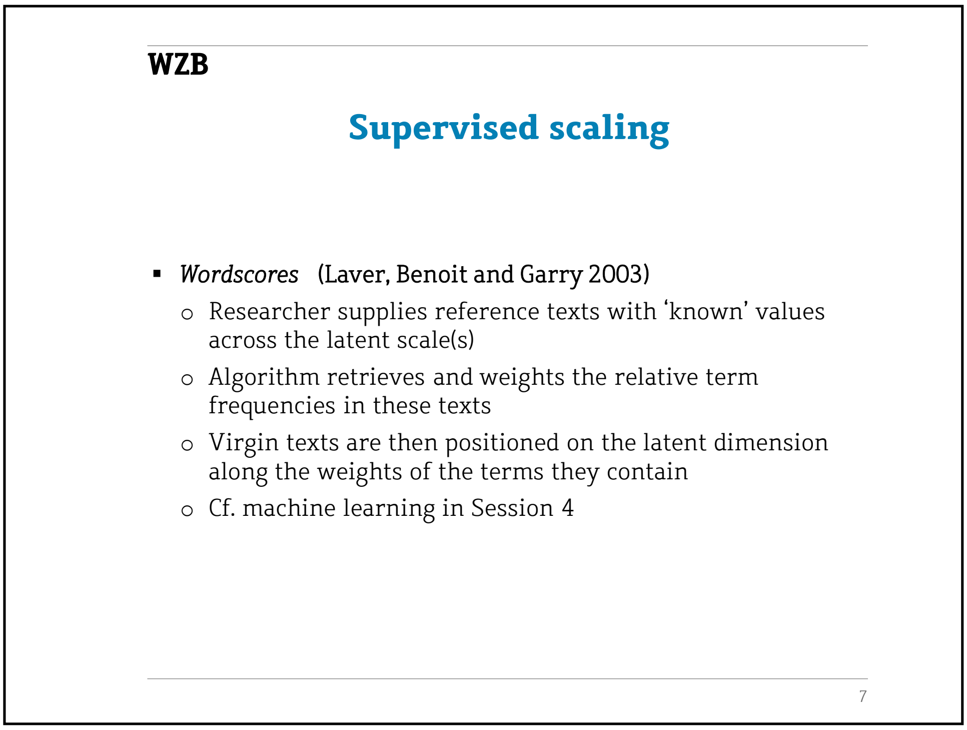


WZB

Wordfish example

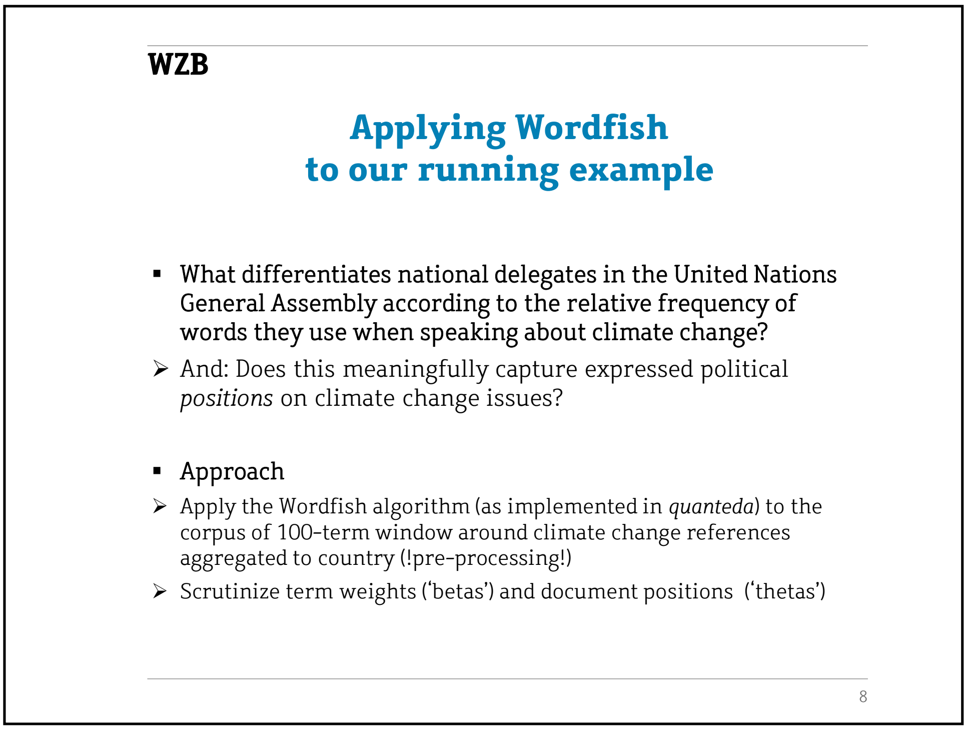
EU Comm proposal on reduction of CO2 emissions from cars

Source: Klüver (2009, European Union Politics 10(4): 543)



Supervised scaling

- *Wordscores* (Laver, Benoit and Garry 2003)
  - Researcher supplies reference texts with 'known' values across the latent scale(s)
  - 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
  - Cf. machine learning in Session 4



Applying Wordfish to our running example

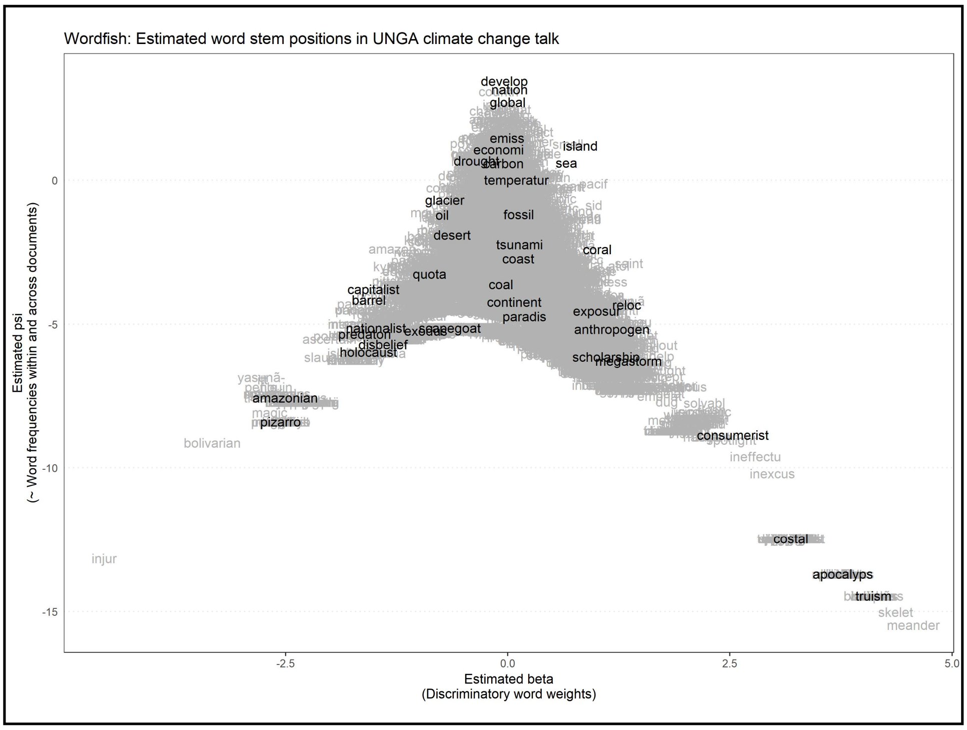
- What differentiates national delegates in the United Nations General Assembly according to the relative frequency of words they use when speaking about climate change?

➢ And: Does this meaningfully capture expressed political *positions* on climate change issues?

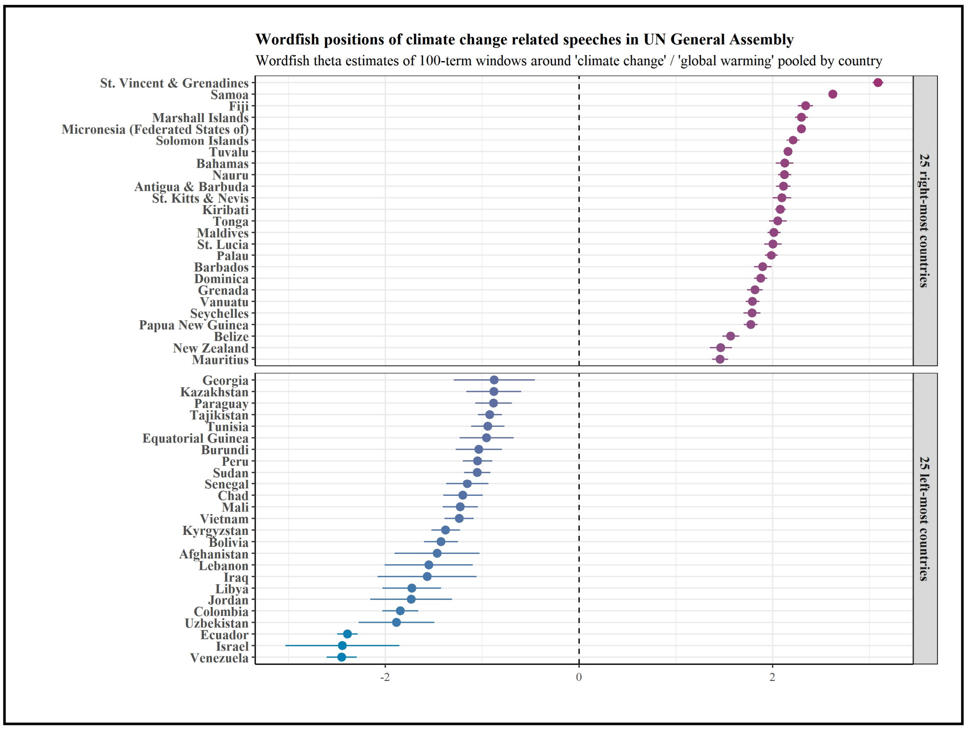
- Approach

➢ Apply the Wordfish algorithm (as implemented in *quanteda*) to the corpus of 100-term window around climate change references aggregated to country (!pre-processing!)

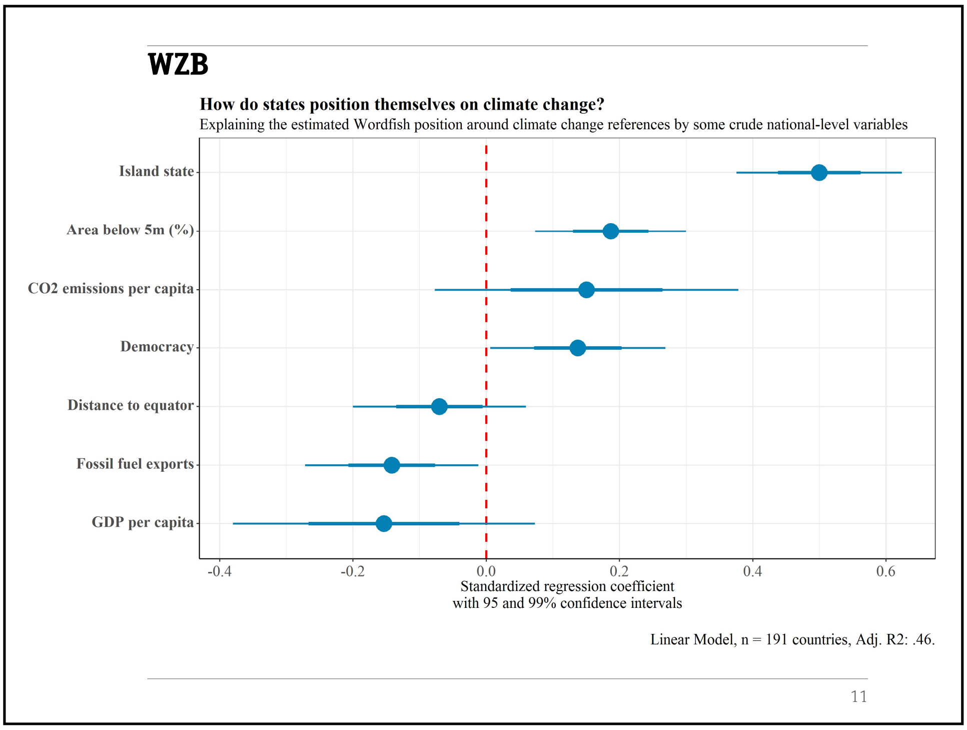
➢ Scrutinize term weights ('betas') and document positions ('thetas')


Wordfish: Estimated word stem positions in UNGA climate change talk
Estimated psi
(~ Word frequencies within and across documents)
Estimated beta
(Discriminatory word weights)
0
-5
-10
-15
-2.5
0.0
2.5
5.0
develop
nation
global
emiss
economi
drought
carbon
island
sea
temperatur
glacier
oil
fossil
desert
tsunami
coast
coral
quota
capitalist
barrel
coal
continent
paradis
exposur
reloc
anthropogen
nationalist
predaton
excess
scapegoat
disbelief
holocaust
scholarship
megastorm
amazonian
pizarro
consumerist
bolivarian
ineffectu
inexcus
injur
costal
apocalyps
truism
skelet
meander

9


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

10

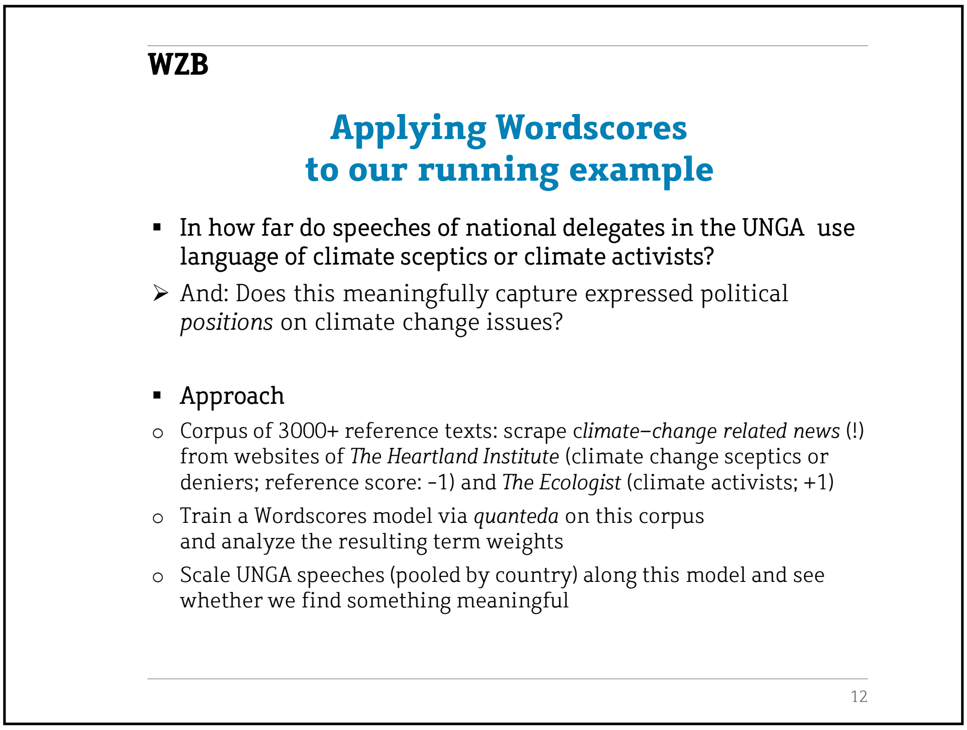


WZB

How do states position themselves on climate change?

Explaining the estimated Wordfish position around climate change references by some crude national-level variables

Linear Model, n = 191 countries, Adj. R2: .46.



WZB

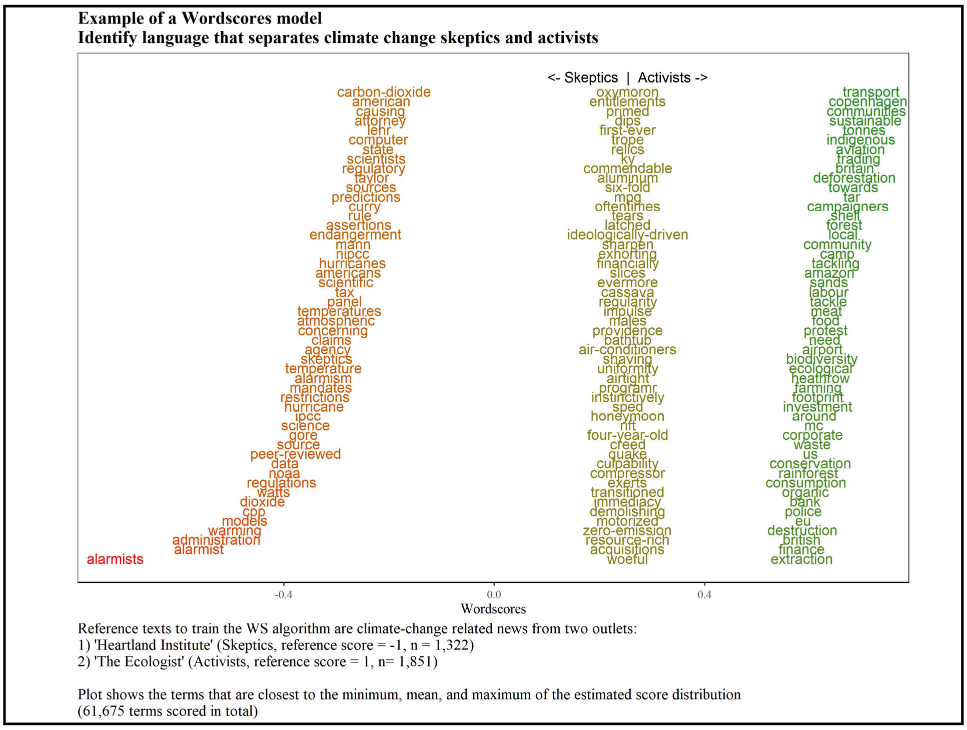
Applying Wordscores to our running example

- 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

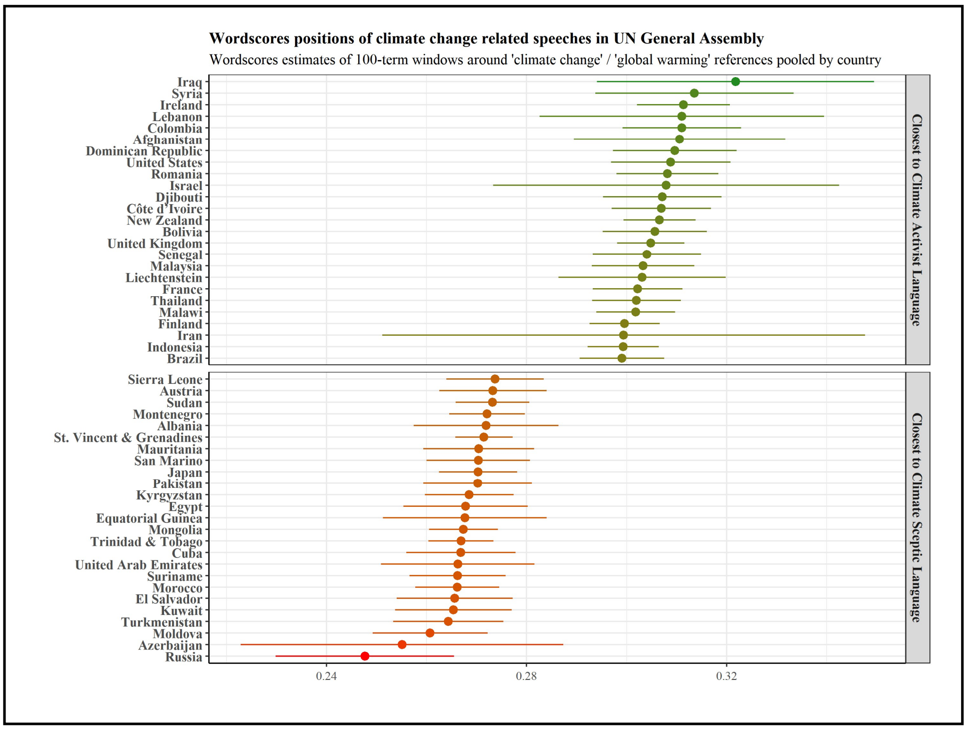


Example of a Wordscores model
Identify language that separates climate change skeptics and activists

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)

13



Wordscores positions of climate change related speeches in UN General Assembly

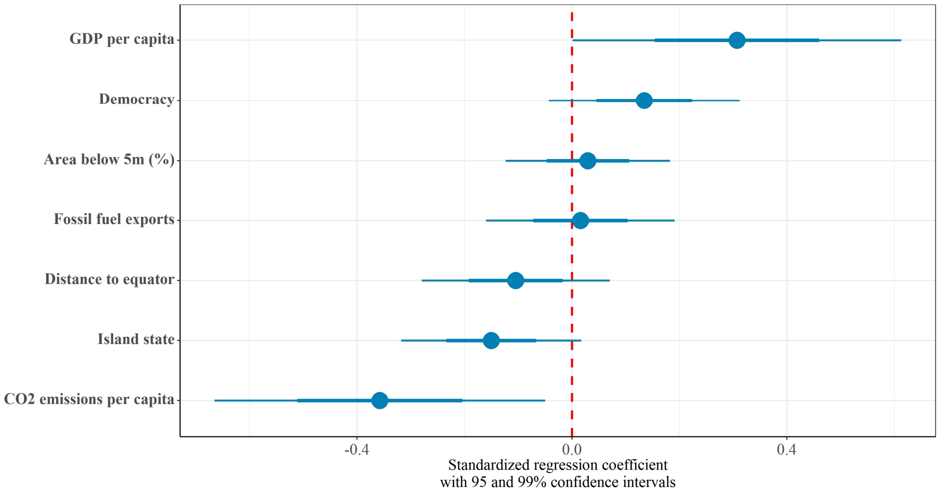
Wordscores estimates of 100-term windows around 'climate change' / 'global warming' references pooled by country

14

WZB

How do states position themselves on climate change?

Explaining the estimated Wordscores position (skeptics/activists) around climate change references by some crude nation



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

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!