WZB

Text-mining international politics

Session 2
Corpus construction and discovery

1

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

2

WZB

Readily available corpora – Examples I

- UNGD corpus
  - Baturo, Dasandi, Mikhaylov (2017, *Research and Politics*)
  - Full texts of all *8,093 speeches* by national representatives in the *United Nations General assembly* between *1970* and *2018* (includes auto-translated versions to English)
  - https://doi.org/10.7910/DVN/0TJX8Y
- EUspeech
  - Schumacher, Schoonvelde, Traber, Dahiya, de Vries (2016)
  - *17,184 speeches* of *national, EU, and IMF executives* during the *2007–2015* period (includes auto-translated versions)
  - https://doi.org/10.7910/DVN/GKABNU

3

3

WZB

Readily available corpora – Examples II

- ParlSpeech
  - Rauh, De Wilde, Schwalbach (2017, Harvard Dataverse)
  - 3.9 million plenary speeches in key parliamentary chambers of seven European states (including CZ/PSP 1993-2016)
  - www.bit.ly/ParlSpeech
- Manifesto corpus
  - Full text of partisan election manifestos from 50 countries
  - https://manifesto-project.wzb.eu/information/documents/corpus
- US SOTU corpus (1790–2018)
- Annual IO reports, UNSC resolutions, EU legislation …. or: *scrape/build your own!*

4

4

WZB

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

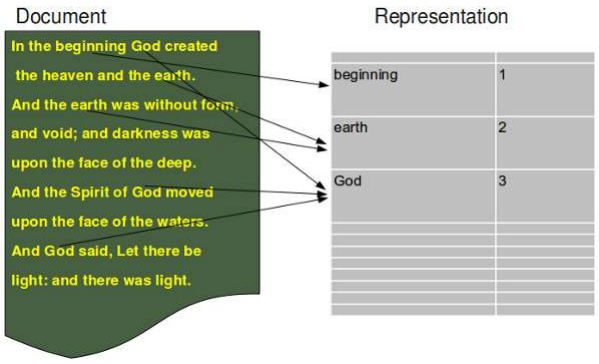
- Remove...
  ... document "boilerplate" (info not part of the analysed message)
  ... punctuation, capitalization, numbers
  ... very common and very uncommon terms ("stop words", <1% of docs)
- Lemmatization / Stemming
  - Words referring to the same concept mapped to a single root
  - {economy, economic , economically} → economi
- Turn documents into "*bags of words*"
  - Discards the order in which words occur!
  - Unigrams, bigrams ... n-grams

⇒ Document frequency matrix

5

WZB

"Bags of words" (illustration w/out stemming)



Source: python-course.eu

6

WZB

Document frequency matrix (illustration)

Label	Title
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

7

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

8

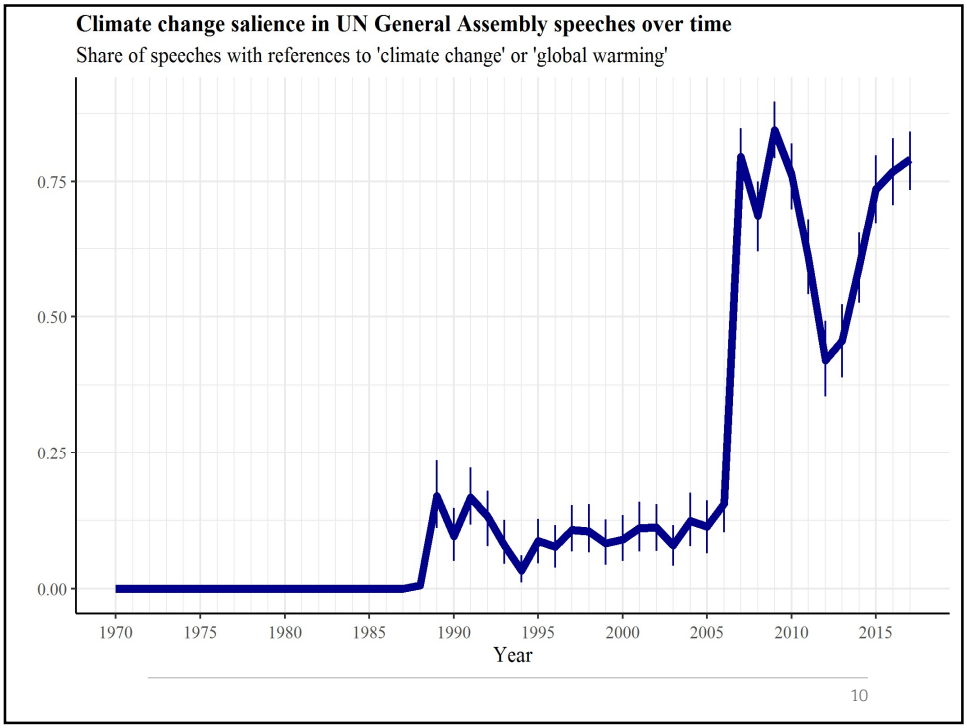
WZB

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...

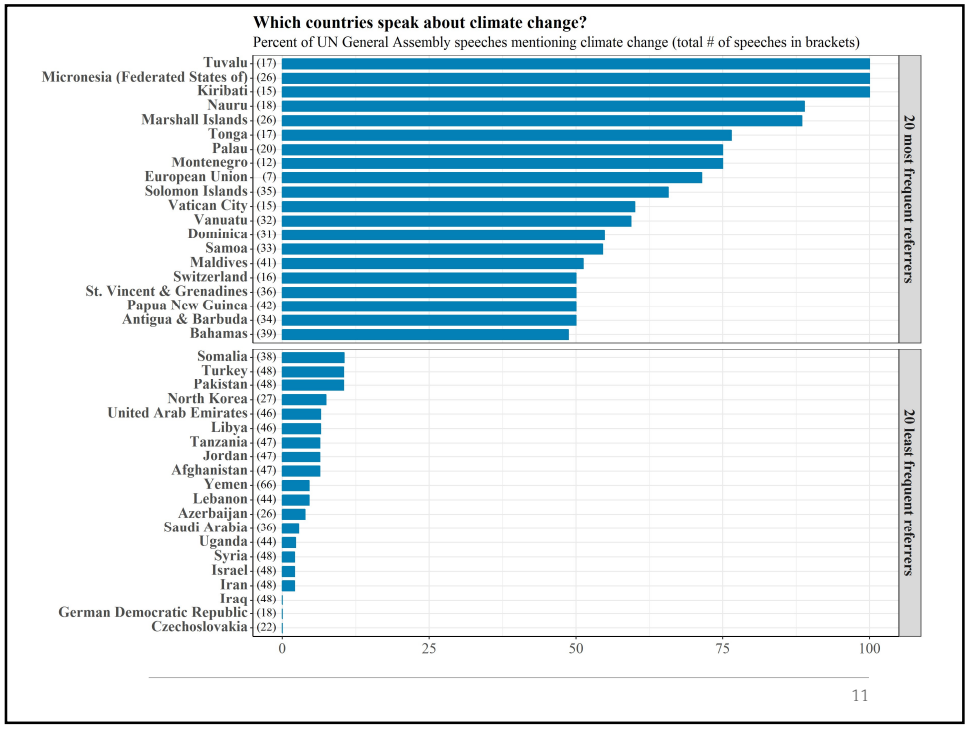
9

9

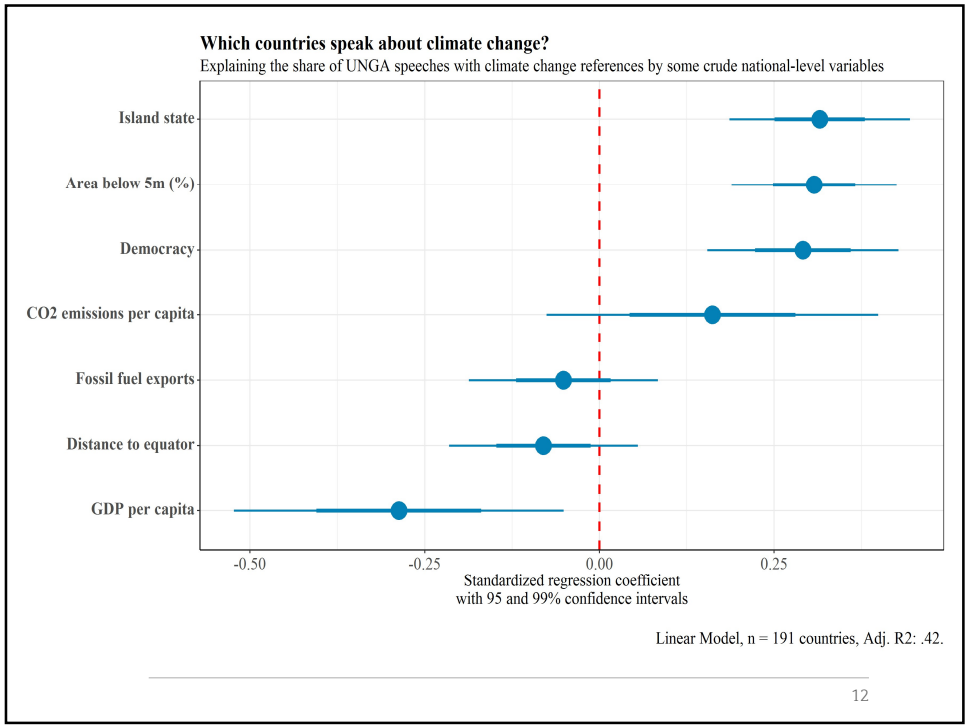


10

10


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
11

11


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.
12

12

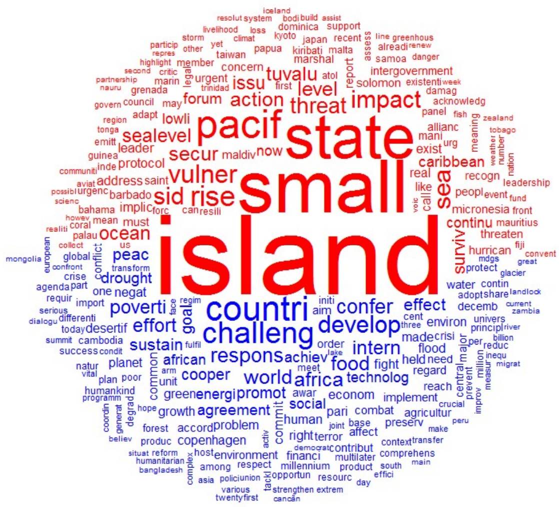
13

After 2006

Until 2006

14

Island states



All others

15