WZB

Berlin Social Science Center

# Text-mining international politics

## Session 1
## Qualitative, quantitative, and automated content analysis

Charles University Prague
Institute of Political Studies
May 7 2019

CHRISTIAN RAUH
www.christian-rauh.eu

1

WZB

## Welcome!

- This block seminar aims to:
  - Enable you to assess studies using text analysis (or: text mining, automated content analysis, corpus analytics, ...)
  - Enable informed methodological and pragmatic choices for conducting text analyses on your own
  - Generate awareness for the promises and the pitfalls of analyzing human language with algorithms
  - Encourage you to play!
- Your instructor
  - Dr. Christian Rauh, senior researcher WZB Berlin Social Science Center
  - Causes and consequences of the public politicization of the EU and other international organizations
  - www.christian-rauh.eu

➢ *Who are you and what are you interested in?*

2

2

WZB

## The plan in detail

- Contrast *promises and pitfalls* of automated text analysis with more general challenges of *social science content analysis*
- *Intuition, pragmatic challenges and exemplary EU/IR applications*
  - Corpus construction and discovery
  - Dictionary-based analyses
  - Machine learning and topic models
  - Text scaling procedures
  - Briefly: other linguistic indicators and advanced NLP approaches
- Running example and *tutorials* implemented in R
  Climate change in United Nations General Assembly speeches
- Please bear *your* research interests in mind, apply the discussed ideas to them, and interrupt me whenever something is unclear!
- Research (design) paper
  Discuss how a freely chosen research question (international politics, EU, or related fields) could be tackled by the approaches we discuss here



3

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?



4

WZB

## Content analysis in between …

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

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



5

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



6

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



7

WZB

## The general methodological challenge

– Analysing content as a measurement problem

$m = t + e$

| Measurement | Truth | Error |
|---|---|---|
| of relevant variables in appropriate units (observation) | The 'true' value of a relevant variable | Sum of *random* and *systematic* errors during observation |

– Scientific measurement

- *m* should equal *t*
- Minimize random and especially systematic error
- Maximize reliability and validity



8

WZB

## Challenge 1: Unitizing

- What is to be observed? How are observations recorded? What are relevant data points?

→unit of analysis

- Different (sub-) units relevant in content analysis

*a) Sampling units*

*b) Recording / Coding units*

*c) Context units*

- Distinguishing units
  From "physical" to "thematic"; reliability and validity



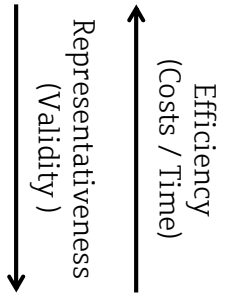


WZB

## Challenge 2: Sampling

- Population, representativeness and sampling bias

- Sampling strategies
  - ~~Convenience sample~~
  - ~~Snowball sampling~~ -
  - Cluster or stratified sampling
  - Random sample
  - Relevance sampling
  - Census




- Sample size and the 'split-half' technique

WZB

## Example: Differing coding and context units

Fictitious speech

Expressed evaluations towards China (pos/neut/neg) ?

"China is a powerful neighbour.
Clearly, the Chinese political system hardly lives up to our ideals about popular democracy.
But China has proven to be a reliable and trustworthy trading partner.
We are concerned with its military strength, though bilateral talks have improved significantly.
Nevertheless, the China's decision as of yesterday is absolutely unacceptable to us."





WZB

## Unitizing and Sampling: Exercises

– Example research questions:

a) *Which policies are transferred across national public administrations in the European Union?*
b) *How does Ethiopia justify human rights violations?*
c) *How present are populist strategies in Czech discourse?*
d) *Your question?*

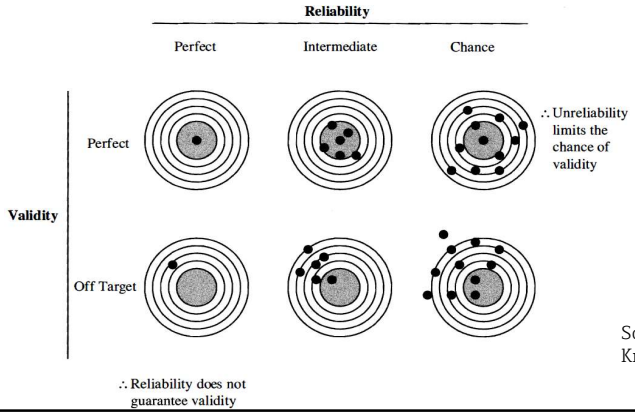
– Define:

- The population/universe of texts
- Sampling units and a sampling strategy
- Suitable coding units
- Suitable context units

WZB

## Challenge 3: Reliability

- What is reliability and why does it matter?
  - Confidence in data/interpretations → confidence in conclusions
  - Positivist and interpretative notions
  - Reliability and validity are related, but not equal




Source:
Krippendorff 2004: 214

13

WZB

## Reliability tests in human coding...

- Always require duplication
- Come in differing levels of strength:

**Table 11.1** Typese of Reliability

| Reliability | Designs | Causes of Disagreements | Strength |
|---|---|---|---|
| **Stability** | test-retest | intraobserver inconsistencies | weakest |
| **Reproducibility** | test-test | intraobserver inconsistencies + interobserver disagreements | medium |
| **Accuracy** | test-standard | intraobserver inconsistencies, + interobserver disagreements, + deviations from a standard | strongest |

Source: Krippendorff 2004: 215

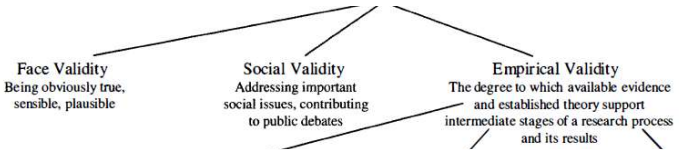
- Require clear *coding schemes* and coder criteria !
- Must take agreement by chance into account
  (→ Krippendorff's alpha)

14

14

WZB

## Challenge 4: Validity




15

WZB

## Why automate?

- Political texts increasingly available in *digitized formats*
  (in IR research often the only continuously available data source!)
- 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!

16

16

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*!**