Universität Münster

Text as Data

Session 4: Basics of Text Analysis

Mirko Wegemann

Universität Münster
Institut für Politikwissenschaft

13 May 2026

Logistics

- Due to cancelled session last week, subsequent sessions will shift by one week
- Final week (22 July) will be our wrap up session now
- Please plan accordingly for your assignments

Regular Expressions

In our R crash course, part of the group focused on R as programming language, the other part on how to implement regular expressions. Jonah and Jonas will briefly introduce you to what we can do with regular expressions in `R`.

Content analysis

> *"One can count the characters, words, or sentences of a text. One can categorize its expressions, analyze its metaphors, describe the logical structure of its compositions, and ascertain its associations, connotations, denotations, and commands. One can also offer psychiatric, sociological, political, or poetic interpretations of that text."*

– Krippendorff (2018, p. 28)

Components of Communication

"**Who** says **what** to **whom** with **what effect**?" (Gerbner 1956)
Accordingly, communication consists of four central components.

1. Sender(s)
2. Content
3. Receiver(s)
4. Effect

Sender

Figure: Individual vs group

Receivers

Figure: To the converted vs. to the skepticists

Example of Content: Framing

When we analyze political communication, we often speak of **frames**.

- A frame emphasizes a specific aspect of a topic; it is a "subset of potentially relevant considerations" (Druckman and K. R. Nelson 2003, p. 730)
- Through targeted framing, political actors attempt to influence how a particular topic should be understood and discussed (T. E. Nelson and Kinder 1996)

Effect

Gerbner (1956) distinguishes between:

1. Effectiveness of a message: is the intended goal of the message achieved?
2. Consequence: effects that go beyond the direct goal

Task I

Split yourselves into two groups. Half of you (those German speaking) will listen to Merz talk in a school, the other half watches his presence in the Oval Office [10 minutes]

- What is the setting of the scene?
- Which frames does Merz set?
- How's his body language?
- Do you notice anything else in the speech?

Merz in School

Figure: Merz in School

Merz in Oval Office

Figure: Merz in Oval Office (watch 23:30-24:30 and 55:00-56:03)

Merz in School vs Merz in Oval Office

	School	Oval Office
Setting		
Frames		
Body Language		
Other		

Meaning and Text

- Text is context-dependent
  - Readers interpret text differently
  - Content is also dependent on the authors of the text
  - Analysts approach text in different ways

How do we need to structure content analysis then? What are its ingredients?

Essential Requirements for Content Analysis

- A priori formulation of a falsifiable **research question**

Essential Requirements for Content Analysis

- A priori formulation of a falsifiable **research question**
- Clarification of the **context** in which the text was created

Essential Requirements for Content Analysis

- A priori formulation of a falsifiable **research question**
- Clarification of the **context** in which the text was created
- **Analytical model** in which causal relationships are adapted to the context of the text

Essential Requirements for Content Analysis

- A priori formulation of a falsifiable **research question**
- Clarification of the **context** in which the text was created
- **Analytical model** in which causal relationships are adapted to the context of the text
- **Validation** (ideally with 'out-of-domain' evidence → other data)

Why Quantitative? I

The biggest problem of qualitative text analysis:

Why Quantitative? II

> *[...] scholars have struggled when using texts to make inferences about politics. The primary problem is volume: there are simply too many political texts. Rarely are scholars able to manually read all the texts in even moderately sized corpora.*

– Grimmer and Stewart (2013)

3 Steps to Successful Text Analysis

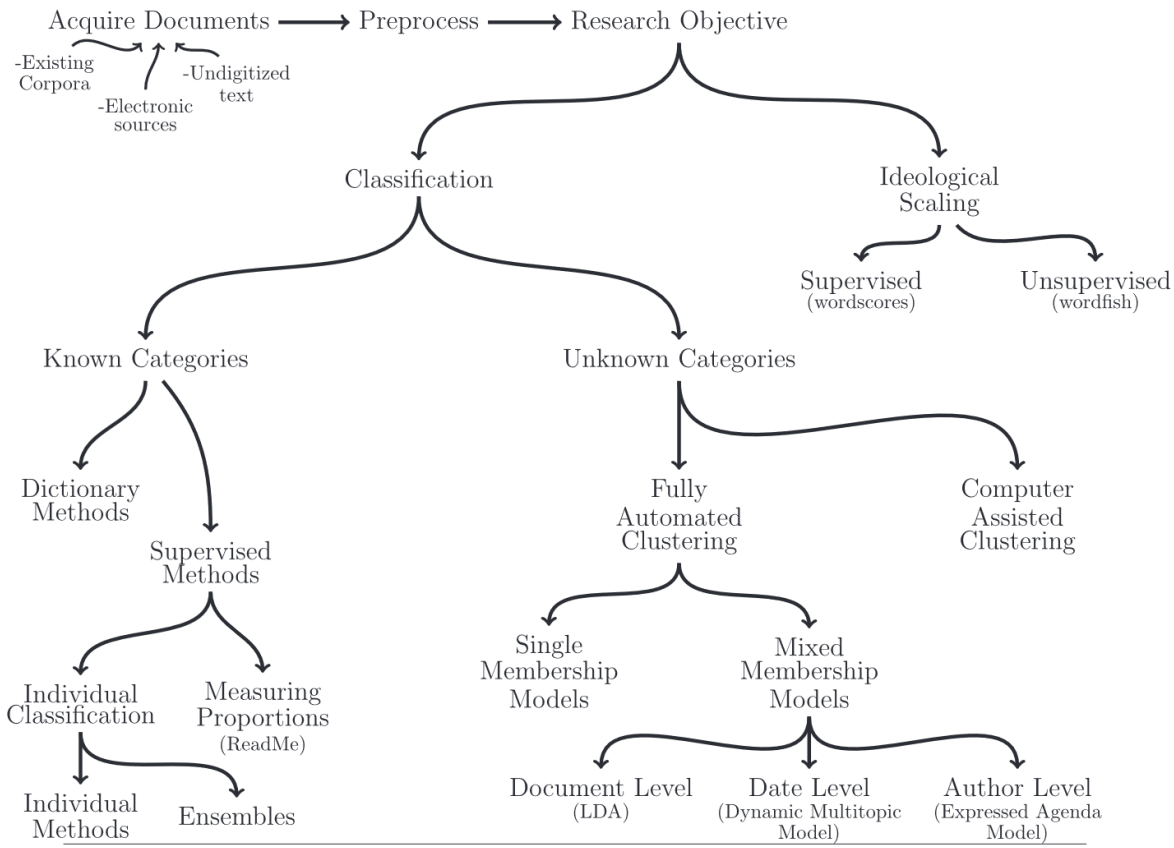


Figure: The process of text analysis according to Grimmer and Stewart (2013, p. 268)

Step 1: Data Collection I

Two options:

- Use of available data
- Acquiring new data (e.g., via scraping techniques)

What could be potential problems in the first step?

Step 1: Data Collection II

- primarily lack of data availability
  - text files not digitized
  - lack of access rights
  - *selection bias* (text only available from certain actors)

Step 2: Text Preparation I

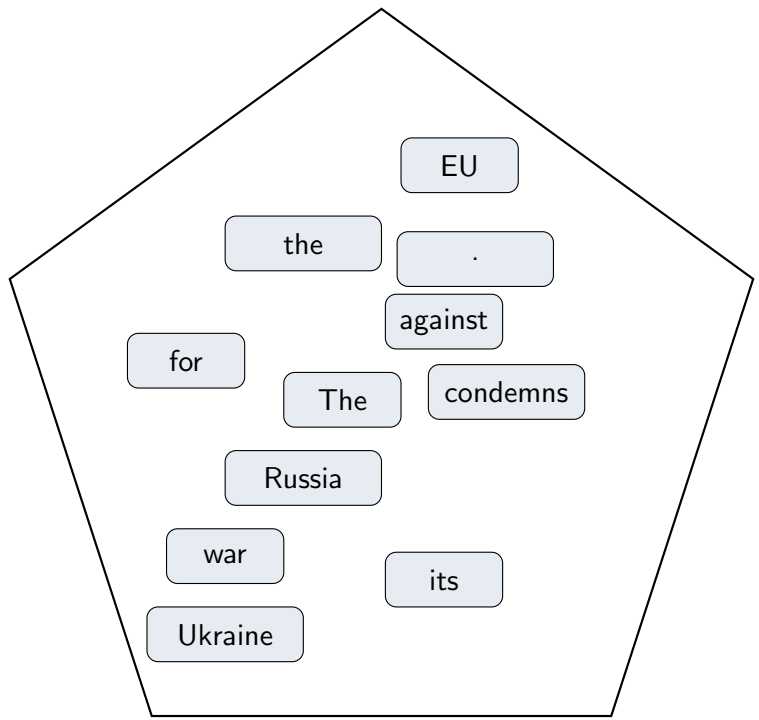
Text is complex and not usable for our quantitative models. Accordingly, we must first transform it.

- Transformation into a representation of numbers
- Simplifying the data structure (e.g., through stemming, lemmatization)
- Exclusion of words (such as stopwords)

Step 2: Text Preparation II

Bags-of-words (BoW) creates a vector for each document D of a corpus containing the tokens t.

The EU condemns Russia for its war against Ukraine.



Step 2: Text Preparation III

The general idea behind this is that we can understand the **meaning** of a text through the **vocabulary** used. A comparison between documents d_1 and d_2 takes place solely on the basis of the **frequency of tokens**.

Step 2: Text Preparation IV

Document D1	*The child makes the dog happy* the: 2, dog: 1, makes: 1, child: 1, happy: 1
Document D2	*The dog makes the child happy* the: 2, child: 1, makes: 1, dog: 1, happy: 1

↓

	child	dog	happy	makes	the	BoW Vector representations
D1	1	1	1	1	2	**[1,1,1,1,2]**
D2	1	1	1	1	2	**[1,1,1,1,2]**

Two documents with different meanings, yet same BoW representation
(Source: AIML.com Research)

Step 2: Text Preparation V

What could be a problem with this structure?

Step 2: Text Preparation VI

Problems with *bags-of-words*:

- We lose information about the order of words
- The grammatical structure of a sentence is neglected.
- Words are not contextualized. Each word can only have one meaning.

Step 2: Text Preparation VII

Embeddings as a way out:

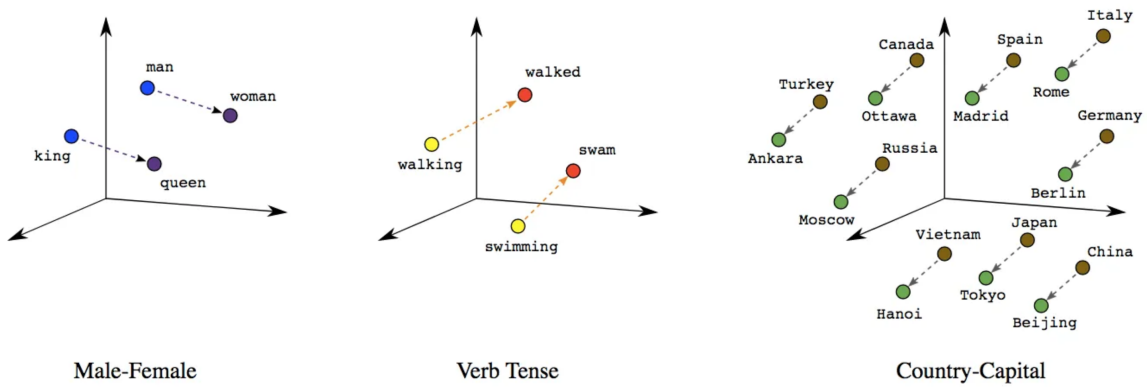


Figure: Source: Towards Data Science

Step 3: The Analysis I

There are various models in quantitative text analysis, which can be classified as follows (cf. Baden et al. 2020):

- **rule-based** analyses: lists of rules or features
  - Dictionary analyses: tone/sentiment of a text, categorization
  - Dependency parsers: sentence structures
- **supervised** analyses: provision of a training dataset, on the basis of which patterns of a text are learned and can be applied to non-classified texts
  - Naive Bayes
  - Support Vector Machine
  - Random Forest
  - etc.

Step 3: The Analysis II

- **unsupervised** analyses: based on statistical procedures in which a function is optimized and texts are classified
  - Topic Models
  - Scaling Models
- **hybrid** approaches: several analysis methods are used
  - e.g., 1.) unsupervised models for the categorization of data
  - these serve as input for the classification, which is refined by 2.) supervised models

Step 3: The Analysis III

What do you think are the biggest difficulties in quantitative text analysis?

Step 3: The Analysis IV

- Not enough data
- Language barriers
- Lack of computing power
- Replicability
- Validation

Context in Quantitative Text Analyses I

Krippendorff (2018) emphasizes especially the **context** from which a text originates.

> *What do you think: how can we include the context of a text when we analyze it quantitatively? Is this even possible? What could be potential difficulties?*

Context in Quantitative Text Analyses II

- **Contextual knowledge about senders** of a message can be captured via metadata
  - Context of a speech (e.g., date, location)
  - Characteristics of the sender (e.g., gender, party affiliation, role)
- **Contextual knowledge about receivers** is often difficult to capture
  - possibly some data in social media channels, letters to the editor, etc.

Context in Quantitative Text Analyses III

- Context about **analysts**
  - Ideally, a text is analyzed by several researchers, for whom potentially relevant meta-information is available

The Four Principles of Quantitative Text Analysis

No.	Principle
1	All quantitative models of language are wrong – but some are useful.
2	Quantitative methods for text amplify resources and augment humans.
3	There is no globally best method for automated text analysis.
4	Validate, Validate, Validate.

Table: Principles of QTA according to Grimmer and Stewart (2013, p. 269)

Validity and Reliability

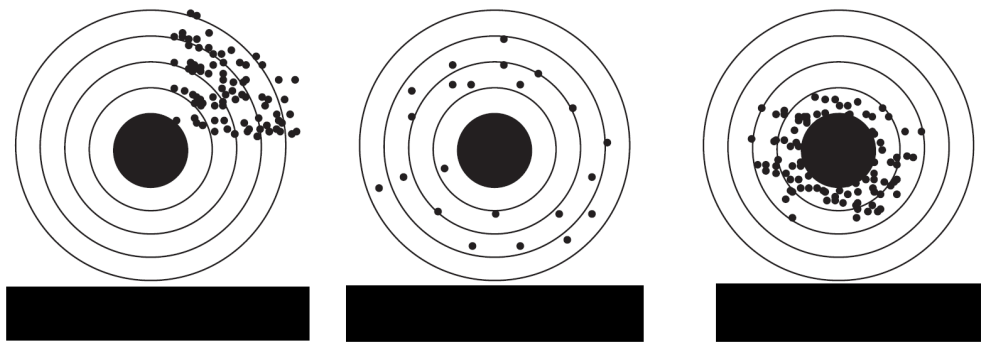


Figure 4.2 Reliability (precision) and validity

1

Which measurements are valid? Which are reliable?

[1] Gerring 2012, p. 82

Validity and Reliability

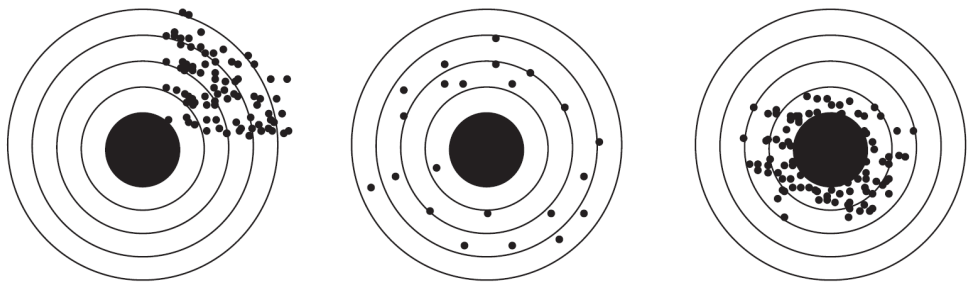


Figure 4.2 Reliability (precision) and validity

2

[2] Gerring 2012, p. 82

Quality Criteria of Content Analyses (and other research designs)

- Can research goals be answered through the research design?
- Validity: Can we derive the correct conclusions with the research design?
  - **internal validity**: high validity of results within our sample
  - **external validity**: results can be extended to cases outside our analysis
- Reliability: Can the results be replicated under the same conditions?

Validation Steps in Quantitative Text Analysis I

Type of validation depends on the chosen design:

- **supervised analyses** are based on statistical prediction models
  - various metrics available: how well can the model predict our already classified data
  - Examples: Confusion Matrix, Accuracy, F1-Score, etc.
- **rule-based** and **unsupervised** approaches are harder to validate

Outlook I

- Next week: Data acquisition with web-scraping techniques
- Background reading:
  Munzert (2015) $\rightarrow$ Chapter 1
- Think about websites whose content you'd like to download!
- Install SelectorGadget for your browser (for Firefox and Safari: bookmark; for Chrome: add-on)

Literature I

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020).Hybrid Content Analysis: Toward a Strategy for the Theory-driven, Computer-assisted Classification of Large Text Corpora. *Communication Methods and Measures*. Retrieved June 2, 2024, from https://www.tandfonline.com/doi/abs/10.1080/19312458.2020.1803247

Druckman, J. N., & Nelson, K. R. (2003).Framing and Deliberation: How Citizens' Conversations Limit Elite Influence. *American Journal of Political Science*, *47*(4), 729–745. https://doi.org/10.1111/1540-5907.00051

Gerbner, G. (1956).Toward a General Model of Communication. *Educational Technology Research and Development*, *4*(3), 171–199. https://doi.org/10.1007/BF02717110

Literature II

Gerring, J. (2012). *Social science methodology: A unified framework* (2nd ed). Cambridge University Press. OCLC: ocn775022701.

Grimmer, J., & Stewart, B. M. (2013).Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Krippendorff, K. (2018). *Content Analysis: An Introduction to its Methodology* (Fourth Edition). SAGE.

Munzert, S. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (1st ed.). Wiley.

Literature III

Nelson, T. E., & Kinder, D. R. (1996).Issue Frames and Group-Centrism in American Public Opinion. *The Journal of Politics*, *58*(4), 1055–1078. https://doi.org/10.2307/2960149

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010).How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.