Text as Data

Session 1: Introduction

Mirko Wegemann

University of Münster Department of Political Science

April 15, 2026

About Me

Mirko Wegemann (he/him)

- Since October 2024: Research Associate at the Chair of Comparative Political Science with Prof. Daniel Bischof
- PhD at European University Institute in Florence
- Research interests
  - political parties
  - political communication
  - gender and political representation
  - social norms

Seminar Goals

- Crash course in the statistical software R and the graphical interface RStudio
- Introduction to basic concepts and methods of quantitative text analysis and their application in R
- Discussion of empirical research articles
- Development of research ideas you could study with text analyses

Our Semester Plan I

Week 1	Introduction
Week 2	No session (problem set in R at home)
Week 3	Basics in R
Week 4	Basics of Text Analysis
Week 5	Automated Data Collection I: Web Scraping
Week 6	Automated Data Collection II: Dynamic Websites and APIs
Week 7	No session (Whitsun Holidays)

Our Semester Plan II

Week 8	Data Preparation
Week 9	Topic Modeling
Week 10	No session (EPSS Conference)
Week 11	Supervised Text Analysis
Week 12	Embeddings
Week 13	LLMs and Classification
Week 14	Wrap-Up and Presentations of Own Project Ideas
Week 15	No session (Exam week)

This is a method seminar with relatively little reading **but** we do have some empirical applications of methods in later sessions.

General Session Structure

1. (Presentation of method application by students)
2. Questions about the previous session
3. Input
4. Application in R
5. (Discussion of applied texts)

Course Organization

- Communication and literature via Learnweb (checking emails is expected); password: **tada_26**
  - Literature search via ULB Münster, Google Scholar or Web of Science
  - Access to most articles via the University of Münster Wi-Fi, from home via VPN
  - For an overview of the university library: Library audio tour
- Slides on my website

Requirements I

Workload

One ECTS credit corresponds to a maximum of 30 hours of 60 minutes of actual workload on the part of the student.[1] Example calculation:

$$\rightarrow\ 30 \times 4 = 120\ \textbf{hours}$$

Attendance of the course:

$$1.5 \times 12 = 18; \text{hours}$$

Preparation, follow-up, and development of assignments

$$120 - 18 = 102; \text{hours}$$

[1] Recht.NRW

Requirements II

- Regular attendance and participation
- Prior engagement with the session literature
- Application of a method using own research data
- Development of a research project

Requirements III

Assignments

- *Studienleistung*: Application and presentation of a method
- *Prüfungsleistung*: Final paper

Studienleistung

Application of a Method

In week 5, 6, 9, 11, 12 and 13 we learn how to collect and analyse data. All students should apply the learned method for **one** of these sessions and present it in a short input during the following session.

- Define the research problem
- Implement the method in R and comment on it
- Short presentation (max. 5-10 minutes) about experiences during implementation

Prüfungsleistung

Final Paper

- 4,000 words ± 10 percent (excl. table of contents, bibliography, and title page), to be submitted by September 30, 2026 at the latest
- Self-chosen topic that is empirically analyzed using at least one learned method
- You can choose between two formats: (1) newspaper report and (2) policy brief
- Develop a socially relevant research question, briefly recap existing work on it and then propose and implement your method to develop an answer to the question
- Pay attention to correct citation style (guidelines for example here)
- Submission of written paper (PDF format) and commented script (in .R or .RMD format)

Plagiarism I

"In terms of examination law, plagiarism constitutes an intentional attempt at deception.

The offence of plagiarism is fulfilled when, in written coursework and examinations, 'texts by third parties are adopted wholly or partially, verbatim or nearly verbatim, and presented as one's own scholarly work'. This practice 'not only contradicts good scholarly practice, it is also a form of intellectual theft and thus a violation of copyright'."

IfPol: Resolution of the German Association of University Professors

Plagiarism II

Types of Plagiarism

- Complete plagiarism
- Quotation without citation
- Translation plagiarism
- Self-plagiarism
- Ghostwriting

Examination submissions are checked for plagiarism.

Inclusion

We maintain an open, inclusive culture in this seminar.

- Gender-inclusive language is used in seminar communication.
  - Studies on why it is useful to move beyond the generic masculine: Stahlberg and Sczesny (2001), Tavits and Pérez (2019), and Vervecken et al. (2013)
  - Guidelines and further information from the University of Münster
- There is no place here for racism, sexism, homophobia, or transphobia.
- We learn together: substantive questions are legitimate and welcome; knowledge gaps are no reason to feel ashamed.

Contact

- After the seminar or during office hours (by prior email appointment)
- Email: m.wegemann@uni-muenster.de
- Address Department of Political Science Chair: Comparative Political Science Room 223, Scharnhorststr. 100 48151 Münster
- Feedback via LamaPoll or by email

Questions?

And You?

A short survey:
https://www.menti.com/ala94oksbrfp

What is Text Analysis? I

Do you have an idea?

What is Text Analysis? II

"Computational text analysis (also called Quantitative Text Analysis, Automated Content Analysis, Text Mining, Text as Data etc.) draws on techniques developed in natural language processing and machine learning to analyse textual documents." (Chun Ting-Ho 2021)

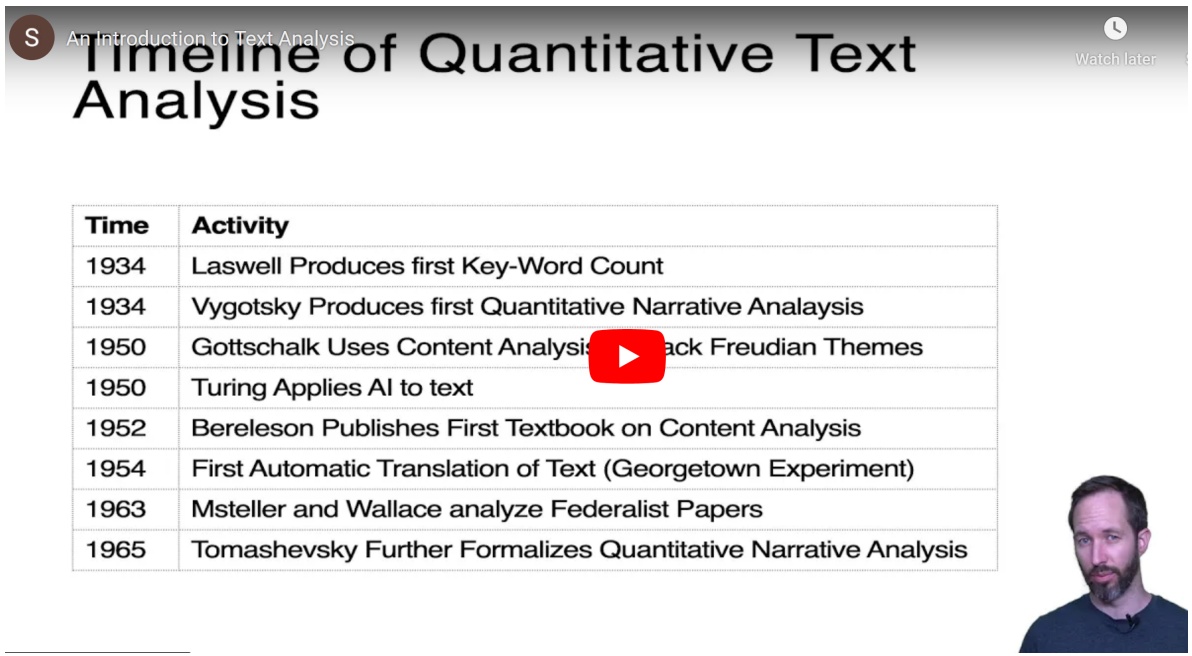
What is Text Analysis? III

It is...

- a form of content analysis in which we convert text or its components into numerical vectors in order to discover relationships and regularities within the text (Benoit 2009)
- QTA can be used both in an exploratory and an explanatory fashion, but it is never atheoretical (Bonikowski and Nelson 2022)

Is This Really New? I

No, quantitative text analysis has been around for a while.



Time	Activity
1934	Laswell Produces first Key-Word Count
1934	Vygotsky Produces first Quantitative Narrative Analaysis
1950	Gottschalk Uses Content Analysis ack Freudian Themes
1950	Turing Applies AI to text
1952	Bereleson Publishes First Textbook on Content Analysis
1954	First Automatic Translation of Text (Georgetown Experiment)
1963	Msteller and Wallace analyze Federalist Papers
1965	Tomashevsky Further Formalizes Quantitative Narrative Analysis

Figure: History of quantitative text analysis

Is This Really New? II

Approaches based on counting words in texts (keyword count) have existed for a long time.

- Breakthrough with the General Inquirer (Stone et al. 1962)
- Idea: categorization into 164 different categories using a dictionary

→ this is the focus of many methods that we will learn about

Is This Really New? III

Can machines think?

> *'I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning.' (Turing 1950, p. 442)*

Is This Really New? IV

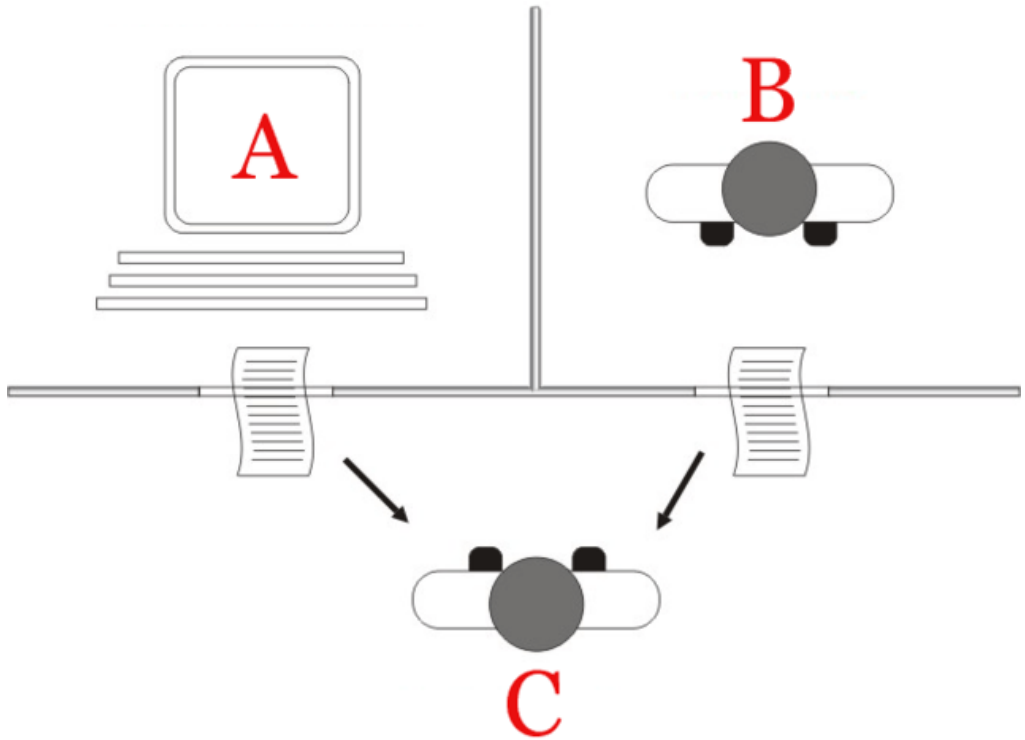


Figure: The Imitation Game

Is This Really New? V

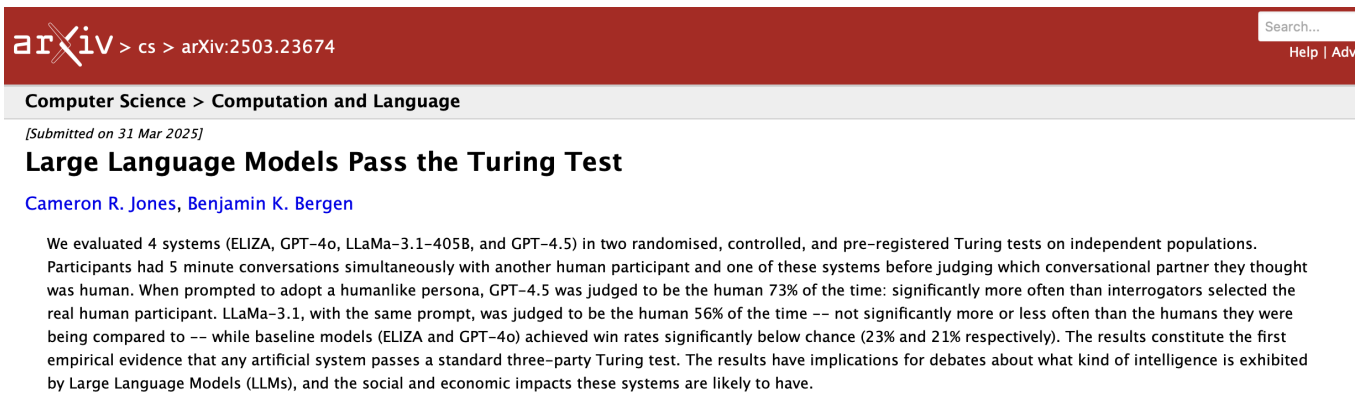
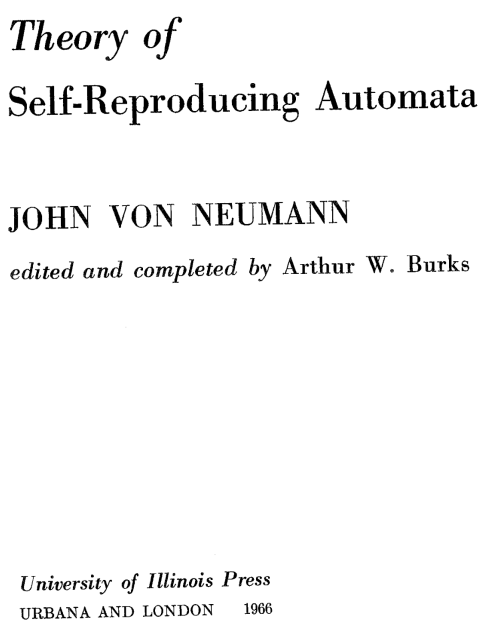


Figure: Have machines taken over? No, according to this article in Nature

Is This Really New? VI


Theory of
Self-Reproducing Automata

JOHN VON NEUMANN

edited and completed by Arthur W. Burks

University of Illinois Press
URBANA AND LONDON 1966

Figure: Neumann's idea of self-replicating machines

Evolution via mutation (similar to DNA), architecture of an encoder and decoder

Is This Really New? VII

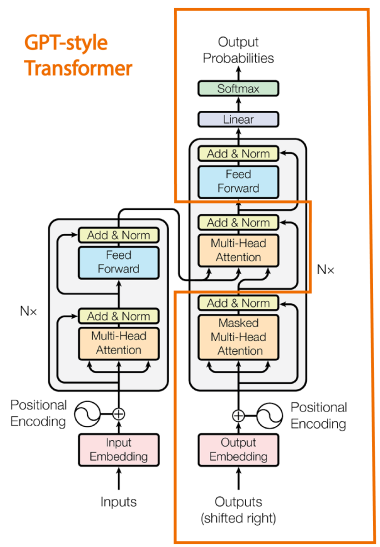


Figure: The Transformer architecture

In sessions 12–13 we will work with neural networks

Why Quantitative?

Figure: Big Data – World Economic Forum

Why Quantitative?

Figure: Big Data – World Economic Forum

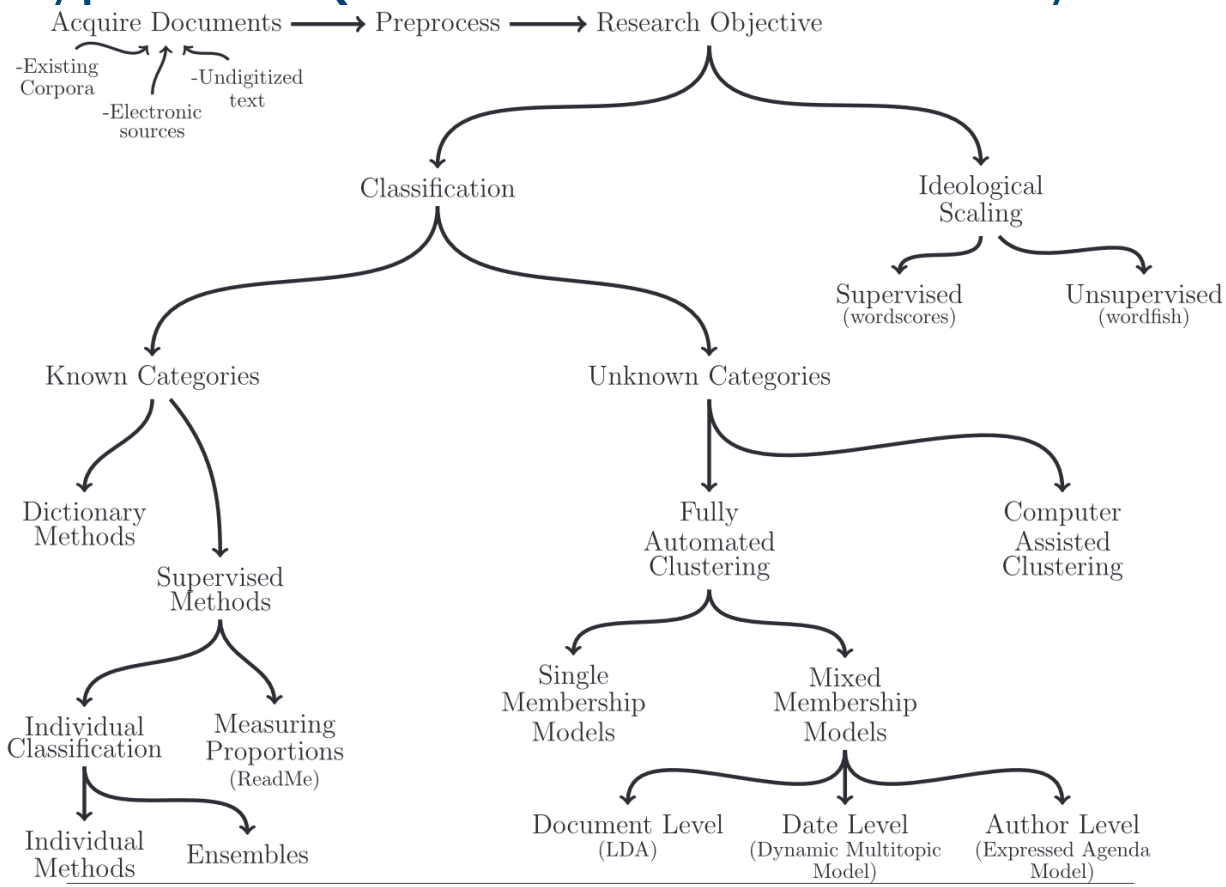
Quantitative analysis allows us to detect patterns in large amounts of data that would otherwise be difficult to uncover.

Types of Quantitative Text Analysis I

There are various classification attempts:

- unsupervised vs. supervised (Grimmer and Stewart 2013)
- rule-based, supervised, unsupervised, and hybrid (Baden et al. 2020)

Types of Quantitative Text Analysis II



Types of Quantitative Text Analysis III

In addition to these types, we also distinguish in the seminar between the way text components are analyzed...

- Bags-of-words approaches $\rightarrow$ no context, words are thrown into a bag
- Embeddings $\rightarrow$ context-dependent

When to Use Quantitative Text Analysis?

No convention, depends on the method...:

- with bag-of-words we often need large amounts of data to obtain valid and reliable results
- with embeddings, even small amounts of data are often sufficient

Until Next Week... I

- Read the syllabus
  - Is there a topic you are missing?
  - Do you have additional literature suggestions?
  - Do you already know which method you would like to use?
  - Do you have any remaining questions about the course requirements?

Until Next Week... II

- Preparation for the R crash course
  - Download R and RStudio
  - Work through the tutorials
    - Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023, June). *R for Data Science*. O'Reilly Media, Inc.
    - Ellis, A., & Mayer, B. (2024). Einführung in R.
  - Complete problem set

Thank you for your attention!
Any questions?

References 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*.

Benoit, K. (2009). Introduction to quantitative text analysis.

Bonikowski, B., & Nelson, L. K. (2022).From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. *Sociological Methods & Research*, *51*(4), 1469–1483. https://doi.org/10.1177/00491241221123088

Chun Ting-Ho, J. (2021). Introduction to Computational Text Analysis and Social Media Research using R.

Ellis, A., & Mayer, B. (2024). Einführung in R.

References II

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

Stahlberg, D., & Sczesny, S. (2001).Effekte des generischen Maskulinums und alternativer Sprachformen auf den gedanklichen Einbezug von Frauen. *Psychologische Rundschau*, *52*(3), 131–140. https://doi.org/10.1026//0033-3042.52.3.131

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962).The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*(4), 484–498. https://doi.org/10.1002/bs.3830070412

References III

Tavits, M., & Pérez, E. O. (2019).Language influences mass opinion toward gender and LGBT equality. *Proceedings of the National Academy of Sciences*, *116*(34), 16781–16786. https://doi.org/10.1073/pnas.1908156116

Turing, A. M. (1950).Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vervecken, D., Hannover, B., & Wolter, I. (2013).Changing (S)expectations: How gender fair job descriptions impact children's perceptions and interest regarding traditionally male occupations. *Journal of Vocational Behavior*, *82*(3), 208–220. https://doi.org/10.1016/j.jvb.2013.01.008

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023, June). *R for Data Science*. O'Reilly Media, Inc.