

Quantitative Textanalyse

Sitzung 4: Was ist quantitative Textanalyse?

Mirko Wegemann

Universität Münster
Institut für Politikwissenschaft

30. Oktober 2024

Logistik und Fragen

- Fragen zu letzter Sitzung?

Inhaltsanalyse I



Figure: Die Vielfalt der Inhaltsanalyse laut GPT-4o

Inhaltsanalyse II

“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)

¹ **Prompt:**

Gebe eine Infografik aus, welche die Vielfalt der Inhaltsanalyse als wissenschaftliche Methode visualisiert. Zeige darin u.a. verschiedene Datenquellen, Analysemethoden und Forschungsfragen. Es sollte verdeutlicht werden, dass es manuelle, aber auch automatisierte Inhaltsanalyse gibt.

Komponenten von Kommunikation

"Who says what to whom with what effect?" (Gerbner 1956)

Kommunikation besteht dementsprechend aus vier zentralen Komponenten.

1. Sender*innen
2. Inhalt
3. Empfänger*innen
4. Wirkung

Sender*innen



Figure: Gruppe vs. Individuum

Empfänger*innen



Figure: To the converted vs. to the skeptics

Beispiel für den Inhalt: Framing

Wenn wir politische Kommunikation analysieren, sprechen wir oft von **Frames**.

- Ein Frame betont einen bestimmten Aspekt eines Themas, er ist ein “subset of potentially relevant considerations” (Druckman and K. R. Nelson 2003, p. 730)
- Durch gezieltes Framing versuchen politische Akteur*innen zu beeinflussen, wie über ein bestimmtes Thema debattiert werden sollte (T. E. Nelson and Kinder 1996)

Wirkung

Gerbner (1956) unterscheidet zwischen

1. Effektivität einer Botschaft: wird das Ziel der Botschaft erreicht
2. Konsequenz: Auswirkungen, welche über direktes Ziel hinausgehen

Obama's Inauguration Speech



Barack Obama inaugural address: Jan. 20, 2009



CBS News
6.13M subscribers

Subscribe

3.2K



Share

Save



Figure: Obama's Inauguration Speech (Video; Text)

Trump's Inauguration Speech



Figure: Trump's Inauguration Speech Video and Text

Aufgabe I

Teilt Euch in zwei Gruppen auf. Die Hälfte von Euch beschäftigt sich mit Obama's Inauguration Speech, die anderen mit der von Trump. [15 Minuten]

- Was sind die wichtigsten Themen im Text?
- Welche Frames versuchen die Präsidenten zu setzen?
- Werden bestimmte Gruppen besonders stark adressiert?
- Fällt Euch sonst noch etwas in den präsidialen Ansprachen auf?

Inaugural Speech Obama vs. Trump

Obama

Trump

Thema

Frames

Gruppen

Weiteres

Bedeutung und Text I

An einer Stelle schreibt Krippendorff (2018):

“There is nothing inherent in a text; the meanings of a text are always brought to it by someone” (28).

Was meint er damit?

Bedeutung und Text II

- Text ist kontextabhängig
 - Leser*innen interpretieren Text unterschiedlich
 - Inhalt ebenfalls abhängig von Verfasser*innen des Texts
 - Analyst*innen gehen unterschiedlich an Text heran

Wesentliche Voraussetzungen für eine Inhaltsanalyse

- A priori Formulierung einer falsifizierbaren **Forschungsfrage**

Wesentliche Voraussetzungen für eine Inhaltsanalyse

- A priori Formulierung einer falsifizierbaren **Forschungsfrage**
- Verdeutlichung des **Kontexts**, in dem Text entstanden ist

Wesentliche Voraussetzungen für eine Inhaltsanalyse

- A priori Formulierung einer falsifizierbaren **Forschungsfrage**
- Verdeutlichung des **Kontexts**, in dem Text entstanden ist
- **Analytisches Modell**, in dem Wirkungsbeziehungen dem Kontext des Texts angepasst werden

Wesentliche Voraussetzungen für eine Inhaltsanalyse

- A priori Formulierung einer falsifizierbaren **Forschungsfrage**
- Verdeutlichung des **Kontexts**, in dem Text entstanden ist
- **Analytisches Modell**, in dem Wirkungsbeziehungen dem Kontext des Texts angepasst werden
- **Validierung** (im Idealfall mit 'out-of-domain' Evidenz → anderen Daten)

Formen der Inferenz I

Krippendorff (2018) unterscheidet zwischen drei verschiedenen Formen der Inferenz:

1. **deduktive** Inferenz: vom Allgemeinen ins Spezifische
2. **induktive** Inferenz: vom Spezifischen ins Allgemeine
3. **abduktive** Inferenz: vom Spezifischen in andere Spezifika

Welche Form der Inferenz können Inhaltsanalysen laut Krippendorff (2018) leisten? Stimmt ihr dem zu?

Warum quantitativ? I

Das größte Problem der qualitativen Textanalyse:

Harvard Dataverse > ParlSpeech Dataverse >

The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies

Version 1.0



Rauh, Christian; Schwalbach, Jan, 2020, "The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies", <https://doi.org/10.7910/DVN/L4QAKN>, Harvard Dataverse, V1

Cite Dataset

Learn about [Data Citation Standards](#)

Access Dataset

Contact Owner Share

Dataset Metrics

14,501 Downloads

Warum quantitativ? 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 Schritte zur erfolgreichen Textanalyse

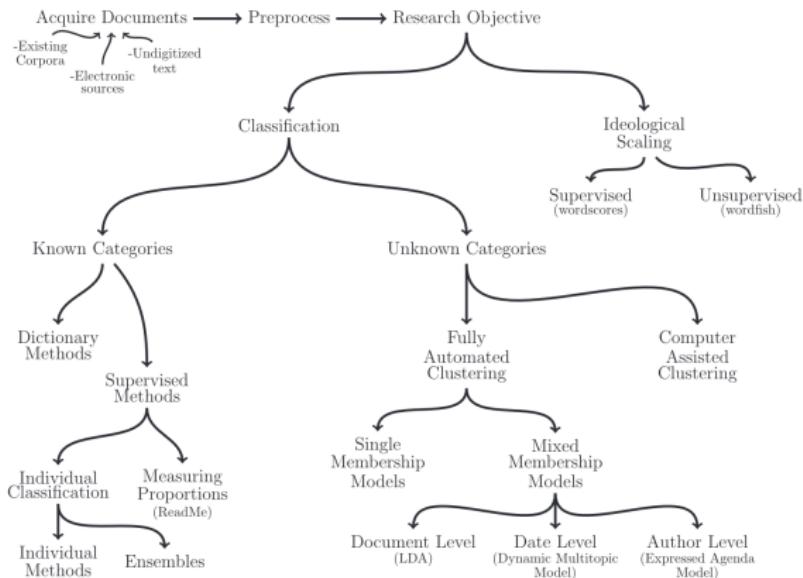


Figure: Der Prozess der Textanalyse laut Grimmer and Stewart (2013, p. 268)

1. Schritt: Datensammlung I

Zwei Optionen:

- Nutzung verfügbarer Daten
- erschließen neuer Daten (bspw. über Scraping-Techniken)

Was könnten potentielle Probleme im ersten Schritt sein?

1. Schritt: Datensammlung II

- vor allem mangelnde Datenverfügbarkeit
 - Textdateien nicht digitalisiert
 - mangelnde Zugriffsrechte
 - *selection bias* (Text nur von bestimmten Akteur*innen verfügbar)

2. Schritt: Vorbereitung des Texts I

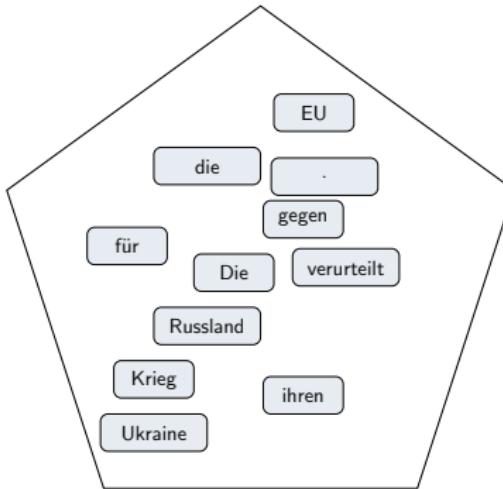
Text ist komplex und für unsere quantitativen Modelle nicht nutzbar. Dementsprechend müssen wir ihn zunächst umwandeln.

- Umwandlung in eine Repräsentation von Zahlen
- Vereinfachen von der Datenstruktur (z.B. durch Stemming, Lemmatisierung)
- Ausschluss von Wörtern (wie Stopwords)

2. Schritt: Vorbereitung des Texts II

Bags-of-words (BoW) erstellt für jedes Dokument D eines Korpus einen Vektor, welcher die Token t enthält.

Die EU verurteilt Russland für ihren Krieg gegen die Ukraine.



2. Schritt: Vorbereitung des Texts III

Die generelle Idee dahinter ist, dass wir die **Bedeutung** von Text durch das genutzte **Vokabular** nachvollziehen können. Ein Vergleich zwischen Dokumenten d_1 und d_2 findet alleinig auf Basis der **Häufigkeiten von Tokens** statt.

2. Schritt: Vorbereitung des Texts IV

Document D1	<i>The child makes the dog happy</i> the: 2, dog: 1, makes: 1, child: 1, happy: 1				
Document D2	<i>The dog makes the child happy</i> 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)

2. Schritt: Vorbereitung des Texts V

Was könnte ein Problem dieser Struktur sein?

2. Schritt: Vorbereitung des Texts VI

Probleme von *bags-of-words*

- Wir verlieren Informationen zur Reihenfolge der Wörter
- Die grammatische Struktur eines Satzes wird vernachlässigt.
- Wörter werden nicht kontextualisiert. Jedes Wort kann nur eine Bedeutung haben.

2. Schritt: Vorbereitung des Texts VII

Embeddings als Ausweg:

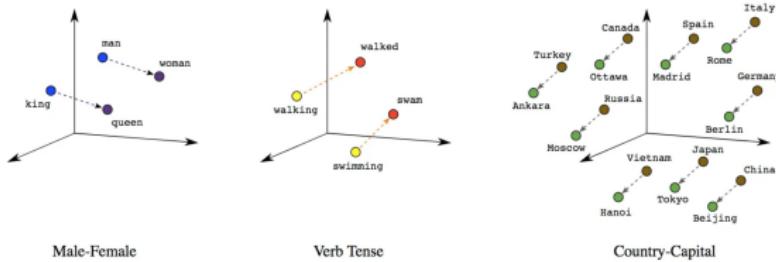


Figure: Quelle: Towards Data Science

3. Schritt: Die Analyse I

Es gibt verschiedenste Modelle in der quantitativen Textanalyse, welche sich folgendermaßen klassifizieren können (vgl. Baden et al. 2020):

- **regelbasierte** Analysen: Listen von Regeln bzw. Features
 - Wörterbuch-Analysen: Ton/Sentiment eines Texts, Kategorisierung
 - Dependency-Parser: Satzstrukturen
- **supervised** Analysen: Bereitstellung eines Trainingdatensatzes, auf dessen Basis Muster eines Textes gelernt werden und auf nicht-klassifizierte Texte angewandt werden können
 - Naive Bayes
 - Support Vectors Machine
 - Random Forest
 - etc.

3. Schritt: Die Analyse II

- **unsupervised** Analysen: basieren auf statistischen Verfahren, in denen eine Funktion optimiert und Texte klassifiziert werden
 - Topic Models
 - Scaling Models
- **hybride** Ansätze: mehrere Analyseverfahren werden genutzt
 - z.B. 1.) unsupervised Modelle zur Kategorisierung von Daten
 - diese dienen als Input für die Klassifikation, welche von durch 2.) supervised Modelle verfeinert werden

3. Schritt: Die Analyse III

Was denkt ihr, was sind die größten Schwierigkeiten bei der quantitativen Textanalyse?

3. Schritt: Die Analyse IV

- Zu wenig Daten
- Sprachbarrieren
- Mangel an Rechenpower
- Replizierbarkeit
- Validierung

Verbindung zwischen qualitativer und quantitativer Textanalyse

“Without human intelligence and the human ability to read and draw inferences from texts, computer text analysis cannot point to anything outside of what it processes. Computers have no environment of their own making; they operate in the contexts of their users' worlds without understanding those contexts”
(Krippendorff 2018, p. 29)

Kontext in quantitativen Textanalysen I

Krippendorff (2018) betont vor allem den **Kontext**, aus dem ein Text entstammt.

Was meint ihr, wie können wir den Kontext eines Textes einbeziehen, wenn wir ihn quantitativ analysieren? Ist dies überhaupt möglich? Was könnten potentielle Schwierigkeiten sein?

Kontext in quantitativen Textanalysen II

- **Kontextwissen über Sender*innen** einer Botschaft über Metadaten erfassbar
 - Kontext einer Rede (bspw. Datum, Ort)
 - Charakteristiken der Sender*in (bspw. Gender, Parteizugehörigkeit, Rolle)
- **Kontextwissen über Empfänger*innen** aber oftmals schwer erfassbar
 - evtl. einige Daten in Social Media Kanälen, Leser*innen-Briefe etc.

Kontext in quantitativen Textanalysen III

- Kontext über **Analyst*innen**
 - Idealerweise wird ein Text von mehreren Forscher*innen analysiert, über die potentiell relevante Meta-Informationen zur Verfügung stehen

Die vier Prinzipien der quantitativen Textanalyse

Nr.	Prinzip
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: Prinzipien der QTA nach Grimmer and Stewart (2013, p. 269)

Validität und Reliabilität

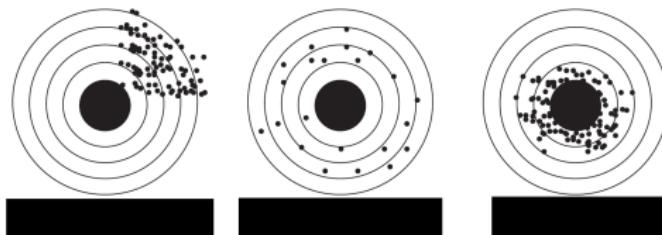


Figure 4.2 Reliability (precision) and validity

2

Welche Messungen sind valide? Welche reliabel?

²Gerring 2012, p. 82

Validität und Reliabilität

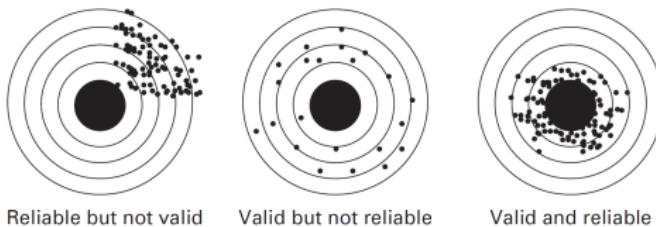


Figure 4.2 Reliability (precision) and validity

3

³Gerring 2012, p. 82

Gütekriterien von Inhaltsanalysen (und anderen Forschungsdesigns)

- Können Forschungsziele durch Forschungsdesign beantwortet werden?
- Validität: Können wir mit dem Forschungsdesign die richtigen Schlussfolgerungen ableiten?
 - **interne Validität:** hohe Validität der Ergebnisse innerhalb unserer Stichprobe
 - **externe Validität:** Ergebnisse lassen sich auch auf Fälle außerhalb unserer Analyse ausweiten
- Reliabilität: Lassen sich die Ergebnisse unter selben Bedingungen replizieren?

Validierungsschritte in der quantitativen Textanalyse I

Art der Validierung abhängig vom gewählten Design

- **supervised Analysen** basieren auf statistischen Vorhersagemodellen
 - verschiedene Metriken verfügbar: wie gut kann Modell unsere bereits klassifizierten Daten vorhersagen
 - Beispiele: Confusion Matrix, Accuracy, F1-Score, etc.
- **regelbasierte** und **unsupervised** Approaches sind schwieriger zu validieren

Validierungsschritte in der quantitativen Textanalyse II

Nach **Quinn.2010** sollten wir vor allem fünf verschiedene Arten von Validität sicherstellen:

1. **semantische Validität**: sind Themen kohärent
2. **konvergente Konstruktvalidität**: stimmt Indikator mit anderen Indikatoren überein, wenn wir Übereinstimmung vermuten
3. **diskriminatorische Konstruktvalidität**: weicht Indikator von anderen Indikatoren ab, wenn wir Unterschiede erwarten
4. **vorhersagende Validität**: korreliert Indikator mit externen Ereignissen, wenn wir es erwarten (Bsp.: Terroranschlag & 9/11)

Validierungsschritte in der quantitativen Textanalyse III

5. **Hypothesen**-Validität: können wir Hypothesen mithilfe des Indikators testen?

Reliabilität und Validität von quantitativer Textanalyse

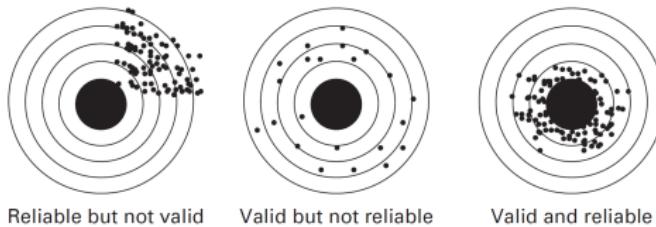


Figure 4.2 Reliability (precision) and validity

4

Eurer Einschätzung nach, sind quantitative Textanalysen valider und/oder reliabler als qualitative Textanalysen?

⁴Gerring 2012, p. 82

Ausblick I

- nächste Woche: Datensammlung durch Web-Scraping
- Literaturempfehlung:
Munzert (2015) → Kapitel 1
- Macht euch Gedanken über Websites, deren Inhalt ihr gerne automatisiert herunterladen wollt
- Ladet Euch den SelectorGadget für Euren Browser herunter
(entweder als Lesezeichen bei Firefox oder als Add-On bei Chrome)

Literatur 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>

Literatur 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.

Literatur 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>