Quantitative Textanalyse

Sitzung 9: Datenanalyse – Scaling

Mirko Wegemann

Universität Münster
Institut für Politikwissenschaft

04.12.2024

Plan für heute

- Präsentationen von Katharina und Elsa
- Rest von letzter Woche (Validierung von Topic Models)
- Latent Semantic Scaling
- Feedback zum bisherigen Kursverlauf

Präsentationen

Zuerst die Kurz-Präsentationen von **Katharina** und **Elsa**.

Validierung

Validierung kann in unterschiedlichem Maße stattfinden (Überblick über die Konsequenzen in Bernhard et al. (2023)):

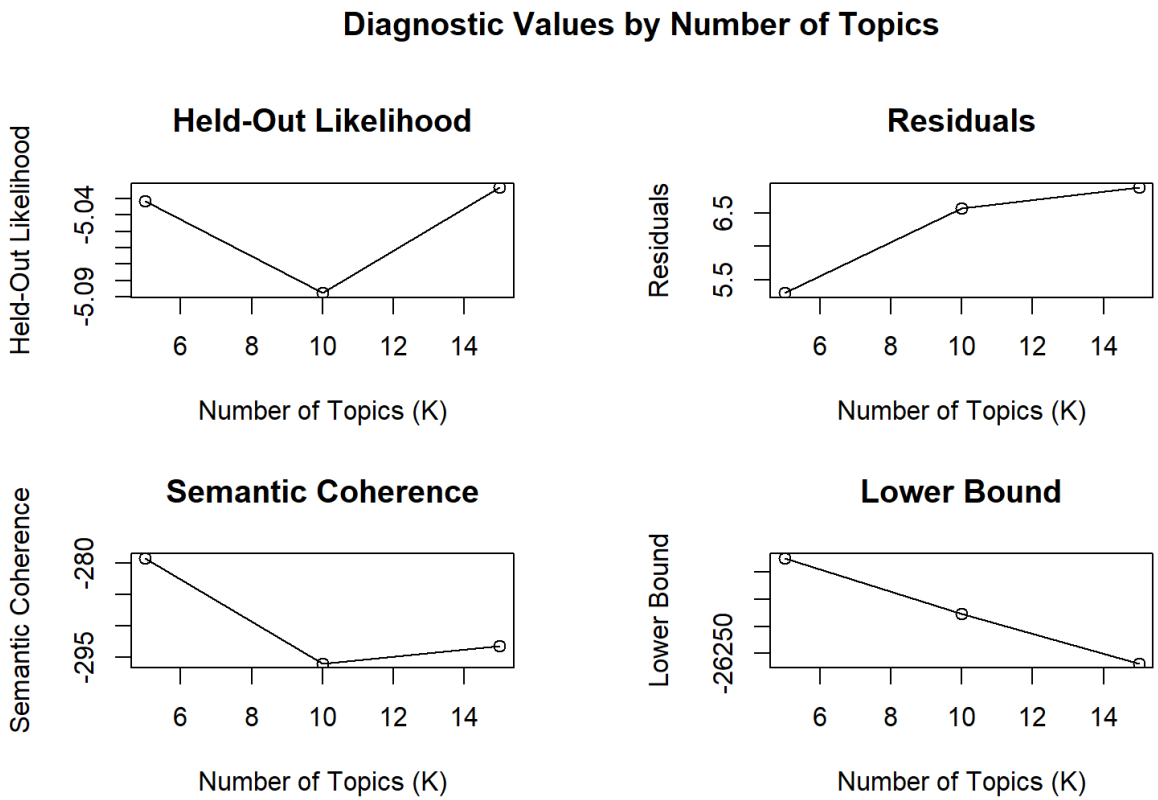
1. **semantische** Validierung (= Interpretation der Topics durch Häufigkeitsmaße)
2. **statistische** Validierung (= geringster statistischer Fehler)
3. **externe** Validierung nach Quinn et al. (2010) (= Vorhersagekraft)
4. anhand **manueller** Kodierung

Statistische Validierung

Wir können die `searchK()`-Funktion von stm verwenden, um die ideale Anzahl an Themen k basierend auf Likelihood, Untergrenze, Residuen und semantischer Kohärenz zu bestimmen.

```
1  # stm-Format fuer diese Funktion notwendig
2  stm_left <- convert(dfm_left, "stm")
3
4  # Vektor mit Zahl der Topics
5  K <- c(5,10,15)
6
7  best_k <- searchK(stm_left$documents,
       stm_left$vocab, K, seed=421, emtol=0.001,
       LDAbeta = F)
```

Statistische Validierung



Statistische Validierung II

Alternativ können wir auch, wenn die Anzahl an Topics klar ist, das Modell auswählen, welches die besten Parameter hat (`selectModel()`)

```
1    best_m <- selectModel(stm_left$documents,
         stm_left$vocab, 4, runs=10, seed=421,
         emtol=0.001, init.type = "Spectral", LDAbeta
         = F)
2    plotModels(best_m)
```

Scaling I

Manchmal haben wir ein klar definiertes Thema (z. B. die Wirtschaft) und möchten bewerten, wie Akteure zu diesem Thema stehen.

- Scaling-Methoden können uns helfen, Positionierungen von Akteur*innen auf der Basis ihrer Texterzeugnisse zu gewinnen
- insbesondere in der Politikwissenschaft ist das häufig ein Ziel

Frühe Ansätze des Scalings I

Scaling erlebte einen Höhepunkt in den 2000er-Jahren, in denen zwei prominente Ansätze entworfen worden sind.

- **Wordscores** (Laver et al. 2003)
  - Wordscores ist eine semi-supervised Methode
  - Wir müssen Referenz-Texte auswählen (deren Position wir kennen; meist in den Extremen)
  - Die Verteilungen der Referenz-Texte dienen dem Algorithmus zur Einschätzung von sogenannten *virgin texts* → out-of-sample Texterzeugnisse
  - **Wordfish** (Slapin and Proksch 2008)
    - Berechnet ein Gewicht für jedes Wort (in der ursprünglichen Anwendung: wie stark es Parteien unterscheidet).
    - Liefert eine **Position** einer Partei auf einer unidimensionalen Skala.
    - Meistens auf einer stärker aggregierten Ebene.

Frühe Ansätze des Scalings II

Beide Ansätze sind etwas in die Jahre gekommen. Wenn Dokumente sehr eindeutig von einem Thema handeln, können sie noch immer gute Ergebnisse liefern. Sie sind aber sehr abhängig von der Datenqualität und sensitiv gegenüber Pre-Processing.,

Wordfish in R

```
1
2     m_wordfish <- textmodel_wordfish(m_dfm2)
          summary(m_wordfish)
3
4     Call: textmodel_wordfish.dfm(x = m_dfm2)
5
6     Estimated Document Positions: theta se 2001.1
          -0.9434 0.02570 [...]
```

Latent Semantic Scaling I

Latent Semantic Scaling (LSS) (Watanabe 2021)

- LSS "creates a **polarity score** of words depending on a certain number of seed words" (Watanabe 2021).
- Polarity scores der Wörter basieren auf *singular value decomposition* (ähnlich der Faktoranalyse)
- LSS ist **semi-supervised**; wir nutzen unser domänenspezifisches Wissen, um Beispielwörter zu bestimmen, welche die Polarität unserer Dimension abbilden
- Wir benötigen dabei **nicht jedes Wort**, welches in Verbindung mit einer Position stehen könnte; stattdessen sucht LSS automatisch nach Wörtern, welche häufig im Kontext unserer Polaritätswörter stehen und gewichtet diese stärker

Latent Semantic Scaling II

- Resultat: jedes Dokument erhält einen document polarity score, welcher auf der Summe der word polarity scores / deren Anzahl basiert
- $\mu = 0$ und $\sigma = 1$
- LSS funktioniert auf **Satzebene**; dementsprechend müssen wir unsere Dokumente zunächst in Sätze aufteilen

Latent Semantic Scaling III

Vorteile von LSS

- benötigt wenig Input; dementsprechend flexibel nutzbar mit geringen Sprachkenntnissen
- Texte sind meistens mehrdimensional; polarity und seed words helfen dabei, die zu untersuchende Dimension zu identifizieren

Latent Semantic Scaling IV

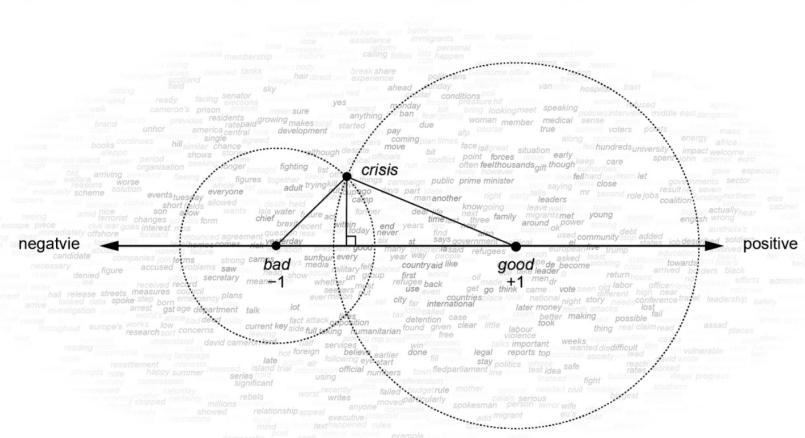
Nachteile von LSS

- idealerweise haben wir recht große Korpora (5,000-10,000 Dokumente mit ca. 200k-400k Sätzen)

Parameter von LSS

Wir müssen verschiedene Parameter setzen...

- **seed words**: negative und positive Wörter, welche unsere Dimension abdecken
- **model terms**: Kontextwörter unserer Dimension; oftmals über keyness scores [nur benötigt, falls wir über keinen kontext-spezifischen Korpus verfügen]
- **Dimensionen der SVG**: anhand wie vieler Dimensionen wird die semantic proximity von Wörtern geschätzt (konventionell meist 200-300)



Cleavage Identities in Voters' Own Words I

- **Forschungsfrage**:
- **Argument**:
- **Daten und Analyse**:
- **Ergebnisse**:
- **Implikationen**:

Cleavage Identities in Voters' Own Words II

- **Forschungsfrage**: Hat sich eine kollektive Identität im Hinblick der Universalismus-Partikularismus Konfliktlinie entwickelt?
- **Argument**: Menschen schaffen sich neue Gruppenidentität entlang politischer Konflikte
- **Daten und Analyse**: Offene Angaben in zwei Schweizer-Umfragen; Latent Semantic Scaling und Keyness-Analysen
- **Ergebnisse**: Wähler*innen haben (positive) Identität zu ihrer kulturellen Bezugsgruppe entwickelt – Identität hängt mit Parteipräferenz für grüne und rechtsradikale Parteien zusammen
- **Implikationen**: Die *kulturelle* Konfliktlinie hat sich in sozialen Identitäten verstetigt.

Forschungsfrage und Motivation

Im politischen Wettbewerb ist eine neue kulturelle Konfliktlinie entstanden, welche verstärkt von politischen Akteur*innen bedient wird. Entsteht dabei auch eine neue Identität in der Bevölkerung entlang dieser Konfliktlinie.

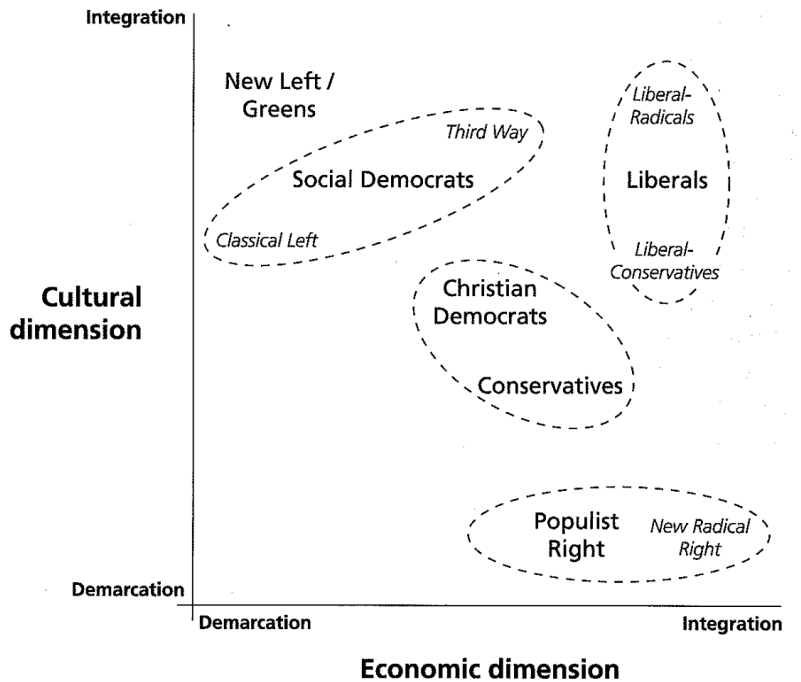


Figure: Cleavages nach Kriesi (2008, p. 15)

Theorie I

Anhaltspunkt Nr. 1 ist die sogenannte **Cleavage**-Theorie von Lipset and Rokkan (1967)

- **Konfliktlinien**, welche durch soziostrukturelle Spannungen entstehen, werden durch politische Akteur*innen aufgegriffen
- Bevölkerung reagiert und **bildet** neue **Identitäten** entlang der Konfliktlinien
- Lipset and Rokkan (1967) identifizieren vier Konfliktlinien:
  1. Stadt-Land
  2. Staat-Kirche
  3. Kapital-Arbeiter*innen
  4. Zentrum-Peripherie

Theorie II

In der Nachkriegszeit hat sich zudem eine Konfliktlinie zwischen **globalization losers and winners** (Kriesi 2008) herausgebildet.

- H_1 Befragte, welche eine Universalismus/Partikularismus-Identität haben, unterstützen vor allem die *Neuen Linken* bzw. die *Neuen Rechten*
- H_2 Befragte verbinden Begriffe, welche mit Universalismus/Partikularismus in Verbindung stehen mit ihrer politischen Identität

Theorie III

Darüber hinaus orientiert sich Zollinger (2024) an die **social identity**-Theorie von Tajfel and Turner (2004).

- H_3 Menschen definieren ihre Identität in Abgrenzung von ihrer *outgroup* ab.
- H_4 Die Ingroup-Identität der Befragten ist positiv konnotiert.

Daten und Methoden I

Die zentralen Fragen in der Umfrage sind...

- "If you imagine people with a lifestyle and opinions similar to your own, what kind of people would these be? How would you describe them?"
- "For outgroups, respondents were asked, 'And someone who is not at all like you? Someone who lives completely differently and who has very different opinions? How would you describe them?' "

Daten und Methoden II

Vorgehen:

- Recht wenige pre-processing Steps (nur stopwords + punctuation removal und lowercasing)
- *seed words* anhand von a) theoretischen Überlegungen und b) keyness-Statistiken
- daraufhin *Latent Semantic Scaling* (vermutlich ohne model terms)

Ergebnisse I

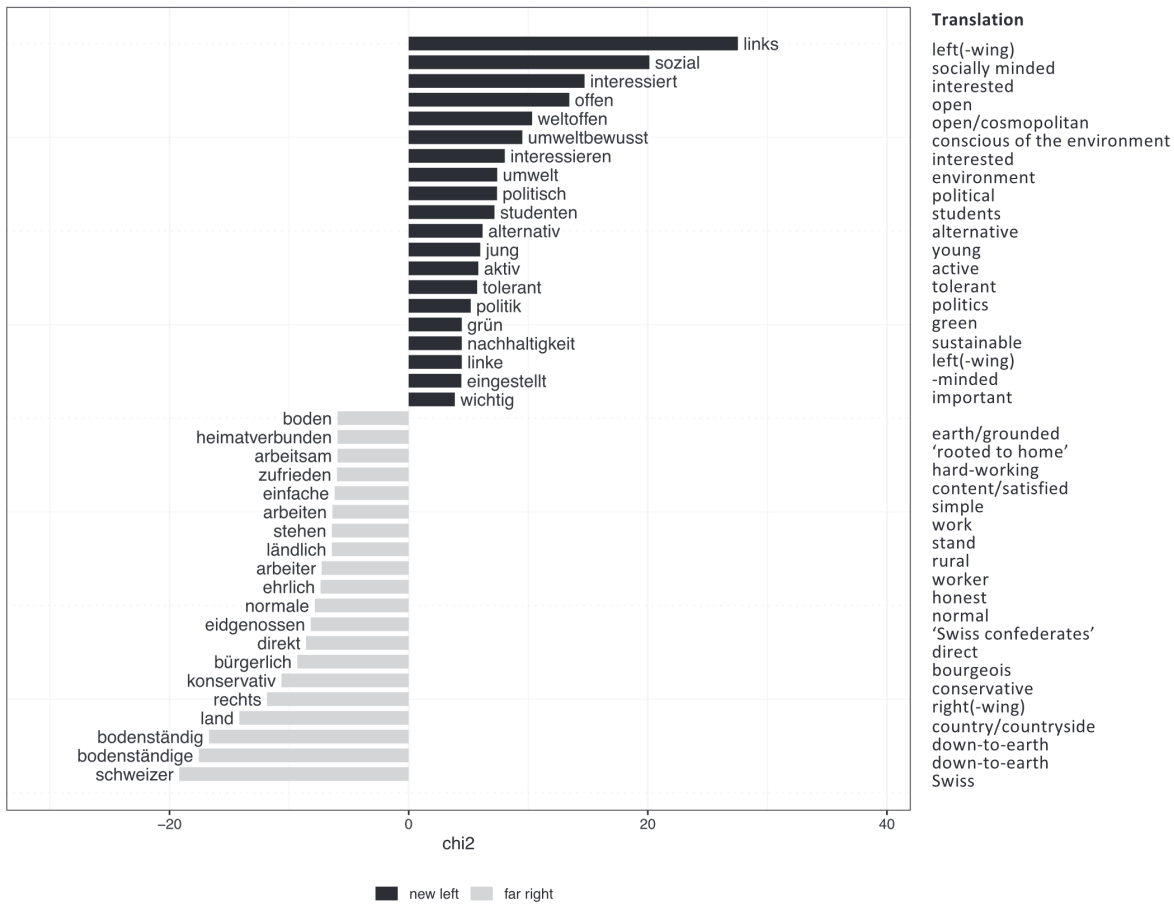


Figure: Wie Wähler*innen ihre Ingroups beschreiben (Zollinger 2024, p. 146)

Ergebnisse II

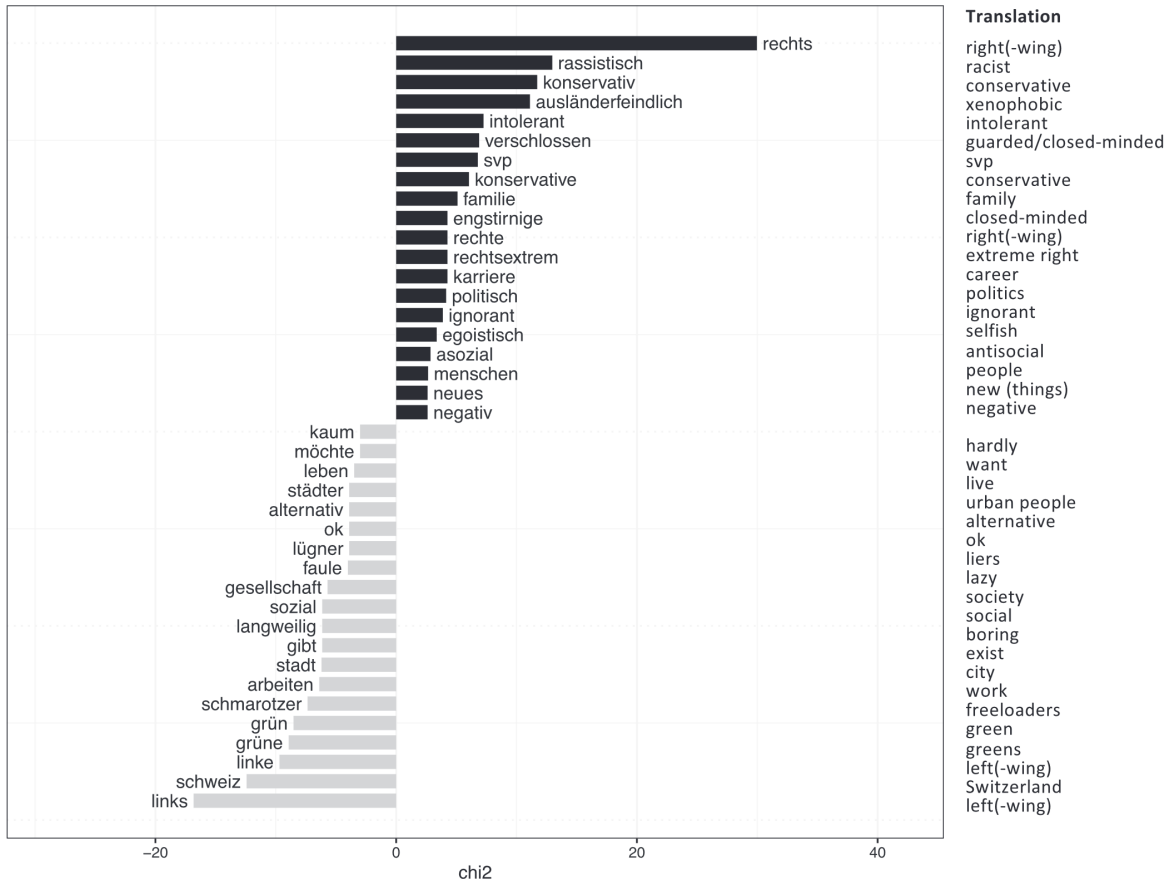


Figure: Wie Wähler*innen ihre Outgroups beschreiben (Zollinger 2024, p. 147)

Ergebnisse III

TABLE 1 Binomial Logistic Regressions: Ingroup Identity Scores and New Left/Far Right Party Preference

	(1) Far Right	(2) Far Right	(3) Far Right	(4) Far Right	(5) New Left	(6) New Left	(7) New Left	(8) New Left
Particularist (vs. universalist)	4.51**	6.41**	4.58**	6.61**	−3.81**	−5.82**	−5.12**	−6.32**
ingroup identity	(0.96)	(1.59)	(0.93)	(0.81)	(0.89)	(1.58)	(1.02)	(0.77)

Figure: Parteipräferenz, erklärt durch Identitäten (Zollinger 2024, p. 151)

Zudem: Identitäten können durch sozio-ökonomische Dispositionen und Einstellungen erklärt werden.

Latent Semantic Scaling in R

1. Definition der Polarity Words
2. Definition der Model Terms
3. Schätzung des LSS-Modells

```
1
2     econ_left <- c("sozial", "gerechtigkeit",
          "armut", "proletariat", "ungleichheit",
          "solidaritaet")
3     econ_right <- c("defizit", "haushalt",
          "stabilitaet", "unternehmer", "steuern")
4
5     econ_dict <- dictionary(list(left = econ_left,
6     right = econ_right))
7     seed <- as.seedwords(econ_dict)
```

Latent Semantic Scaling in R II

```
1  m_terms <- char_context(toks_parl2, "wirtschaft*")
```

Latent Semantic Scaling in R III

```
1   lss_model <- textmodel_lss(dfm_parl2, seeds =
        seed, terms = m_terms, k = 300)
```

$\rightarrow$ k ist die Anzahl der Singulärwerte (Dimensionenreduktion).

Let's do it in R

Ausblick

- letzte Wochen: unsupervised topic modelling und dem semi-supervised scaling approach
- **nächste Woche:** Grundlagen des supervised machine learnings
- **Literatur**
  - James, G., Witten, D., Hastie, T., & Tibshirani. (2021). *An Introduction to Statistical Learning*. Retrieved September 19, 2024, from https://www.statlearning.com
  - Gessler, T., & Hunger, S. (2022). How the Refugee Crisis and Radical Right Parties Shape Party Competition on Immigration. *Political Science Research and Methods*, *10*, 524–544. https://doi.org/10.1017/psrm.2021.64