Semantische Kontextanalyse von COVID-19-Diskursen mittels Association Rule Mining

Dominik Kremer
13.03.2021

Institut für Geographie

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

1 Theorie und Modellierung

2 Forschungsstand Text Mining

3 Diskursanalyse und COVID-19

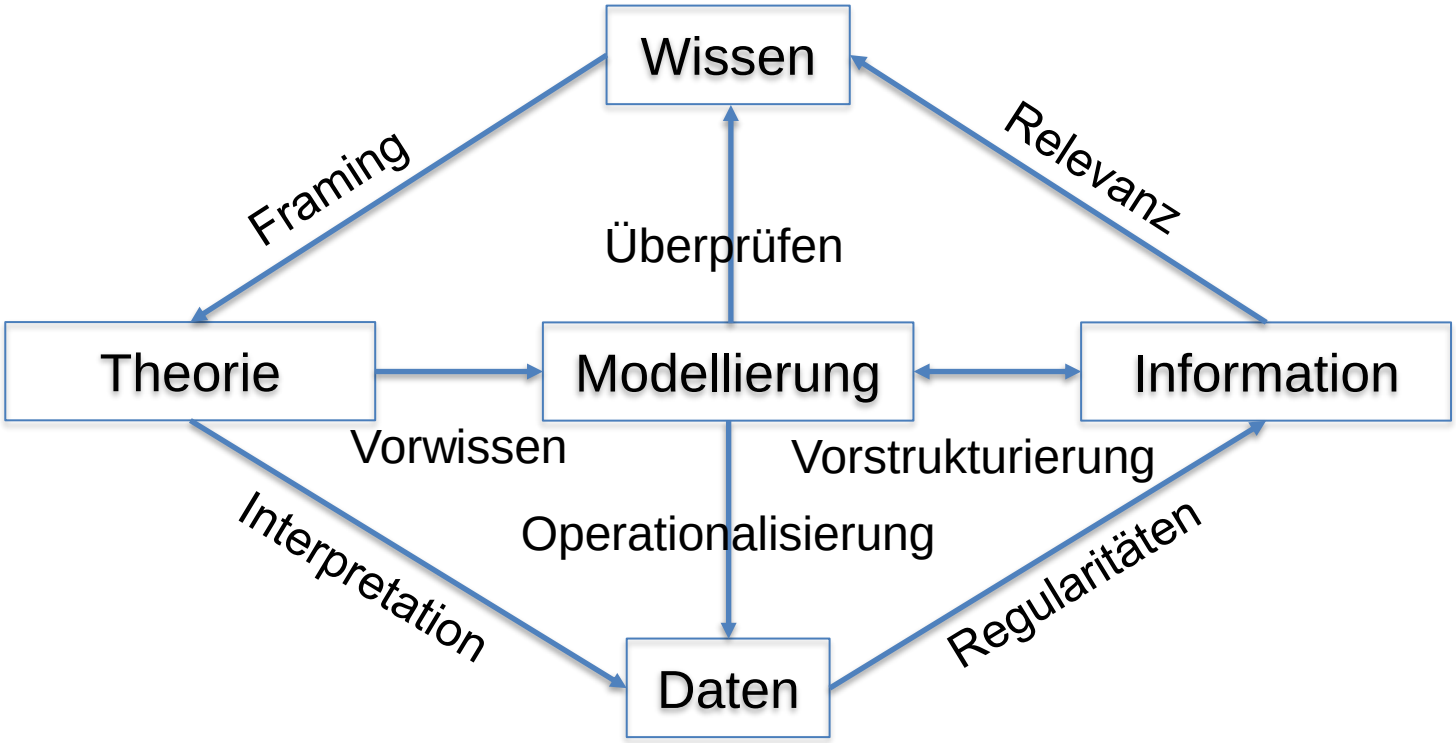
4 Association Rules als Benchmark für Relation Extraction

5 Limitationen/Nächste Schritte

Theorie der Digital Humanities

- Ausgangspunkt
  - kulturwissenschaftliche Disziplinen mit ähnlichen Hintergrundtheorien und Methoden
  - Geschichte, Kunstgeschichte, Bildwissenschaft, Linguistik, Philosophie, Kulturgeographie, …
- Spannungsfeld
  - Digitale Transformation der Forschungsmethodik: Erweiterung und Angleichung
- Hauptfrage
  - Wie kann eine Rückbindung an etablierte Theorien erhalten bleiben?
  - Gemeinsamer Theorierahmen oder Jeweiligkeit der Fächer?
- Digital Humanities
  - Spezialisierung innerhalb oder außerhalb der Fächer?
  - Methodeninnovation außerhalb der Informatik?

Spannungsfeld Problemrepräsentation



DH – (Digitale) Methodikinnovation im Fach?

- Methodikanwendung
  - Klassische Empirische Sozialforschung
  - Data Science, z.B. mittels R oder Python
- Methodikentwicklung
  - für bestimmte fachliche Fragestellung
  - z.B. Indikatoren oder Scores in der Data Science
- Methodikadaption
  - Evaluation der Methoden anderer Bereiche oder der Werkzeug der Informatik
  - z.B. Topic Modelling
- Methodikinnovation
  - Eigenständige strukturelle Weiterentwicklung der Methodik des Fachs
  - Nicht zwingend digital, aber vielfach durch digitale Transformation ausgelöst

Ein Text-Mining-Framework für Grounded Theory?

- Grounded Theory
  - Glaser/Strauss 1998
  - Entwickelt zur Theoriegenerierung
  - Iteratives, exploratives Verfahren
- Wesentliche Elemente
  - In-Vivo-Codes: wörtlich vorkommende Begriffe
  - Spätere schrittweise Abstraktion
- Knowledge discovery
  - Fayyad et al. 1996
  - Data Mining als exploratives Vorgehen zum Entdecken interessanter Information?
- Natural language processing
  - z.B. Jurafsky/Martin 2019
  - Stabiles Tooling zur Lemmatisierung von Texten

1 Theorie und Modellierung

2 Forschungsstand Text Mining

3 Diskursanalyse und COVID-19

4 Association Rules als Benchmark für Relation Extraction

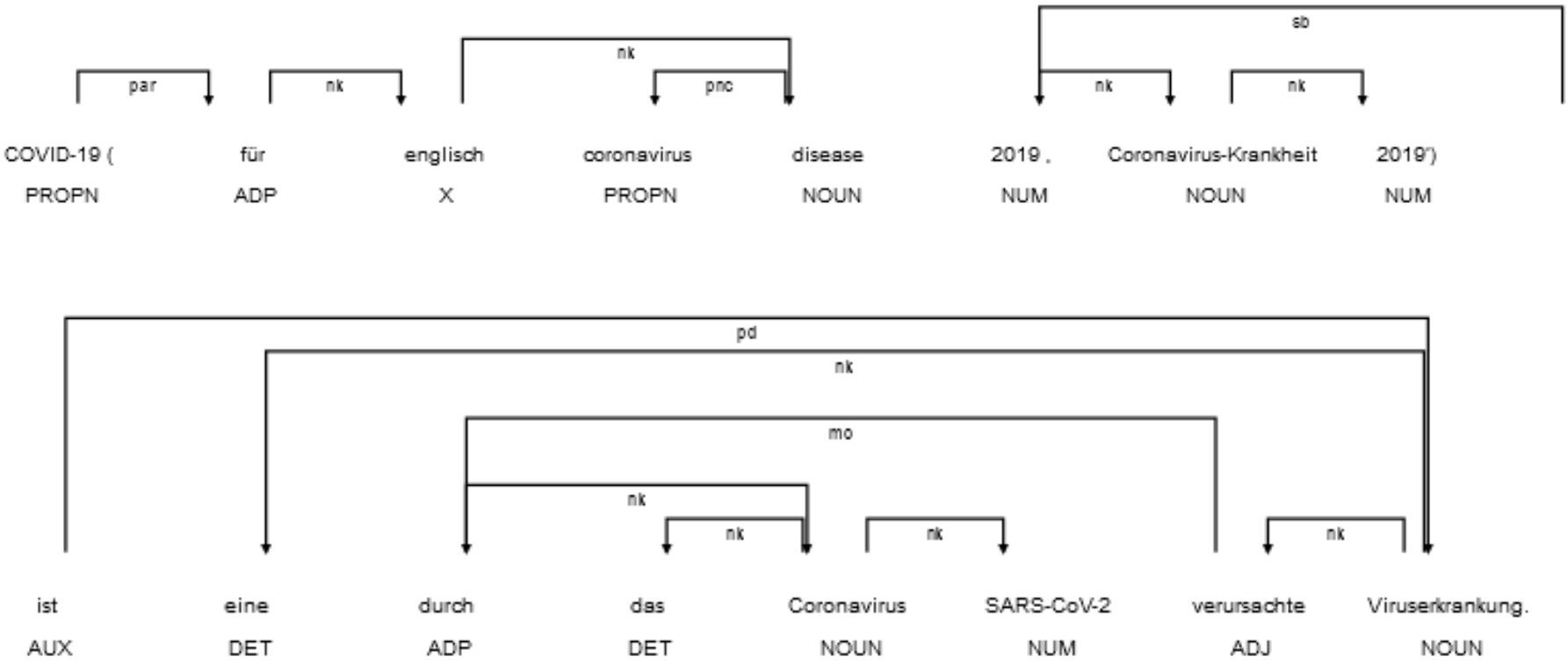
5 Limitationen/Nächste Schritte

Kollokation

(Bubenhofer 2015+, Evert 2009+)

- N-Gramme
  - Häufiges direktes Folgen von Begriffen in Phrasen
  - Lexikalisiert: Phraseologie
  - Begriff-Verb-Kopplungen
- Kookkurrenzen
  - Bag of word (z.B. Satz)
  - Kontextbindung zweier Lemmata in einem bestimmten Fenster (z.B. Satz)
- Relation extraction?
  - Kookkurrenzen sind noch kein Zusammenhang
  - Exploratives, Hypothesengenerierendes Verfahren
  - Art des Zusammenhangs muss qualitativ näher bestimmt werden, z.B. über keyword in context

Linguistische Zusatzinformation



1 Theorie und Modellierung

2 Forschungsstand Text Mining

3 Geographische Diskursanalyse von COVID-19

4 Association Rules als Benchmark für Relation Extraction

5 Limitationen/Nächste Schritte

Raum als Argument

(Entrikin 1991, Felgenhauer 2009, Kremer 2013)

- Abgrenzung
  - Container sind endlich und erfassen die Welt vollständig
- Einheitlichkeit
  - Container sind homogen und widerspruchsfrei
- Unveränderlichkeit
  - Container sind dauerhaft diskursiv gebunden (es ändert sich eher die Verortung)
- Letztbegründung
  - Körperlich erfahrbare Orte dienen als Begründung in Argumenten
- Anthropomorphismus
  - Regionen werden personalisiert
- Zentrum-Peripherie-Metapher
  - Erwünschtes im Nahbereich, Unerwünschtes im Außenbereich

#stayathome!

- „*Was macht denn der Bus aus Bielefeld hier? Der soll daheim bleiben, in Niedersachsen ist doch Corona!*“

https://commons.wikimedia.org/wiki/File:M%C3%BCnchen_%E2%80%94_Sophienstra%C3%9Fe_7_(Park_Caf%C3%A9_Biergarten).jpg

Qualitative Vorstudie

- Teilkorpus des DWDS-Corona-Corpus
  - BR
  - Blog TÜV Rheinland
  - Apothekenumschau
  - nordbayern.de
  - TaZ
- Insgesamt 30 Texte

- Axiales Kodieren
  - Verankerung im Raum über Toponyme
  - Synonyme/Metonyme von „COVID-19“
  - Relata im Satz: reasoning means balancing (Johnson 1987)
- Idee
  - Automatisiertes Test Mining kontrastiert qualitative Arbeit

Qualitative Ergebnisse

- Raum
  - als Verortung von Gefahr (z.B. Wuhan)
  - Als kognitives Containment (z.B. Italien)
  - als Verortung politischer Entscheidungen (z.B. Berlin)
- Syn-/Metonyme
  - Corona, COVID-19, Sars-CoV, …
  - Pan-/Epidemie
  - Aktuelle Krise
  - …
- Relata: gesellschaftliche Transformation
  - Digitalisierung, Datensouveränität
    - HomeOffice etc.
  - Entschleunigung, Innehalten
    - Nachhaltiger Alltag?
  - Lockdown
    - Resilienz
    - Anspannung

1 Theorie und Modellierung

2 Forschungsstand Text Mining

3 Diskursanalyse und COVID-19

4 Association Rules als Benchmark für Relation Extraction

5 Limitationen/Nächste Schritte

Association Rule Mining: einfache Relationsextraktion

Hipp et al. 2000

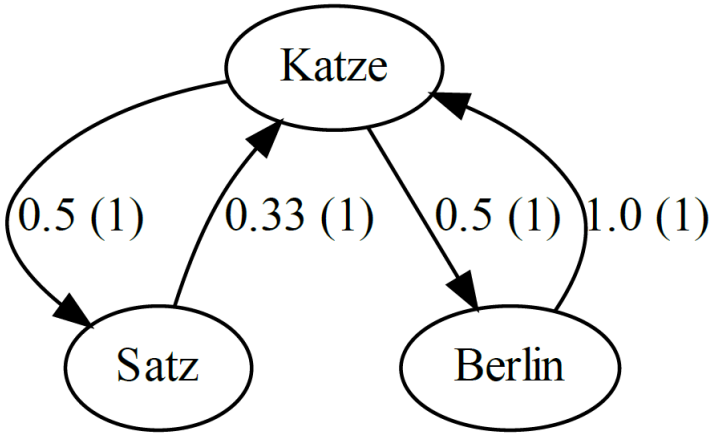
```
Dies ist ein Satz. Dies ist noch ein Satz. Der Satz gehört der Katze. Die Katze wohnt in Berlin.
```

{'Satz': 3, 'Katze': 2, 'Berlin': 1}

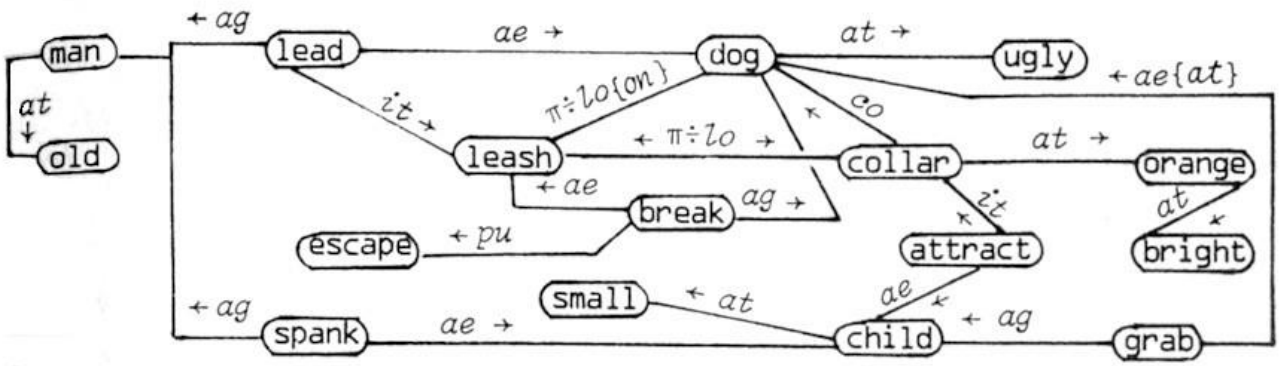
[Katze -> Satz(0.5), Satz -> Katze(0.3333333333333333), Katze -> Berlin(0.5), Berlin -> Katze(1.0)]

Preprocessing

- Lemmatisierung
- Named entity recognition
- Part of speech tagging
- Bag of words



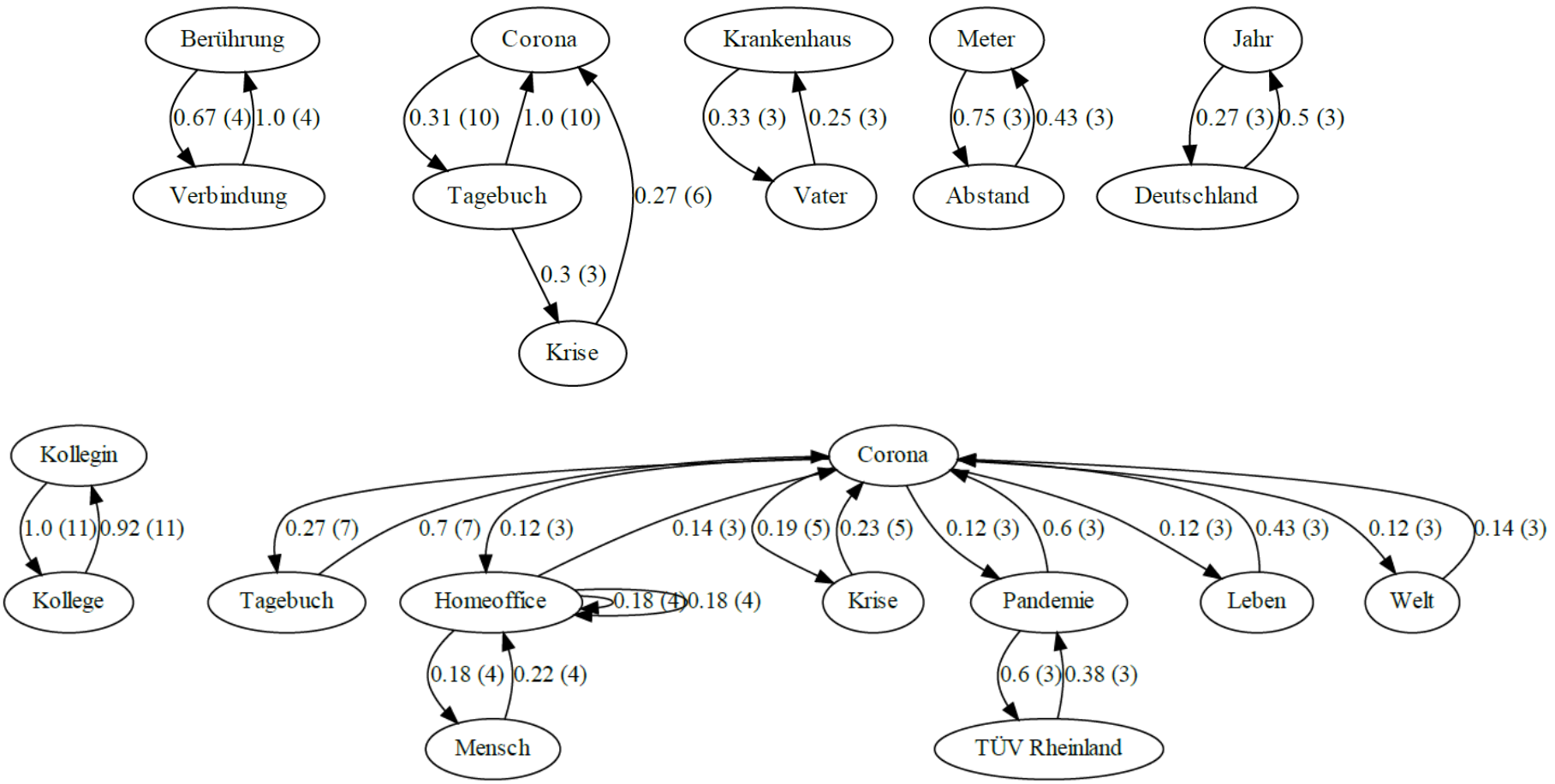
Graphvisualisierungen von Text – methodisch neu?



Key: *ae*: affected entity; *ag*: agent of; *at*: attribute of; *co*: containment of; *it*: instrument of; *lo*: location of; *pu*: purpose of; *π*: proximity

Beaugrande 1980

Teilkorpus BR bzw. TÜV



Berührung <-> Verbindung (4)

- **„Ohne Berührung gibt es keine wirkliche Verbindung mehr zwischen uns“**
- „Ohne Berührung gibt es keine wirkliche Verbindung mehr zwischen uns“
- „Ohne Berührung gibt es keine wirkliche Verbindung mehr zwischen uns.“
- „Aber keine Berührung mehr, keine Verbindung, das wäre ja jetzt schon die Isolierstation. Für ihn schwer erträglich“

1 Theorie und Modellierung

2 Forschungsstand Text Mining

3 Diskursanalyse und COVID-19

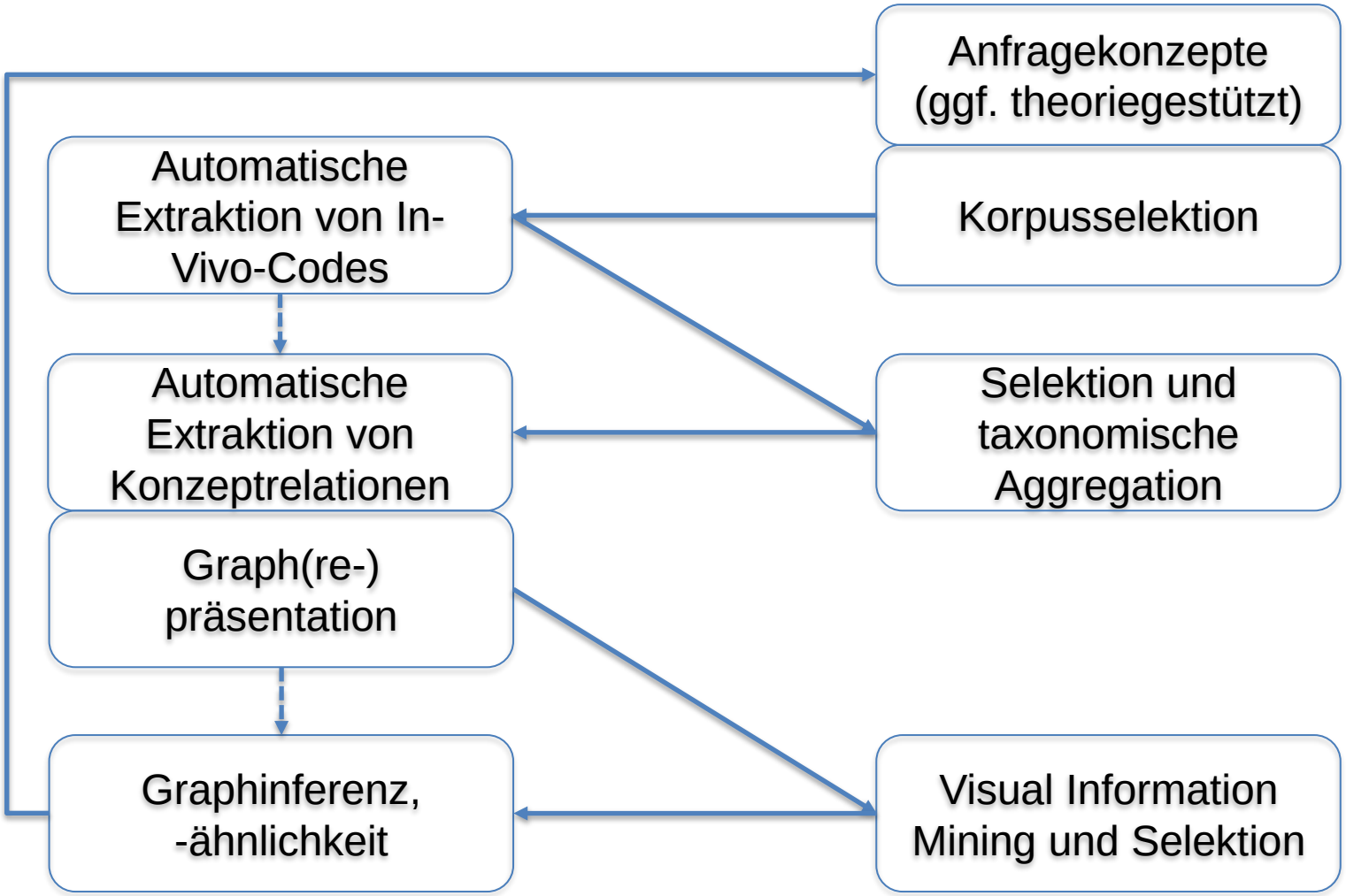
4 Association Rules als Benchmark für Relation Extraction

5 Limitationen/Nächste Schritte

Diskussion

- Limitationen
  - Rein statistische Indikation
  - Nur Nomina(lphrasen)
  - Isolierte exemplarische Betrachtung
  - Sehr einfache Relationsextraktion
  - Heterogene Daten
  - Keine Möglichkeit zur Inferenz
- Nächste Schritte
  - Nutzung linguistischer Dependenz: weitere Attribuierungen
  - Ausrollen auf größeren Teilkorpus
  - Informiertere Verfahren der Relationsextraktion
  - Individuelles Sprechen
  - Berechnung semantischer Ähnlichkeit?

Ein Text-Mining-Framework für Grounded Theory?



Vielen Dank für Feedback und Fragen!

Institut für Geographie