Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Dr. Gregor Wiedemann, gwiedemann@informatik.uni-hamburg.de
21.06.2018 – TDM in Recht, Wissenschaft und Gesellschaft, Trier

Textkorpora in den Sozialwissenschaften: Erhebung, (Nach-)Nutzung und Archivierung

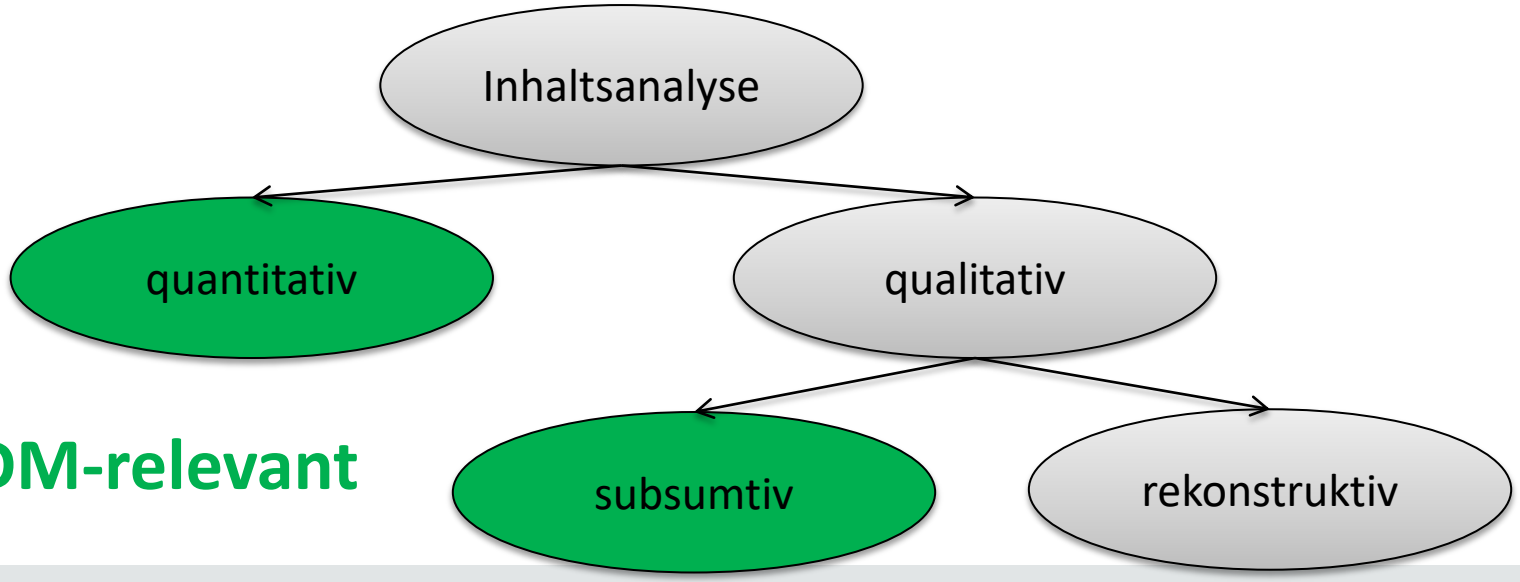
Fragen

- In welchen Forschungsdesigns der empirischen Sozialforschung spielen Textkorpora eine Rolle?
- Welche Datenquellen/-arten werden für TDM in den Sozialwissenschaften erschlossen?

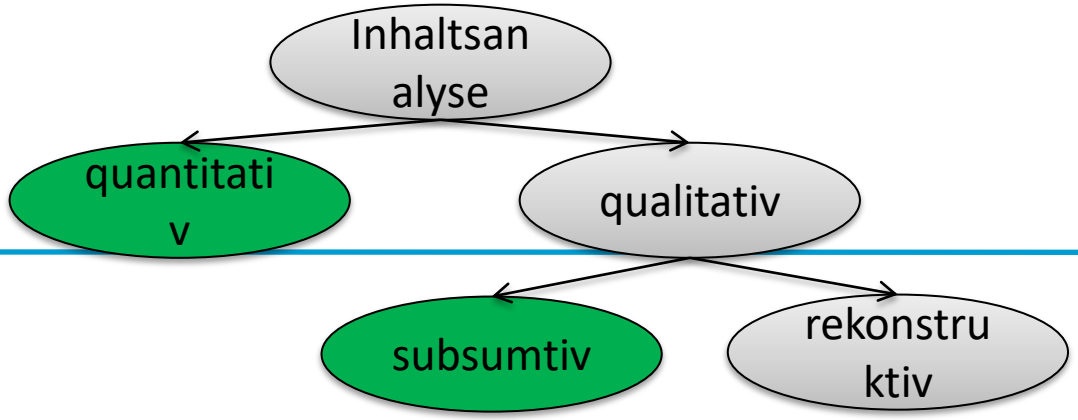
→ **Welche Daten werden wie und wofür geteilt?**

TDM in der empirischen Sozialforschung

- „Text as digital trace data“: zunehmend wichtige Rolle in der computergestützten Inhaltsanalyse



TDM-relevant



Inhaltsanalyse

- Kaum standardisierte Forschungsfragen
- Vielfalt an methodischen Vorgehensweisen
- Projektspezifische Datenerhebung, z.B.
  - Interview-Transkripte
  - Ausgewählte Medientexte
  - Webdaten anhand bestimmter Crawl-Strategien (z.B. Blogs, Social Media Kanäle)
- Ergebnisse: Kategorien/Codes, Annotationen/Codierungen

Archivierung, Nachnutzung, Wiederverwendbarkeit

- Wiederverwendbarkeit von Korpora und Annotationen **aus spezifischen Projektergebnissen** nur sehr eingeschränkt sinnvoll
  - BMBF Förderung verlangt mittlerweile zwar (anonymisierte) Archivierung von qualitativen Studien
  - Sekundäranalysen und „Big Qual“ potenziell möglich (Chance f. TDM)
  - aber Mehrwert davon in der Community noch umstritten
  - bislang kaum praktische Erfahrungen damit, v.a. im Bereich qualitativer Forschung

Datenquellen

- **ABER:** Korpora und Annotationen **als geteilte Ausgangsdatenbasis** während Datenerhebungsphase extrem sinnvoll
- Sehr große, themenübergreifende **Vollkorpora** → Auswahl relevanter Volltexte als Ausgangsdaten
  - Zeitungstext-Archive: Volltext + Zeitung, Ressort , Datum, AutorIn, enthaltene Entitäten (Eigennamen), ggf. Thema/Schlagworte (schon schwierig)
  - Parlamentsdokumente (Protokolle, Drucksachen etc.): Volltext + Datum, TOP, RednerIn, Partei, gehalten/zu Protokoll gegeben, Zwischenrufe
    Bsp. GermaParl (Blätte 2018), ParlSpeech (Rauh et al. 2017)
  - Social Media Posts öffentlicher Personen: Bsp. Gesis BT-Wahl 2017 Twitter-Korpus
  - Juristische Texte: Gerichts-Urteile + Begründungen (coming soon)

Datenquellen

- Leider bislang kaum öffentliche Ressourcen dieser Art verfügbar
  - v.a. Copyright Issues, auch Datenschutz
- Derzeitige Praxis: individuelle Neubeschaffung speziell zusammengestellter Artikelsammlungen aus vorgenannten Vollkorpora (teilw. mit erheblichen finanziellen Aufwänden)
  - keine Veröffentlichung der Forschungsdaten
  - keine Reproduzierbarkeit / Sekundäranalysen

Utopie

- gepflegte „wachsende“ Vollkorpora aller wichtigen dt. Tages-/Wochenzeitungen in einem gemeinsamen Repository
  - Zentrale Verwaltung / Pflege durch die relevanten Fachinformationsdiensten der DFG
  - Nationallizenzen für freie wissenschaftliche Nutzung (Anbindung an DFN-User-Authentifizierung)
  - Vergütung für Verlage aus einem Fonds aus BMBF-Mitteln