Schreiben über NLP-Experimente (Fortsetzung) / Qualitätssicherung

HS Sprachtechnologie für eine bessere Welt (Winter semester 2021/22)

Nils Reiter,
nils.reiter@uni-koeln.de

January 18, 2022

Section 1

Schreiben über NLP-Experimente (Fortsetzung)

Schreiben über NLP-Experimente

- Schreiben ist Arbeit und braucht Zeit
- Iterativer Prozess aus Schreiben, lesen, überarbeiten, lesen, überarbeiten, ...
- (Wiss.) Schreiben ist eine Fähigkeit, die man lernen kann und muss
- Schreiben ist individuell: Mit der Zeit weiß man wie man funktioniert
- Kompetenzzentrum Schreiben: `https://schreibzentrum.phil-fak.uni-koeln.de`

Schreiben über NLP-Experimente

- Schreiben ist Arbeit und braucht Zeit
- Iterativer Prozess aus Schreiben, lesen, überarbeiten, lesen, überarbeiten, ...
- (Wiss.) Schreiben ist eine Fähigkeit, die man lernen kann und muss
- Schreiben ist individuell: Mit der Zeit weiß man wie man funktioniert
- Kompetenzzentrum Schreiben: `https://schreibzentrum.phil-fak.uni-koeln.de`

Verschiedene Aspekte

- Formalia (Layout, Zitate, Rechtschreibung)
- Stil
- Inhalt

Inhalt

- Leitlinien: Reproduzierbarkeit und Transparenz
- Experimente sollen so dokumentiert sein, dass sie überprüfbar sind
- Regelfall: Nicht alles was wir gemacht haben, landet im Artikel

Inhalt

Forschungsstand

- NLP-Papiere berichten über Fortschritt für eine bestimmte Aufgabe
- Forschungsstand gibt wieder, was man vor dem vorliegenden Papier wusste
- Konkret genannt werden Arbeiten:
  - die sich mit exakt dem gleichen Problem beschäftigt haben
  - die sich mit einem strukturell ähnlichen Problem beschäftigt haben
  - die eine Methode präsentieren, die für unser Problem einsetzbar sein könnte
- Fokus auf Methoden, nicht Tools

Inhalt

Häufige Probleme

- Unvollständige/ungenaue Informationen
- Abweichungen von der Gliederung
  - Z.B.: Im Abschnitt ‚Forschungsstand' geht es nicht um die eigene Arbeit
- Falsch verwendete Fachbegriffe
- Fehlende Konzentration auf das, was relevant ist
  - Nebenbei werden Fässer aufgemacht, die gar nicht nötig sind
- Zu wenig Abstraktion: Implementierungsdetails haben in NLP-Texten nichts verloren
- Selten: Experimente die nichts zeigen können, weil der Aufbau nicht durchdacht wurde

(Meine) Tipps

- Überarbeiten – lesen – überarbeiten – lesen – überarbeiten – lesen – ...
- Bis zur Deadline ist alles im Fluss
- Man muss nicht auf Anhieb perfekte Sätze hinschreiben
- Nicht: Vor ein leeres Dokument setzen und Text hinschreiben
- Erst Notizen (was will ich eigentlich sagen?) machen, dann ausformulieren
- Den Text mal eine Woche liegen lassen und dann wieder lesen
- Am Anfang Kapitel schreiben, bei denen man weiß was man schreiben muss (Experimente)
- Einleitung und Schluss als letztes schreiben
- Denglish: Ein ML/NLP-Text ist durchsetzt von englischen Begriffen
- Motivation erhalten, indem man kleine Ziele definiert

Abbildung: PhD Comics `https://phdcomics.com/comics.php?f=1785`

Organisatorisches

- Nächste Woche kein Seminar!
- 01.02.: Letzte Sitzung des Semesters
  - Abschlussdiskussion
  - Weise Worte
  - Feedbackrunde

Section 2

How to Measure Quality in NLP?

Motivation

- Impact of NLP methods/applications on the world
- How to systematically consider that in research?

Established Evaluation in NLP

Intrinsic

- Comparison with a gold standard
- Precision/recall/f-score/accuracy/...
- Goal: Replicate the gold standard

Established Evaluation in NLP

Intrinsic

- Comparison with a gold standard
- Precision/recall/f-score/accuracy/...
- Goal: Replicate the gold standard

Extrinsic

- Integrate your method into a larger system
- Evaluate this larger system
  - Against a gold standard
  - Post-hoc with human evaluators

Established Evaluation in NLP

Intrinsic

- Comparison with a gold standard
- Precision/recall/f-score/accuracy/...
- Goal: Replicate the gold standard

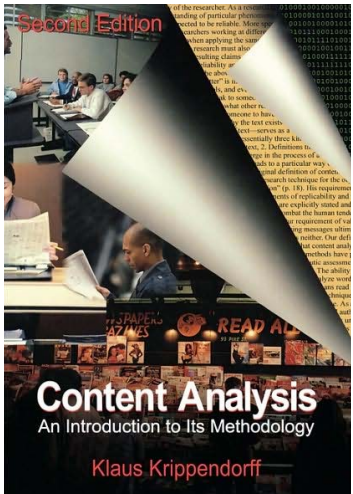
Extrinsic

- Integrate your method into a larger system
- Evaluate this larger system
  - Against a gold standard
  - Post-hoc with human evaluators

Further considerations

- Is this method *significantly* better than an alternative?
- Comparison against one or more baseline(s)
- Cost-benefit-analysis (e.g., learning curve)

Quality Assurance in Social Sciences



Klaus Krippendorff (2004). *Content Analysis: An Introduction to its Methodology*. 2nd. Los Angeles, California, USA: Sage

Two chapters available in Ilias

Two aspects of quality:

- Reliability
- Validity

Assumptions

- Data: Measurement results
- No strict separation between manual annotation and automatic prediction
- Content analysis: Analyzing large volumes of text (or data) with an interest in (some aspect of) the content
  - I.e.: Not purely methodological interest
  - Not an interest in the words, but the meaning of the words

Reliability

[...] content analysts must be confident that their data (a) have been generated with all conceivable precautions in place against known pollutants, distortions, and biases, intentional or accidental, and (b) mean the same thing for everyone who uses them. Reliability grounds this confidence empirically.
[...] a research procedure is reliable when it responds to the same phenomena in the same way regardless of the circumstances of its implementation. (Krippendorff, 2004, 211)

Reliability Designs

- Three types of reliability (and three ways to test it)
- Generate ‚reliability data' – in addition to the data whose reliability is in question

Reliability	Designs	Causes of Disagreements	Strength
Stability	test-retest	intraobserver inconsistencies	weakest
Reproducibility	test-test	intra+inter	medium
Accuracy	test-standard	intra+inter+deviations from standard	strongest

Table: Types of Reliability (Krippendorff, 2004, 215)

Validity

> *A measuring instrument is considered valid if it measures what its user claims it measures.* *(Krippendorff, 2004, 313)*

- Validity is not a quantitative test we can apply
- Validity is the result of a process of validation, an argumentation
  - A researcher makes arguments for the validity
  - An audience may be skeptical about some or all of them
- Krippendorff differentiates several sources of validity or ways of validating

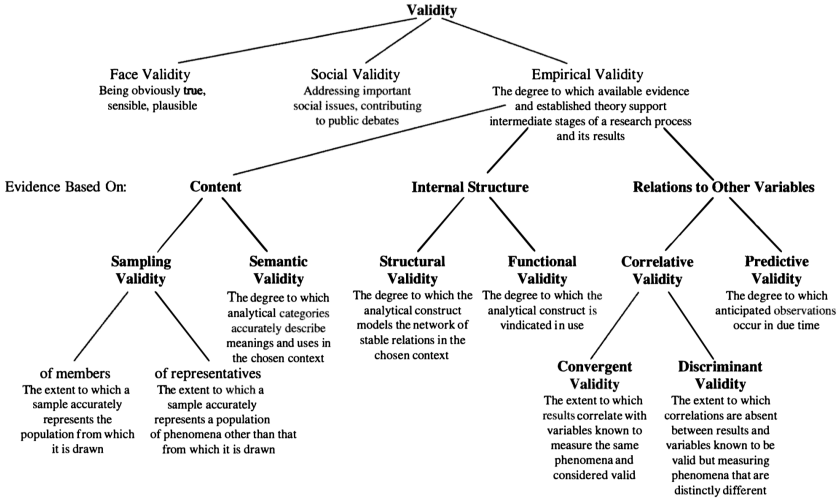
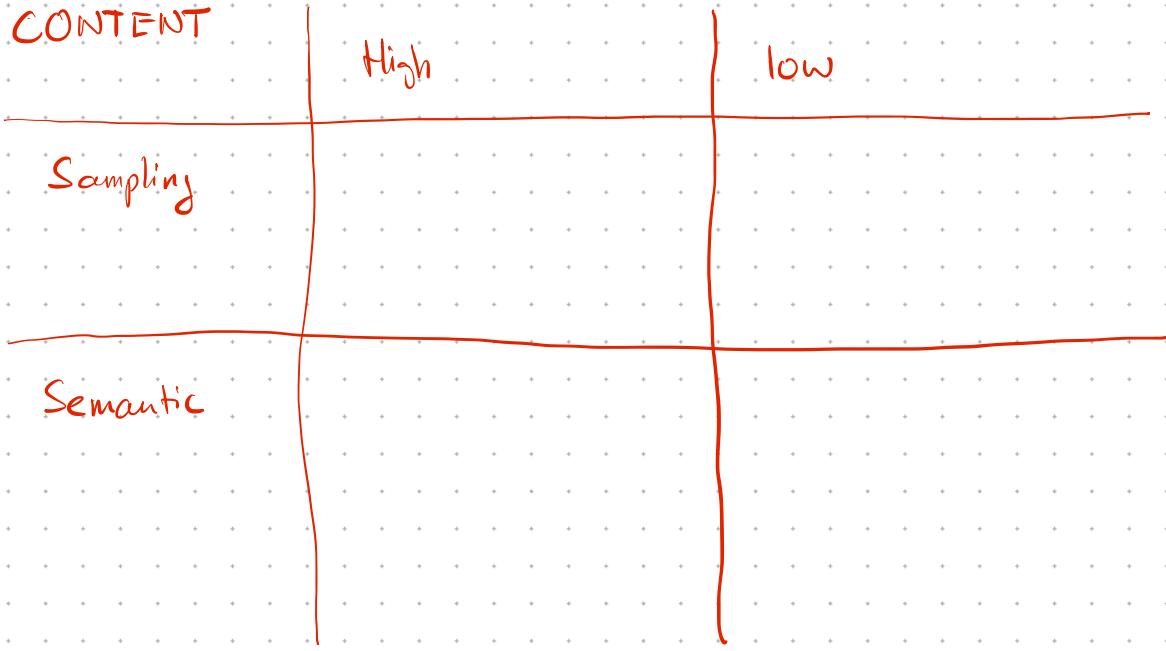


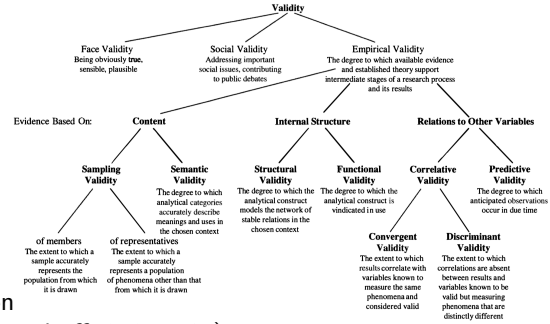
Figure: A Typology of Validation Efforts (Krippendorff, 2004, 319)


CONTENT
High
low
Sampling
Semantic

STRUCTURE	HIGH	LOW
Structural		
Functional		

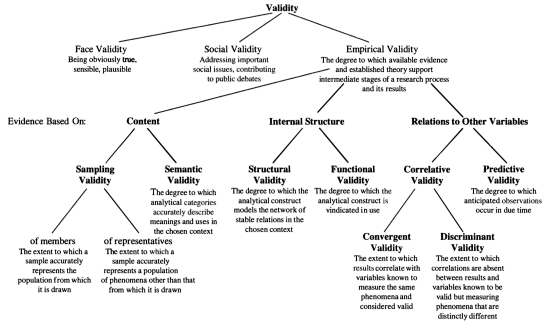
OTHER VARIABLES	HIGH	LOW
Correlative		
Predictive		

Social and Face Validity



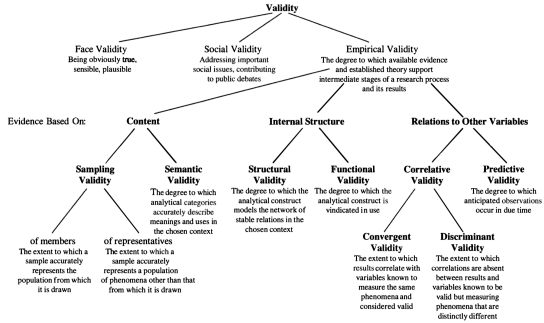
- Face Validity
  - Obvious or common truth
  - „an individual's judgment with the assumption that everyone else would agree with it“ (Krippendorff, 2004, 314)
- Social Validity
  - Research is valuable for the society
  - „[S]ocial validity of content analysis studies is often debated, negotiated, and a matter of public concern“ (Krippendorff, 2004, 314)

Evidence Based on Content



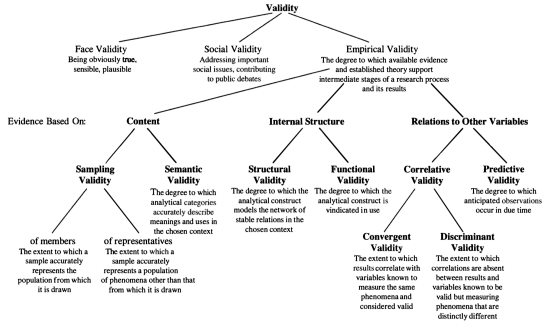
- Sampling validity
  - Does the sample we investigate accurately represent the population we want to say something about?
  - Ideally: Use sampling methods that ensure representativeness
  - Reality: Not always controllable
- Semantic validity
  - To which extent do the categories we investigate correspond to the meanings of the text?
  - Do our categories represent the meaning of interest at all?

Evidence Based on Internal Structure



- Structural Validity
  - Does the structure of the content analysis (i.e., the different components together) accurately represent the domain?
  -
- Functional Validity
  - Is the analytical construct established and useful?
  - „[O]ne must demonstrate that its analytical constructs [...] are useful over time and in many empirical situations.“(Krippendorff, 2004, 332)

Evidence Based on Relations to Other V



- Correlative Validity
  - Validity ‚travels' along high correlations
  - If two variables are highly correlated, and one's validity is established, the other is considered valid as well
- Predictive Validity
  - To which degree do content analysis methods accurately predict events, identify properties etc.?
  - How well can we actually predict previously unknown things?

Reliability and Validity

> *whereas reliability provides assurances that particular research results can be duplicated [...] validity provides assurances that the claims emerging from the research are borne out in fact.* *(Krippendorff, 2004, 212)*

- Unreliability limits the chance of validity

Reliability and Validity

> *whereas reliability provides assurances that particular research results can be duplicated [...] validity provides assurances that the claims emerging from the research are borne out in fact.* *(Krippendorff, 2004, 212)*

- Unreliability limits the chance of validity
- Reliability does not guarantee validity

Reliability and Validity

> *whereas reliability provides assurances that particular research results can be duplicated [...] validity provides assurances that the claims emerging from the research are borne out in fact.* *(Krippendorff, 2004, 212)*

- Unreliability limits the chance of validity
- Reliability does not guarantee validity
- In the pursuit of high reliability, validity tends to get lost
  Example: Merritt (1966)
  - Study about rising national consciousness in 13 American colonies
  - No operationalization of „national sentiment“
  - Instead: Enumerate and count American place-names

Group Exercise

- Let's make this more concrete
- What are examples related to NLP with low and high validity?
- Find examples for as many validity types as possible

Validity Types

Sampling, semantic, structural, functional, correlative, predictive

References I

- Krippendorff, Klaus (2004). *Content Analysis: An Introduction to its Methodology*. 2nd. Los Angeles, California, USA: Sage.
- Merritt, Richard L (1966). *Symbols of American Community, 1735-1775*. Yale University Press.