Statistik
01. Inferenz

Roland Schäfer

Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

stets aktuelle Fassungen: `https://github.com/rsling/VL-Deutsche-Syntax`

Inhalt

Probative Wissenschaft

Empirisch, objektiv, realistisch, exakt(?)

- beobachtbare Phänomene
- Beobachtungen reproduzierbar
- messbar = beobachtbar (Sinneswahrnehmung an sich irrelevant)

- Realismus | wirkliche Phänomene und ihre Mechanismen
- keine postmoderne Realitäts- und Objektivitätsverweigerung

- kontrolliertes Experiment

Empirie | Gründe für Reproduzierbarkeitsbedingung

- intrinsische Ungenauigkeiten der Messung (Wirkung plus Störeinflüsse)
- potentiell inadäquate Messung des theoretischen Konstrukts

→ Vermeidung von Fehlschluss auf unechte Ursachen

→ relevante Ursachen

→ insgesamt Stärkung der Validität

Beispiel | Fehlgeleitete Generative Grammatiker

- Gegenstand: interne (mentale) Grammatik (I-Grammatik)
  universeller und individueller Teil
- I-Grammatik bei jedem Sprecher (leicht) verschieden
- I-Grammatik erlaubt immer binäre Grammatikalitätsentscheidung

→ Linguisten können eigene I-Grammatik untersuchen (Introspektion)!

Das Ergebnis ist die aktuelle Krise der Linguistik.

(Logischer) Positivismus

Formale Ableitung von Wissen (= Theorien) aus Beobachtbarem und irgendeiner Logik. Induktion. Keine Metaphysik. Keine Kreativität erwünscht. (Carnap 1928, ...)

Aber suchen wir wirklich nur nach Mustern, z. B. in Korpusdaten?

- Was ist der zugrundeliegende Mechanismus?
- Wie kommen wir zu erklärenden Theorien von Mustern in Daten?
- Datenaufbereitung (z. B. im Korpus) kann dann nicht theoriegeleitet sein.
- Die ART folgt auch nicht einfach so aus Daten!

Rationalistischer Probativismus

Theorien werden aufgestellt von Menschen, die die Welt beobachten. Theorien werden getestet an Daten, aber nicht logisch aus Daten abgeleitet. (Popper 1962, Mayo 1996, ...)

Unter dieser Philosophie werden plötzlich Dinge wichtig ...

- Ist eine Stichprobe repräsentativ für das, was man zeigen will?
- Welche Methode der statistischen Analyse wird verwendet?
- Für eine Korpusstudie muss die Datenaufbereitung damit theoriegeleitet sein!
- Liefert die Studie *a serious Argument from Error*?

  There is evidence an error is absent to the extent that a procedure with a very high capability of signalling the error, if and only if it is present, nevertheless detects no error. (Mayo 2018: 16)

Hypothesen

Die konkrete Hypothesen, die in einem Experiment getestet werden,
sind nie die Primärhypothesen der Theorie.

- abgeleitete Partikularhypothesen über konkrete Erwartungen im Experiment
- Einfluss zahlreicher Auxiliarhypothesen, z. B. über Messprozeduren
  Duhem (1914), Quine (1951), Laudan (1990)

- „interessante“ Hypothesen
  - Formulierung relevanter Kausationsbedingung (wenn, dann)
  - universelle Gültigkeit | ein Sprecher vs. alle Sprecher
  - also z. B. uninteressant | *Welchen Kasus nimmt wegen?*

Kann die Hypothese weiter angenommen werden,
oder liefert das Experiment starke Evidenz gegen sie?

- Probleme bei Prüfung
  - falsch abgeleitete Partikularhypothese
  - falsche Sekundärhypothesen
  - Störeinflüsse, intrinsische Messungenauigkeit
  - mangelhafte Operationalisierung
  - zu wenige Daten (oder zu viele Daten?)

Elemente der Empirie

Grundgesamtheit

- von Interesse | allgemeine Gesetzmäßigkeiten
- also Untersuchungsgegenstand: alle x (Sprecher, Sätze, ...)
- untersuchbar | kleine Menge von x

Grundgesamtheit | alle x

datengenerierender Prozess (DGP) | Prozess, der alle x hervorbringt

Stichprobe | eine kleine Menge x, aus der auf Grundgesamtheit bzw. DGP geschlossen werden soll

Stichprobe

uniform zufällige Stichprobe

jedes Element der Grundgesamtheit hat die gleiche Chance beim Ziehen

stratifizierte Stichprobe

Stichprobe so zusammengesetzt, dass wichtige Eigenschaften proportional repräsentiert sind

- Problem bei Letzterem: haufenweise Auxiliarhypothesen

Operationalisierung und Auxiliarhypothesen

- Operationalisierung | präzise Formulierung der Messmethode für ein theoretisches Konstrukt

- Bsp. Konstrukt „Satzlänge“: Wortanzahl? Phonemanzahl? Phrasenanzahl?
- Bsp. Konstrukt „Satztopik“: Oha!?! (Cook & Bildhauer 2013)

- alle genannten Beispiele: abhängig von Auxiliarhypothesen bzw. anderen theoretischen Konstrukten (Wort, Phonem, Phrase, ...)

Variablen I

- uninteressanter Typ Fragestellung | „Wieviel Prozent X haben Eigenschaft A?“
- Fehlen jeglicher Aussagen über kausale Zusammenhänge
- Bsp. | Wie oft wird *wegen* mit Dat bzw. Gen verwendet?

- Besser | „Wie bedingt Eigenschaft B die Wahrscheinlichkeit von A bei X?“
- Bsp. | Per Hypothese nehmen denominale Präpositionen eher den Gen als den Dat.

konzeptuell:

	denominale P	andere P
Dat	x_1	x_2
Gen	x_3	x_4

Variablen II

Operationalisierte und gemessene Eigenschaften sind Variablen.

- im Experiment:
  - kontrolliere für Theorie irrelevante Variablen (Störvariablen)
  - variiere „Ursachen-Variablen“ (unabhängige Variablen)
  - beobachte „Wirkung-Variablen“ (abhängige Variablen)

Experiment und Quasi-Experiment

- Problem in Astronomie, Korpuslinguistik usw. | keine Experimente möglich
- unabhängige Variablen nicht variierbar
- Daten liegen bereits vor bzw. fallen vom Himmel
- Auswahl von Datensätzen, so dass von den unabhängigen Variablen die zur Theorieprüfung nötigen Permutationen im Datensatz vorkommen
- dabei Zusatzproblem bei Korpuslinguistik: Korpus meist nicht das eigene, wenig Informationen über mögliche Verzerrungen

- Was ist die Grundgesamtheit bzw. der DGP?

Validität

Statistische Validität

Gefahren für statistische Schlussverfahren

- mathematische Vorbedingungen für das Testverfahren nicht
- zu viele Partikulartests einer übergeordneten Hypothese aus denselben Daten
- zu kleine Stichprobe
- zu große Variation in der Grundgesamtheit

- bei Korpora | schlechte Zusammensetzung des Korpus

Interne Validität

- Irrtum beim Herstellen des Kausalzusammenhangs
- Fiktives Bsp.:
  - Hypothese | Im DECOW2012 kommt öfter das Pronomen *son* vor als im DWDS Kernkorpus, weil es erst nach 2000 zum eigenständigen Pronomen wurde.
  - Die Hypothese wird bestätigt anhand von Stichproben aus den beiden Korpora.
  - Die wirkliche Ursache sind aber Registerunterschiede.

Validität des Konstrukts

- Korrektheit des theoretischen Konstrukts

- eigentlich aus der Psychologie
- aber riesiges Problem in der Linguistik

- Echtes Bsp.
  - Beobachtung | Das Deutsche bewahrt genus-typische Pluralflexion am Substantiv.
  - Konstrukt | Nominalklammer/Klammerprinzip (NP-Kongruenzklammer Art – Subst) (Ronneberger-Sibold 2010)
  - Hypothese zu Beobachtung | Flexionserhalt stärkt Klammerprinzip
  - Das Konstrukt ist hochgradig beliebig und unterdefiniert, damit nicht testbar.
  - Abhilfe: nur Konstrukte/Hypothesen, die starke Vorhersagen generieren

Externe Validität

- Generalisierbarkeit der Ergebnisse (über Raum, Zeit usw.)
- Problem | zu große Homogenität der Stichprobe
  (was für statistische Validität wiederum gut ist)
- Bezug auf Korpora:
  - zu spezifische Stratifikation (DeReKo)
  - verzerrte Stichprobe (Webkorpora)

Ronald A. Fisher, Wahrscheinlichkeit, Ereignisraum, Teetassen

Ronald A. Fisher (1890–1962)

- Statistik als Teil der rationalen wissenschaftlichen Argumentation, der Interpretation von Experimenten
- daher: möglichst kein Mathematik-Jargon
- eingeschränkte Induktion als theoriegeleitete Dateninterpretation
- Kontrolle aller unabhängigen Variablen
- alle anderen (Stör-)Variablen konzeptuell zufallsgebunden

Das Tassen-Problem

- Behauptung: Dame X kann am Geschmack erkennen, ob der Tee oder die Milch zuerst in die Tasse gegossen wurde.
- prä-fishersches Konzept: alle Störvariablen kontrollieren und gleich machen, sonst keine valide Inferenz möglich
- Fisher: Das ist prinzipiell unmöglich, umständlich, teuer, unnötig!
- wenn alle irrelvanten Stör-Faktoren zufällig, dann:
  - Variiere die relevante unabhängige Variable.
  - Vergleiche das Ergebnis mit zufällig erwartbaren Ergebnissen.
  - Wie unwahrscheinlich ist das erzielte Ergebnis unter der Zufallsannahme?

Wahrscheinlichkeiten, richtig zu raten

- acht Tassen (zwei Milch zuerst, zwei Tee zuerst)
- Mit wie vielen richtigen Treffern wären Sie zufrieden?
- Es muss die Wahrscheinlichkeit errechnet werden, eine, zwei, drei oder vier Tassen richtig zu raten.

- Typischerweise schätzen Menschen solche Kombinatorikprobleme intuitiv falsch ein.

GENAU Richtig durch Zufall

$$P(\text{richtig per Zufall}) = \frac{\text{Anzahl richtiger Zuweisungen}}{\text{Anzahl aller potentiellen Zuweisungen}} \tag{1}$$

- Anzahl richtiger Zuweisungen: 1
- mögliche Zuweisungen: einfaches kombinatorisches Problem

mögliche Zuweisung von acht Tassen zu Milch/Tee zuerst

- erste MZ-Tasse: eine von 8
- zweite MZ-Tasse: eine von 7
- dritte MZ-Tasse: eine von 6
- vierte MZ-Tasse: eine von 5
- fünfte MZ-Tasse: STOPP (automatisch TZ)
- Möglichkeiten, 4 Tassen aus 8 auszuwählen:
  $8 \cdot 7 \cdot 6 \cdot 5 = 1680$

Reihenfolge egal

- bisher: jedes Set von 4 MZ-Tassen ist in verschiedenen Reihenfolgen in der Menge der Möglichkeiten
- Möglichkeiten, vier Tassen zu ordnen:
  $4 \cdot 3 \cdot 2 \cdot 1 = 24$

- Reihenfolge hier egal, also:

$$\text{Anzahl aller potentiellen Zuweisungen} = \frac{1680}{24} = 70 \tag{2}$$

Wenn Dame X also genau richtig liegt

- per Zufall genau richtig:
  in einem von 70 Fällen, $p = 0.014$
- α-Niveau von 0.05 also erreicht

Binomialkoeefizienten

- eigentlich Binomialkoeffizient
- „Lotto-Kombinationen“: k aus n
  ohne Zurücklegen und ohne Beachtung der Reihenfolge

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{3}$$

Kombinatorik für 3 richtige MZ

- die drei richtigen: $\binom{4}{3}$
- die eine falsche aus vier TZ: $\binom{4}{1}$
- Anzahl der Möglichkeiten drei richtige aus vier MZ
  und dann eine falsche aus vier TZ zu ziehen: $\binom{4}{3} \cdot \binom{4}{1} = 4 \cdot 4 = 16$

$$P(\textit{drei richtig per Zufall}) = \frac{16}{70} = 0.229 \tag{4}$$

Problem zu kleiner Stichproben

- Bei $\alpha = 0.05$ reichen also drei richtige nicht im Ansatz!
- alle schlechteren Ergebnisse: folglich auch nicht ausreichend
- Und bei 30 von 40 richtigen (also insgesamt 80 Tassen)?

Nächste Woche | Überblick

Einzelthemen

1. Statistik, Inferenz und probabilistische Grammatik
2. Deskriptive Statistik
3. Nichtparametrische Verfahren
4. z-Test und t-Test
5. ANOVA
6. Freiheitsgrade und Effektstärken
7. Power
8. Lineare Modelle
9. Generalisierte Lineare Modelle
10. Gemischte Modelle

Literatur I

Carnap, Rudolf. 1928. *Der logische Aufbau der Welt*. Berlin: Weltkreis Verlag.

Cook, Philippa & Felix Bildhauer. 2013. Identifying “aboutness topics”: two annotation experiments. *Dialogue and Discourse* 4(2), 118–141.

Duhem, Pierre. 1914. *La Théorie Physique: Son Objet et sa Structure*. Marcel Riviera & Cie.

Laudan, Larry. 1990. Demystifying Underdetermination. In C. Wade Savage (Hrsg.), *Scientific Theories*, 267–297. Minneapolis: University of Minnesota Press.

Mayo, Deborah G. 1996. *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Popper, Karl Raimund. 1962. *Conjections and Refutations: The Growth of Scientific Knowledge*. New York: Basic Books.

Quine, Willard Van Orman. 1951. From a Logical Point of View. In 2. Aufl. Cambridge: Harvard University Press. Kap. Two Dogmas of Empiricism, 20–46.

Ronneberger-Sibold, Elke. 2010. Der Numerus – das Genus – die Klammer : die Entstehung der deutschen Nominalklammer im innergermanischen Vergleich. In Antje Dammel, Sebastian Kürschner & Damaris Nübling (Hrsg.), *Kontrastive Germanistische Linguistik. Teilband 2*, Bd. 206/209 (Germanistische Linguistik), 719–748. Hildesheim: Olms.

Autor

Kontakt

Prof. Dr. Roland Schäfer
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena
Fürstengraben 30
07743 Jena

https://rolandschaefer.net
roland.schaefer@uni-jena.de

Lizenz