Müssen Texte statistisch anders ausgewertet werden als Menschen?

Stella Neumann

Grundüberlegung I

- Empirische Methode am besten für Soziologie und Psychologie (also Humanwissenschaften) entwickelt
- Über Sprachlehrforschung und Psycholinguistik in die Sprachwissenschaft übernommen, dabei ebenfalls Mensch als Untersuchungsgegenstand
- In der Korpuslinguistik Texte (= Produkt menschlicher sprachlicher Äußerung) als Untersuchungsgegenstand

→ Wie sehr sind die Methoden, die für die Untersuchung von Menschen entwickelt wurden, auf die Untersuchung von Texten übertragbar?

Grundüberlegung II

Unterschiedliche Datengrundlage in Humanwissenschaften und in Korpuslinguistik:

	Humanwissenschaften	**Korpuslinguistik**
Untersuchungs-gegenstand	Menschl. Verhalten	Sprachliche Äußerungen
Grundgesamt-heit	z.B. im Telefonbuch verzeichnete Bewohner oder auf dem Campus erreichbare Menschen	z.B. im Verzeichnis Lieferbarer Bücher verzeichnete Bücher oder im Internet abrufbare Texte
Auswahl der Stichprobe aufgrund…	Alter, Bildung, Einkommen, Muttersprache etc.	Textsortenzugehörigkeit, Erscheinungsdatum, Autor, Muttersprache, etc.
Daten	Äußerungen/Einschätzungen von Versuchspersonen	Häufigkeiten in sprachlichen Einheiten, Verhältnismaßzahlen

Grundgesamtheit

- [Humanwissenschaft: im Telefonbuch verzeichnete Bürger]
- Bücher aus dem Verzeichnis Lieferbarer Bücher
  - Vorteil: gut eingrenzbare, sinnvolle Menge an aktuellen Texten
  - Nachteil: enthält nur Bücher; viele oft gelesene Texte (Gebrauchsanweisungen, Zeitungsartikel, Rechtstexte) sind nicht darin enthalten
- Im Internet verfügbare Texte
  - Vorteil: große Bandbreite an verfügbaren Texten
  - Nachteil: Ungeklärte Herkunft und Qualität der Texte; mengenmäßige Eingrenzung der Grundgesamtheit → Ziehung einer Zufallsstichprobe nicht möglich

→ Wie repräsentativ sind die Texte, die Korpuslinguisten verwenden?

Was ist ein Fall?

- [Humanwissenschaften: Mensch]
- Wort?
- Satz?

} Gegenstand der Annotation

- Text bzw. Textausschnitt? → Einheit, aus der sich das Korpus zusammensetzt
- Subkorpus (im Vergleich zu anderen Subkorpora)? → Einheit, über die etwas ausgesagt werden soll

→ Was ist nun ein Fall?

Skalenniveau?

Bezeichnung	Erläuterung	Beispiel
Nominalskala	willkürliche Vergabe von Zahlen zur Unterscheidung verschiedener Fälle; die Abstände zwischen den Zahlen sagen nichts aus	Registerzuordnung: Essay = 1, Fiction = 2, etc.
Ordinalskala	Das Objekt mit der größeren Merkmalsausprägung erhält die größere Zahl; Zahlen bilden eine Rangordnung	Schulnoten: 1 ist besser als 2, ist besser als 3…, aber der Abstand zwischen den Noten ist unklar
Intervallskala	Information über Abstände zwischen den Zahlen, aber keinen „echten“ Nullpunkt	Temperatur: Abstand zwischen 0 und 10° Celsius ist genauso groß wie der zwischen 10 und 20°
Verhältniskala	Der Abstand zwischen den Zahlen ist gleich und es gibt einen sinnvoll interpretierbaren Nullpunkt	Längenmessung: Ein 10 cm langes Brett plus ein 20 cm langes Brett ergibt ein 30 cm langes Brett ☺

Trifft auf Häufigkeitsauswertungen in einem Korpus zu.
Was aber wenn zwei Register verglichen werden?

Konsequenzen für korpusbasierte Untersuchungen

- Skalenniveau bestimmt Auswahl an Signifikanztests
  Bei Tests für niedrigere Skalenniveaus wie dem Chi^2-Test gehen sonst nutzbare Informationen verloren

→ Wird das in der Korpuslinguistik für jede Untersuchung geklärt?

Unabhängige Gruppen oder wiederholte Messungen?

- Unabhängige Gruppen (= between subject)
  Es werden zwei verschiedene Stichproben ausgewertet

  Einsatz unterschiedlicher Lehrmethoden:
  Lerngruppe 1 übt Leseverständnis mit einer traditionellen Methode,
  Gruppe 2 mit einer neuartigen Methode
- Wiederholte Messungen (= within subject)
  Die gleiche Stichprobe wird in verschiedenen Zuständen ausgewertet

  Einfluß von Alkohol auf die Fahrtüchtigkeit:
  Die gleichen Versuchspersonen müssen in nüchternem Zustand und in alkoholisiertem Zustand Verkehrshütchen umfahren

→ Sind Übersetzungen andere Zustände von Originaltexten?

Beispiele aus Albert & Koster 2002

Schlußfolgerung

- Liegt das alles nur daran, dass wir nicht genug über Statistik wissen?
- Oder brauchen wir für Auswertungen von Texten einen anderen Ansatz?