Korpuslinguistik und Statistik

(050370)

Fabian Barteld, M.A.
Fabian.Barteld@ruhr-uni-bochum.de

7. Sitzung

Werkzeugkasten Korpuslinguistik

Korpusaufbau

- Theorie: Korpuserstellung
  - repräsentativ
  - ausgewogen
- Metadaten und Annotation
- Struktur, Größe
- Persistenz

Werkzeugkasten Korpuslinguistik

Korpusauswahl

- wichtige deutschsprachige Korpora
  - DeReKo
  - DWDS-Kernkorpus
  - Bonner Frnhd. Korpus
  - Falko
  - (Archiv für Gesprochenes Deutsch)

Werkzeugkasten Korpuslinguistik

Korpuserstellung

- Texte einfach durchsuchen
  - AntConc
- Texte aufbereiten
  - Tokenizer
  - Tree-Tagger/ STTS
  - Tabellen (CSV)

Werkzeugkasten Korpuslinguistik

Korpusverwendung

- Programme zur Korpusverwaltung
  - Cosmas II
  - Annis
- Erstellen von einfachen Suchanfragen

Literatur

Lüdeling, A. & Kytö, M. (Hrsg.). (2009). *Corpus Linguistics. An International Handbook*. Handbücher zur Sprach- und Kommunikationswissenschaft 35. Berlin: Mouton de Gruyter

Statistik

Bedeutung von *Statistik*

1. Teilgebiet der Mathematik, das sich mit Daten beschäftigt
2. **ein mathematisches Objekt**

Definition von Statistik

Statistik: Anweisung, was mit Daten gemacht werden soll

Beispiel: Ordnungsstatistik

Würfelergebnisse

$$1,3,5,1,2,4,2,3,1,5,6,1$$

$$\downarrow$$

$$1,1,1,1,2,2,3,3,4,5,5,6$$

Einteilung von Statistiken

Grundlage: Wozu verwendet man Statistiken?

1. Daten übersichtlich aufbereiten
   Deskriptive Statistik
   (auch: beschreibende Statistik)
2. Fragen über die Grundgesamtheit beantworten
   Induktive Statistik
   (auch: schließende Statistik)

Beispiel: Wortlänge (I)

	Wort	Länge
1	Wortlänge	9
2	wird	4
3	danach	6
4	bestimmt	8
	...	
76	ein	3
77	keineswegs	10
78	triviales	9
79	Problem	7

(08.12.2012)

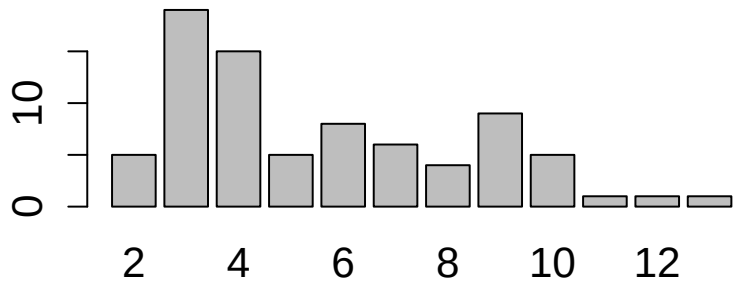
http://de.wikipedia.org/wiki/Wortlänge

Deskriptive Statistik

- **Basis**: nur die vorhandenen Daten
- **Methoden**: Tabellen, Graphiken, Kennzahlen

Beispiel: Wortlänge (II)

Länge	2	3	4	5	6	7	8	9	10	11	12	14
Häufigkeit	5	19	15	5	8	6	4	9	5	1	1	1



durchschnittliche Wortlänge:
5
(Median)
Streuung:
5
(Interquartilsabstand)

Induktive Statistik

- **Basis**:
  - die vorhandenen Daten
  - ein Modell der Grundgesamtheit
    ← Wahrscheinlichkeitstheorie
- **Methoden**: Schätzer, Tests, (Modelle)

Beispiele: Wortlänge (III)

- Was ist die durchschnittliche Wortlänge in der deutschen Sprache?
  $\longrightarrow$ Schätzer
- Machen Wortformen, die kürzer als 6 sind, 50% der Wortformen in deutschen Texten aus?
  $\longrightarrow$ Test

Was sind Daten?

- als Daten für Statistiken dienen **Beobachtungen**
  diese entsprechen den **Belegen** in der Korpuslinguistik
- die interessierenden Eigenschaften der Belege werden in **Variablen** festgehalten
- Beispiele: Wortart, Länge, ...
- die Werte, die eine Variable annehmen kann, werden **Ausprägungen** genannt
- Ausprägungen:
  Substantiv, Verb, Adjektive, ...
  1, 2, 3, ...

Gruppenarbeit

Welche Variablen sind für Ihre Untersuchung relevant?
Welche Ausprägungen dieser Variablen sind möglich?

10 Minuten

Klassifikation von Variablen

... nach Art der Ausprägungen

- nominale Daten
  Wortart, Muttersprache
- komparative Daten
  Akzeptanz, Schulnoten
- metrische Daten
  Wortlänge, Worthäufigkeit

Klassifikation von Variablen

... nach Rolle in der Untersuchung

- unabhängige Variable(n)
- abhängige Variable(n)

Ausblick

- Datenaufbereitung
- Deskriptive Statistik
- Wahrscheinlichkeitstheorie
- Schließende Statistik
- Arbeit mit der Software *SOFA* (Statistics Open For All)
  http://www.sofastatistics.com/home.php