Korpuslinguistik und Statistik

(050370)

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

2. Sitzung

Materialien zur Übung

http://homepage.ruhr-uni-bochum.de/fabian.barteld/korpus/

Benutzername:	korpus
Passwort:	korpus

Was ist ein Korpus

- Sammlung von Texten ...
- ... für die sprachwissenschaftliche Auswertung
  - zusammengestellt für eine spezifische Untersuchung
  - mit Metadaten/ Annotation versehen
  - persistent

Mögliche Strategien zur Korpuserstellung

- Die Sprache von Thomas Manns *Der Zauberberg*
  $\longrightarrow$ vollständig
- Die Sprache der *Bild*-Zeitung von 1952 bis 2012
  $\longrightarrow$ repräsentativ
- Die deutsche Sprache
  $\longrightarrow$ ausgewogen (balanced)

Größe von Korpora

„more data is better data“
(Church, K.W./ Mercer, R.L. (1993). Introduction to the special issue on computational linguistics. Using large corpora. *Computational linguistics*, 19 (1), 1–24)
Probleme:

- Korpuserstellung aufwändig ⟶ Referenzkorpora
- genaue Kenntnis der Daten unmöglich

Welche Daten enthält ein Korpus?

- Primärdaten
- Metadaten
- Annotation

Arten von Metadaten

- Autor/ Urheber + biographische Angaben
- Entstehungszeit
- Entstehungsort (Herkunft des Autors)
- ...
- vgl. Stratifizierung

Arten von Annotation

	Part of Speech	**Flexionsmorphologie**	**Lemma**
Ein	ART	Nom.Sg.Neutr.	ein
kurzes	ADJA	Nom.Sg.Neutr.stark	kurz
Beispiel	NN	Nom.Sg.Neutr.	Beispiel
hilft	VVFIN	3.Sg.Präs.Ind.	helfen
...			

Ausblick

- Arbeiten mit existierenden Korpora/ Textsammlungen
- Konkrete Hinweise zur Erstellung eigener Korpora (wie speichern, durchsuchen, annotieren?)

Literatur zur Wiederholung und Vertiefung

- *Linguistische Korpora* und *Linguistische Annotation und ihre Nutzung*
  in Lemnitzer und Zinsmeister, 2010.
- *Sprache sammeln* und *Korpustexte anreichern*
  in Perkuhn, Keibel und Kupietz, 2012.

Literatur zum Korpusaufbau

- Literary and Linguistic Computing, 8(4), 1993.
  - Biber, D.: Representativeness in corpus design.
  - Atkins, S./ Clear, J./ Ostler, N.: Corpus design criteria.
- Leitner, G. (Hrsg.) (1992). New directions in english language corpora. Methodology, results, software developments Topics in English Linguistics 9. Berlin und New York: Mouton de Gruyter.
  - de Haan, P: The optimum corpus sample size?
  - Clear, J.: Corpus sampling.

Literatur zu bestehenden Korpora (I)

- Leitner, G. (1992). International corpus of english: Corpus design – problems and suggested solutions. In: G. Leitner (Hrsg.) (1992) (S. 33–64).
- Burnard, L. (2002). Where did we go wrong? A retrospective look at the british national corpus. In B. Kettemann & G. Marko (Hrsg.), *Teaching and learning by doing corpus analysis. Proceeding of the fourth international conference on teaching and language corpora, Graz 19–24 july, 2000* (S. 51–71). Language and Computers: Studies in Practical Linguistics 42. Amsterdam und New York: Rodopi.

Literatur zu bestehenden Korpora (II)

- Wegera, K.-P. (2000). Grundlagenprobleme einer mittelhochdeutschen Grammatik. In W. Besch, A. Betten, O. Reichmann & S. Sonderegger (Hrsg.), Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung (2. vollständig neu bearbeitete und erweiterte Auflage, Bd. 2, S. 1304–1320). HSK 2. Berlin und New York: de Gruyter.
- Wegera, K.-P. (2012). Language data exploitation. Design and analysis of historical language corpora. In P. Bennett, M. Durrell, S. Scheible & R. J. Whitt (Hrsg.), *New methods in historical corpus linguistics* (Bd. 3). Corpus Linguistics and Interdisciplinary Perspectives on Language. Tübingen: Gunter Narr [im Druck].

Hausaufgabe

Registrieren Sie sich bei den folgenden Seiten:
http://www.ids-mannheim.de/cosmas2/projekt/registrierung/
http://www.dwds.de/user/login/

2-Minuten Frage

Schreiben Sie 2 Stichwörter auf einen Zettel:

1. Was haben Sie besonders gut verstanden?
2. Was haben Sie nicht/schlecht verstanden?