Korpuslinguistik und Statistik

(050370)

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

4. Sitzung

DWDS-Kernkorpus (www.dwds.de)

- Datengrundlage des *Digitalen Wörterbuchs der Deutschen Sprache*
- Korpus zur Sprache des 20. Jhd.
- ca. 100 Millionen Tokens
- ausgewogen
  (Belletristik, Zeitung, Wissenschaft, Gebrauchsliteratur)
- außerdem ca. 5% gesprochene Sprache
- http://www.dwds.de/ressourcen/kernkorpus/

DWDS-Kernkorpus – Annotation

- strukturell u.a.
  - Seitenumbrüche
  - Fußnoten
  - Kapitel
  - Absätze
- linguistisch u.a.
  - Lemma
  - PoS (STTS)

DWDS-Kernkorpus – Literatur

- Geyken, A. (2007). The DWDS corpus. A reference corpus for the German language of the twentieth century. In C. Fellbaum (Hrsg.), *Idioms and collocations. Corpus-based linguistic and lexicographic studies* (S. 23–41). London: Continuum Publishing Group (http://www.dwds.de/media/publications/text/DWDS-Corpus_Desc4_draft.pdf).
- **Anleitung zur Suche**
  http://retro.dwds.de/HilfeSuche/index

Simples Aufbereiten von Texten (I)

Text mit Annotationen und Metadaten als Tabelle aufarbeiten:
ein Token pro Zeile, eine Annotation pro Spalte

Token	PoS	Lemma	..., z.B. Quelle
Der	ART	die	
Ausdruck	NN	Ausdruck	
Textkorpus	NN	Textkorpus	

Tipp zum Erstellen der Tabelle
In einem Texteditor (oder Word) Leerzeichen durch Zeilenumbrüche ersetzen, die entstandene (Text-)Datei in eine Tabellenkalkulation (z.B. Excel) laden

Simples Aufbereiten von Texten (II)

Hilfsprogramme

- **Tokenisierer**
  (trennt u.a. Klammern und Punkte von den Wortformen)
  z.B. http:
  //www.linguistics.ruhr-uni-bochum.de/~dipper/tokenizer.html
- **Tagger**
  (führt eine Annotation der einzelnen Token durch, meistens PoS)
  z.B. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
  (Part-of-Speech nach STTS; führt auch Tokenisierung und Lemmatisierung durch; Ausgabe ist eine CSV-Datei)

Simples Aufbereiten von Texten (III)

Standards für Annotationen

- strukturelle Annotation: TEI
  http://www.tei-c.org/index.xml
- PoS-Tags: STTS
  http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html

Standardformat für Tabellen

CSV: https://de.wikipedia.org/wiki/CSV_%28Dateiformat%29

Ausblick – weitere Annotationsmöglichkeiten

- Parser
  führt eine syntaktische Analyse von Texten durch
  http://kitt.cl.uzh.ch/kitt/parzu/
  - TIGERCorpus
    (Ein Beispiel für ein geparstes (d.h. syntaktisch annotiertes) Korpus)
    http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
- Exmaralda (v.a. für gesprochene Sprache)
  http://www.exmaralda.org/

Auswerten eines Textes ohne Aufbereitung

AntConc
http://www.antlab.sci.waseda.ac.jp/antconc_index.html

- mit diesem Programm lassen sich Texte durchsuchen und einige einfache Statistiken erzeugen
  (z.B. Worthäufigkeiten)
- Eingabedaten sind der rohe Text oder auch HTML-Dateien
- Als Suchergebnis können Konkordanzen
  (auch: **K**ey-**W**ords **i**n **C**ontext)
  ausgegeben werden

Literatur

Eigenes Korpus in Bubenhofer, 2006–2011

Vorschläge zur Vertiefung

- weitere Programme
  Corpus Workbench, *DB: Filemaker* und *Anhang* in Bubenhofer, 2006–2011
- Reguläre Ausdrücke
  2.4.1 *Reguläre Ausdrücke* in Perkuhn, Keibel und Kupietz, 2012
- XML
  3.5 *Von Rohdaten zum Korpus* in Perkuhn u. a., 2012