Tutorium der DGfS-Sektion Computerlinguistik

KORPUSLINGUISTIK MIT ONLINE-RESSOURCEN

EINE INTERAKTIVE EINFÜHRUNG FÜR LINGUISTEN

Repräsentativität

Stefanie Dipper
Stefan Evert
Heike Zinsmeister

München, 28.1.2011

Repräsentativität

- Sprache besteht aus einer potenziell unendlichen Menge von Sätzen, aber jedes Korpus ist endlich.
- unklarer Status als Stichprobe für eine Sprache
  - Durch externe Faktoren bedingte Frequenzverhältnisse

a. I live in Dayton Ohio. (unwahrscheinlich)

b. I live in New York. (wahrscheinlich)

  - Einfluss von Konventionen und Tabus

Unvollständigkeit

- Korpora decken nicht alle Phänomene ab.
  - Parasitic Gap:
    - Which book did she review without reading?
  - Lange w-Bewegung
    - Mit wem glaubst du, dass Max meint, dass Jonas gesprochen hat?

Relevanz

- Ein Korpus enthält viele Phänomene, die für die zu beschreibende Sprache irrelevant sind.

Verlässlichkeit

- Korpora beinhalten ungrammatische Äußerungen
  - Ich habe fertig. (Trappatoni)
  - immmer (251.000 Google-Treffer, 03/2008)

Umgang mit der methodischen Kritik

- Beschränkung auf das Korpus
  - ist meistens nicht das Ziel!
- Erstellung eines ausgewogenen Korpus
- Berücksichtigung statistischer Methoden zum Verhältnis Stichprobe zu Gesamtheit
  - Überprüfung an mehreren Stichproben

Literaturangaben

Brants, Thorsten (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, p. 224–231.

Brill, Eric (1995). Unsupervised learning of disambiguation rules for part of speech tagging. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, p. 1–13.

Church, Kenneth Ward (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, p. 136–143.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, p. 79- 86.

Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, IMS, University of Stuttgart and SfS, University of Tübingen. http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf

Schmid, Helmut (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop.*