Tutorium der DGfS-Sektion Computerlinguistik

KORPUSLINGUISTIK MIT ONLINE-RESSOURCEN

EINE INTERAKTIVE EINFÜHRUNG FÜR LINGUISTEN

Einführung

Stefan Evert
Stefanie Dipper
Heike Zinsmeister

Berlin, 23.2.2010

Korpus

- eine Sammlung gesprochener oder geschriebener Äußerungen
- typischerweise digitalisiert und maschinenlesbar
- Ebenen eines Korpus
  - Sprachdaten
  - Metadaten
  - Annotationen

(vgl. Lemnitzer & Zinsmeister 2010: 8)

Bestandteile eines Korpus

- Rohdaten
  - die "authentischen" Sprachdaten
  - Video-/Audiodaten, Transkriptionen von gesprochener Sprache, Text
- Metadaten
  - "Daten über Daten"
  - Informationen über Sprache, Datenumfang, Datenformat, Sprecher, Autoren, Erstellungszeit, Annotatoren, Annotationstagsets, ...
- Annotationen
  - linguistische und außerlinguistische Interpretationen
  - Markierung von Wortgrenzen, Wortart, syntaktischen Phrasen und Funktionen, Zeitrelationen, Diskursrelationen, Mimik, Torstand...

Beispiel: Europarl-Korpus

Genehmigung des Protokolls

Das Protokoll der gestrigen Sitzung wurde verteilt.

Wenn keine E ... ch es als genehmig

Herr Präsider ... lten, daß sich Ibrahim R ... könnten Sie den Parlament ... einer Familie im Namen de ... übermitteln.

Sie werden sich erinnern, daß wir im April-Plenum in diesem Haus eine Entschließung verabschiedeten, in der wir Herrn Rugova einluden, vor dem Ausschuß für auswärtige Angelegenheiten zu sprechen.

Rohdaten

redigierte Transkriptionen von Seminardebatten des Europäischen Parlaments und Übersetzungen in 11 offizielle EU-Sprachen

Beispiel: Europarl-Korpus

Genehmigung des Protokolls
<SPEAKER ID=1 NAME="Der Präsid
Das Protokoll der gestrigen Sitzung w

Wenn keine Einwände vorgebracht w
genehmigt.

<SPEAKER ID=2 LANGUAGE="EN" NAME="Spencer">
Herr Präsident, wir haben über Nacht die Meldung erhalten, daß sich Ibrahim Rugova jetzt in Rom befindet, und vielleicht könnten Sie den Parlamentspräsidenten bitten, Herrn Rugova und seiner Familie im Namen des Parlaments unsere besten Wünsche zu übermitteln.
Sie werden sich erinnern, daß wir im April-Plenum in diesem Haus eine Entschließung verabschiedeten, in der wir Herrn Rugova einluden, vor dem Ausschuß für auswärtige Angelegenheiten zu sprechen.

(ep-99-05-06.txt)

Metadaten
- Datum der Debatte
- Sprechername
- teilweise Sprache des Sprechers = Originalsprache des Beitrags (‘turn’)

Beispiel: Europarl-Korpus

- <words>
- <word id="word_1" pos="NN">Genehmigung</word>
- <word id="word_2" pos="ART">des</word>
- <word id="word_3" pos="NN">Protokolls</word>
- <word id="word_4" pos="ART">Das</word>
- <word id="word_5" pos="NN">Protokoll</word>
- <word id="word_6" pos="ART">der</word>
- <word id="word_7" pos="ADJA">gestrigen</word>
- <word id="word_8" pos="NN">Sitzung</word>
- <word id="word_9" pos="VAFIN">wurde</word>
- <word id="word_9" pos="VVPP">verteilt</word>
- <word id="word_9" pos="$.">.</word>

(ep-99-05-06.txt)

Beispiel: Europarl-Korpus

- <words>
- <word id="word_1" po
- <word id="word_2" po
- <word id="word_3" po
- <word id="word_4" po
- <word id="word_5" po
- <word id="word_6" po
- <word id="word_7" po
- <word id="word_8" po

Terminologie
- Token
- Wortart ‘part of speech’
- Wortartenlabel ‘pos tag’
- Definiertes Inventar ‘tag set’
- Quasi-Standard zur Beschreibung der Wortarten des Deutschen: STTS (Stuttgart-Tübingen-TagSet)

Annotationen
- Satzsegmentierung
- Wortsegementierung
- Wortart

(ep-99-05-06.txt)

Kommentar: Primärdaten

- Die Rohdaten des Korpus sind oft nicht mit den Primärdaten identisch, sondern eine Vereinfachung / Abstraktion / Interpretation davon.
- Primärdaten sind
  - das Video-/ Audiosignal einer Sprachaufnahme
  - der (publizierte) Text in seiner physischen Form

Kommentar: Primärdaten

- Audio-/Videosignal
  - Segmentierung
    - kontinuierliches Signal vs. diskrete Zeichen
  - Transkription
    - Pausen, Intotation, Lautstärke
- (Publizierter) Text
  - Font, Farbe
  - Segmentierung
    - Mehrwort-Token (en passent), Trennstrich vs. Bindestrich am Zeilenende (zwei-teilig)
  - Illustrationen
  - Seitenaufteilung
    - z.B. Artikelbeginn auf der Titelseite einer Zeitung, Artikelfortführung auf einer anderen Seite

Kleine Typologie

- Medium
  - gesprochen
  - geschrieben
  - multimodal
- Sprachenauswahl
  - monolingual
  - bi- oder multilingual

Sprachenauswahl

- Paralles Korpus
  - Texte in Originalsprache und Übersetzungen
  - Alignierung
    - Paragraphen
    - Sätze
    - Wörter
- Vergleichskorpus
  - vergleichbare Texte in verschiedenen Sprachen
  - Genre, z.B. politische Texte, Zeitungskorpora

Europarl

- Paralleles Korpus (Philipp Koehn)
  - http://www.statmt.org/europarl/
- Protokolle von Seminardebatten des Europäischen Parlaments
  - in 11 offizielle EU-Sprachen übersetzt
  - Romanisch (Französisch, Italienisch, Spanisch, Portugiesisch), Germanisch (Englisch, Niederländisch, Deutsch, Dänisch, Schwedisch), Griechisch and Finnisch.
- Bis zu 44 Millionen Wörter pro Sprache (Release 3)
  - 110 Sprachpaare
  - satzweise aligniert
- Zweck
  - Trainingdaten für die maschinelle Übersetzung

Kleine Typologie (Forts.)

- Größe
  - BROWN-Korpus
    - Amerikanisches Englisch von 1961
    - 1 Millionen Token
  - British National Corpus (BNC)
    - Britisches Englisch der 1990er Jahre
    - 100 Millionen Token
  - deWaC
    - Deutschsprachige Internetseiten
    - 1,7 Milliarden Token

Kleine Typologie (Forts.)

- Zeitbezug
  - synchron
  - diachron
- Sprachbezug
  - Allgemein: Referenzkorpus
    - Designkriterien (‘sampling criteria’)
  - Terminologisch: Spezialkorpus
  - Menge: Opportunistische Sammlung

DWDS-Kernkorpus

- Repräsentatives Korpus
- Deutsch des 20. Jahrhunderts
- 100 Millionen Token
  - 90 % geschrieben
  - 10 % gesprochen
  - 79 830 Dokumente
- verschiedene Genres
- Lemma und Wortart
- Datenbasis für das Digitale Wörterbuch der Deutschen Sprache