Tutorium der DGfS-Sektion Computerlinguistik

KORPUSLINGUISTIK MIT ONLINE-RESSOURCEN

EINE INTERAKTIVE EINFÜHRUNG FÜR LINGUISTEN

Einführung

Stefanie Dipper
Stefan Evert
Heike Zinsmeister

München, 28.1.2011

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 Einwände vorgebracht werden, betrachte ich es als genehmigt.

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.

Beispiel: Europarl-Korpus

Genehmigung des Protokolls

Das Protokoll der gestrigen Sitzung wurde verteilt.

Wenn keine Einwände vorgebracht werden, betrachte ich es als genehmigt.

Herr Präsident, wir haben über
Ibrahim Rugova jetzt in Ror
Parlamentspräsidenten bitte
Namen des Parlaments uns

Sie werden sich erinnern, daß
Entschließung verabschied
dem Ausschuß für auswärti

Rohdaten

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

Beispiel: Europarl-Korpus

Genehmigung des Protokolls

<SPEAKER ID=1 NAME="Der Präsident">

Das Protokoll der gestrigen Sitzung wurde verteilt.

<P>

Wenn keine Einwände vorgebracht werden, betrachte ich es als genehmigt.

<P>

<SPEAKER ID=2 LANGUAGE="EN" NAME="Spencer">

Herr Präsident, wir haben über Nacht die Meldung erhalten, daß sich Ibrahim Rugova jetzt in Rom befind
Parlamentspräsidenten bitte
des Parlaments unsere bes

Sie werden sich erinnern, daß
Entschließung verabschied
Ausschuß für auswärtige A

(ep-99-05-06.txt)

Metadaten

- Datum der Debatte
- Sprechername
- 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" 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>
- FIN">wurde</word>
- P">verteilt</word>
- .</word>

Annotationen

- Satzsegmentierung
- Wortsegementierung
- Wortart

(ep-99-05-06.txt)

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
- <word id="word_5" pos
- <word id="word_6" pos
- <word id="word_7" pos
- <word id="word_8" pos

Annotationen

- Satzsegmentierung
- Wortsegementierung
- Wortart

Terminologie

- Token
- Wortart: 'part of speech'
- Wortartenlabel: 'pos tag'
- Gesamtinventar: 'tag set'
- Quasi-Standard zur Beschreibung der Wortarten des Deutschen: STTS (Stuttgart-Tübingen-TagSet)

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, Intonation, 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 2005)
  - http://www.statmt.org/europarl/
- Protokolle von Plenardebatten des Europäischen Parlaments
  - in 11 offizielle EU-Sprachen übersetzt
  - Romanisch (Französisch, Italienisch, Spanisch, Portugiesisch), Germanisch (Englisch, Niederländisch, Deutsch, Dänisch, Schwedisch), Griechisch und Finnisch
- Bis zu 44 Millionen Wörter pro Sprache (Release 3; aktuell: Release 5)
  - 110 Sprachpaare
  - satzweise aligniert
- Zweck
  - Trainingsdaten 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
- Annotation: Lemma und Wortart
- Datenbasis für das Digitale Wörterbuch der Deutschen Sprache

Referenzen

- Philipp Koehn (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation*. In: Proceedings of the 10th Machine Translation Summit (MT Summit), 79-86.
- Lothar Lemnitzer und Heike Zinsmeister (2006/2010): *Korpuslinguistik: Eine Einführung*. Tübingen: Narr.