Korpuslinguistik

Jürgen Spitzmüller

Vorlesungsreihe »Methoden der Linguistik«
(WS 2004/05)
Universität Zürich, 7.12.2004

Überblick

1. Teil: Theorie – Grundlegende theoretische Fragestellungen:
   - Was sind überhaupt Korpora?
   - Wozu Korpora?
   - Was sollen Korpora leisten/was können sie *nicht* leisten?
   - Welche Korpustypen gibt es?
   - Wie müssen linguistische Korpora aufbereitet sein?
2. Teil: Praxis – Konkrete Arbeit mit einem vorgefertigten Korpus
   - Vorstellung der Korpora des IDS und des COSMAS-Systems
   - Versuch, einige kleine lexikologische, semantische und syntaktische Fragestellungen korpusanalytisch zu lösen

Linguistische Datenerhebung

Linguistische Datenerhebung

1. Introspektion

Linguistische Datenerhebung

1. Introspektion
2. Experiment

Linguistische Datenerhebung

1. Introspektion
2. Experiment
3. Datenerhebung

Linguistische Datenerhebung

1. Introspektion
2. Experiment
3. Datenerhebung
4. Korpora

»Corpus linguists« vs. »armchair linguists«

»Armchair linguistics«: Bilden die Daten selbst (Introspektion) und beurteilen ihre »Richtigkeit« mit Hilfe eigener Sprachkompetenz

»Empirical linguistics«: Sammeln die Daten im konkreten »Sprechalltag« und beurteilen sie auf der Grundlage ihres Vorkommens

»Corpus linguists« vs. »armchair linguists«

»Armchair linguistics«: Bilden die Daten selbst (Introspektion) und beurteilen ihre »Richtigkeit« mit Hilfe eigener Sprachkompetenz

- Daten entsprechen nicht dem tatsächlich Geäußerten; Linguist verlässt sich zu stark auf sein eigenes »Sprachgefühl« und auf die Varietäten, die er verwendet.
- Wichtige Daten werden übersehen, andere überschätzt

»Empirical linguistics«: Sammeln die Daten im konkreten »Sprechalltag« und beurteilen sie auf der Grundlage ihres Vorkommens

- Die Performanz ist unvollständig
- Die Performanz enthält Fehler
- Die Performanz ist irrelevant (gg. der Kompetenz)

»Corpus linguists« vs. »armchair linguists«

»These two don't speak to each other often, but when they do, the corpus linguist says to the armchair linguist, ›Why should I think that what you tell me is true?‹ and the armchair linguist says to the corpus linguist, ›Why should I think that what you tell me is interesting?‹«

> Charles Fillmore (1992): »Corpus linguistics« or »Computer-aided armchair linguistics«. In: Jan Svartvik (Hg.): Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82, Stockholm, 4.–8. August 1991, Berlin/New York, S. 35–60.

Korpora: Sinn und Unsinn

Was sollen/können Korpora leisten?

Korpora: Sinn und Unsinn

Was sollen/können Korpora leisten?

- Empirische Überprüfung linguistischer Fragen

Korpora: Sinn und Unsinn

Was sollen/können Korpora leisten?

- Empirische Überprüfung linguistischer Fragen
- Möglichst realitätsnaher »Ausschnitt« aus einer bestimmten Sprechpraxis

Korpora: Sinn und Unsinn

Was sollen/können Korpora leisten?

- Empirische Überprüfung linguistischer Fragen
- Möglichst realitätsnaher »Ausschnitt« aus einer bestimmten Sprechpraxis
- Vermeidung der Fehl- und Überinterpretation eigener Sprechweisen des Forschers und künstlich zu Stande gekommener Daten

Korpora: Sinn und Unsinn

Was sollen/können Korpora leisten?

- Empirische Überprüfung linguistischer Fragen
- Möglichst realitätsnaher »Ausschnitt« aus einer bestimmten Sprechpraxis
- Vermeidung der Fehl- und Überinterpretation eigener Sprechweisen des Forschers und künstlich zu Stande gekommener Daten
- Quantitative (z. B. statistische) Aussagen sind (mit Einschränkungen) möglich

Korpora: Sinn und Unsinn

Was können Korpora nicht leisten?

Korpora: Sinn und Unsinn

Was können Korpora nicht leisten?

- Niemals »vollständiges« Abbild einer Sprache oder Varietät

Korpora: Sinn und Unsinn

Was können Korpora nicht leisten?

- Niemals »vollständiges« Abbild einer Sprache oder Varietät
- Auf eine Auswahl von Textsorten beschränkt

Korpora: Sinn und Unsinn

Was können Korpora nicht leisten?

- Niemals »vollständiges« Abbild einer Sprache oder Varietät
- Auf eine Auswahl von Textsorten beschränkt
- Aussagen beziehen sich *immer* zunächst nur auf das Korpus selbst

Korpora: Sinn und Unsinn

Was können Korpora nicht leisten?

- Niemals »vollständiges« Abbild einer Sprache oder Varietät
- Auf eine Auswahl von Textsorten beschränkt
- Aussagen beziehen sich *immer* zunächst nur auf das Korpus selbst
- Quantität ist nicht alles!

Korpustypen

Korpustypen

Grundsätzlich: Korpuswahl abhängig vom Forschungsziel

Korpustypen

Grundsätzlich: Korpuswahl abhängig vom Forschungsziel

1. Zusammensetzung
   - Homogen: thematisch, textsortenspezifisch, medial eingeschränkt

Korpustypen

Grundsätzlich: Korpuswahl abhängig vom Forschungsziel

1. Zusammensetzung
   - Homogen: thematisch, textsortenspezifisch, medial eingeschränkt
   - Heterogen: möglichst breite Auswahl verschiedenartigster Texte

Korpustypen

Grundsätzlich: Korpuswahl abhängig vom Forschungsziel

1. Zusammensetzung
   - Homogen: thematisch, textsortenspezifisch, medial eingeschränkt
   - Heterogen: möglichst breite Auswahl verschiedenartigster Texte
2. Metasprachliche Informationen

Korpustypen

Grundsätzlich: Korpuswahl abhängig vom Forschungsziel

1. Zusammensetzung
   - Homogen: thematisch, textsortenspezifisch, medial eingeschränkt
   - Heterogen: möglichst breite Auswahl verschiedenartigster Texte
2. Metasprachliche Informationen
3. Genese
   - Vorgefertigtes Korpus

Korpustypen

Grundsätzlich: Korpuswahl abhängig vom Forschungsziel

1. Zusammensetzung
   - Homogen: thematisch, textsortenspezifisch, medial eingeschränkt
   - Heterogen: möglichst breite Auswahl verschiedenartigster Texte
2. Metasprachliche Informationen
3. Genese
   - Vorgefertigtes Korpus
   - Selbst erstelltes Korpus

Korpustypen

Grundsätzlich: Korpuswahl abhängig vom Forschungsziel

1. Zusammensetzung
   - Homogen: thematisch, textsortenspezifisch, medial eingeschränkt
   - Heterogen: möglichst breite Auswahl verschiedenartigster Texte
2. Metasprachliche Informationen
3. Genese
   - Vorgefertigtes Korpus
   - Selbst erstelltes Korpus
4. Frage der »Repräsentativität«

Korpustypen

Grundsätzlich: Korpuswahl abhängig vom Forschungsziel

1. Zusammensetzung
   - Homogen: thematisch, textsortenspezifisch, medial eingeschränkt
   - Heterogen: möglichst breite Auswahl verschiedenartigster Texte
2. Metasprachliche Informationen
3. Genese
   - Vorgefertigtes Korpus
   - Selbst erstelltes Korpus
4. Frage der »Repräsentativität«
5. Analysemethode
   - quantitativ (computergestützt)
   - qualitativ (hermeneutisch)

Anforderungen an ling. Korpora

Anforderungen an ling. Korpora

Annotationen (metasprachliche Informationen und Informationen zu den Dokumenten):

Anforderungen an ling. Korpora

Annotationen (metasprachliche Informationen und Informationen zu den Dokumenten):

- Meta-Informationen: Textsorte, Verfasser, Quelle, Entstehungszeit etc.

Anforderungen an ling. Korpora

Annotationen (metasprachliche Informationen und Informationen zu den Dokumenten):

- Meta-Informationen: Textsorte, Verfasser, Quelle, Entstehungszeit etc.
- Morphosyntaktische Annotationen: Wortart, Flexionsform, Stellung im Satz, Stellung des Satzes im Text, Phrasenstruktur, satzübergreifende Verweisen etc.

Anforderungen an ling. Korpora

Annotationen (metasprachliche Informationen und Informationen zu den Dokumenten):

- Meta-Informationen: Textsorte, Verfasser, Quelle, Entstehungszeit etc.
- Morphosyntaktische Annotationen: Wortart, Flexionsform, Stellung im Satz, Stellung des Satzes im Text, Phrasenstruktur, satzübergreifende Verweisen etc.
- Semantische Annotationen: Bedeutung, zum Wortfeld etc.

Anforderungen an ling. Korpora

Annotationen (metasprachliche Informationen und Informationen zu den Dokumenten):

- Meta-Informationen: Textsorte, Verfasser, Quelle, Entstehungszeit etc.
- Morphosyntaktische Annotationen: Wortart, Flexionsform, Stellung im Satz, Stellung des Satzes im Text, Phrasenstruktur, satzübergreifende Verweisen etc.
- Semantische Annotationen: Bedeutung, zum Wortfeld etc.
- Kontext- und Kotextannotationen: Wortumfeld, Verwendungsspezifik etc.

Anforderungen an ling. Korpora

Annotationen (metasprachliche Informationen und Informationen zu den Dokumenten):

- Meta-Informationen: Textsorte, Verfasser, Quelle, Entstehungszeit etc.
- Morphosyntaktische Annotationen: Wortart, Flexionsform, Stellung im Satz, Stellung des Satzes im Text, Phrasenstruktur, satzübergreifende Verweisen etc.
- Semantische Annotationen: Bedeutung, zum Wortfeld etc.
- Kontext- und Kotextannotationen: Wortumfeld, Verwendungsspezifik etc.
- *Treebanks*: Möglichkeit, Strukturbäume zu generieren

Annotationsverfahren

Auszeichnungssprachen (Markup Languages):

- HTML (Hypertext Markup Language)
- SGML (Standard Generalized Markup Language)
- XML (Extensible Markup Language)

Methode (bei allen): Setzen von »Tags« (Marken), die ein Programm (»Parser«) lesen und verarbeiten kann, die aber im Text unsichtbar bleiben.

Annotationsverfahren: Beispiele

(1) Chomskys Bücher sind gerade wieder <ADJ>modern</ADJ>.

Annotationsverfahren: Beispiele

(1) Chomskys Bücher sind gerade wieder <ADJ>modern</ADJ>.

(2) Chomskys Bücher <VRB>modern</VRB> in der Bibliothek vor sich hin.

Annotationsverfahren: Beispiele

(1) Chomskys Bücher sind gerade wieder <ADJ>modern</ADJ>.

(2) Chomskys Bücher <VRB>modern</VRB> in der Bibliothek vor sich hin.

(3) <NOU CAS=GEN>Chomskys</NOU>Bücher <VRB NUM=Plu PERS=3 MOD=Ind GV=akt>modern</VRB>

Das COSMAS II-System

Ort: IDS Mannheim (*http://www.ids-mannheim.de/cosmas2/*)

Größe: ca. 1,52 Milliarden Wortformen

Davon annotiert: 26 Mio. Textwörter

Struktur: Modular (156 Teilkorpora)

Zugangsmöglichkeit: Online (passwortgeschützt) und vor Ort, jeweils mit Hilfe eines „Client-Programms“ (derzeit nur Windows)

Textsorten: Zeitungstexte, Belletristik, Fachtexte, transkribierte Gespräche, auch historische Texte

Besonderheiten: Sehr mächtige Recherchemöglichkeiten; Recherche in Gesprächstranskripten; sehr umfangreich; modulare Zusammenstellung; Teilkorpora sehr gut dokumentiert; Belege jederzeit verifizierbar

Einschränkungen: Texte werden (aus urheberrechtlichen Gründen) nur auszugsweise angezeigt (nicht geeignet für diskursanalytische Studien)

Das Leipziger Wortschatz-Projekt

Ort: Uni Leipzig (*http://wortschatz.uni-leipzig.de*)

Zugang: Online einsehbar (keine Anmeldung erforderlich), Recherche mit Hilfe des Browsers (Eingabeformular)

Größe: ca. 518 Mio. Wortformen

Textsorten: v. a. Zeitungstexte, z. T. auch spez. Wortlisten und Fachtexte (Lexika) sowie Belletristik (Projekt Gutenberg)

Besonderheiten: Kollokationen (grafisch), Suche nach Anagrammen

Einschränkungen: Keine detaillierte Auskunft über die Korpuszusammensetzung, Ausgabe nur einiger ausgewählter Textstellen

Das TIGER-Korpus

Ort: Institut für Maschinelle Sprachverarbeitung Stuttgart (*http://www.ims.uni-stuttgart.de/projekte/TIGER/*)

Zugang: Muss komplett heruntergeladen und mit einem speziellen Programm (*http://www.tigersearch.de*) eingesehen werden. Anmeldung (für Programm und Korpus jeweils separat) erforderlich.

Größe: 700.000 Wortformen

Textsorten: Zeitungstexte (*Frankfurter Rundschau*)

Besonderheiten: *Treebanks* (Generierung von Strukturbäumen)

Einschränkungen: relativ klein, relativ homogen.

Strukturbäume mit dem TIGER