

Die Repräsentation und Auflösung von ambigen Wortbedeutungen in der Computerlinguistik

Korpora und Annotation

PD Dr. Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

15. Januar 2010

Korpora und Annotation

- **Korpus**: (große) Sammlung von Äußerungen
- Empirischer Ansatz zur Linguistik
- **Annotation**: linguistische Beschreibung der Korpusdaten

↪ Ressource zu lexikalischem Wissen

Empirischer Ansatz

- Linguistik: Beschreibung und Erklärung natürlicher Sprache
- **Rationalismus** (Gültigkeit erfahrungsunabhängiger Aussagen beruht auf vernünftiger Einsicht) vs. **Empirismus** (alle Erkenntnis wurzelt in der sinnlichen Anschauung und beruhen auf Beobachtung)
- **Kompetenz** (Fähigkeit sich sprachlich zu äußern) vs. **Performanz/Sprachverwendung** (konkrete sprachliche Äußerungen)
- Kritik am empirischen Ansatz: Gesamtheit der Äußerungen ist eine fiktive Größe

Empirischer Ansatz

- Objektive, reproduzierbare Behauptungen zur natürlichen Sprache
- Quantitative Analysen durch Sprach-“Muster”
- Erstellung robuster Werkzeuge für die maschinelle Sprachverarbeitung
- Mängel:
 - nicht wohlgeformte Äußerungen im Korpus
 - mögliche wohlgeformte Äußerungen nicht im Korpus

Korpus-Definition

- Any collection of more than one text (McEney & Wilson, 2001)
- Large body of linguistic evidence typically composed of attested language use (McEney, 2003)
- Collection of electronic texts built according to explicit design criteria for a specific purpose (Atkins et al., 1992)
- Collection of pieces of language that are selected and ordered according to explicit linguistic criteria, in order to be used as a sample of the language (Sinclair, 1996)

Korpus-Definition

- **Korpus**: Sammlung schriftlicher oder gesprochener Äußerungen
- Daten sind typischerweise digitalisiert, also auf Rechnern gespeichert und maschinenlesbar
- Bestandteile: Primärdaten, Metadaten, Annotation
- Sammlung ist zufällig oder geplant entstanden
- Repräsentativität: Textausschnitt ist groß genug um ein bestimmtes Phänomen herum
- Medien: Text, Ton, Bilder, Videos etc.
- **Korpuslinguistik**: wissenschaftliche Tätigkeit auf der Grundlage von Analysen authentischer Texte zur Beschreibung von Äußerungen natürlicher Sprachen, ihrer Elemente und Strukturen, und die darauf aufbauende Theoriebildung

Stichprobe einer Sprache

- Korpora stellen (nur) eine Stichprobe einer Sprache dar.
- Die Stichprobe sollte entsprechend von Design-Kriterien so erstellt werden, dass sie sowohl **ausgewogen** als auch **repräsentativ** ist für einen bestimmten Verwendungszweck.

Korpus, Metadaten und Annotation

- **Primärdaten**: Daten, die in einem Korpus erfasst wurden
- **Metadaten**: Daten, die verschiedene Aspekte der Informationsressource beschreiben, also über die Primärdaten Auskunft geben, z.B. Herkunft, Entstehungszeit, beteiligte Personen, Inhalt, Trägermedium, Art der Kodierung
- **Annotation**: linguistische Beschreibung der Primärdaten, z.B. Wortarten, syntaktische Kategorien, semantische Rollen

Typologie

Designkriterien:

- **Funktionalität:** Zweck
- **Sprachenauswahl:** **monolingual** (mit/ohne Varietäten); **bilingual, multilingual** (Parallelkorpora vs. Vergleichskorpora)
- **Medium:** geschriebene/gesprochene Sprache, multimodale Korpora
- **Größe:** Brown (1 Million Wortformen) \ll WaC; abhängig von Fragestellung und Annotationsaufwand
- **Sprachbezug:** Referenzkorpus (allgemein abdeckend) vs. Spezialkorpus (spezielle Varietät)

Korpusaufbereitung:

- **Annotation:** Ist das Korpus überhaupt annotiert?
Welche Ebenen der Annotation sind vorhanden?

Physische Aspekte:

- **Persistenz:** statisch vs. dynamisch (Monitorkorpus)
- **Verfügbarkeit:** frei vs. (kostenlos) lizenziert

Deutschsprachige Korpora: Beispiele

- Chat-Korpus

F: empirische Grundlage für Forschung in der computervermittelten Kommunikation, **S:** Deutsch, **M:** geschrieben, **G:** 0,6 Millionen Token, **P:** statisch, **SB:** Logfiles aus Chatkommunikation, **V:** online und frei

- DWDS-Kernkorpus

F: Teil der Textbasis für Digitales Wörterbuch der Deutschen Sprache des 20. Jahrhunderts, **S:** Deutsch, **M:** gesprochen, geschrieben, **G:** 100 Millionen Token, **P:** statisch, **A:** Lemma, Morphosyntax, **SB:** Referenzkorpus, **V:** online

- Europarl

F: Textbasis für Maschinelle Übersetzung, **S:** D, Dä, E, Fr, Gr, Fi, It, Nd, Port, Sp, Sw, **M:** gesprochen, **G:** 11x28 Millionen Token, **P:** statisch, **A:** satzaligniert, **SB:** Spezialkorpus: Europäische Parlamentsdebatten 1996-2003, **V:** frei

Deutschsprachige Korpora: Beispiele

- Fehler-Annotiertes Linguistisches Korpus (FALKO)

F: Untersuchungen zu typischen Fehlern von Zweitsprachlern , **S:** Deutsch von Nichtmuttersprachlern, **M:** geschrieben, **G:** im Aufbau, **P:** statisch, **A:** geplant, **SB:** Lernersprache, **V:** frei

- Huge German Corpus (HGC)

F: Datengrundlage für Projekte der maschinellen Sprachverarbeitung, **S:** Deutsch, **M:** geschrieben, **G:** 204,5 Millionen Token, **P:** statisch, **A:** Morphosyntax, **SB:** opportunistische Sammlung diverser Tageszeitungen, **V:** auf Anfrage, teilweise Primärkorpus-Lizenzen

- Projekt Deutscher Wortschatz

F: sprachstatistische Analysen, **S:** Deutsch, **M:** geschrieben, **G:** ca. 500 Millionen Token, **P:** statisch, **A:** syntaktisch (Unterscheidung der Grundwortarten), semantisch (semantische Primitive und Relationen zwischen Wörtern), **SB:** opportunistisches Korpus aus Webquellen, **V:** frei, mit Webservice

Deutschsprachige Korpora: Beispiele

- Saarbrücken Lexical Semantics Acquisition Project (SALSA)

F: Materialbasis für computerlinguistische Anwendungen, **S:** Deutsch, **M:** geschrieben, **P:** statisch, **A:** TIGER plus semantische Frames, **SB:** Zeitungssprache: Frankfurter Rundschau, **V:** frei auf Anfrage

- TIGER

F: Materialbasis für computerlinguistische Forschung und Anwendungen, **S:** Deutsch, **M:** geschrieben, **G:** 0,9 Millionen Token, **P:** statisch, **A:** Morphosyntax, Morphologie, Syntax (Konstituenten und funktionale Information), Semantik (Eigennamen), **SB:** Zeitungssprache: Frankfurter Rundschau, **V:** kostenlose Lizenz

Web as Corpus (WaC)

- Ist das World Wide Web ein Korpus? **Ja**.
- zwischen Texten in verschiedenen Sprachen unterscheiden
- fortlaufenden Text von textähnlichen Artefakten wie Tabellen oder Teilen von Programmcode trennen
- kaum Daten über Herkunft, Entstehungszeitpunkt, Autorschaft etc. (Metadaten)
- [Web-as-Corpus kool ynitiative \(WaCKy\)](#)

WaC-Korpus-Beispiele

- ukWaC: Korpus mit 2 Milliarden Wörtern
- deWaC: Korpus mit 1,7 Milliarden Wörtern
- itWaC: Korpus mit 2 Milliarden Wörtern

Annotation

- linguistische Anreicherung der Primärdaten, z.B. Wortarten, syntaktische Kategorien, semantische Rollen
- konsumiert linguistisches Wissen und stellt linguistisches Wissen bereit
- potentielle Desambiguierung durch Annotation
- zeitaufwändig und teuer
- manuell vs. (semi-)automatisch
- Adjudikation und Evaluierung sinnvoll bzw. notwendig

Annotationsebenen

Ebene	Annotation(sbeispiele)
Morpho-Syntax	Wortart
Morphologie	Flexionsmorphologie, Lemmatisierung
Syntax	Konstituenten, Abhängigkeiten, topologische Felder
Semantik	Lesarten, semantische Rollen, Eigennamen
Pragmatik	Koreferenz, Informationsstruktur, Diskursstruktur
Weitere	Orthographie, Zeit, Emotion, Gestik, Mimik

Annotationsrichtlinien

- **Annotationsrichtlinien:** Spezifikation der Annotation
- Liste der Symbole, die in der Annotation verwendet werden
- Definition der Symbole
- Beschreibung für die Anwendung der Symbole auf das Korpus

Annotationsbeispiele

- Tokenisierung: implizite, vorbereitende Annotation
- Tagging: Annotation mit Wortarten
- Syntax und Semantik in TIGER und SALSA

Tokenisierung

- **Tokenisierung**: Segmentierung eines Textes in Einheiten (Sätze und Wörter)
- Was ist ein **Token**?
weil Haus . ? in rannte 40. usw.
- Hauptprobleme: Erkennung von Satzgrenzen, Abkürzungen und Mehrwortausdrücken; Normalisierung von Großschreibung
- Lösungswege: reguläre Ausdrücke, Heuristiken, automatische Klassifikation

Tokenisierung: Beispiel

Dunkel\swar's,\sder\sMond\sschien\shelle,\n\nschneebedeckt\sdie\sgrüne\
sFlur,\n\nals\sein\sAuto,\sblitzeschnelle,\n\nlangsam\sum\sdie\sEcke\sfuhr.\n\nDritten\ssaßen\sstehend\sLeute,\n\nschweigend\sins\sGespräch\svertieft,\n\nals\sein\stotgeschoss'ner\sHase\n\nauf\sder\sSandbank\sSchlittschuh\sli-
e f.\n\nUnd\saufl's'ner\sgrünen\sBank,\n\nndie\srot\sangestrichen\swar,\n\nnsaß\sei-
n\sblondgelockter\sJüngling\n\nmit\skohlrabenschwarzem\sHaar.\n\n...

Tokenisierung: Beispiel

Dunkel war's, der Mond schien helle,
schneebedeckt die grüne Flur,
als ein Auto, blitzeschnelle,
langsam um die Ecke fuhr.
Drinne saßen stehend Leute,
schweigend ins Gespräch vertieft,
als ein totgeschoss'ner Hase
auf der Sandbank Schlittschuh lief.
Und auf 'ner grünen Bank,
die rot angestrichen war,
saß ein blondgelockter Jüngling
mit kohlrabenschwarzem Haar.

...

Tagging

- Part-of-Speech Tagging = Wortarten-Annotation
- wichtiger morpho-syntaktischer Vorverarbeitungsschritt für weitere Analysen
- nützlich für maschinelle Sprachverarbeitung, z.B. Text-Indexierung (Nomen sind wichtiger als Verben); Prosodie in der Sprachsynthese
- Tagging erfordert bedingt syntaktische Desambiguierung, vgl. englisch *saw*
- Tagset: Menge von Wortarten für die Annotation, sprachabhängig und zweckbedingt

Beispiel: Wortarten

Penn Treebank Tagset (English) - 37 tags		STTS Tagset (German) - 54 tags	
JJ	adjective, positive	ADJA	adjective, attributive
JJR	adjective, comparative	ADJD	adjective, predicative
JJS	adjective, superlative	NN	common noun
NN	non-plural common noun	NE	proper name
NNS	plural common noun	APPR	preposition
NNP	non-plural proper name	APPRART	preposition incorporating article
NNPS	plural proper name	APPO	postposition
IN	preposition	VVFIN	base verb, finite
VB	base verb	VVIMP	base verb, imperative
VBD	base verb, past tense	VVINFIN	base verb, non-finite
VBG	base verb, gerund or participle I	VVIZU	base verb incorporating <i>zu</i>
VBN	base verb, participle II	VVPP	base verb, participle II
VBP	base verb, non-3rd person	PPOSS	possessive pronoun, substituting
VBZ	base verb, 3rd person	PPOSAT	possessive pronoun, attributive
POS	possessive pronoun	PRF	personal pronoun, reflexive

Beispiel: Wortarten

Under/IN

[the/DT proposal/NN]

,/,

[Delmed/NNP]

would/MD issue/VB about/IN

[123.5/CD million/CD additional/JJ Delmed/NNP common/JJ shares/NNS]

to/TO

[Fresenius/NNP]

at/IN

[an/DT average/JJ price/NN]

of/IN about/IN

[65/CD cents/NNS]

[a/DT share/NN]

,/, though/IN under/IN

[no/DT circumstances/NNS]

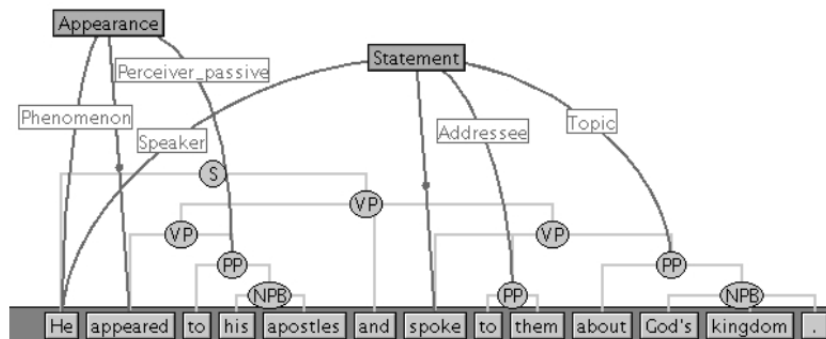
more/JJR than/IN

[75/CD cents/NNS]

[a/DT share/NN]

./.

Beispiel: TIGER/SALSA-Annotation



Korpora und Annotation: Zusammenfassung

- Korpora sind empirische Quelle für lexikalisches Wissen
- automatische Methoden können lexikalisches Wissen aus Korpora extrahieren
- Grundlage: Kontext
- Annotation stellt zusätzliches linguistisches Wissen zu Daten zur Verfügung

Referenzen



Tony McEnery and Andrew Wilson.

Corpus Linguistics.

Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press, Edinburgh, 1996.



Lothar Lemnitzer and Heike Zinsmeister.

Korpuslinguistik – Eine Einführung.

narr studienbücher. Gunter Narr Verlag, Tübingen, 2006.



Marco Baroni and Adam Kilgarriff.

Large Linguistically-processed Web Corpora for Multiple Languages.

In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006.