

Einführung in die Computerlinguistik

WS 2003/2004

Manfred Pinkal, Sebastian Pado

Inhalt der Vorlesung

Die Vorlesung besteht inhaltlich aus drei Komponenten:

- Übersicht über Phänomene und Fragestellungen der Computerlinguistik
- Exemplarische Einführung in computerlinguistische Methoden anhand ausgewählter Probleme
- Vorstellung und Diskussion sprachtechnologischer Anwendungssysteme

Die drei Komponenten werden nicht nacheinander abgearbeitet, sondern greifen ineinander.

Technisches

Zur Vorlesung gehören:

- Das Vorlesungsskript (Vorlesungs-Homepage)
- Ausgewählte Kurztexte in englischer und deutscher Sprache (Ordner Bibliothek)
- Übungsaufgaben (ca. alle 2 Wochen)

Scheine werden in folgender Weise erworben:

- Prüfungsvoraussetzung: Schriftliche Bearbeitung der Übungsaufgaben. Alle Aufgabenblätter (mit höchstens einer Ausnahme) müssen bearbeitet sein. Aufgabenblatt zählt als bearbeitet, wenn für jede Aufgabe ein Lösungsversuch vorliegt. Aufgaben können in Gruppen mit bis zu drei Studierenden bearbeitet werden. – Übungsaufgaben werden nicht benotet.
- Prüfungsleistung: Klausur über den Stoff der Vorlesung, der im Vorlesungsskript und den Lektüretexten vorkommt. Klausurtermin: voraussichtlich der Vorlesungstermin in der letzten Semesterwoche.

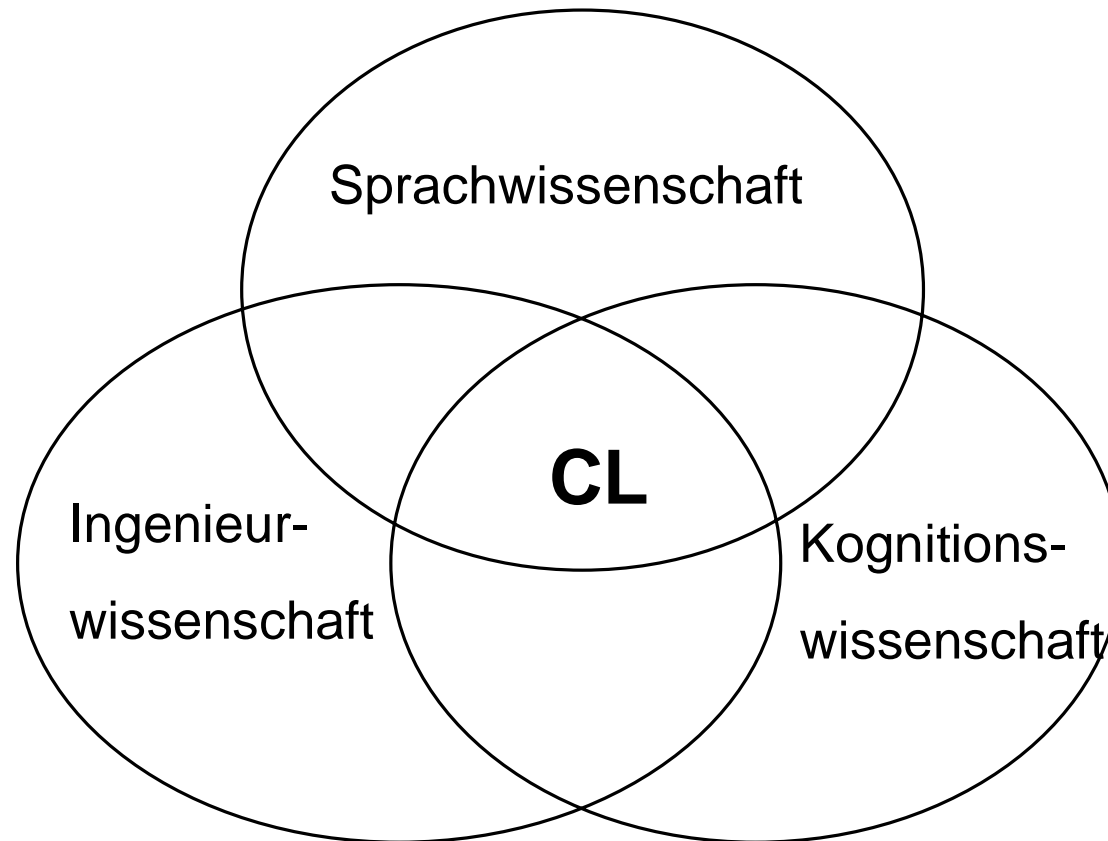
Einführungsliteratur und andere Informationsquellen

- Eine ausgezeichnetes englisch-sprachiges Einführungswerk: Jurafsky, Daniel/ Martin, James H. 2000. Speech and Language Processing. Prentice-Hall.
- Ein aktuelles deutsches Handbuch zur Computerlinguistik. Carstensen, Kai-Uwe et al. 2001. Computerlinguistik und Sprachtechnologie - Eine Einführung. Heidelberg: Spektrum Akademischer Verlag.
- Ein stärker anwendungsorientierter Überblick über die Sprachtechnologie, leider schon nicht mehr ganz aktuell: Cole, R.A. et al. 1996 Survey of the State of the Art in Human Language Technology, gibt es auch unter <http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html>
- Ein gutes deutsches Glossar zur Computerlinguistik: <http://www.ifi.unizh.ch/CL/Glossar/glossary.html>

Gegenstände der Computerlinguistik

- Die Entwicklung von Formalismen und Werkzeugen für die Repräsentation, Verarbeitung und Akquisition von linguistischem Wissen der verschiedenen Ebenen:
 - Phonetik und Phonologie
 - Morphologie und Syntax
 - Semantik
 - Pragmatik und Diskursstruktur
- Die Modellierung und Implementierung der komplexen Zusammenhänge und Abläufe bei:
 - Sprachverstehen
 - Sprachproduktion
 - Spracherwerb
- Die Entwicklung von natürlich-sprachlichen Anwendungssystemen.

Grundlegende Aspekte der Computerlinguistik



Computerlinguistik als Sprachwissenschaft

Eine wesentliche Voraussetzung für die Computerlinguistik ist die systematische und einheitliche Beschreibung von sprachlichem Wissen und sprachlichen Strukturen. Umgekehrt stellt die Computerlinguistik für die Erhebung und Erfassung komplexer sprachlicher Struktur Theorien und Werkzeuge zur Verfügung. Insofern gehört Computerlinguistik zu den sprachwissenschaftlichen Disziplinen, zusammen mit

- Theoretischer Linguistik / allgemeiner Sprachwissenschaft
- Historischer und vergleichender Sprachwissenschaft
- Phonetik
- Germanistischer, romanistischer, japanischer ... Sprachwissenschaft

Computerlinguistik als Kognitionswissenschaft

Das übergeordnete Erkenntnisziel der Computerlinguistik ist die Erforschung der menschlichen Sprachfähigkeit: Wie ist sprachliches Wissen beim Menschen organisiert, und wie wird Sprache produziert, verstanden, und gelernt? Insofern gehört die Computerlinguistik zu den Kognitionswissenschaften, die die "kognitiven" Fähigkeiten des Menschen erforschen, zusammen mit den Fächern und Forschungsbereichen:

- Psycholinguistik
- kognitive Psychologie
- Neuropsychologie
- Künstliche Intelligenz (in der Informatik)
- Sprachphilosophie

Computerlinguistik als Ingenieurwissenschaft

Die praktische Zielsetzung der Computerlinguistik ist die Realisierung von Computersystemen, die sprachliches Wissen und sprachliche Fertigkeiten einsetzen, um den Menschen in der Kommunikation, beim Verwenden von Sprache und beim Umgang mit sprachlichen Dokumenten zu unterstützen. Computerlinguistik als Sprachtechnologie gehört in den Bereich der Informationstechnologie, zusammen mit den Fächern und Forschungsbereichen

- Angewandte Informatik
- Informationswissenschaft
- Signalverarbeitung (in der Elektrotechnik)

Computerlinguistische Anwendungen: Sprachassistentz

Geschriebene Sprache:

- Rechtschreib-/Grammatik-/Stilkorrektur

Unter Einschluss gesprochener Sprache:

- Vorlese- und Diktiersysteme
„Text-to-Speech“ und „Speech-to-Text“

Unter Einschluss von Mehrsprachigkeit
(Multilingualität):

- Maschinelle Text- und Dialogübersetzung
- Sprachlernsysteme

Computerlinguistische Anwendungen: Dialog

- Auskunfts- und Buchungssysteme: Bahn-
auskunft, Telebanking, e-Commerce, Call Center
- Dialogschnittstellen für Rechner, Roboter,
Geräte, Automobil
- Persönliche Dialogassistenten

Computerlinguistische Anwendungen: Informationsmanagement

- Linguistisch gesteuerte Suche nach Dokumenten (Information Retrieval)
- Klassifikation von Dokumenten
- Informations-Extraktion
- Frage-Beantwortung („Question Answering“)

Anwendungsbereiche:

- Archive, Textdatenbanken, Nachschlagewerke
- World-Wide Web
- Einsprachige und multilinguale Anwendungen