

Einführung in die Computerlinguistik

Überblick

Alex Fraser / Robert Zangenfeind

Center for Information and Language Processing

2019-10-14

Die Grundfassung dieses Foliensatzes wurde von Dr. Benjamin Roth erstellt. Fehler und Mängel sind ausschließlich meine Verantwortung.

- 1 Was ist Computerlinguistik?
- 2 Organisation
- 3 Linguistik
- 4 CL-Methoden
- 5 Sprachtechnologie
- 6 Allgemeines

Definition

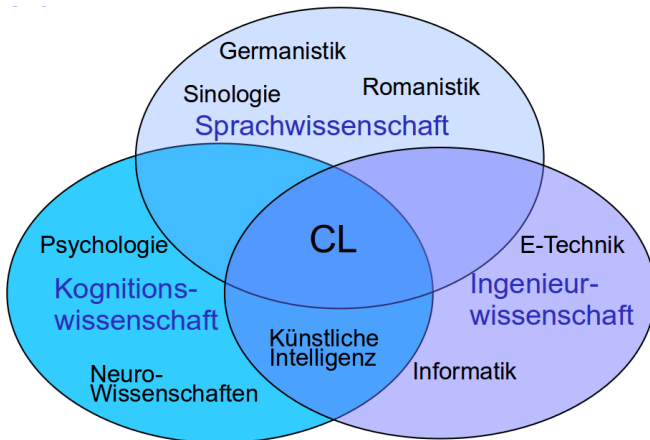
Computational linguistics is the scientific study of models and methods for **automatic processing of natural language**.

Computational linguistics is an interdisciplinary field that **shares a large part of its subject matter with computer science and linguistics**. However, computational linguists also work on theories, models and methods that are **not part of core linguistics or core computer science**.

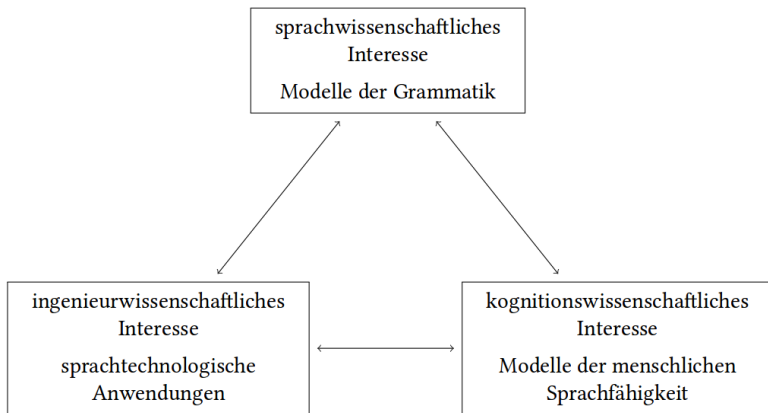
Traditionell zwei Teilbereiche:

- 1 Theoretische Computerlinguistik: Sieht sich als Teildisziplin der Linguistik, die formale, berechenbare Modelle natürlicher Sprache entwickelt, implementiert und untersucht.
- 2 Angewandte Computerlinguistik: interdisziplinäres Forschungsgebiet (Linguistik, Informatik), das konkrete Algorithmen für die maschinelle Sprachverarbeitung entwickelt (maschinelle Übersetzung, Spracherkennung, ...)

Disziplinen: Eine leicht andere Sicht (1)



Disziplinen: Eine leicht andere Sicht (2)



Flying systems in nature



- Häufigkeitsanalysen von Vorkommen von Wörtern und linguistischen Phänomenen
- Lexikographie (Thesauri, Wörterbücher)
- Suche: Google, Facebook etc. Sehr große Menge an Information, aber hochgradig unstrukturiert → direkter Zugang zu relevanten Daten ist schwierig.
- Dialoganwendungen: Zugang zu komplexen Systemen, z.B. Bestellung eines Bahn- oder Flugtickets, Interaktion mit Bank, auch mit natürlichsprachlichen Anwendungen
- Übersetzungssysteme: fremdsprachige Web-Seiten, Gebrauchsanweisungen, Social Media etc.
- automatische Silbentrennung, Rechtschreibprüfung und -korrektur
- automatische Spracherkennung
- Informationsextraktion, z.B. relevante Qualifikationen aus Bewerbungsschreiben und Lebensläufen maschinell extrahieren

- Verarbeitung gesprochener Sprache für die Interaktion mit Computern
- Verarbeitung von Texten (suchen, bearbeiten und verwalten)
- Einsatz sprachtechnologischer Software und Ressourcen (in Verlagen, Übersetzungsbüros, Verwaltungen etc.): Maschinelle Übersetzung, elektronische Wörterbücher, Spracherkennung, Sprachgenerierung, Optical-Character-Recognition-Verfahren (OCR)
- akademischer Bereich
- Bedarf an Experten steigt tendenziell

- Entwicklung von Methoden (Theorie)
- Entwicklung realistischer Anwendungen (Praxis)
- Aufbau und Verwaltung großer wiederverwendbarer Korpora (Daten)
- Konzeption effektiver Evaluierungsmechanismen (Experimente)

- Linguistik
 - Die Wissenschaft, die sich mit menschlicher Sprache beschäftigt
 - Grundinventar linguistischer Termini
 - Teilgebiete: Phonetik/Phonologie, Morphologie, Syntax, Semantik, Pragmatik; Korpuslinguistik
- Informatik (Algorithmen, Datenstrukturen, Software Engineering)
- Kognitionswissenschaft (Sprachbeherrschung ist spezieller Teilbereich der kognitiven Fähigkeiten des Menschen)
- Künstliche Intelligenz (knowledge representation, reasoning, learning)

Nachbardisziplinen (2)

- Philosophie (Verbindung von Sprache, Denken und Handeln; Relation zu außersprachlichen Gegebenheiten)
- Mathematik
 - Insbesondere: Logik, Wahrscheinlichkeitstheorie, Statistik, Graphentheorie
- Sprache ist oft nicht logisch:
 - (1) *Ein großer Berg* vs. *Eine große Ameise* → Vagheit des Adjektivs (kein Problem für Menschen) → Logik ist nicht der geeignete Formalismus?
 - (2) *Vögel fliegen.* / *Pinguine sind Vögel.* / *Pinguine fliegen.*
→ scheinbar widersprüchliche Aussagen (Mensch hat wenig Probleme damit)

- Vorlesung/Übung
 - Prof. Dr. Alex Fraser
 - Dr. Robert Zangenfeind
- Tutorium/Übungsaufgaben (Moodle)
 - Cornelia von Grafenstein (Tutorium ab 23.10.)
 - Irine Dmitrieva
 - Philip Baader
 - Andreas Wassermayr
- Sie erreichen uns unter:
`EIC11920 (at) cis.lmu.de`

- Zur Klärung von Fragen zu Übungsaufgaben und zur Vorlesung:
- Übung: Mittwoch, 11–11:45h
- Tutorat: Mittwoch, 12:15–13:45 (Raum 033)

- Bearbeitung in Moodle
- Freischaltung: Jeweils Mittwoch nach der Übung.
- Abgabetermin: Mittwoch (eine Woche später) vor der Vorlesung.
- Übungsblätter müssen von den Teilnehmern **eigenständig** bearbeitet werden.
- Klausurbonus: In Abhängigkeit der erreichten Übungspunkte wird ein Klausurbonus von bis zu 10% der maximal erreichbaren Klausurpunkte gewährt, **wenn die Klausur auch ohne die Bonuspunkte als bestanden gewertet würde.**

- Lektüre: für die meisten Vorlesungen wird es einen ergänzenden kurzen (je ca. 10 Seiten) Abschnitt aus einem Lehrbuch geben. Dieser wird auf der Kursseite bekanntgegeben und sollte bis zum angegebenen Datum gelesen werden.
- Inverted classroom: ergänzende Internet-Tutorials, die auf der Kursseite verlinkt und bis zum angegebenen Datum angesehen werden sollen.

Teil 1: Sprachwissenschaft

Klassische Aufteilung der linguistischen Teilbereiche in Bezug auf das Verhältnis von sprachlicher “Form” zu kommunikativer “Funktion”

Teil 2: Computerlinguistische Methoden

Computerlinguistische Techniken, die in verschiedenen Kontexten genutzt werden

Teil 3: Computerlinguistische Anwendungen

Praktische Anwendungen, wie z.B. automatische Übersetzungssysteme

Teil 1: Sprachwissenschaft

- 1 Phonetik / Phonologie
Merkmale sprachlicher Laute
Lautsystem, Lautstrukturen
- 2 Morphologie
Wortbildung, Flexion, Wortarten
Wortstrukturen
- 3 Syntax
Größere sprachliche Einheiten und deren Zusammenhang
Satzstrukturen
- 4 Semantik
Bedeutung sprachlicher Einheiten
Bedeutungsstrukturen
- 5 Pragmatik
Sprache im kommunikativen Kontext
Kommunikative Bedeutung

Teil 2: Computerlinguistische Methoden

- 1 Reguläre Sprachen, Endliche Automaten
Beschreibungsmittel für einfache Zeichenketten.
- 2 Hidden Markov Models, Wortartenzuweisung
Statistische Verfahren, Wortarten zu bestimmen.
- 3 Kontextfreie Grammatiken, Parsing
Automatische syntaktische Analyse.

Teil 3: Computerlinguistische Anwendungen

- 1 Maschinelle Übersetzung.
- 2 Suchmaschinen.

Moodle

Kursseite

Erstsemestereinführung der Fach-
schaft, heute 16h, Raum 151.

Fragen?

geschrieben	gesprochen
Buchstabe	Laut
Silbe	Silbe
Wort	Wort
Phrase	Phrase
Satz	Äußerung
Paragraph	Diskurs

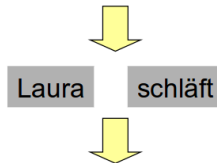
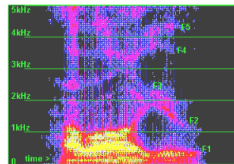
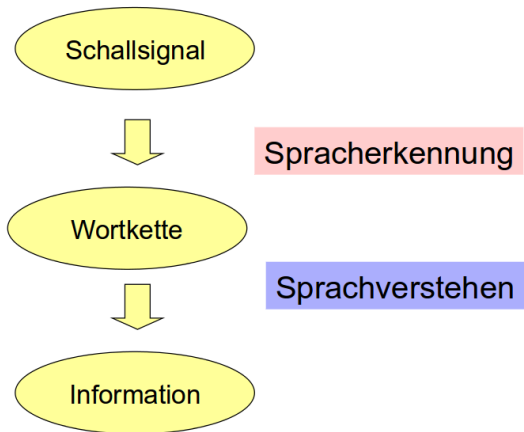
beschreibt Strukturen der Sprache(n) auf den Ebenen

- des Lautsystems
- des Aufbaus von Wörtern
- der Kombination von Wörtern zu Phrasen und Äußerungen
- der Bedeutung dieser Äußerungen, insbesondere im (außersprachlichen) Kontext (\Rightarrow Pragmatik)
- erklärt das Funktionieren dieser Strukturen als Kommunikationsmittel
- zunächst einzelsprachlich, aber durch Vergleich und Generalisierungen auch sprachübergreifend (universell)

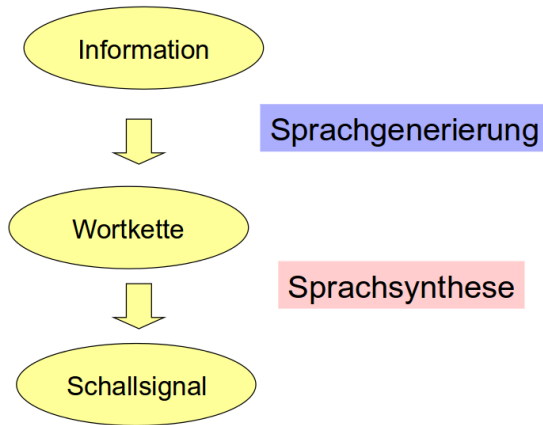
Zuordnung sprachlicher Einheiten zu ihren linguistischen Teildisziplinen

Einheit	Teildisziplin
Laut	Phonetik, Phonologie
Silbe	Phonetik, Phonologie
Wort	Morphologie
Phrase	Syntax
Satz, Äußerung	Syntax, Semantik
Diskurs	Pragmatik

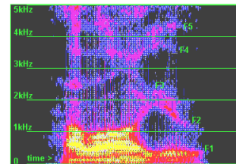
Verstehen von Sprache (gesprochen)



Erzeugen von Sprache (gesprochen)



Laura schläft



- Phonetik und Phonologie
- Morphologie
- Syntax
- Semantik
- Pragmatik
- Jedes dieser Teilgebiete hat auch eine Entsprechung in der Computerlinguistik.

- artikulatorische Merkmale
- Lautstruktur natürlicher Sprachen
- Spracherkennung: Erkennung und Produktion gesprochener Sprache
- modellieren, welche Segmente ein Wort enthält und wie sich deren Struktur auf die Aussprache auswirkt
- z.B. wenn ein im Prinzip stimmhafter Konsonant am Wortende stimmlos wird (“Auslautverhärtung”):

(3) *Dieb* /Di:p/ vs. *Diebe* /Di:be/

- Bildung und Struktur von Wörtern
- Was ist die lexikalische Wurzel einzelner Wörter?
- Welche Prozesse sind verantwortlich für unterschiedliche Erscheinungsformen an der Oberfläche?
- Veränderung der Verwendung und Bedeutung des Wortes durch Oberflächenmodifikationen
- z.B. Suffix -e als Pluralmarkierung:

(4) *Dieb-e* → *Dieb-pl* → “Mehr als ein Dieb”

- Struktur von Sätzen
- traditionell am stärksten vertretene Teildisziplin der Computerlinguistik
- Erkennung von Grammatikalität und darauf folgende Bedeutungserschließung
- z.B.

(5) *Der gewitzte Dieb stahl das Geld.*

vs.

**Der Dieb gewitzte stahl das Geld.*

- Bedeutung sprachlicher Einheiten (Wort, Satz etc.)
- z.B.

(6) *Die Polizei beschlagnahmte das Diebesgut.*

vs.

Das Diebesgut beschlagnahmte die Polizei.

→ gleiche Bedeutung

- Zweck einer Äußerung in der Welt, z.B.
Wissen Sie, wie spät es ist?
- Bestimmung des Bezugs von Wörtern: Antezedens eines Pronomens, z.B.:
Die Katze schnurrt. Sie hat Hunger.
- implizite Annahmen (Präsuppositionen), z.B.:
Welche Drogen hat Peter genommen?
Präsupponiert: Peter hat Drogen genommen.

- Methode, die auf alle Beschreibungsebenen angewandt werden kann
- seit Anfang 1980er-Jahre

Definition

A corpus (plural corpora) or text corpus is a large and structured set of texts, nowadays usually electronically stored and processed.

- Corpora are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.
- A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus).

(from Wikipedia)

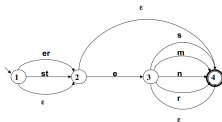
- Erkennung gesprochener Sprache
- Wortartendisambiguierung (Tagging)
- syntaktische Analyse (Parsing)
- semantische Lesartendisambiguierung (z.B. *Bank 1* vs. *Bank 2*)
- maschinelle Übersetzung

Für jede sprachliche Beschreibungsebene gibt es passende computerlinguistische Methoden

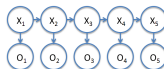
- Phonetik – Signalverarbeitung (Spracherkennung – Sprachsynthese)
- Morphologie – Wortzerlegung, Wortartenbestimmung
- Syntax – computerlesbare Grammatiken, automatische syntaktische Analyse
- Semantik – Wissensdatenbanken, automatische semantische Analyse
- Pragmatik – Koreferenzresolution, Kontextmodellierung (Dialogsysteme, sprachliche Schnittstellen z.B. in der Robotik)

Computerlinguistische Methoden zu den Beschreibungsebenen

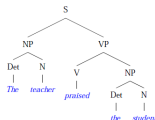
- **Wortzerlegung** - Endliche Automaten



- **Wortartenbestimmung** – HMMs



- **computerlesbare Grammatiken** - CFGs



- **automatische syntaktische Analyse** - Parsing

Verarbeitung	Modalität	Mensch	Maschine
Produktion	Lautsprache Schriftsprache	Sprechen Schreiben	Synthese Generierung
Rezeption	Lautsprache Schriftsprache	Hören Lesen	Erkennung Analyse

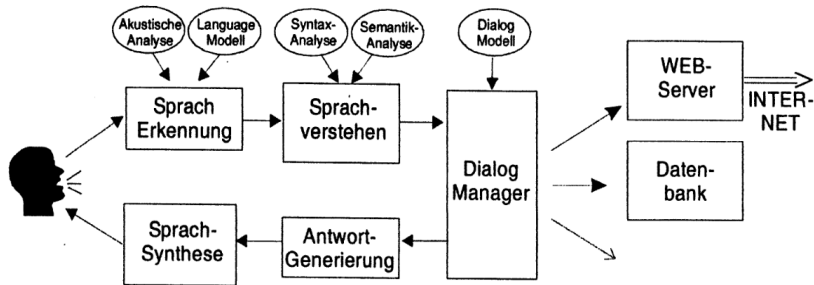
Anwendungen kombinieren oft viele verschiedene Methoden, um eine bestimmte Aufgabe zu lösen (1)

- Spracherkennung (Diktiersysteme, Transkription z.B. Untertitel)
- Sprachsynthese
- Rechtschreibkorrektur
- Maschinelle Übersetzung
DeepL: <https://www.deepl.com/translator>
Google translate: <https://translate.google.com/>
- Automatische Zusammenfassung (Summarisation)
- Suchmaschinen / Information Retrieval

Anwendungen kombinieren oft viele verschiedene Methoden, um eine bestimmte Aufgabe zu lösen (2)

- Dokumentklassifikation
- Strukturierte Gliederung von Information / Relations-Extraktion
- Frage-Beantwortung (Question Answering)
Start-System (MIT): <http://start.csail.mit.edu/>
- Sentiment-Analyse
Bsp.: Stellt der folgende Satz eine gute oder eine schlechte Bewertung dar? *Der Film hat mich ja nicht so richtig begeistert, auch wenn manche behaupten, er wäre ganz toll.*
- Dialogsysteme
 - Telefonie-Systeme: Telefonbanking, Fahrplanauskunft
 - Gerätebedienung
 - Interaktion mit virtuellen Agenten und Robotern

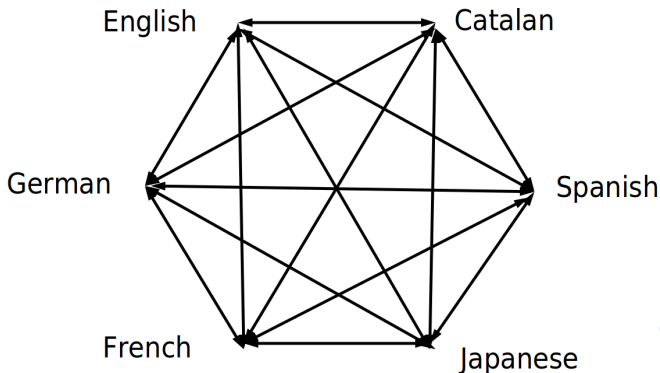
Anwendungen kombinieren oft viele verschiedene Methoden, um eine bestimmte Aufgabe zu lösen (3)



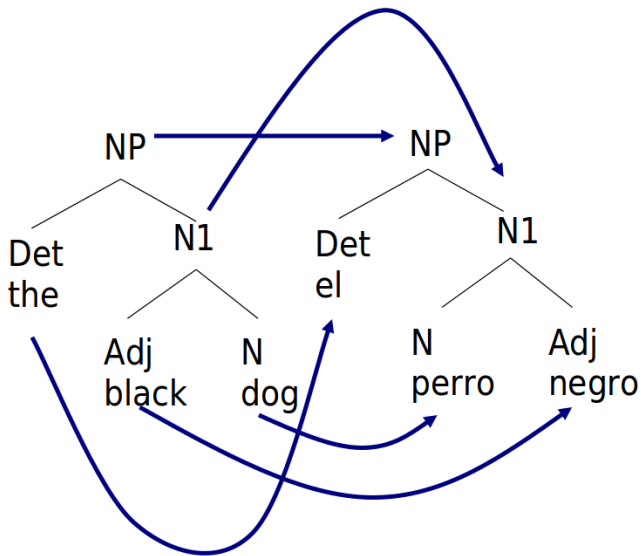
Geschichte der Methoden der Maschinellen Übersetzung (Machine Translation, MT)

- Wörter nachschlagen und übersetzen, aneinanderreihen
- Morphologische Anpassungen
- Syntaktische Umstellungsregeln
- Volle syntaktische Analyse (“parsing”) ⇒ Transfer
- Semantische Analyse (Disambiguierung) ⇒ Interlingua
- Unterstützung durch Welt-Wissen
- Übersetzen mit Statistik/Wahrscheinlichkeiten

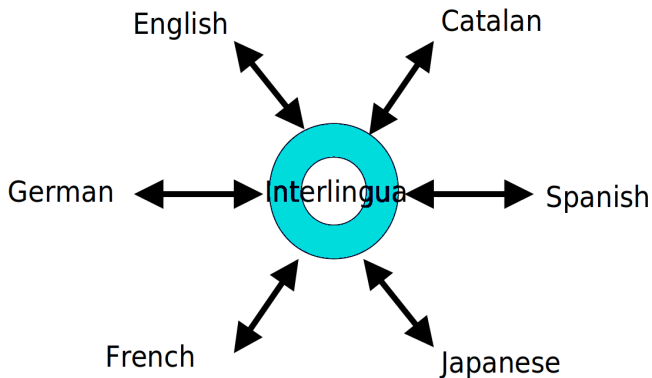
Transfer: Direkte Übersetzung von sprachlichen Elementen, ohne Bedeutungsambiguitäten notwendigerweise aufzulösen.



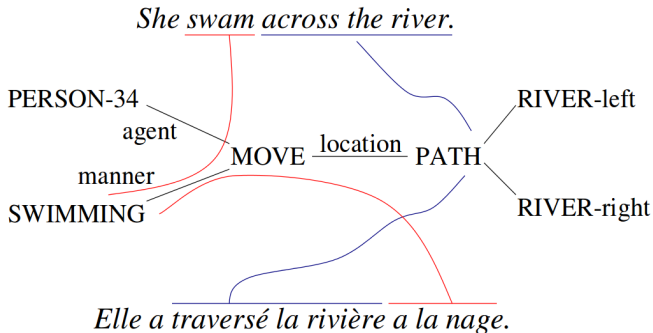
MT Syntaktischer Transfer



Interlingua: Sprachunabhängige Repräsentation von Bedeutung, in die Sprache überführt werden kann und umgekehrt.



MT Interlingua (Beispiel)



- frühe Entwicklung der Computertechnologie (1930er-, 40er-Jahre): numerische Problemstellungen (“Berechnungen”, z.B. ballistische Kurven), auch symbolische Verarbeitungsaufgaben (Dechiffrierung verschlüsselter Nachrichtentexte → maschinelle Übersetzung (MÜ) als Spezialfall einer Dekodierungsaufgabe)
- frühe Ansätze der MÜ haben gemeinsame Wurzel: stochastische Informationstheorie (Betrachtung des fremdsprachlichen Textes als Ergebnis der Übertragung einer Nachricht über gestörten Kanal → Aufgabe: Rekonstruktion des ursprünglichen Nachrichtentextes)
- Statistische Verfahren wurden dann für Jahrzehnte aufgegeben.

- Abkehr von statistischen Verfahren weil:
- Chomsky die Unzulänglichkeit der statistischen Verfahren der 50er- und 60er-Jahre für Sprachmodellierung nachweist.
- die Leistungsfähigkeit der damaligen Hardware nicht ausreichte (Beschränkungen bevorzugen symbolische Ansätze)
- nicht genügend digitalisierte mehrsprachige Textdaten zur Verfügung standen

Herausforderungen für die Computerlinguistik: Variabilität und Ambiguität (Mehrdeutigkeit)

Schwierigkeiten für Sprachverarbeitungssysteme:

- Variabilität: Dieselbe Bedeutung kann durch viele verschiedene sprachliche Formen ausgedrückt werden.
- Ambiguität: Dieselbe sprachliche Form kann verschiedene Informationen ausdrücken (erst durch den Kontext kann erschlossen werden, was gemeint ist).

Typen von Ambiguität (1)

- **Phonetische Ambiguität (Homophone):**
Miene – Mine, Meer – mehr, viel – fiel
⇒ Unterschiedliche Wörter haben dieselbe lautliche Form.
- **Orthographische Ambiguität (Homographen):**
übersétzen – úbersetzen, umfáhren – úmfahren
⇒ Unterschiedliche Wörter werden gleich geschrieben.
- **Morphologische Ambiguität:**
Staub-ecken – Stau-becken
⇒ Eine Wortform kann auf unterschiedliche Arten in seine Bestandteile zerlegt werden.
- **Lexikalische Ambiguität (Homonyme):**
Maria geht zum Ball.
⇒ Ein Wort hat mehrere verschiedene Bedeutungen.

Typen von Ambiguität (2)

- Strukturelle/syntaktische Ambiguität:

- ① *Visiting relatives can be boring.*
- ② *Peter fuhr seinen Freund sturzbetrunknen nach Hause.*
- ③ *Sie traf die Nachbarin mit ihrer Tochter.*

⇒ Die Grammatikregeln lassen verschiedene Analysen zur Kombination der Satzelemente zu.

- Kompositionell-semantische Ambiguität bzw. Skopusambiguität:

- ① *Die zwei Mitarbeiter müssen vier Sprachen beherrschen.*
- ② *Some student likes every course.*
- ③ *Alle Politiker sind nicht korrupt.*

⇒ Quantifikatoren (z.B. *alle, jeder, zwei*) und Negationen können sich auf verschieden große Satzteile beziehen.

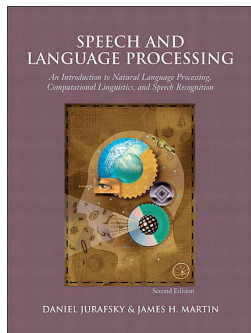
- **Pragmatische Ambiguität:**

- ① *Könnten Sie die Aufgabe lösen?*
- ② *Haben Sie eine Uhr?*

⇒ Der Bezug einer Aussage zum außersprachlichen Kontext kann auf mehrere Arten hergestellt werden.

Wie mit Ambiguität umgehen?

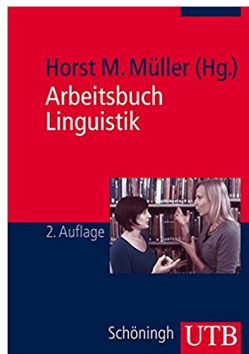
- Alle Lesarten berechnen/aufzählen.
Ist in der Regel nicht praktikabel, manchmal aber von theoretischem Interesse.
- Unterspezifizierte Repräsentation verwenden, die alle möglichen Lesarten in einer kompakten Darstellung zusammenfasst.
- Nur die aufgrund des Kontextes präferierten Lesarten berechnen/aufzählen.
Erfordert ein geeignetes gewichtetes/probabilistisches Modell, oder zusätzliche Information (Weltwissen).
- **Probabilistisches Modell**: Statistisches Modell, das verschiedenen Möglichkeiten Wahrscheinlichkeiten zuweist.
Ein System kann sich dann für die wahrscheinlichste Variante entscheiden. Die Wahrscheinlichkeiten können z.B. durch Auswertung von durch Menschen annotierte Trainingsdaten gewonnen werden.



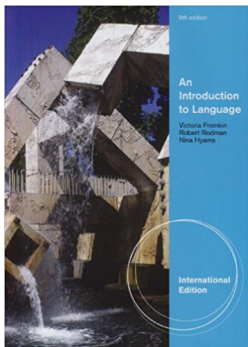
- D. Jurafsky & J. Martin: Speech and Language Processing. Pearson. Prentice Hall. 2009.
(Elektronische Version, 3. Auflage 2018: <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>)
(Lehrbuchsammlung)



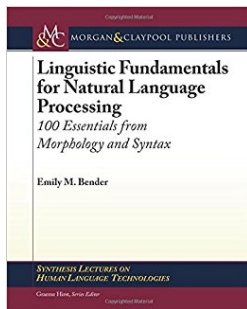
- K.U. Carstensen et al.: Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg 2010 (3. Auflage)
(Elektronische Version: <https://login.emedien.ub.uni-muenchen.de/login>)



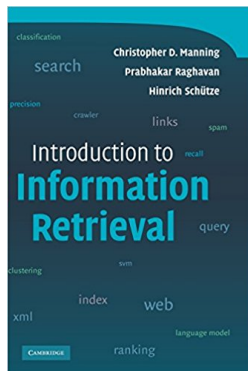
- H. Müller: Arbeitsbuch Linguistik. Schöningh / UTB. 2009.
(Elektronische Version: s.o.)



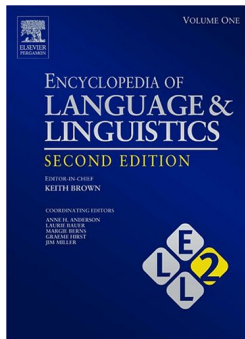
- V. Fromkin, R. Rodman, N. Hyams: An Introduction to Language. Boston et al. 2011.



- E. Bender: Linguistic Fundamentals for Natural Language Processing. Morgan & Claypool. 2013.



- Chr. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval. Cambridge University Press. 2009. <https://nlp.stanford.edu/IR-book/>



- K. Brown (ed.): Encyclopedia of Language & Linguistics. Elsevier. 2006.

- Mit welchen Themen beschäftigt sich CL?
- Wie verhält sich CL zu ihren Nachbardisziplinen insbesondere Linguistik und Informatik?
- Für welche technischen Anwendungen wird CL gebraucht?

- 1 Kursseite
- 2 Anmeldung im Moodle
- 3 Aufgabenbearbeitung im Moodle
- 4 Melden Sie sich für den Email-Verteiler für Studis an (optional)
 - Ankündigung von Vorträgen und Events am CIS (wissenschaftliche Vorträge und Firmenvorträge)
 - Jobs (als Tutor oder wissenschaftliche Hilfskraft)
 - <http://www.cis.uni-muenchen.de/ba/erstsemester-infos/index.html#verteiler>