Einführung in die Computerlinguistik

Hinrich Schütze

Center for Information and Language Processing

2018-10-15

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

Outline

Outline

Was ist Computerlinguistik?

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.

Was ist Computerlinguistik?

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.

Was ist Computerlinguistik?

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.

Was ist Computerlinguistik?

Definition

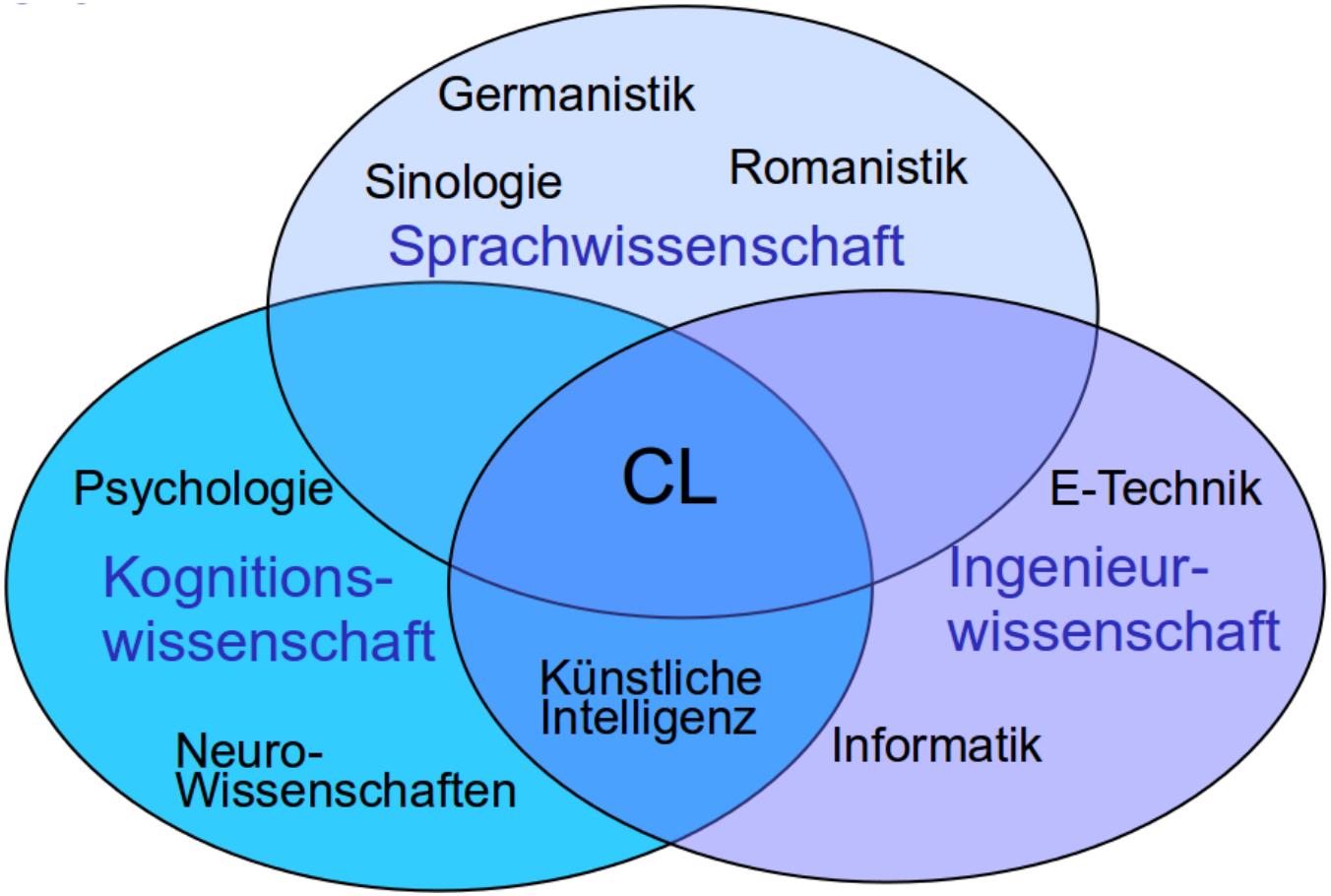
Computational linguistics is the scientific study of models and methods for automatic processing of natural language.
Computational linguistics is an interdisciplinary field that shares a large part of its subject matter with computer science and linguistics. However, computational linguists also work on theories, models and methods that are not part of core linguistics or core computer science.

Teilbereiche der Computerlinguistik

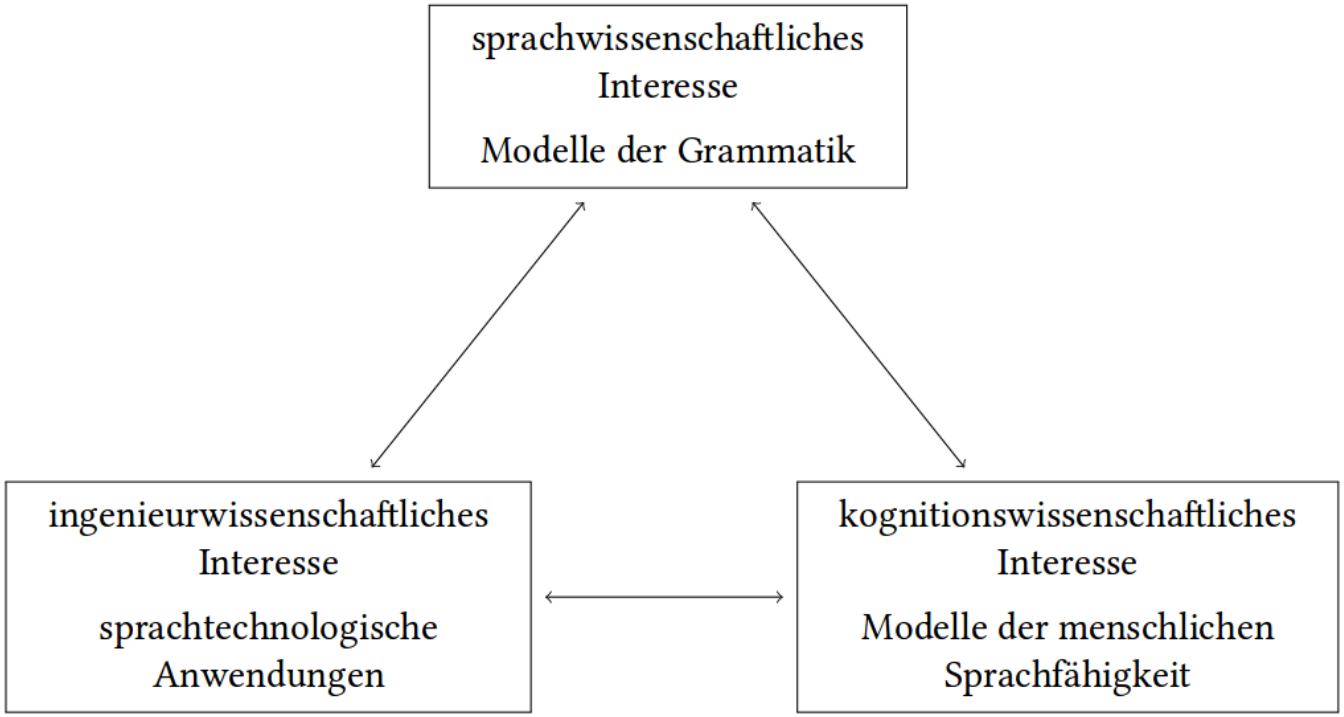
Traditionell zwei Teilbereiche:

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



Disziplinen: Eine leicht andere Sicht



Flying systems in nature

Noch eine andere Sicht

Anwendungen der Computerlinguistik

- 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

Berufsfelder für Computerlinguistik

- 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

Typische Forschungsgegenstände

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

Nachbardisziplinen (1)

- 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*
  $\rightarrow$ Vagheit des Adjektivs (kein Problem für Menschen) $\rightarrow$ Logik ist nicht der geeignete Formalismus?

  (2) *Vögel fliegen. / Pinguine sind Vögel. / Pinguine fliegen.*
  $\rightarrow$ scheinbar widersprüchliche Aussagen (Mensch hat wenig Probleme damit)

Outline

Personal

- Vorlesung / Übung
  - Prof. Dr. Hinrich Schütze
  - MSc Alena Moiseeva
- Tutorien / Aufgaben
  - Ivana Daskalovska (Tutorium ab 2018-10-23)
  - Falk Spellerberg
  - Johanna Strebl
  - Jannis Vamvas
- Sie erreichen uns unter:
  `eicl1819 (at) cis.lmu.de`

Vorstellung

Übungstermine

- Zur Klärung von Fragen zu Übungsblättern und Vorlesung.
- Vorlesung/Übung am Freitag:
  - Freitags 10:15-11:45
- Extra-Tutorat (Ivana Daskalovska):
  - Dienstags 12:15-13:45, U127

Übungsaufgaben

- Bearbeitung in Moodle
- Freischaltung: Jeweils Freitags nach der Vorlesung.
- Diese Woche: Ausnahme
- Bearbeitungsfrist: Freitags (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.

Sonstiges

- Für die meisten Vorlesungen wird es einen zu lesenden kurzen (je ca. 10 Seiten) Abschnitt aus einem Lehrbuch geben.

Vorlesungsgliederung

Teil 1: Sprachwissenschaft
Klassische Aufteilung 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.

Vorlesungsgliederung

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

Vorlesungsgliederung

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

Startseite

Vertiefung

Fachschaftsführung?

Fragen?

Outline

Sprachliche Einheiten

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

Sprachwissenschaft = Linguistik

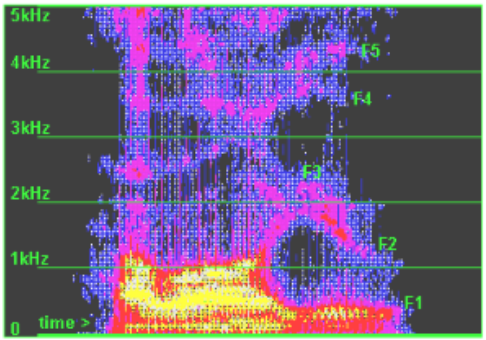
beschreibt Strukturen der Sprache(n) auf den Ebenen

- 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)

Sprachliche Einheiten

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

Verstehen von Sprache (gesprochen)



Schallsignal

Spracherkennung

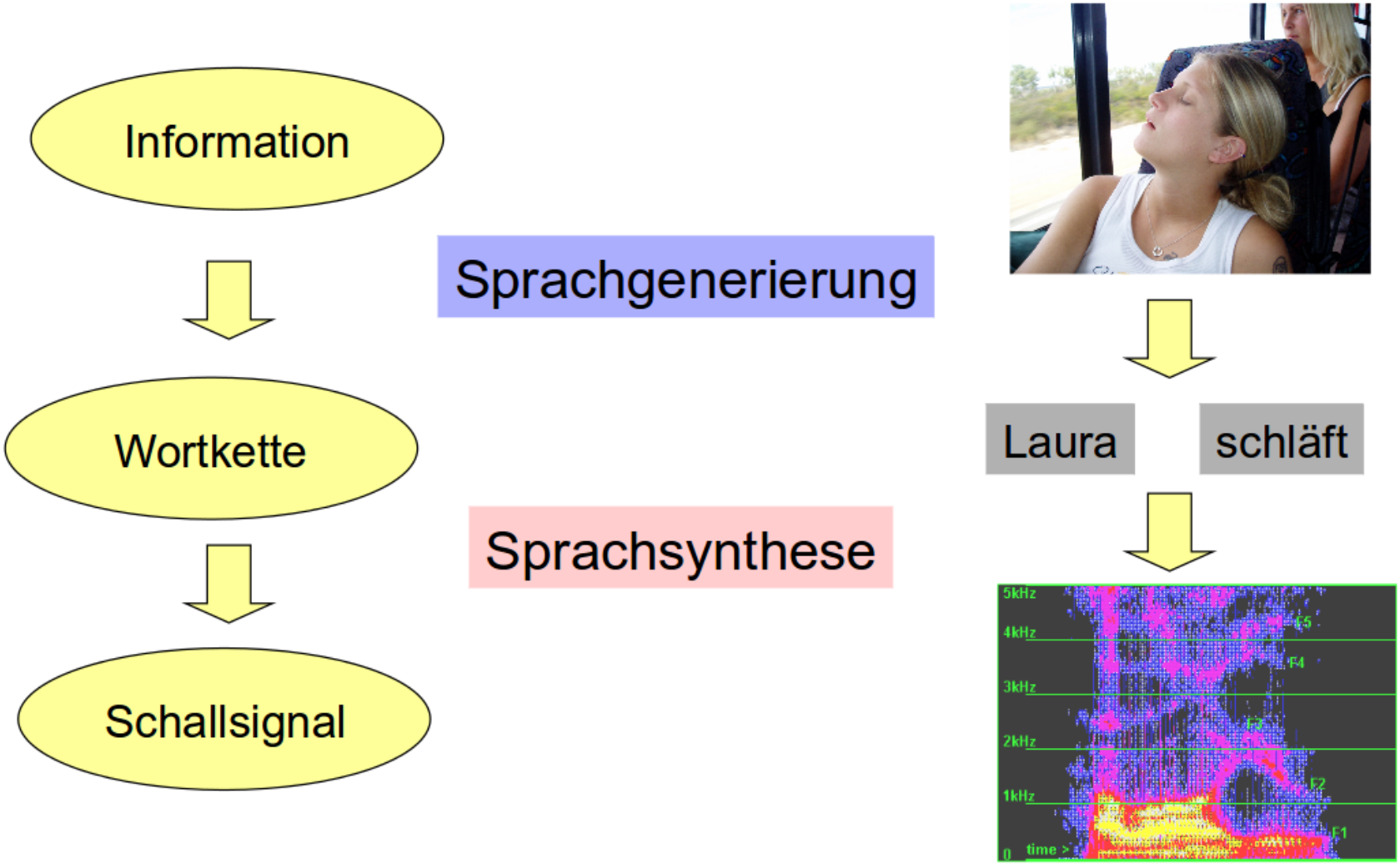
Wortkette

Laura schläft

Sprachverstehen

Information

Erzeugen von Sprache (gesprochen)



Teilgebiete der Linguistik

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

Phonetik und Phonologie

- 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/

Morphologie

- 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* $\rightarrow$ Dieb-pl $\rightarrow$ "Mehr als ein Dieb"

Syntax

- Strukturbildung 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.*

Semantik

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

(6) *Die Polizei beschlagnahmte das Diebesgut.*
vs.
Das Diebesgut beschlagnahmte die Polizei.

→ gleiche Bedeutung

Pragmatik

- 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.

Korpuslinguistik

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

Text corpus

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)

Verwendung von Korpora

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

Slido

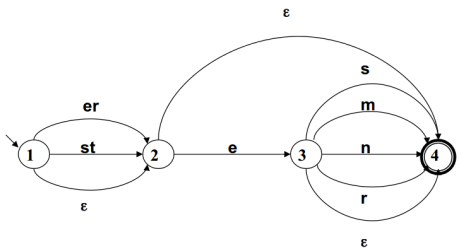
Outline

Für jede Sprachliche Beschreibungsebene gibt es passende Computerlinguistische Methoden

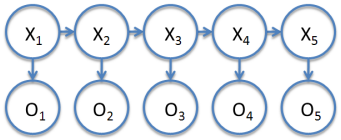
- Phonetik - Signalverarbeitung (Spracherkennung - Sprachsynthese)
- Morphologie - Wortzerlegung, Wortartenbestimmung
- Syntax - computerlesbare Grammatiken, automatische Syntaktische Analyse
- Semantik - Wissensdatenbanken, automatische semantische Analyse
- Pragmatik - Koreferenzresolution, Kontextmodellierung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

Computerlinguistische Methoden zu den Beschreibungsebenen

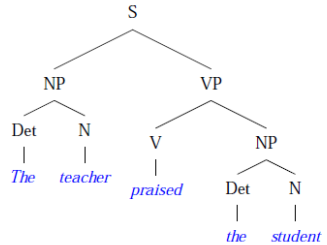
- Wortzerlegung - Endliche Automaten



- Wortartenbestimmung - HMMs



- computerlesbare Grammatiken - CFGs



- automatische Syntaktische Analyse - Parsing

Mehr Terminologie

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

Outline

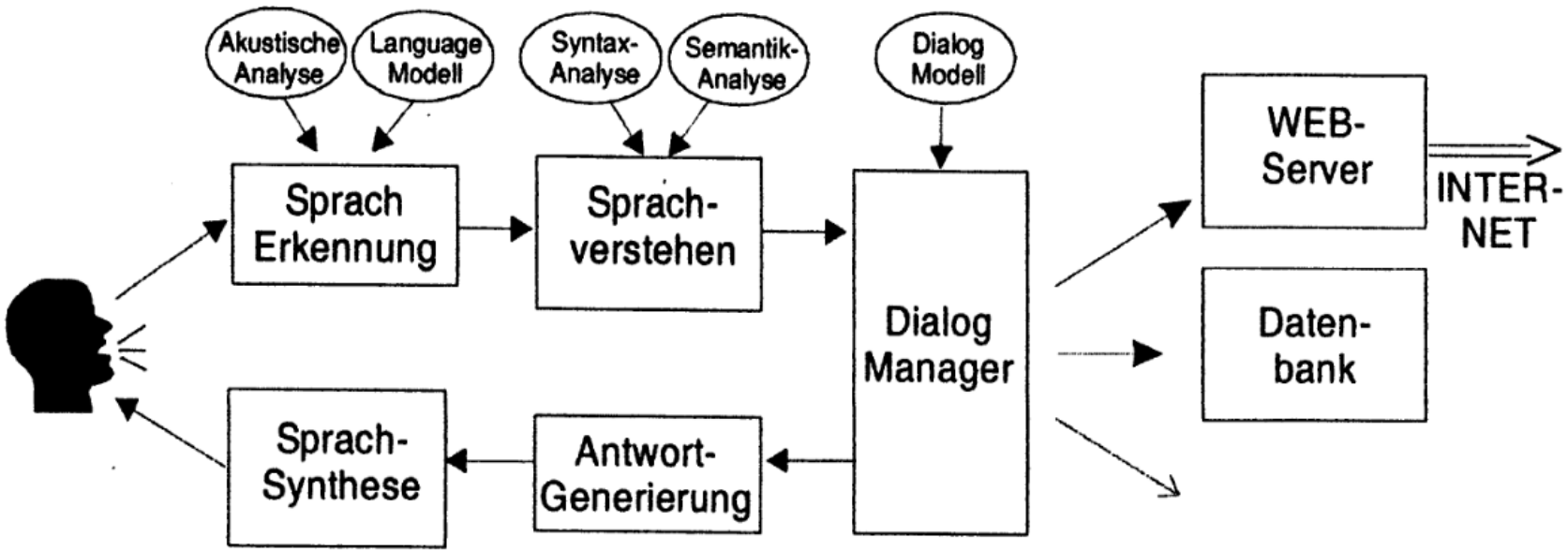
Anwendungen kombinieren oft viele verschiedene Methoden, um eine bestimmte Aufgabe zu lösen

- 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

- Dokumentklassifikation
- Strukturierte Gliederung von Information / Relations-Extraktion
- Frage-Beantwortung (Question Answering)
  Start-System (MIT): `http://start.csail.mit.edu/`
- Sentiment-Analyse
  Bsp.: Ist dies eine gute oder eine schlechte Bewertung? "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



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

- Wörter nachschlagen, aneinanderreihen
- Morphologische Anpassungen
- Syntaktische Umstellungsregeln
- Volle syntaktische Analyse ("parsing")

$\Rightarrow$

Transfer

- Semantische Analyse (Disambiguierung)

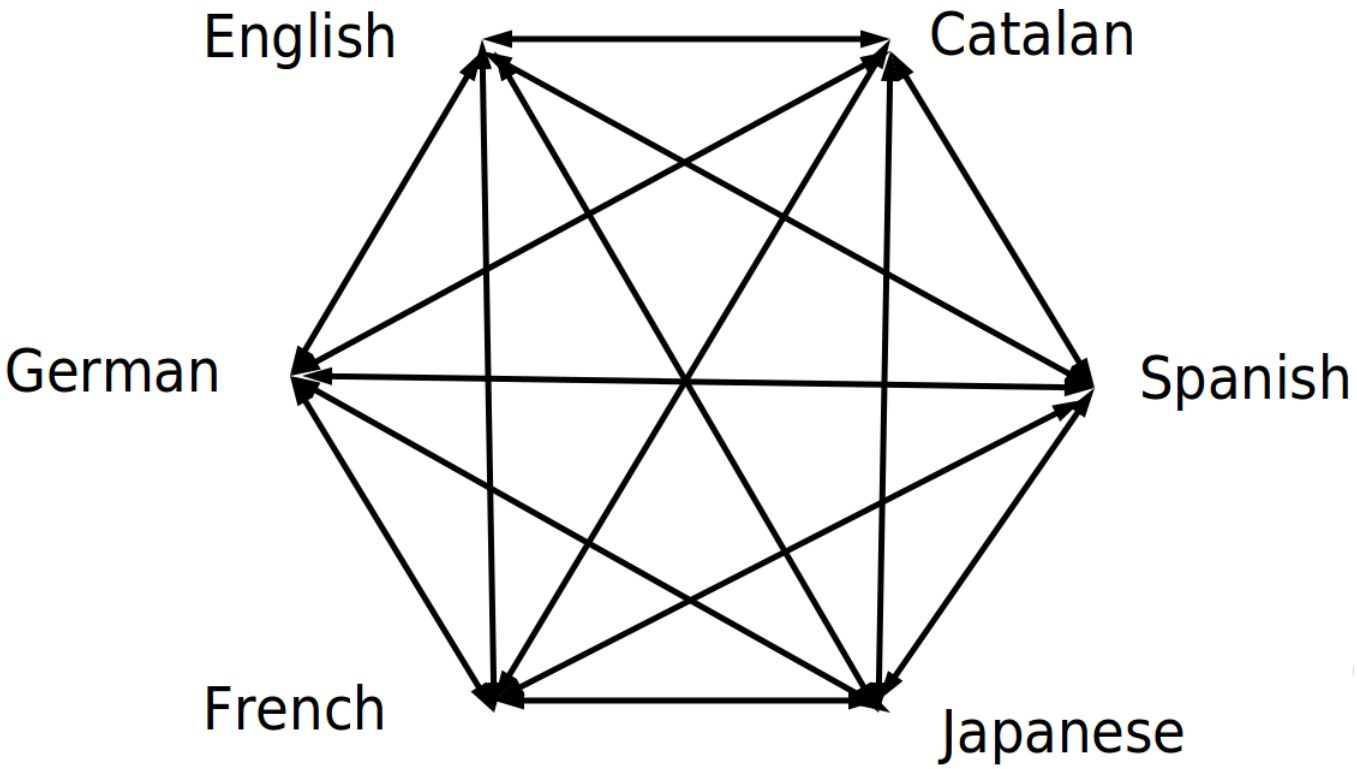
$\Rightarrow$

Interlingua

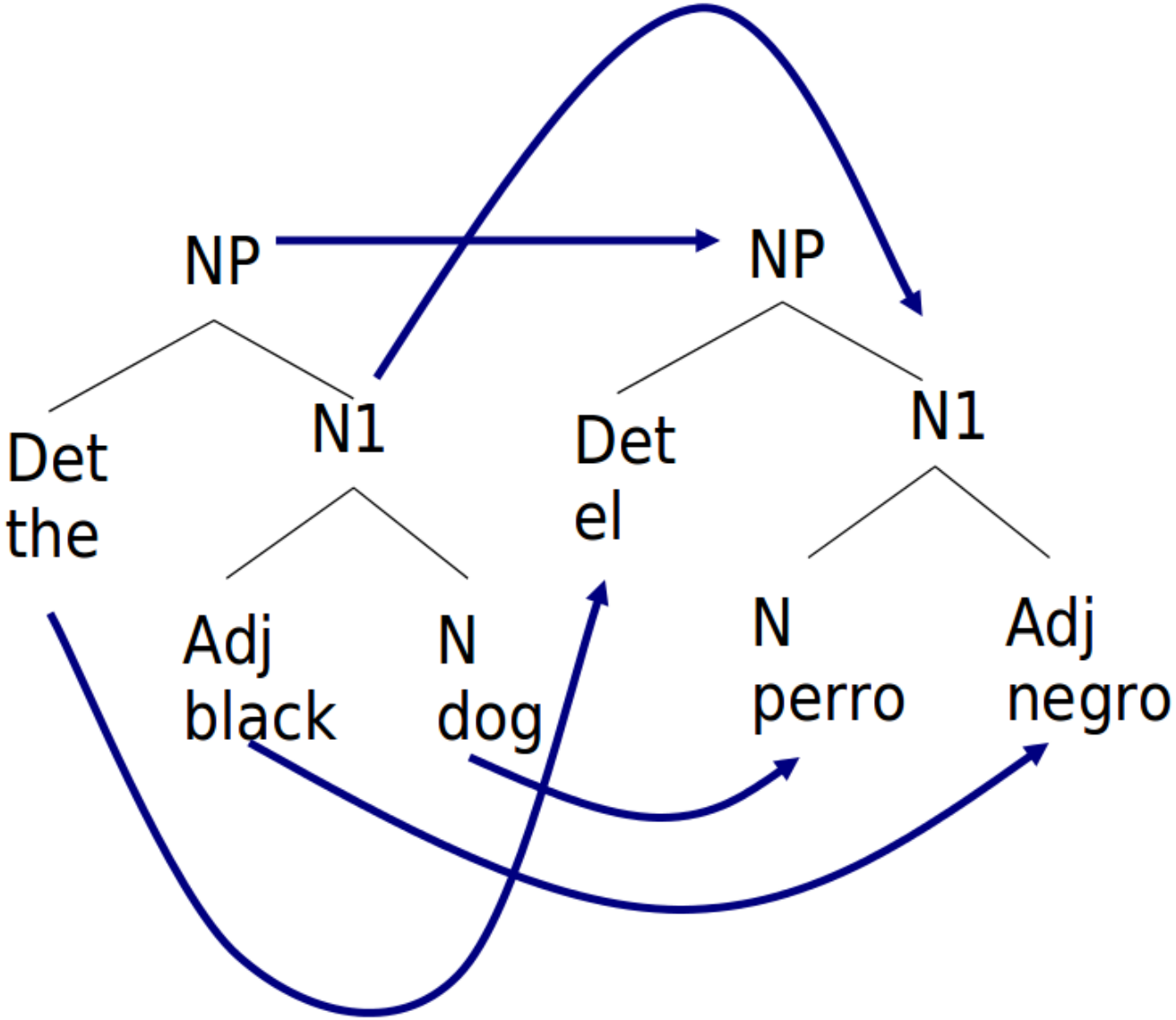
- Unterstützung durch Welt-Wissen
- Übersetzen mit Statistik/Wahrscheinlichkeiten

MT Transfer

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

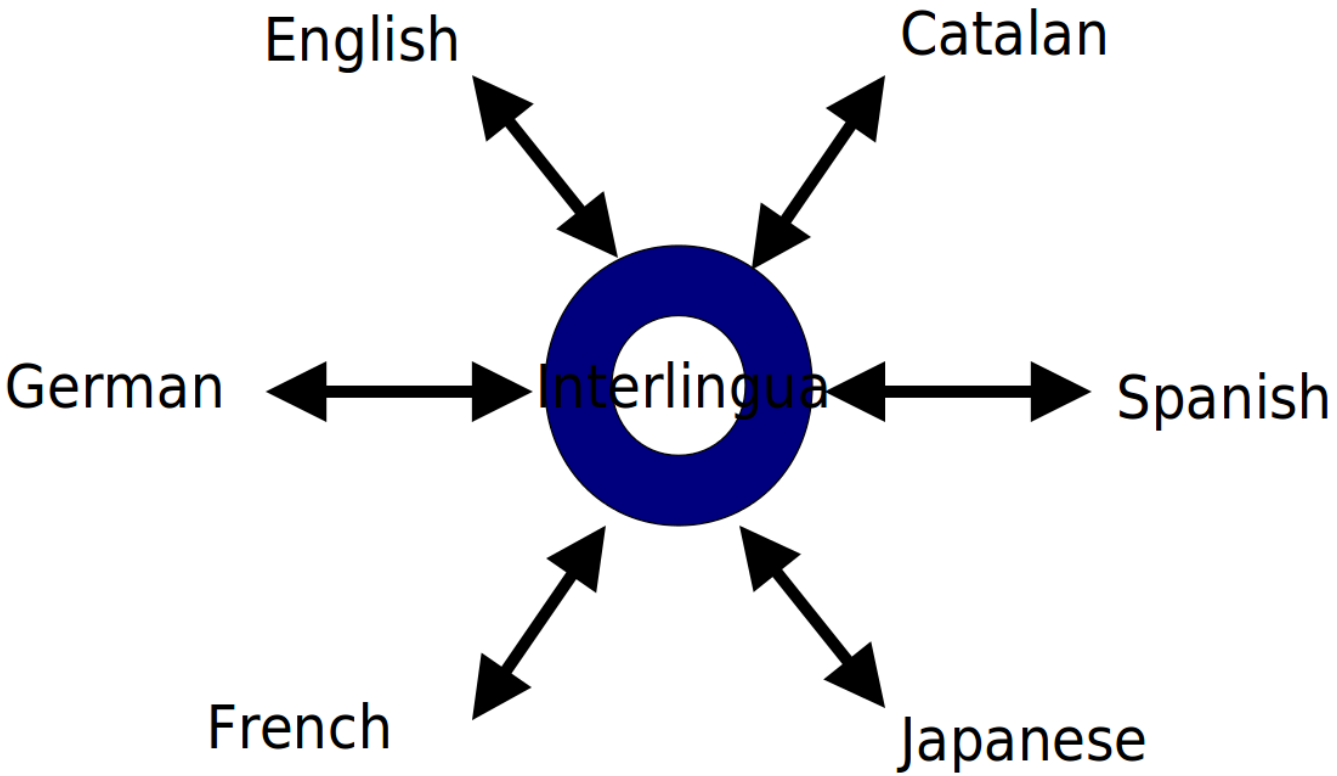


MT Syntaktischer Transfer

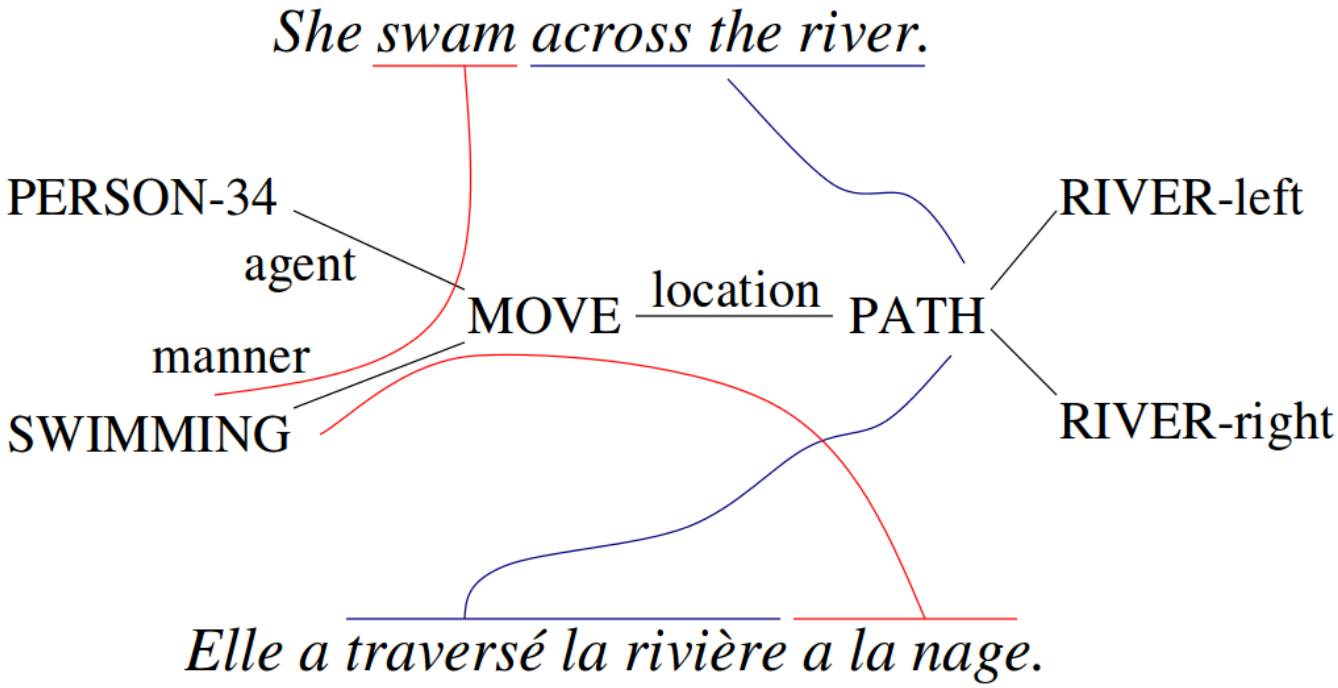


MT Interlingua

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



MT Interlingua (Beispiel)



Outline

Kleine Geschichte der Computerlinguistik (1)

- frühe Entwicklung der Computertechnologie (1930er-, 40er-Jahre): numerische Problemstellungen ("Berechnungen", z.B. ballistische Kurven), auch symbolische Verarbeitungsaufgaben (Dechiffrierung verschlüsselter Nachrichtentexte $\rightarrow$ 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 $\rightarrow$ Aufgabe: Rekonstruktion des ursprünglichen Nachrichtentextes)
- Statistische Verfahren wurden dann für Jahrzehnte aufgegeben.

Kleine Geschichte der Computerlinguistik (2)

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

Herausforderungen der Computerlinguistik: Variabilität und Ambiguität (Mehrdeutigkeit)

Schwierigkeiten für Sprachverarbeitungsysteme:

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

Typen von Ambiguität

- Phonetische Ambigität (Homophone):
  Miene - Mine, Meer - mehr, viel - fiel
  $\Rightarrow$ Unterschiedliche Wörter haben dieselbe lautliche Form.
- Orthographische Ambiguität (Homographen):
  übersetzen - über-setzen, umfahren - um-fahren
  $\Rightarrow$ Unterschiedliche Wörter werden gleich geschrieben.
- Lexikalische Ambiguität (Homonyme):
  Maria geht zum *Ball*.
  $\Rightarrow$ Ein Wort hat mehrere verschiedene Bedeutungen.
- Morphologische Ambiguität:
  Staub-ecken - Stau-becken
  $\Rightarrow$ Eine Wortform kann auf unterschiedliche Arten analysiert werden.

Typen von Ambiguität

- Strukturelle/syntaktische Ambiguität:
  1. Visiting relatives can be boring.
  2. Peter fuhr seinen Freund sturzbetrunken nach Hause.
  3. Ich traf den Sohn des Nachbarn mit dem Gewehr.

  $\Rightarrow$ 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.

  $\Rightarrow$ Quantifikatoren (alle, jeder, zwei) und Negationen können sich auf verschieden große Satzteile beziehen.

Typen von Ambiguität

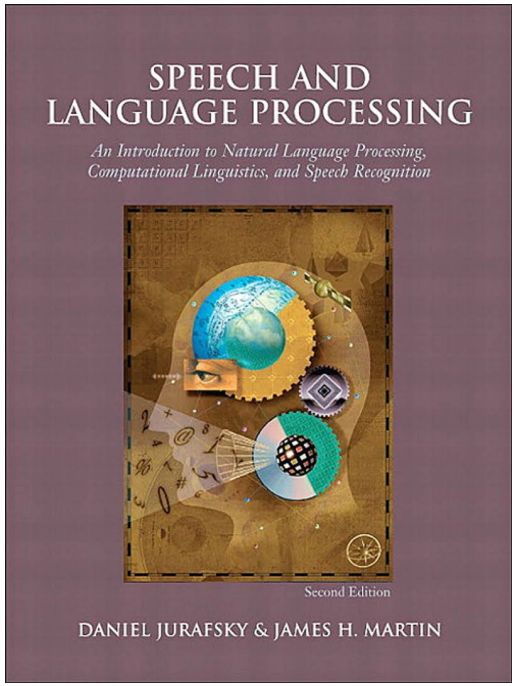
- Pragmatische Ambiguität:
  1. Könnten Sie die Aufgabe lösen?
  2. Haben Sie eine Uhr?

  $\Rightarrow$ Der Bezug einer Aussage zum außerlinguistischen 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 präferierte(n) 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.

Literatur

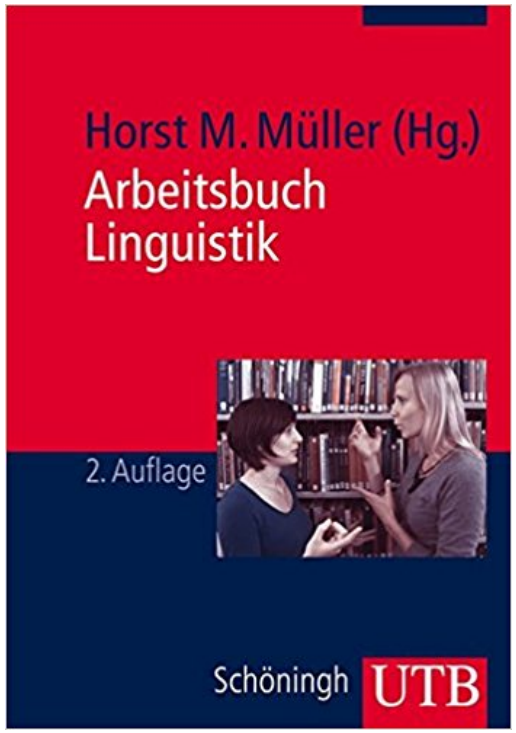


- Jurafsky & Martin: Speech and Language Processing. Pearson Prentice Hall. 2008. (Lehrbuchsammlung)

Literatur

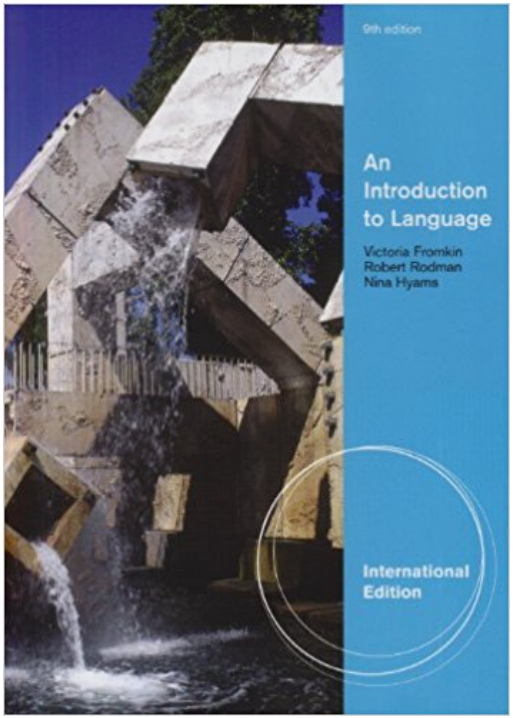
- Carstensen et al.: Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg 2010 (3. Auflage) (Elektronische Version: `https://login.emedien.ub.uni-muenchen.de/login`)

Literatur



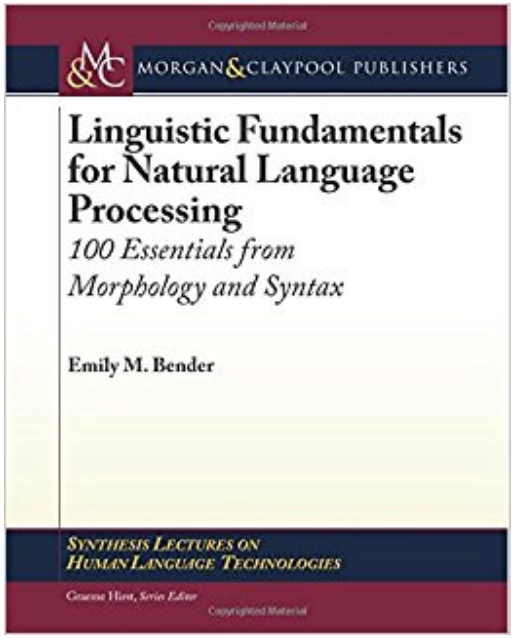
- Müller: Arbeitsbuch Linguistik. Schöningh / UTB. 2009. (Elektronische Version: s.o.)

Literatur



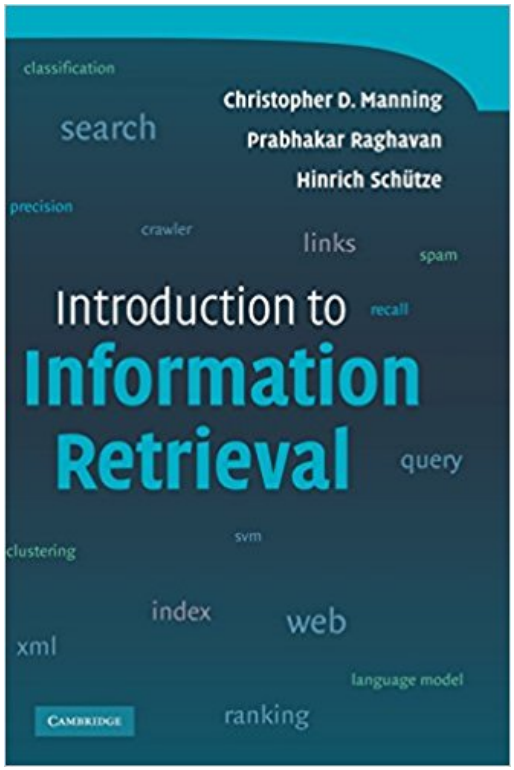
- Fromkin, Rodman, Hyams: An Introduction to Language. 2011.

Literatur



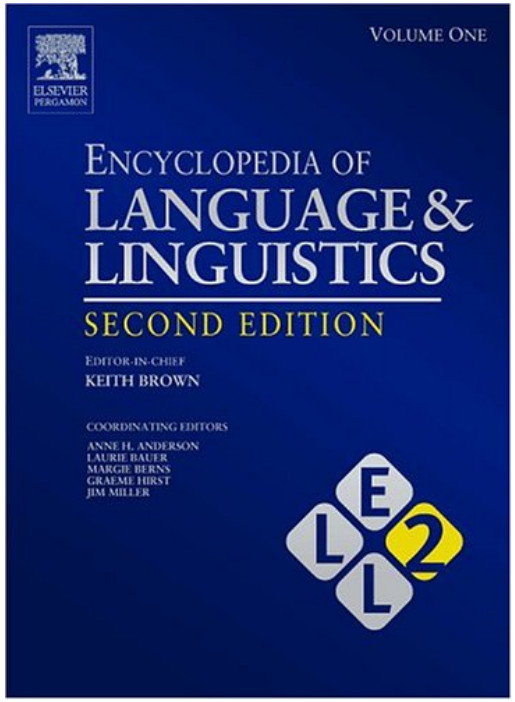
- Bender: Linguistic Fundamentals for Natural Language Processing. Morgan & Claypool. 2013.

Literatur



- Manning, Raghavan, Schütze: Foundations of Introduction to Information Retrieval. Cambridge University Press. 2008. `https://nlp.stanford.edu/IR-book/`

Literatur



- Keith Brown (ed.): Encyclopedia of Language & linguistics. Elsevier. 2006.
  (Elektronische Version: s. Homepage)

Zum Schluss: Besonders klausurrelevant

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

Ihre Todos

1. Startseite
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`