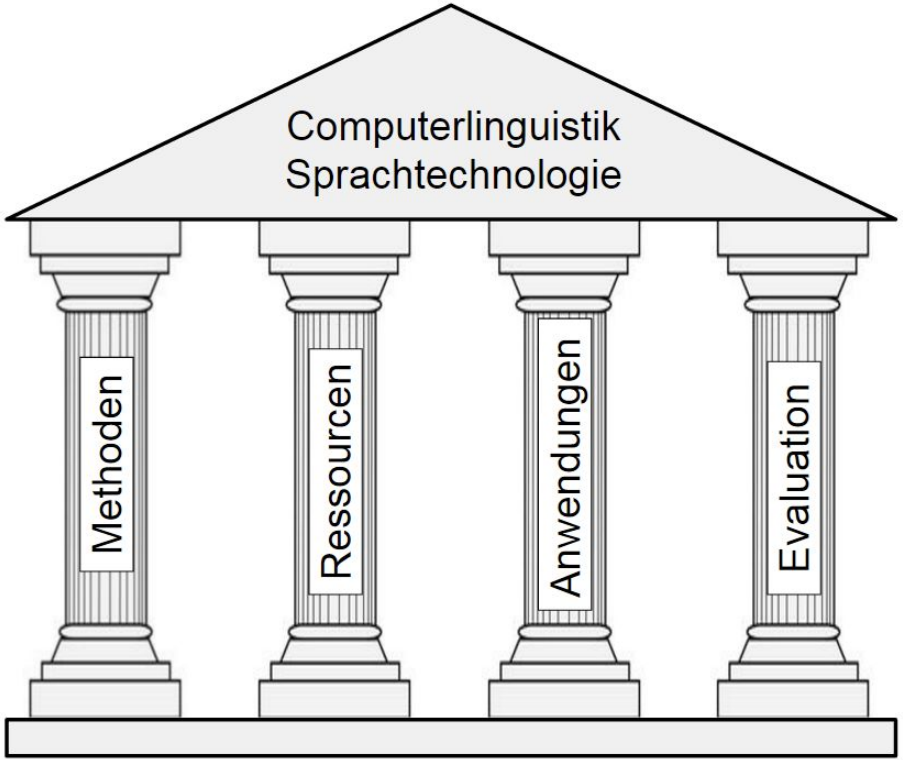
Computerlinguistik

P07: Ressourcen

IDH INSTITUT FÜR DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Hausbau der Computerlinguistik



Methoden

- **Linguistische Methoden:** Identifikation von sprachlichen Einheiten auf unterschiedlichen Ebenen, Untersuchung von deren Zusammenspiel (Phonologie, Morphologie, Syntax) und von Bedeutungen (Semantik)
- **Informatische Methoden:** Nutzung etablierter Methoden zur Datenspeicher- und -verarbeitung (Datenstrukturen und Algorithmen), z.B. Automaten- und Graphentheorie, Netzwerkanalysen, Vektorrepräsentationen, statistische und konnektionistische Methoden.

Anwendungen

- Anwendungsfeld Maschinelle Übersetzung
  - Übersetzungsspeicher, Abgleichwerkzeuge, Terminologie-Datenbanken
- Anwendungsfeld Information Retrieval
  - Text Mining, Information Extraction, Text Classification & Summarization
- Anwendungsfeld Mensch-Maschine-Kommunikation
  - Spracherkennung, Sprachsynthese, Dialogsysteme
- Anwendungsfeld Aufbau von Ressourcen
  - Korpora, Lexika, Wortnetze, Baumbanken

Ressourcen

- Korpora
- Lexika
- Wortnetze
- Baumbanken

Was ist ein Korpus (Lemnitzer/Zinsmeister 2006)

- Sammlung schriftlicher oder gesprochener Äußerungen
  - Typischerweise digitalisiert und maschinenlesbar
- Bestandteile des Korpus (Texte) bestehen aus
  - Den Daten selbst
  - Möglicherweise beschreibenden Metadaten
  - Möglicherweise (linguistischen) Annotationen

Korpustypologie (Scherer 2014)

Merkmal	Ausprägung 1	Ausprägung 2
Speichermedium	Computerlesbar	Nicht computerlesbar
Hierarchie	Gesamtkorpus	Teilkorpus
Vollständigkeit	Volltextkorpus	Probenkorpus
Abgeschlossenheit	Statistisches Korpus	Monitorkorpus
Aufbereitung	Annotiert	Nicht annotiert
Sprachmedium	Geschriebene Sprache	Gesprochene Sprache
Zeitbezug	Gegenwartssprache	Historisches Korpus
Geltungsbereich	Referenzkorpus	Spezialkorpus
Sprachen	Einsprachig	Mehrsprachig

Beispiele für Korpora

- British National Corpus - http://www.natcorp.ox.ac.uk/
- Wortschatz (Uni Leipzig) - http://wortschatz.uni-leipzig.de/
- Deutsches Referenzkorpus (IDS Mannheim) - http://www.ids-mannheim.de/kl/projekte/korpora/
- Projekt Gutenberg - https://www.projekt-gutenberg.org/
- Europarl Parallel Corpora - http://www.statmt.org/europarl/

Erstellung eines Korpus

- Auswahl eines Standards (z.B: TEI)
- Organisation von Metadaten
- Annotation der Daten
  - Wortgrenzenerkennung (Tokenisierung)
  - Satzgrenzenerkennung
  - Lemmatisierung
  - Part-of-Speech-Tagging
  - Parsing

Das Lexikon

- **Lexikon** einer Sprache besteht aus ihrem **Wortschatz**
  - explizit realisierte lexikalische Einträge
  - Menge möglicher Wörter (Wortbildungsregeln)
- **Lexikalische Information:** phonologische,morphologische, syntaktische, semantische Information einzelner Lexeme
- **Lexem:** lautliche und/oder schriftliche Form, die Gegenstand eines lexikalischen Eintrags ist
- **Schnittstelle** zwischen den grammatischen Komponenten
- **Schnittstelle** zum nichtsprachlichen Wissen

Lexikalische Ressourcen

- Beispiel für ein klassisches Wörterbuch: Duden
- Beispiel für ein Übersetzungswörterbuch: dict
- Beispiel für einen Thesaurus: openthesaurus
- Beispiel für eine Enzyklopädie: wikipedia
- Beispiel für ein Projekt des IDH: Pledari Grond

Lexikon und Wissenschaft

- **Lexikographie:** Schaffung von Archiven, Produktion von Datenbanken, Büchern usw.
- **Lexikologie:** linguistisch fundierte Beschreibung der Eigenschaften von Lexikoneinträgen
- **Lexikontheorie:** Rahmen für konsistente Forschungsergebnisse, u.a. kognitive Eigenschaften des menschlichen mentalen Lexikons

Lexikalisch-semantische Wortnetze

- **Konzeptknoten**: Abbildung der (wichtigsten) Wörter einer Sprache und deren bedeutungstragenden Beziehungen zu anderen Wörtern
- **Synset**: zugrundeliegende Repräsentationseinheit, die Synonyme zu Konzeptknoten zusammenfasst
- **Beispiele**: GermaNet (deutsch) und englisch: WordNet - http://wordnet.princeton.edu/
- **Anwendungsperspektiven**: Lesartendisambiguierung, Informationserschließung, Semantische Annotation

Baumbanken

- **Grundlegende Einheit**: In Baumstrukturen kodierte Sätze
- **Erstellung**: Durch Parser, Nachbearbeitung nötig
- **Anwendung**: Training statistischer Parser, phänomenbasiertes Retrieval
- **Qualitätsmerkmale**: Annotation, Dokumentation, Wiederverwertbarkeit, Korrektheit, Konsistenz
- **Beispiele**:
  - Penn-Treebank (englisch) und deutsch:
  - TIGER-Korpus - http://www.ims.uni-stuttgart.de/projekte/TIGER/
  - Liste: https://en.wikipedia.org/wiki/Treebank

Literatur / Hausaufgabe

- **Zur Nachbereitung**:
  - Lesen Sie: Carstensen et al. (2010): Kapitel 4 (S. 405-460)
- **Zur Vorbereitung**:
  - Naumann, Langer (1994): Kapitel 1 (S. 1-18)
- Die Texte (bzw. Links) finden Sie im Ilias-Seminarordner.