Computerlinguistische Grundlagen

Jürgen Hermes

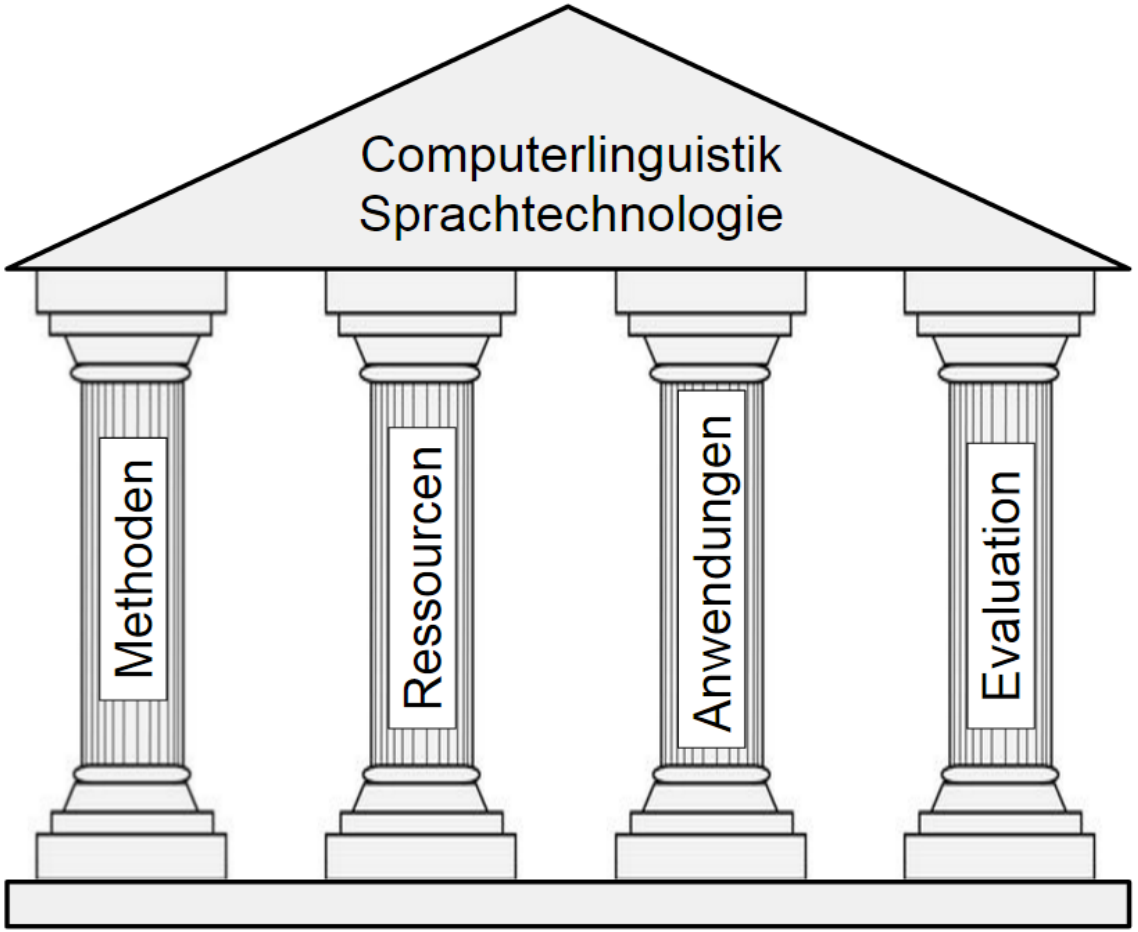
Sommersemester 19

Sprachliche Informationsverarbeitung

Institut für Digital Humanities

Universität zu Köln

Hausbau der Computerlinguistik



Ressourcen

- **Korpora**
- **Lexika**
- **Wortnetze**
- **Baumbanken**

Korpustypen

- **Textkorpora**: geschriebene oder transkribierte gesprochene Texte; Grundeinheit Token
- **Sprachkorpora**: Audioaufnahmen evtl. mit phonetischen und linguistischen Annotationen
- **Multimodale Korpora**: Sprachkorpora mit Annotationen von Prosodien, Mimik, Gestik u.a.
- **Baumbanken**: syntaktisch analysierte Sätze; Grundeinheit: Satz

Korpora - Eigenschaften

- Maschinenlesbare Textsammlung
- Ausgewogen und repräsentativ (?)
- Metainformation / Annotation
- Begrenzte Größe
- Zusammensetzung: Textsorte / Domäne / Alter; homogen vs. heterogen; fest vs. wechselnd
- Das Web als Korpus?

Korpora - Beispiele

- 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 - http://www.gutenberg.org (Leider von Deutschland aus nicht mehr zugänglich)
- Europarl Parallel Corpora - http://www.statmt.org/europarl/

Erstellung eines Korpus

- Struktur- und Metainformationen erkennen
- Tokenisierung (Segmentierung): Aufspaltung des Textes
- Satzgrenzenerkennung: Disambiguierung von Satztrennzeichen
- Hinzufügen linguistischer Information (Annotation): Part of speech (POS) tagging – Lemmatisierer – Chunking – Parsing
- Umwandlung in definiertes Format (evtl. Verwendung von Standards wie Standards wie TEI oder zumindest Dublin Core Metadata)

Abfrage eines Korpus

- **Konkordanzsuche**: KWIC-Format (key word in context)
- **Musterbasierte Suche**: Abfrage über reguläre Ausdrücke
- **Statistische Analyse**: Suche nach wiederholt auftretenden Wortformen (Kookkurenzen, Kollokationen), Wortarten (Kolligationen), Wortclustern (verwandte Lexeme)

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

Links zu lexikalischen Ressourcen

- Beispiel für ein klassisches Wörterbuch: canoonet
- Beispiel für ein Übersetzungswörterbuch: leo
- Beispiel für einen Thesaurus: openthesaurus
- Beispiel für eine Enzyklopädie: wikipedia
- Beispiel für ein Projekt der Spinfo: 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 - http://www.sfs.uni-tuebingen.de/lsd/ WordNet - http://wordnet.princeton.edu/
- **Anwendungsperspektiven**: Lesartendisambiguierung, Informationserschließung, Semantische Annotierung

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 - http://www.cis.upenn.edu/~treebank/ TIGER-Korpus - http://www.ims.uni-stuttgart.de/projekte/TIGER/
- **Liste**: https://en.wikipedia.org/wiki/Treebank

Literatur / Hausaufgabe

Zur Nachbereitung:

Carstensen et al. (2004): Kapitel 4 (S. 405-460)

Zur Vorbereitung:

Naumann, Langer (1994): Kapitel 1 (S. 1-18)

Die Texte finden Sie im ILIAS-Ordner zum Kurs
https://www.ilias.uni-koeln.de/ilias/goto_uk_crs_207163.html