Computerlinguistik

E01: Einführung in die CL

IDH INSTITUT FÜR DIGITAL HUMANITIES UNIVERSITÄT ZU KÖLN

Was ist Computerlinguistik?

- Definition
- Anwendungen
- Fragestellung
- Nachbardisziplinen
- Wissensbereiche
- Semesterprogramm

Was ist Computerlinguistik?

- Computerlinguistik aus *wissenschaftlicher* Sicht:
  Schnittstelle zwischen Linguistik und Informatik
- Computerlinguistik aus *anwendungsbezogener* Sicht:
  Schnittstelle zwischen Sprache und Rechnersystemen

Computerlinguistik in der Anwendung

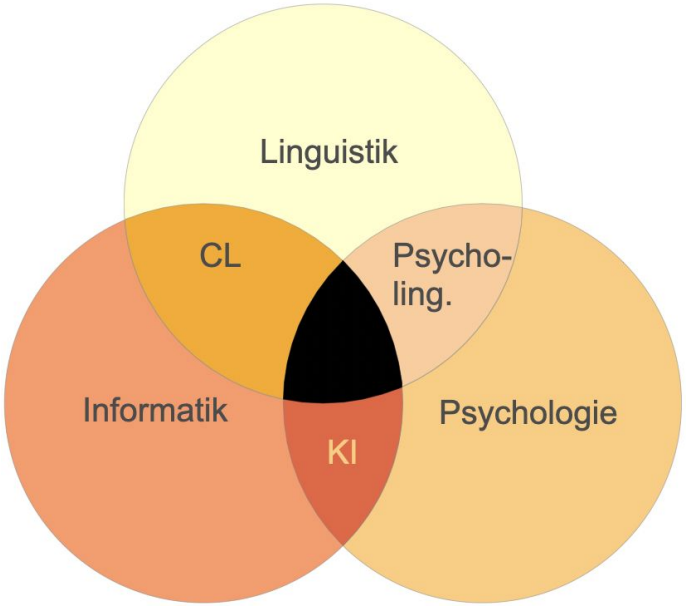
- Mensch-Maschine-Kommunikation
  - Spracherkennung, Sprachsynthese, Dialogsysteme
- Maschinelle bzw. computergestützte Übersetzung
  - Vollautomatische MÜ, Übersetzungsspeicher, Terminologie-Datenbanken
- Information Retrieval & Text Mining
  - Text Classification, Text Summarization, Information Extraction
- Aufbau und Nutzung von Ressourcen
  - (Annotierte) Korpora, Lexika, Wortnetze, Baumbanken

Fragestellungen

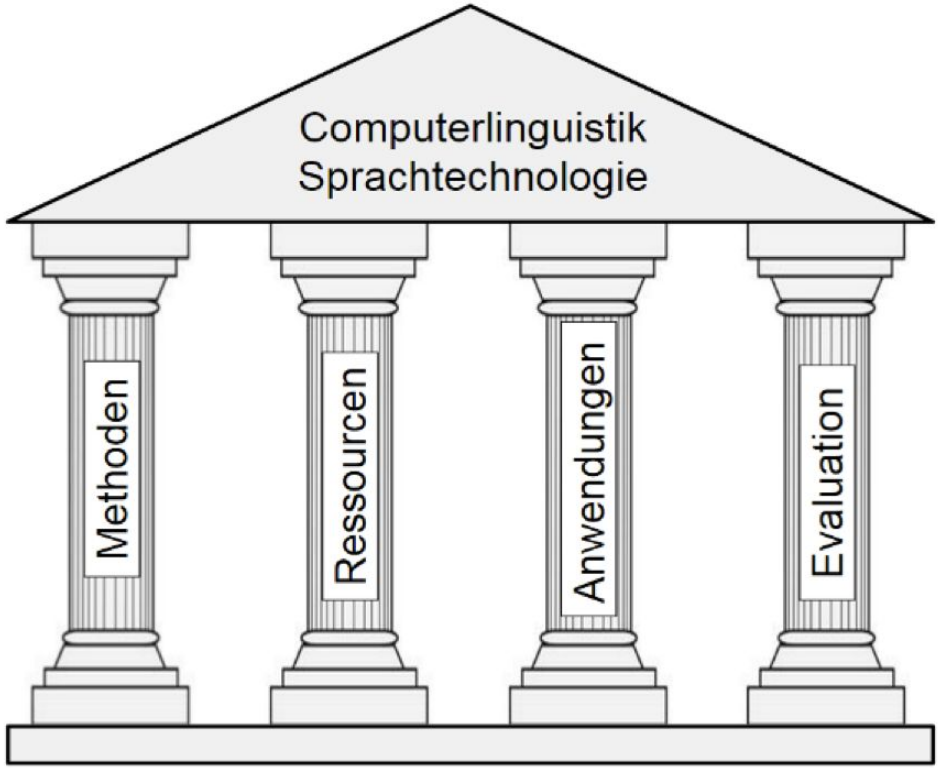
- Computerlinguistik als…
  - Teildisziplin der Linguistik
  - Linguistische Datenverarbeitung
  - Maschinelle Sprachverarbeitung
  - Sprachtechnologie

Nachbardisziplinen

- Linguistik (Gegenstand, Terminologie)
- Informatik (Algorithmen, Datenstrukturen)
- Philosophie (Verhältnis Sprache-Denken)
- Künstliche Intelligenz (Wissensrepräsentation und -verarbeitung)
- Kognitionswissenschaft (Sprache-Denken, Architekturen)
- Mathematik (Logik, Graphen, Statistik, Optimierung)



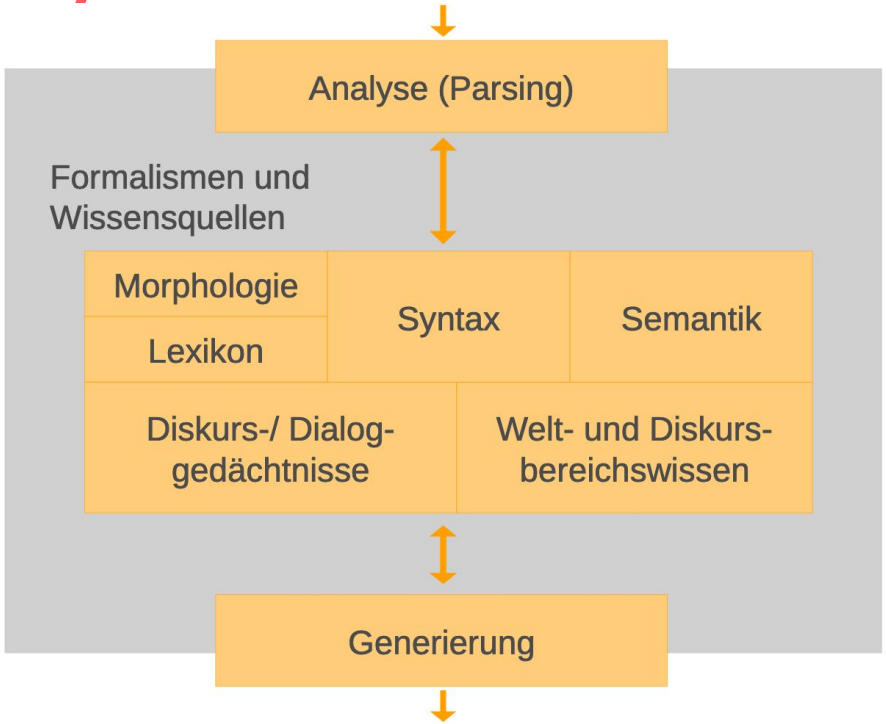
Grundpfeiler der Computerlinguistik



Methoden

- Methoden der Linguistik
  - Identifikation sprachlicher Einheiten auf unterschiedlichen Ebenen
    - Zeichenebene (Graphemik / Phonetik / Phonologie)
    - Wortebene (Morphologie)
    - Satzebene (Syntax)
  - Untersuchung des Zusammenspiels der Ebenen mit der Ebene der Bedeutung (Semantik)
- Methoden der Informatik
  - Nutzung etablierter Methoden zur Datenspeicherung und - verarbeitung
  - Darunter Automaten- und Graphentheorie, formale Sprachen, Netz- und Vektorrepräsentationen, maschinelle Lernverfahren
  - Aktuell vor allem Deep-Learning-Architekturen

Architektur eines natürlichsprachlichen Systems – Symbolischer Ansatz



Architektur eines natürlichsprachlichen Systems – Transformer-based

Architektur eines natürlichsprachlichen Systems – Transformer-based

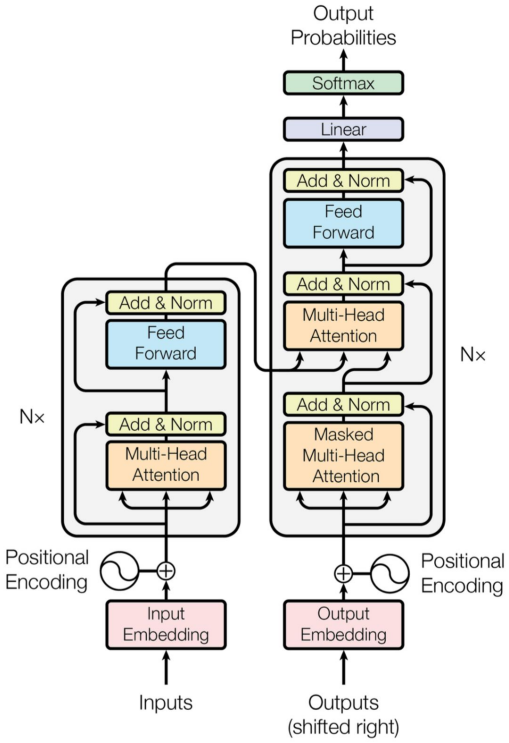


Figure 1: The Transformer - model architecture.

Forschung und Anwendung

- Theoretische Computerlinguistik
  - Suche nach komplexen Beschreibungs-formalismen, die gleichzeitig handhabbar und effizient sind.
  - Abwägung über das Verhältnis von linguistischer Beschreibung und Berechnungsoperationen.
- Praktische Computerlinguistik
  - Entwicklung und Erforschung realitätsnaher Anwendungen.
  - Natürlichsprachliche Kommunikation mit oder auf der Maschine

Semesterprogramm

- Allgemeine Einführung in die Grundbegriffe von Computerlinguistik, Sprachwissenschaft und Annotation
- Beschreibungsebenen der (Computer)Linguistik: Zeichenebene, Wortebene, Satzebene, Textebene
- Annotation zur Erstellung von Ressourcen für computerlinguistische Anwendungen

Literatur / Hausaufgabe

- **Zur Nachbereitung**:
  - Lesen Sie: Carstensen et al. (2009): Kapitel 1 (S. 1-23)
- **Zur Vorbereitung**:
  - Lesen Sie: Vater (2002): Kapitel 1 (S. 2-24)
- Die Texte finden Sie im Ilias-Seminarordner.