Einführung in die Computerlinguistik

Einführung

Laura Kallmeyer

Heinrich-Heine-Universität Düsseldorf

Summer 2018

Anwendungen der Computerlinguistik

Carstensen et al. (2010); Jurafsky and Martin (2009, 2017)

- Dialogsysteme *conversational agent (CA)*
  Bsp.: Anwendungen im Automobilbereich, Roboter, Telefon-Dialogsysteme, Sprachassistenten

- Maschinelle Übersetzung
  Bsp.: "This is my first course in Computational Linguistics, and I am quite excited about it."
  - Systran `http://www.systranet.com`
    "Dieses ist mein erster Kurs in der Computerlinguistik, und ich bin über sie ziemlich aufgeregt."
  - Google `https://translate.google.com/`
    "Dies ist mein erster Kurs in Computerlinguistik, und ich bin ziemlich aufgeregt."
  - DeepL `https://www.deepl.com/translator`
    "Dies ist mein erster Kurs in Computational Linguistics, und ich bin sehr gespannt darauf."

Anwendungen der Computerlinguistik

- Web-basiertes question answering
  Bsp.: START `http://start.csail.mit.edu/`
  "Who was the president of the United States in 1940?"
  "Franklin Delano Roosevelt: March 4, 1933 to April 12, 1945"
- Automatisches Zusammenfassen von Texten.
- Sentiment Analysis.
  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."
- Text Klassifikation und Topik Identifikation.

Bereiche der Computerlinguistik (1)

Carstensen et al. (2010)

- Computerlinguistik als Teilbereich der Linguistik
  - theoriegeleitet
  - Entwicklung formaler Sprachmodelle
  - berechnungsrelevante Aspekte von Sprache und Sprachverarbeitung
  - unabhängig von konkreter Realisierung

  → theoretische Computerlinguistik
- Computerlinguistik als Disziplin für die Verarbeitung linguistischer Daten
  - Korpora

  → linguistische Datenverarbeitung

Bereiche der Computerlinguistik (2)

- Computerlinguistik als Realisierung natürlichsprachlicher Phänomene auf dem Computer
  - Nachbardisziplinen: Kognitionswissenschaft, Künstliche Intelligenz

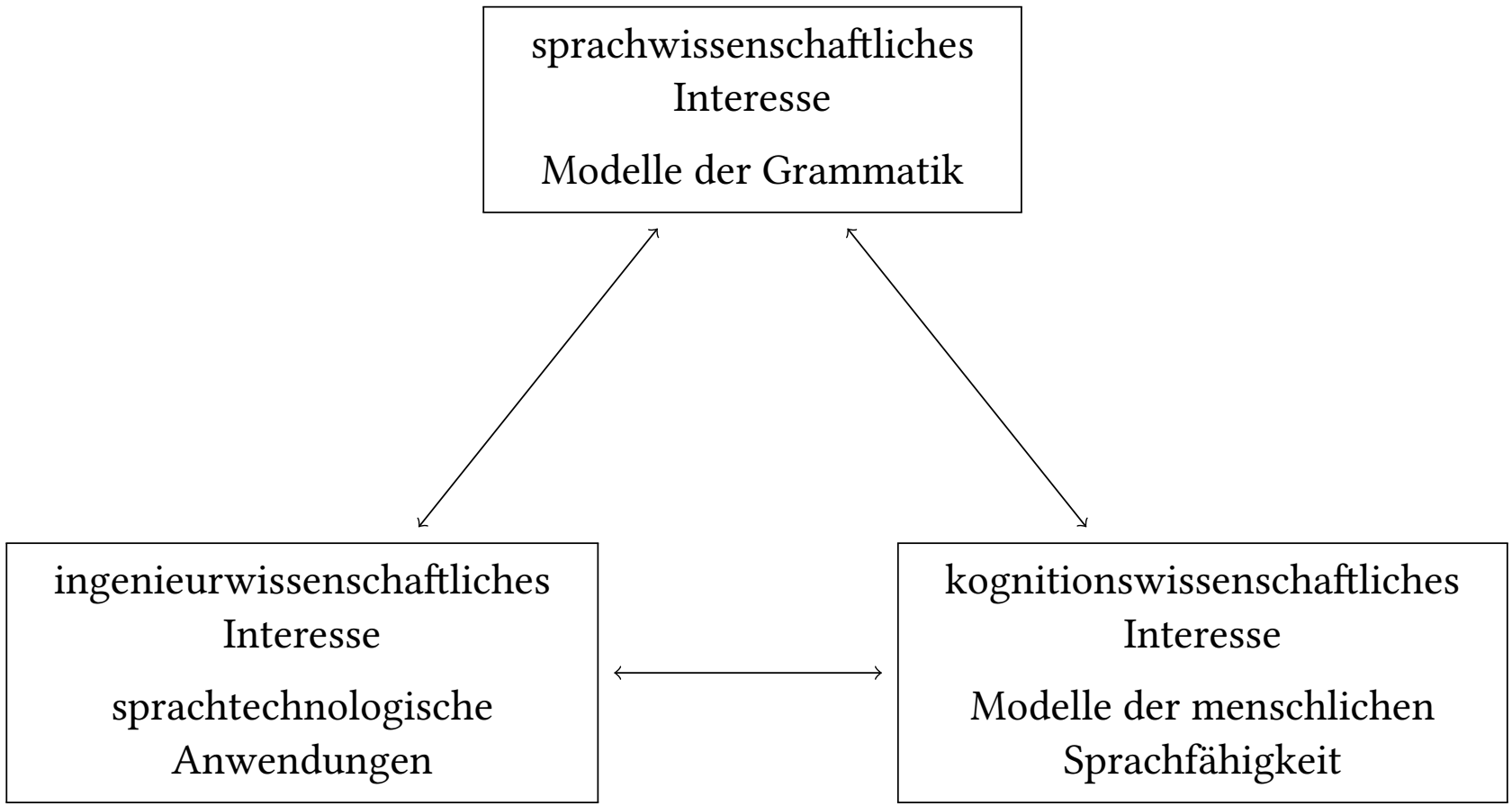
  → maschinelle Sprachverarbeitung
- Computerlinguistik als praxisorientierte, ingenieursmäßig konzipierte Entwicklung von Sprachsoftware
  → Sprachtechnologie

Bereiche der Computerlinguistik (3)

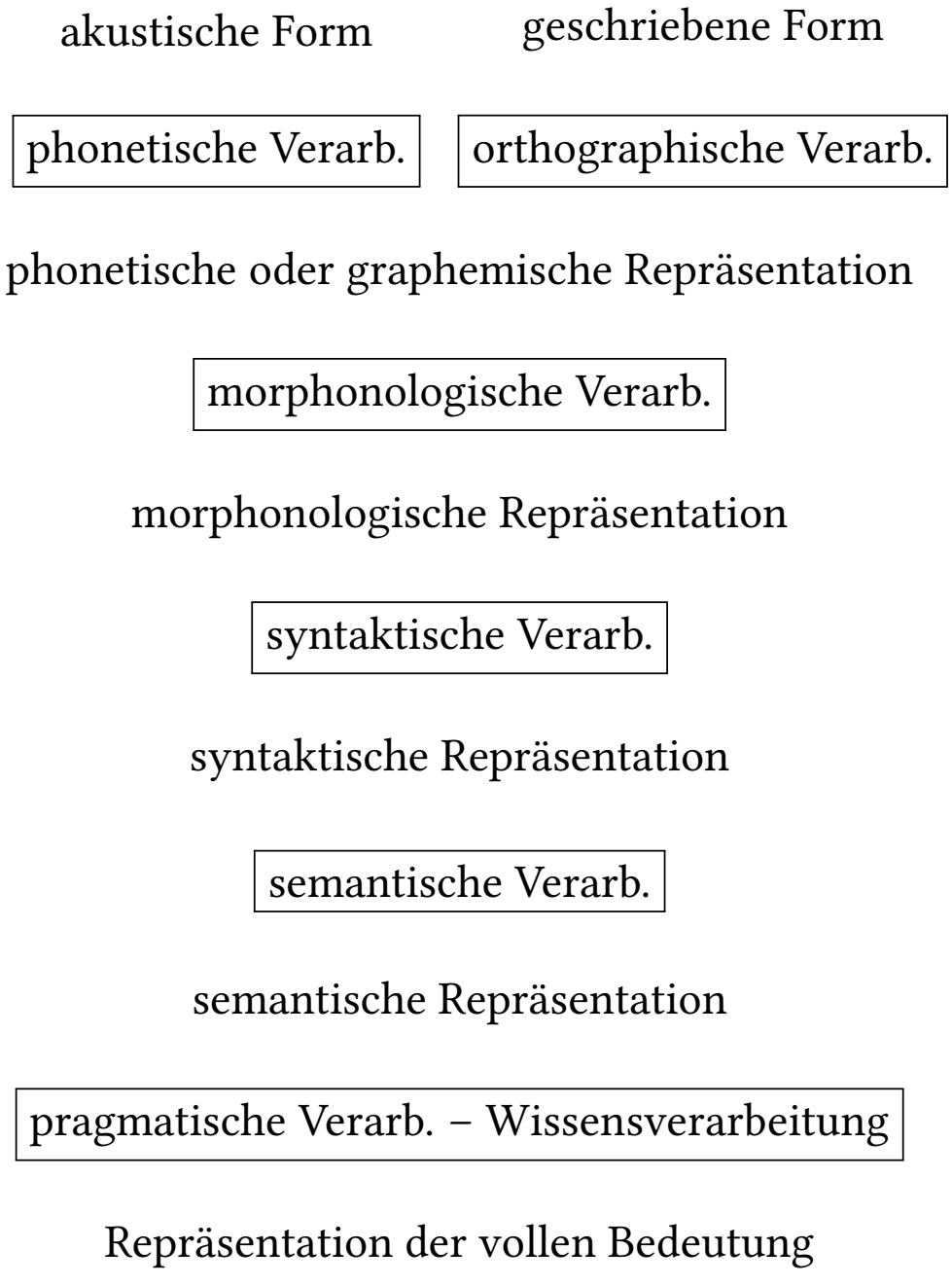
Zwei große Teilbereiche:

- angewandte Computerlinguistik: interdisziplinäres Forschungsgebiet (Linguistik, Informatik), das konkrete Algorithmen für die maschinelle Sprachverarbeitung entwickelt (maschinelle Übersetzung, Spracherkennung ...)
- theoretische Computerlinguistik: Teildisziplin der Linguistik, die formale berechenbare Modelle natürlicher Sprache entwickelt, implementiert und untersucht.

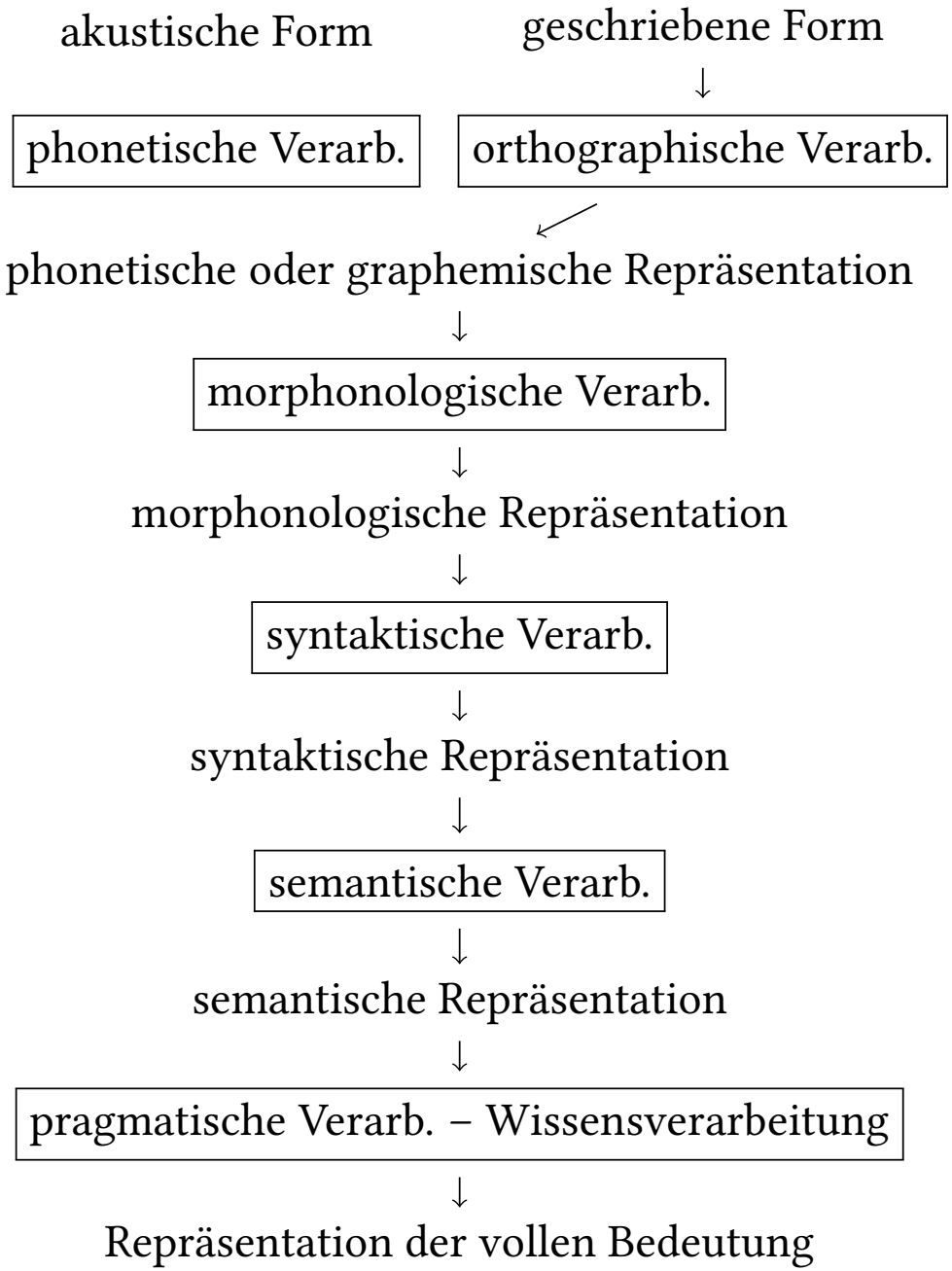
Bereiche der Computerlinguistik (4)



Komponenten eines Sprachmodells



Komponenten eines Sprachmodells: Textverstehen



Ambiguität (1)

Schwierigkeit bei der Sprachverarbeitung: Sprache ist hochgradig ambig.
Wir unterscheiden verschiedene Typen von Ambituitäten:

- *phonetische* Ambiguität (Homophone)
  Miene – Mine, Meer – mehr, viel – fiel
- *orthographische* Ambiguität (Homographen)
  übersetzen – übersetzen, umfahren – umfahren
- *lexikalische* Ambiguität (Homonyme)

  (1) Hans geht zur *Bank*

- *morphologische* Ambiguität
  Staubecken, Hauptpostsekretär

Ambiguität (2)

- *strukturelle/syntaktische* Ambiguitäten

  (2) Visiting relatives can be boring.

  (3) Peter fuhr seinen Freund sturzbetrunken nach Hause.

  (4) Ich traf den Sohn des Nachbarn mit dem Gewehr.

- *kompositionell-semantische* Ambiguität bzw. *Skopusambiguität*

  (5) Die zwei Mitarbeiter müssen vier Sprachen beherrschen.

  (6) Some student likes every course.

  (7) Alle Politiker sind nicht korrupt

- *pragmatische* Ambiguität

  (8) Könnten Sie die Aufgabe lösen?

  (9) Haben Sie eine Uhr?

Ambiguität (3)

Wege, mit Ambiguität umzugehen:

- Alle Lesarten berechnen.
  → Ist in der Regel nicht praktikabel, manchmal aber von theoretischem Interesse.
- Unterspezifizierte Repräsentationen verwenden, die alle möglichen Lesarten in einer kompakten Darstellung zusammenfassen.
  → Meistens Verwendung von *Constraints*.
- Nur die aufgrund des Kontextes präferierte(n) Lesarten berechnen.
  → Erfordert ein geeignetes gewichtetes oder probabilistisches Modell.

Inhalt dieses Kurses

(ist noch ein bisschen under construction …)

- Endliche Automaten, reguläre Ausdrücke, reguläre Grammatiken
- Anwendung: Morphologie
- N-Grams und Sprachmodelle
- POS-Tagging, Hidden Markov Models
- Kontextfreie Grammatiken, Kellerautomaten
- Merkmalsstrukturen und Unifikation
- Symbolisches Parsing
- Probabilistische Grammatiken und probabilistisches Parsing
- Distributionelle Semantik

Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S., Langer, H., and Klabunde, R., editors (2010). *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Spektrum Akademischer Verlag. 3. überarbeitete und erweiterte Auflage.

Jurafsky, D. and Martin, J. H., editors (2009). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Articial Intelligence. Pearson Education International. Second Edition.

Jurafsky, D. and Martin, J. H. (2017). Speech and language processing. an introduction to natural language processing, computational linguistics, and speech recognition. Draft of the 3rd edition. Available here: `https://web.stanford.edu/ jurafsky/slp3/`.