Einführung in die Computerlinguistik – Einführung

Dozentin: Wiebke Petersen

15.4.2010

Computerlinguistik: Die Wissenschaft

Carstensen et. al. (2004)

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

  → theoretische Computerlinguistik

Computerlinguistik: Die Wissenschaft

Carstensen et. al. (2004)

- 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

Computerlinguistik: Die Wissenschaft

Carstensen et. al. (2004)

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

  → maschinelle Sprachverarbeitung

Computerlinguistik: Die Wissenschaft

Carstensen et. al. (2004)

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

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

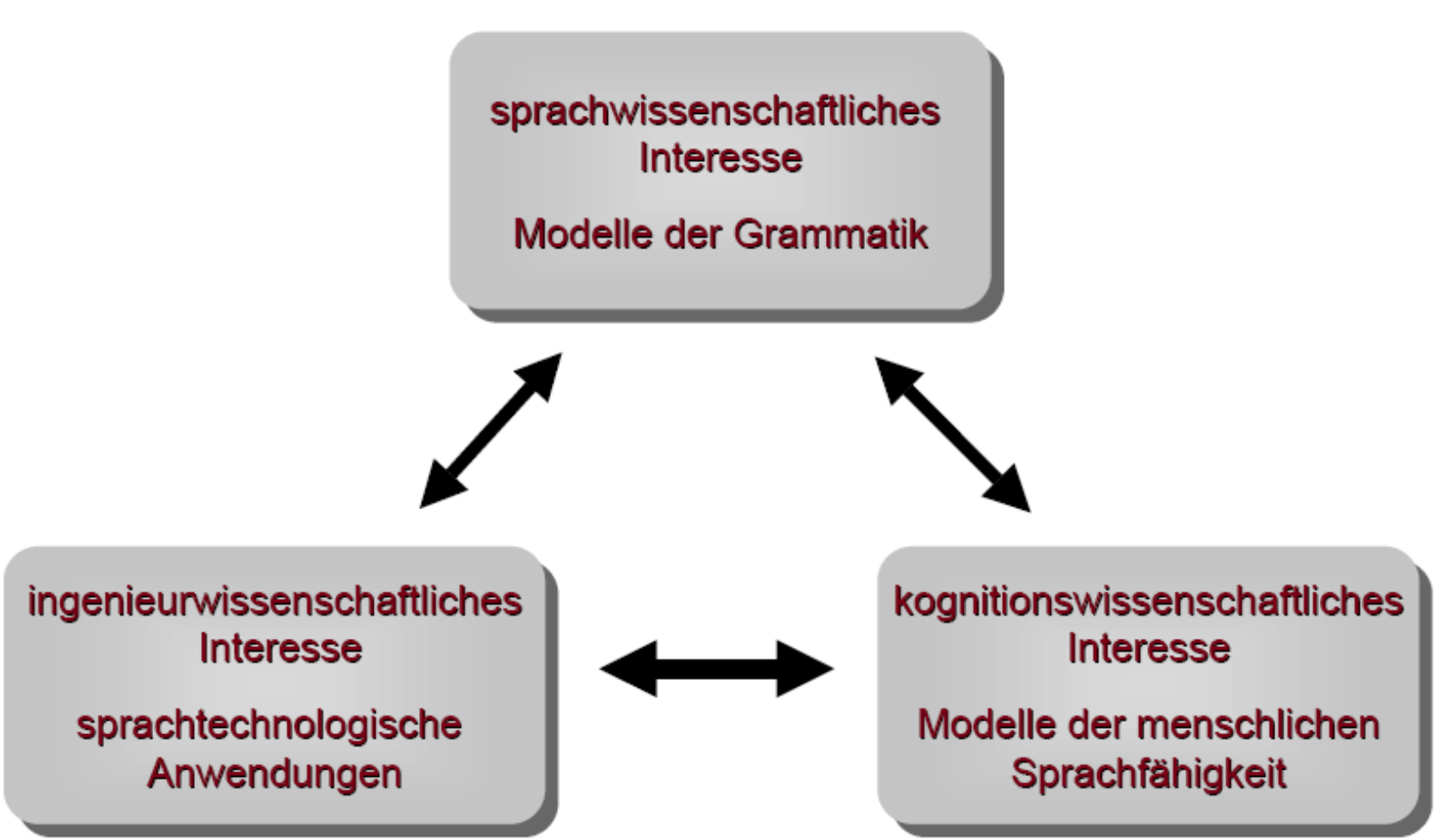
  → Sprachtechnologie

theoretical/applied CL

applied computational linguistics: interdisciplinary research field (between linguistics and computer science) which develops concrete algorithms for natural language processing (machine translation, machine speech recognition ...)

theoretical computational linguistics: discipline in modern linguistics which develops, implements and investigates computational models of human language.

Motivation



Häufige Abkürzungen

- Computational Linguistics (CL)
- Natural Language Processing (NLP)
- Language Engineering
- Human Language Technology (HLT)

Applications

advanced NLP applications

- dialogue systems / conversational agents
  - simplifies human-computer interaction
- machine translation
  - simplifies human-human interaction
- question answering
  - simplifies usage of the web

Applications

advanced NLP applications

- dialogue systems / conversational agents
  - simplifies human-computer interaction
- machine translation
  - simplifies human-human interaction
- question answering
  - simplifies usage of the web

simpler NLP applications

- spell checking
- grammar checking
- word count

machine translation

state of the art

Langenscheidt T1.

source Much older than communication problems between human beings and machines are those between people with different mother tongues. One of the original aims of applied computational linguistics has always been fully automatic translation between human languages. (aus Uszkoreit: What is Computational Linguistics?)

target Viel älter als Kommunikationsprobleme zwischen Menschen und Maschinen sind jene zwischen Leuten mit unterschiedlichen Muttersprachen. Eins der ursprünglichen Ziele von angewandter Rechnerlinguistik ist immer vollautomatische Übersetzung zwischen menschlichen Sprachen gewesen.

machine translation

Gladly to be little, and who is glad needs
it, that one is king.

machine translation

Gladly to be little, and who is glad needs it, that one is king.

Froh zu sein bedarf es wenig, und wer froh ist, der ist König.

machine translation

Gladly to be little, and who is glad needs it, that one is king.

Froh zu sein bedarf es wenig, und wer froh ist, der ist König.

Wenn you were and the degree of the hindrance at least 70 the actual costs were or existed at a degree of the hindrance of at least 50 simultaneously a considerable going-hindrance, also in the case of utilization of your own PASSENGER CAR are recognized the return journey or without individual record 60 cent per distance kilometer (30 cent per driven kilometer).

machine translation

Gladly to be little, and who is glad needs it, that one is king.

Froh zu sein bedarf es wenig, und wer froh ist, der ist König.

Wenn you were and the degree of the hindrance at least 70 the actual costs were or existed at a degree of the hindrance of at least 50 simultaneously a considerable going-hindrance, also in the case of utilization of your own PASSENGER CAR are recognized the return journey or without individual record 60 cent per distance kilometer (30 cent per driven kilometer).

Wenn Sie behindert waren und der Grad der Behinderung mindestens 70 betragen hat oder bei einem Grad der Behinderung von mindestens 50 gleichzeitig eine erhebliche Gehbehinderung bestand, werden auch bei Benutzung Ihres eigenen PKW die tatsächlichen Kosten der Hin- und Rückfahrt oder ohne Einzelnachweis 60 Cent je Entfernungskilometer (30 Cent je gefahrenen Kilometer) anerkannt. (Elster-Formular 2008)

machine translation

Can be credited to a final test on two different courses, it can only be credited to the extent of a degree program in store audits. If one of the subjects required in an audit is already stored in another compartment, in accordance with the choice of the subject also store another audit. The same applies for participation certificates.

machine translation

Can be credited to a final test on two different courses, it can only be credited to the extent of a degree program in store audits. If one of the subjects required in an audit is already stored in another compartment, in accordance with the choice of the subject also store another audit. The same applies for participation certificates.

(`http://translate.google.com`)

Lässt sich eine Abschlussprüfung auf zwei verschiedene Studiengänge anrechnen, kann sie nur auf den Umfang der in einem Studiengang abzulegenden Abschlussprüfungen angerechnet werden. Falls eine in einem der Fächer geforderte Abschlussprüfung bereits in einem anderen Fach abgelegt ist, ist nach Maßgabe der Wahlmöglichkeiten des Faches zusätzlich eine andere Abschlussprüfung abzulegen. Entsprechendes gilt für Beteiligungsnachweise. (BA Prüfungsordnung)

Sometimes human "translations" go wrong too!

Welsh text reads: "I am not in the office at the moment. Send any work to be translated."

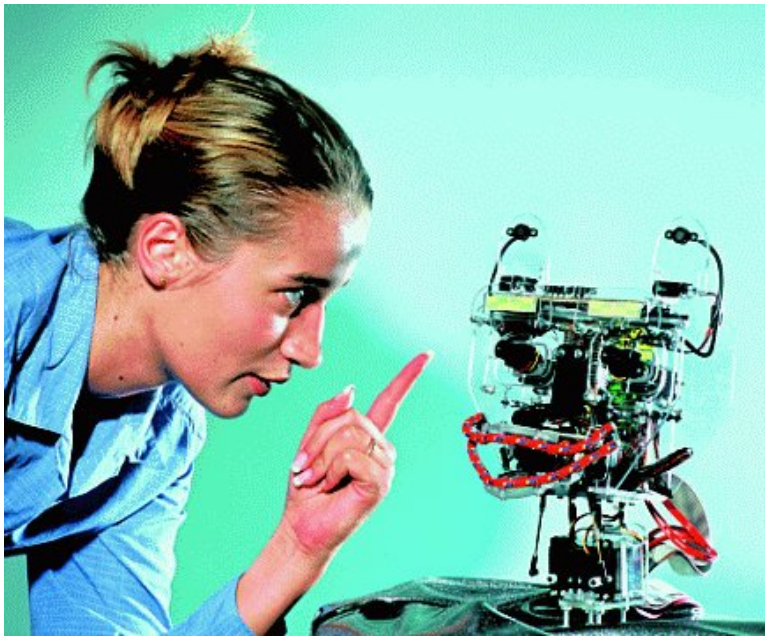
Sometimes human “translations” go wrong too!

Question answering

Mögliche Fragen

- Was bedeutet "homophon"?
- Wann wurde Heinrich Heine geboren?
- Wer regierte damals in Deutschland?
- Was denken Wissenschaftler über das menschliche Klonen?
- Wie verhalten sich CL und NLP zueinander?
- Wer ist der Rektor der HHU?
- An welcher Universität hat er zuvor gelehrt?
- Wie weit ist Düsseldorf von Gießen entfernt?
- Zu welcher Sprachfamilie gehört Zulu?

conversational agents



conversational agents

Interaction with HAL 9000 the computer in Stanley Kubrick's film "2001: A Space Odyssey":

Dave Bowman: Open the pod bay doors, HAL.

HAL: I'm sorry Dave, I'm afraid I can't do that.

required language knowledge

- speech recognition
- natural language understanding
- natural language generation
- speech synthesis

`http://www-306.ibm.com/software/pervasive/tech/demos/tts.shtml`

Knowledge needed to build HAL?

- Speech recognition and synthesis
  - Dictionaries (how words are pronounced)
  - Phonetics (how to recognize/produce each sound of English)
- Natural language understanding
  - Knowledge of the English words involved
    - What they mean
    - How they combine (what is a `pod bay door'?)
  - Knowledge of syntactic structure
    - I'm I do, Sorry that afraid Dave I'm can't

What's needed?

- Dialog and pragmatic knowledge
  - "open the door" is a REQUEST (as opposed to a STATEMENT or information-question)
  - It is polite to respond, even if you're planning to kill someone.
  - It is polite to pretend to want to be cooperative (I'm afraid, I can't...)
  - What is `that' in `I can't do that'?
- Even a system to book airline flights needs much of this kind of knowledge

Komponenten eines Sprachmodells

akustische Form

geschriebene Form

phonetische Verarbeitung

orthographische Verarbeitung

phonetische o. graphemische Repräsentation

morphonologische Verarbeitung

morphonologische Repräsentation

syntaktische Verarbeitung

syntaktische Repräsentation

semantische Verarbeitung

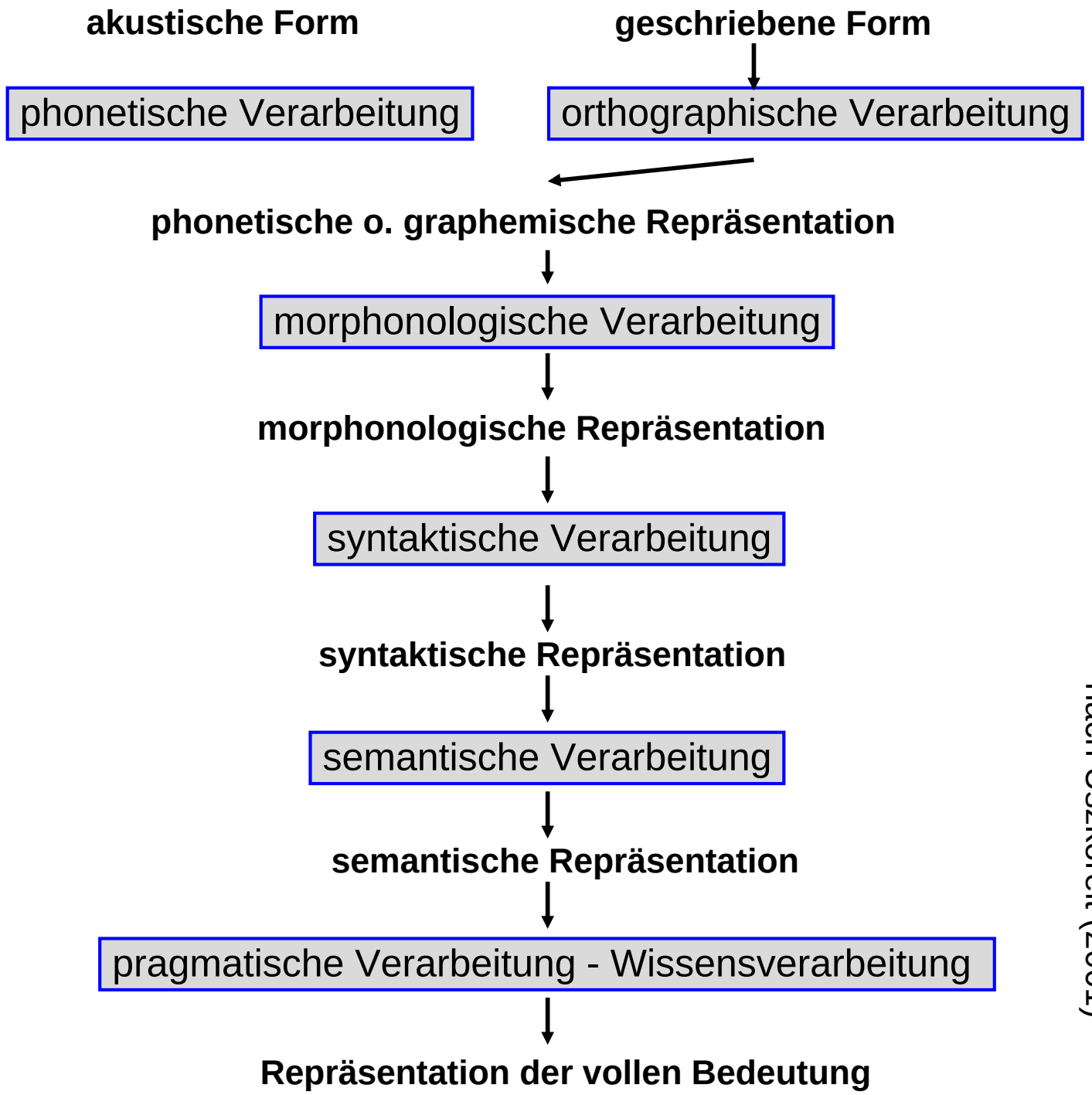
semantische Repräsentation

pragmatische Verarbeitung - Wissensverarbeitung

Repräsentation der vollen Bedeutung

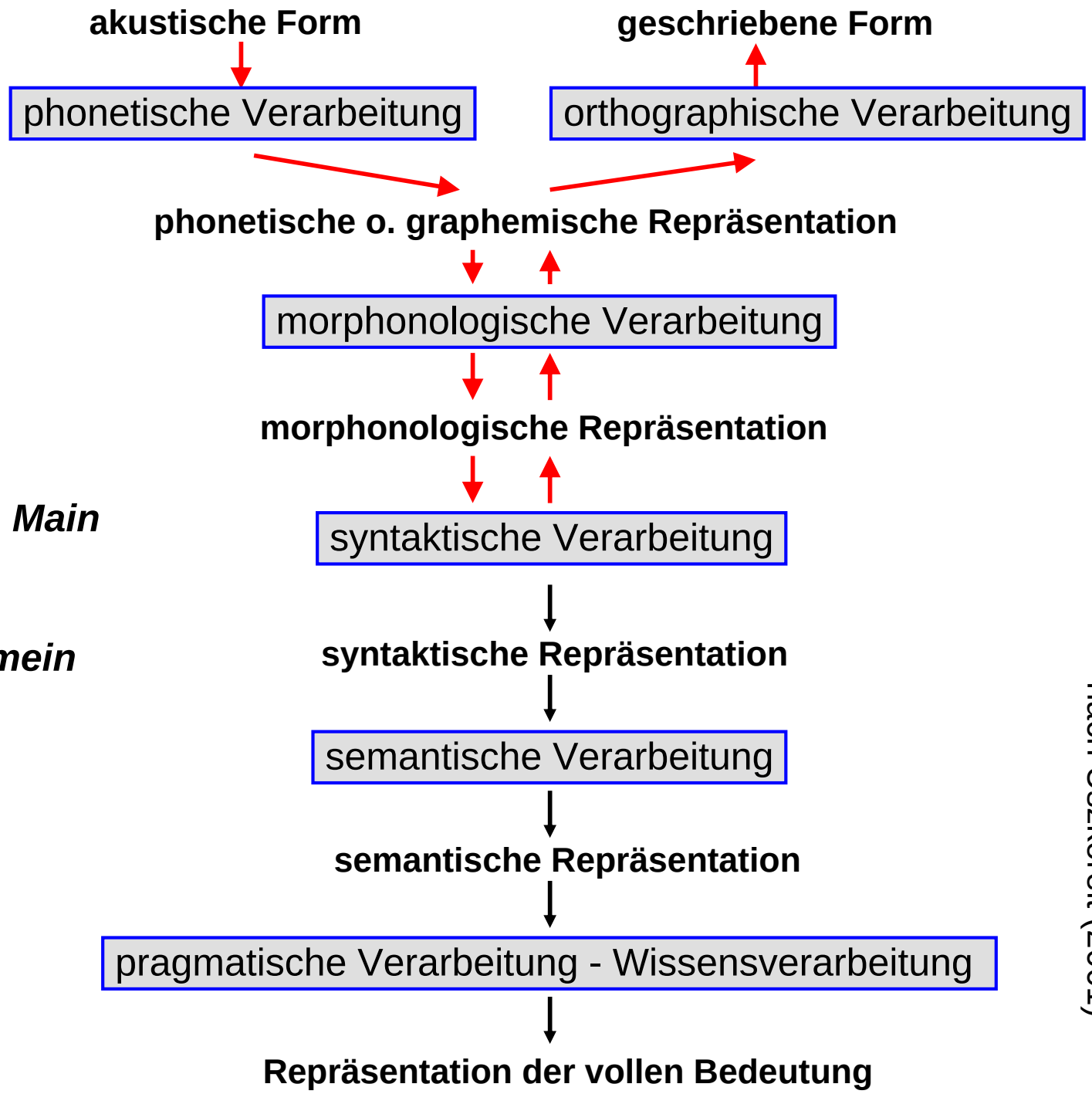
nach Uszkoreit (2001)

Textverstehen



nach Uszkoreit (2001)

Diktat



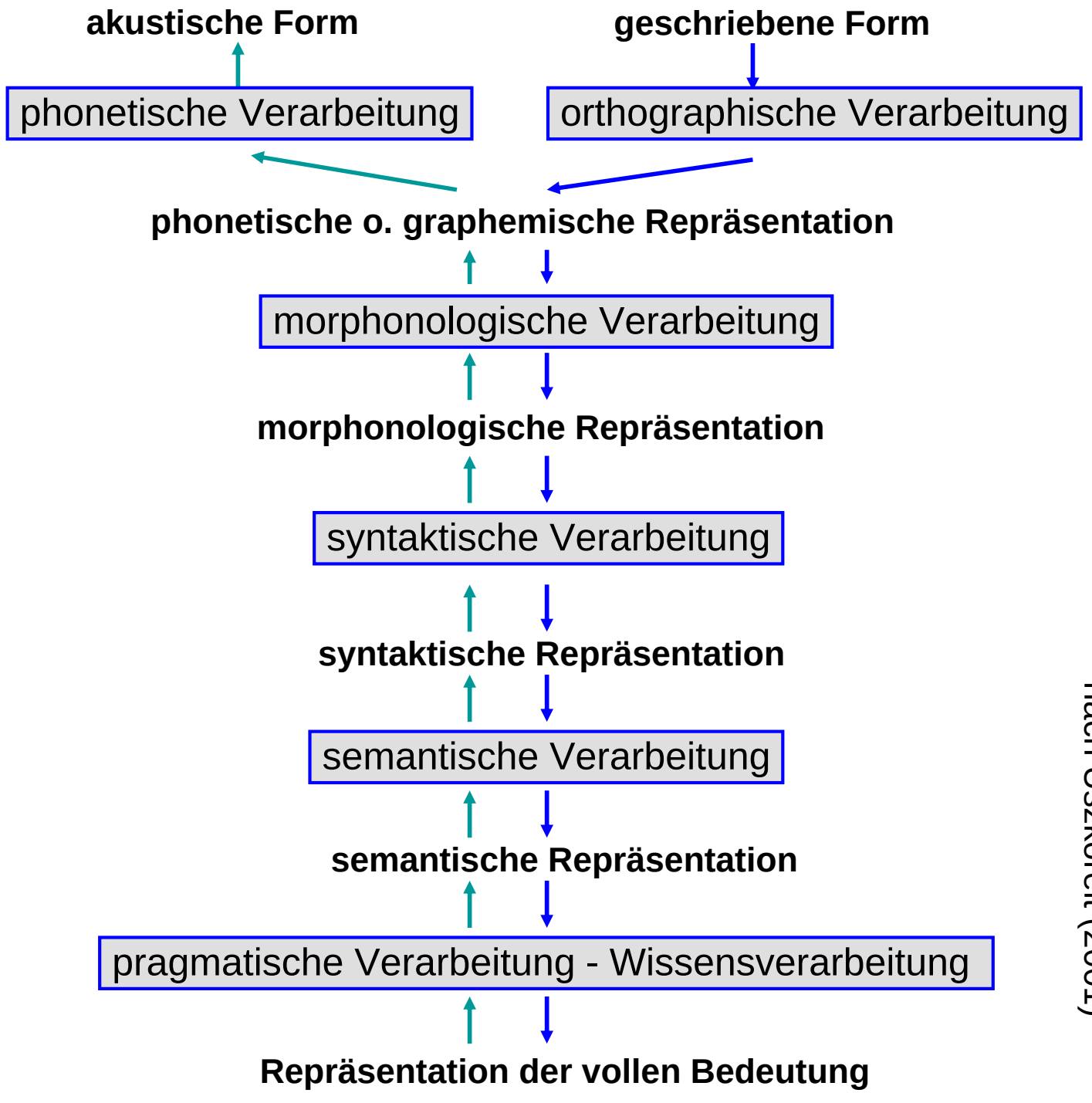
das Boot auf dem Main

oder

daß bot auf dem mein

nach Uszkoreit (2001)

Maschinelle Übersetzung



nach Uszkoreit (2001)

Ambiguität

phonetische Ambiguität (Homophone)

Miene - Mine

orthographische Ambiguität (Homographen)

übersetzen - übersetzen

lexikalische Ambiguität (Homonyme)

Ball - Ball

morphologische Ambiguität

Staubecken - Staubecken

Hauptpostsekretär

Lexikalische Ambiguität

Gewisse Lesarten sind weniger stark präferiert:

Auf dem Tisch lag ein Heft. *Auf der Werkbank lag ein Heft.*

Die Präferenz für eine Lesart kann durch den Kontext beeinflußt werden:

Der Mittelstürmer eröffnete den Ball.	versus	*Der Präsident eröffnete den Ball.*
Der Gärtner sprengte das Schloß.	versus	*Der Einbrecher sprengte das Schloß.*
The astronomer married a star.	versus	*The movie director married a star.*

Ambiguität II

syntaktische Ambiguität

Peter fuhr seinen Freund sturzbetrunken nach Hause.

Visiting relatives can be boring.

Ich traf den Sohn des Nachbarn mit dem Gewehr.

kompositionell-semantische Ambiguität

Die zwei Mitarbeiter müssen vier Sprachen beherrschen.

pragmatische Ambiguität

Könnten Sie die Aufgabe lösen.

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Der Satz weist lexikalische (L), syntaktische (S) und anaphorische (A) Ambiguitäten auf, die uns nicht auffallen.

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Der Satz weist lexikalische (L), syntaktische (S) und anaphorische (A) Ambiguitäten auf, die uns nicht auffallen.

Wieviele Lesarten besitzt dieser Satz?

258.048

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Das berechnet sich so:

- **L** ***Früher*** **kann sowohl eigenständiges Adverb als auch Komparativ von** ***früh*** **sein (2);**
- **L** **die Verbform** ***stellten*** **is ambig zwischen Präteritum und Konjunktiv (2);**
- **S** **die Nominalphrase** ***die Frauen*** **kann sowohl Subjekt als auch Objekt des Satzes sein (2);**
- **S** ***am Wochenende*** **kann die Insel, die Frauen oder das Verb modifizieren (3);**
- **S** ***mit Blumenmotiven*** **kann sich auf die Kopftücher beziehen, ein Instrument der Herstellung sein oder ein Adjunkt im Sinne von** ***gemeinsam mit Blumenmotiven*** **(3);**
- **L** ***her*** **hat auch eine direktionale Bedeutung (2);**

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Und weiter:

- **S** der Relativsatz könnte jede der vier Nominalphrasen im Plural modifizieren (4);
- **S** sowohl ***die*** als auch ***ihre Männer*** kann Subjekt des Relativsatzes sein (2);
- **A** das Possessivpronomen ***ihr*****e** kann auf jede der Nominalphrasen referieren (4);
- **L** ***Montagen*** hat eine zweite Lesart als Nominalisierung von ***montieren*** (2);
- **S** ***die Hauptinsel*** kann im Genitiv zu der vorangegangenen NP gehören oder im Dativ die Käuferin bezeichnen (2);
- **S** die drei Präpositionalphrasen des Relativsatzes können sich in insgesamt sieben Kombinationen mit den jeweils vorhergehenden NPs oder mit dem Verb verbinden (7);
- **L** ***verkauften*** zeigt wieder die Ambiguität zwischen Präteritum und Konjunktiv auf (2).

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Durch Multiplikation ergibt sich die Gesamtambiguität:

2 x 2 x 2 x 3 x 3 x 2 x 4 x 2 x 4 x 2 x 2 x 7 x 2 = 258.048

Hausaufgaben

Abgabe bis zum 22.4.2010
Hinweis: Für den BN muß nur eine der beiden Aufgaben bearbeitet werden.

1. Suchen Sie drei verschiedene Definitionen für den Term "Computerlinguistik" und vergleichen Sie diese (höchstens 300-400 Wörter ohne die Definitionstexte). Sie können die Definitionen entweder Büchern oder dem Internet entnehmen.
2. Erstellen Sie eine Liste aller Veranstaltungen, die für die Module C1, C2, C3 und C4 in diesem Semester angeboten werden, und ordnen Sie diese den vier Bereichen der Computerlinguistik nach Carstensen et. al. zu (Folie 2). Begründen Sie ihre Zuordnung jeweils in 1-2 Sätzen.

Referatsthemen: Anwendungen der CL (1)

Die Referatsthemen sind dem 5. Kapitel von Carstensen et. al. entnommen.

1. **Korrekturprogramme:** Rechtschreibkorrektur, Grammatikkorrektur Korrekturvorschläge, Verbesserung von Texterfassung mittels OCR.
2. **Computergestützte Lexikographie und Terminologie**: Hilfe bei der Erstellung und Pflege von Lexika; Akquisition lexikalischer Information, Repräsentation lexikalischer Information, Bereitstellung der lexikalischen Information für Anwendungen, Extraktion von Fachwortschatz und Identifikation von Synonymen, Extraktion von Relationen zwischen Konzepten.
3. **Volltextsuche und Textmining**: Indexkonstruktion, Auswertung von Suchanfragen, Retrievalmodell, Strukturierung großer Textkollektionen, Textklassifikation, Schlüsselwortextraktion, Analyse von Einzeltexten und Textkollektionen, Aufbau von Taxonomien.
4. **Textklassifikation**: Erlernen von Klassenprofilen anhand von Trainingsdaten, Klassifikationsalgorithmen
5. **Informationsextraktion**: Identifizierung relevanter Information in Texten, Instantiierung von Templates.

Referatsthemen: Anwendungen der CL (2)

6. **Textzusammenfassung**: Reduktion / Verdichtung, Textproduktion.
7. **Sprachsynthesesysteme**: Produktion gesprochener Sprache aus geschriebener Sprache, Computerarbeitsplatz für Blinde, telefonische Auskunftsysteme, Navigationsysteme.
8. **Spracherkennungssysteme**: Diktiersysteme, telefonische Auskunftsysteme; Signalanalyse, Geräuschfilterung, Adaption an verschiedene Sprecher.
9. **Dialogsysteme**: ELIZA, automatische Auskunftsysteme, natürlichsprachliche Benutzerschnittstellen.
10. **Sprachlehr- und -lernsysteme**: Hilfe bei dem Erwerb von Fremdsprachen; Anpassung an das individuelle Arbeitstempo und den Kenntnisstand, ortsungebunden, zeitlich flexibel, objektiv, nicht ermüdend

Referatsthemen: Anwendungen der CL (3)

12. **Elektronische Kommunikationshilfen**: Wort- und Satzvervollständigung (SMS), Texttelephone, Textvereinfachungswerkzeuge, ...
13. **Angewandte natürlichsprachliche Generierungs- und Auskunftssysteme**: Wettervorhersagen, Gesundheitswesen, technische Dokumentationen, Computerspiele, ...
14. **Maschinelle Übersetzung**: Vollautomatische Übersetzung, Computergestützte Übersetzung.