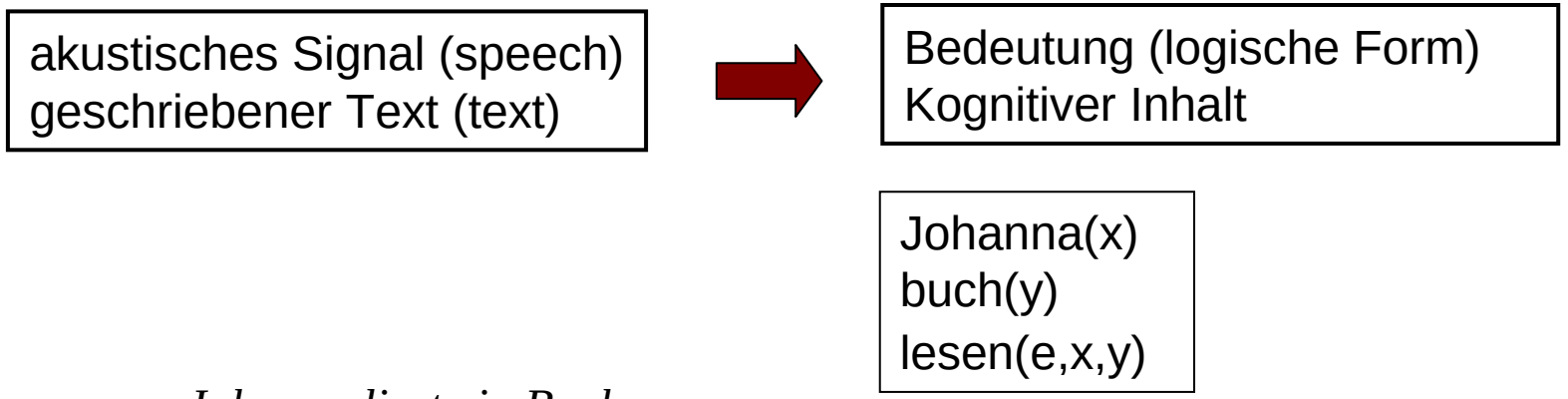
Computerlinguistik

– Einführung und Überblick

Was ist Computerlinguistik?

- Computerlinguistik beschäftigt sich mit der maschinellen Verarbeitung natürlicher Sprache.
- Sie beschäftigt sich mit den *strukturellen Eigenschaften* und den *Verarbeitungsmechanismen* natürlicher Sprache.



Johanna liest ein Buch.

Ziele der Computerlinguistik

- Erklärung / Simulation von Sprachverständnis
  - Analyse (Sprachverstehen)
  - Generierung (Sprachproduktion)
  - Übersetzung
  - Modellierung menschlicher Sprachperformanz
- Studium der formalen Eigenschaften von Sprache
  - Formale Modellierung und algorithmische Umsetzung
- Verarbeitung sprachlicher Daten/Information
  - Sprache als Träger von Information
  - Automatische Analyse des Informationsgehalts für praktische Anwendungen in der Informationsgesellschaft

Aspekte und Anwendungen der Computerlinguistik

Computational Linguistics
Human Language Technology
Natural Language Processing

Aspekte und Anwendungen der Computerlinguistik

Natural language understanding: Simulation von Sprachverstehen

- Bedeutung
- Kommunikation, Dialog und Interaktion
- Sprachliches Handeln: *zeigen* (referieren), *befehlen, ..*
- Gemeinsames Wissen – „privates Wissen“ (private beliefs)

Computational Linguistics
Human Language Technology
Natural Language Processing

Nachbardisziplinen

- Künstliche Intelligenz

  Planen, Wissensrepräsentation, Schlussfolgern (Inferieren), Mensch-Maschine-Interaktion
- Robotik

  Mechanik und Steuerung, Bildverarbeitung, Navigation

Aspekte und Anwendungen der Computerlinguistik

Sprache und Gestik

– Bedeutung

– Situative Sprache

  – Deixis („*dieses Buch*“)

  – Sprachliches Handeln (nicken, winken, ..)

Computational Linguistics
Human Language Technology
Natural Language Processing

Hier:

• **Übersetzung** natürlicher Sprache in *American Sign Language*

Weiter:

• Sprachlernsysteme (CALL: Computer Aided Language Learning)

• Rechtschreibkorrekturprogramme

• Übersetzungshilfen

Aspekte und Anwendungen der Computerlinguistik

Verarbeitung gesprochener Sprache:
z.B. Benutzerportale *(Reiseauskunft, Tourismus, ...)*

- Spracherkennung und –synthese
- Abbildung auf Worte (Wortgraphen)
- Prinzipien der Kommunikation und Dialogsteuerung

Computational Linguistics
Human Language Technology
Natural Language Processing

Probleme und Herausforderungen

- Sprecher(un)abhängigkeit
- Segmentierung: *recognize speech – wreck a nice beach*
- Unbekanntes Vokabular
- Prosodie: Sprache als Emotionsträger (*emotional speech*)

Aspekte und Anwendungen der Computerlinguistik

Computationelle Psycholinguistik

- Simulierung der Sprachverarbeitung durch statistische Modelle
- Messung von Blickbewegung zur Erforschung kognitiver Prozesse bei der Sprachverarbeitung
  - Garden Path Effekte: *A horse raced past the barn fell*
- *Empirische Validierung* (computer)linguistischer Theorien

Computational Linguistics
Human Language Technology
Natural Language Processing

Nachbardisziplinen
- Neurolinguistik
- Klinische Linguistik

Aspekte und Anwendungen der Computerlinguistik

Sprache als Informationsträger

- *WWW: Verfügbarkeit großer Mengen sprachlicher Daten*
  - *Wikipedia, Blogs, Literatur, Tabellen, Listen, ...*
- Sprache: Informationen in „unstrukturierter“ Form
- Informationen in multiplen Sprachen

Computational Linguistics
Human Language Technology
Natural Language Processing

Anwendungen

- Dokumentensuche (Information Retrieval): Suche nach relevanten *Dokumenten*
- Textzusammenfassung: Erstellung eines Exzerpts aus (einem/mehreren) Texten
- Informationsextraktion (Information Extraction): Extraktion relevanter *Fakten*
- Fragebeantwortung: Suche nach Antworten auf spezielle Fragen
  - Faktoid: *Wann wurde Leonardo da Vinci geboren?*
  - Komplex: *Wie alt war Leonardo da Vinci als Michelangelo geboren wurde?*
- Automatisches Lernen von Wissensbasen aus Texten (Ontology Learning)

Aspekte und Anwendungen der Computerlinguistik

Verarbeitung domänenspezifischer sprachlicher Information

- *Information Management*
- *Strukturierte und unstrukturierte (sprachliche) Daten*
- *Ontologien (Begriffshierarchien)*

Computational Linguistics
Human Language Technology
Natural Language Processing

Anwendungen

- Automatische Suche in wissenschaftlichen Artikeln (eScience)
- Suche in firmeninternen Informationsportalen
- Automatische Email-Beantwortung in Call-Centern:
  Sortierung, Beantwortung

Aspekte und Anwendungen der Computerlinguistik

Maschinelle Übersetzung *(machine translation, MT)*

- *Gesprochene – geschriebene Sprache*
  - *– Dokumente (Produkte, Legislation, ...)*
  - *– Simultanübersetzung (Konferenzen)*
- *Sprachübergreifende Informationssuche*

Computational Linguistics
Human Language Technology
Natural Language Processing

Probleme und Herausforderungen

- Unterschiedliche linguistische Eigenschaften verschiedener Sprachen
  - Lexik, Konzepte, Wortreihenfolge, Auslassungen (Determiner, Subjekt, ...)
- Interpretation, Wissen und Übersetzung
- Varianten
  - Vollübersetzung (wissensbasiert – beispielbasiert – statistisch (SMT))
  - Unterstützte Systeme (HAMT) – human-aided MT
  - Unterstützende Systeme (MAMT) – machine-aided MT

Aspekte und Anwendungen der Computerlinguistik

- Computerlinguistik in der Informationsgesellschaft
  - **Informationssuche und –management**
  - **Überwindung von Sprachbarrieren**
    - Übersetzung
    - Sprachübergreifende Informationssuche
  - **Mensch-Maschine-Interaktion**
- Relativ neu:
  Computerlinguistik als Hilfswissenschaft in Geistes- und Sozialwissenschaften
  - Cultural Heritage
  - Wissensmodellierung
  - Extration von Wissen/Informationen aus textuellen Daten

Computerlinguistik als Wissenschaft

- Zugrundeliegende Wissensgebiete
  - Linguistik
  - Informatik
  - ➢ Computerlinguistik: neue, eigenständige Methoden

Computerlinguistik als Wissenschaft

- **Linguistisch orientiert:** Formalisierung und Implementierung linguistischer Theorien
  - Kohärente Beschreibung sprachlicher Daten eines Teilbereichs der Linguistik
  - Formalisierung erlaubt Implementierung
  - Implementierung erlaubt Evaluierung/Validierung
- **Empirisch orientiert:** Systematische Sammlung und Aufbereitung linguistischer Daten
  - Aufbau von Sprachkorpora
  - Abfragewerkzeuge
  - Statistische Auswertung (→ Sprachmodellierung)

Computerlinguistik als Wissenschaft

- **Psychologisch orientiert:** Modellierung menschlicher Sprachperformanz (→ Kognitionswissenschaft)
  - Unterspezifikation *(Vier Männer trugen zwei Klaviere nach oben)*
  - Verarbeitungdauer und –strategien
  - Unterbestimmtheit, Vagheit, Präferenzen
- **Informatik-orientiert:** Effiziente maschinelle Verarbeitung natürlicher Sprache
  - Mittels Methoden aus Informatik und der Künstlichen Intelligenz
  - Für bestimmte Anwendungszwecke

Computerlinguistik und Nachbardisziplinen

- Linguistik
  - Gemeinsamer Untersuchungsgegenstand
    - Struktur und Funktion natürlicher Sprache(n)
  - Orientierung an linguistische Teilbereichen und Untersuchungsmethoden
    - Phonologie, Morphologie, Syntax, Semantik und Pragmatik
- Informatik
  - Datenstrukturen, effiziente Verfahren, Systemarchitektur
    - Modellierung, Algorithmik, Implementierung
  - Theoretische Informatik
    - Berechenbarkeit, Komplexitätstheorie, formale Sprachen
  - Praktische Informatik
    - Parallelverarbeitung, Cluster Computing

Computerlinguistik und Nachbardisziplinen

- Philosophie
  - Sprachphilosophie
    - Sprache in Relation zu Denken, Handeln und sozialer Gemeinschaft
    - Formale Logik: präzise Darstellung sprachlicher Phänomene
- Künstliche Intelligenz
  - Such- und Planungsverfahren
    - Spracherkennung, grammatische Analyse, Generierung
  - Wissensrepräsentation und Inferenzsysteme
    - Informationsverarbeitung, „intelligente“ Systeme (Expertensysteme)
    - Formalisierung enzyklopädischen Wissens (Cyc – Cycorp.)
  - Logikbasierte Programmierung: LISP, PROLOG

Computerlinguistik und Nachbardisziplinen

- Kognitionswissenschaft
  - Sprachfähigkeit als kognitiver Prozess:
    Menschliche Sprachverarbeitung und ihre Beziehung zu allgemeinen kognitiven Prozessen
    - Garden-Path Effekte: *A horse raced past the barn fell.*
- Mathematik und Logik
  - Mathematisch-logische Basistheorien
    - Prädikatenlogik, Mengenlehre, Funktionen, etc.
  - Automatentheorie, formale Sprachen, Komplexitätstheorie
    - Komplexität von Sprache und adäquate Formalisierung (‚Ausdrucksstärke')
  - Graphentheorie: Merkmalstrukturen, Unifikation, Wortgraphen
  - Statistik: Frequenzbasierte Sprachmodellierung

Theoretische Computerlinguistik

Untersuchung prinzipieller Fragestellungen der maschinellen Sprachverarbeitung

- Berechenbarkeit, Adäquatheit, Erlernbarkeit
- Eignung eines Formalismus zur Beschreibung bestimmter Phänomene

- Komplexität natürlicher Sprache und Modellierbarkeit durch bekannte Formalismen und Berechnungsverfahren
- Adäquatheit von Formalismen für unterschiedliche Teilaspekte natürlicher Sprachen
  - Unterschiedliche Formalismen für Phonetik, Phonologie, Morphologie, Syntax, Semantik, Pragmatik
  - Möglichkeiten (pros und cons) für ebenenübergreifende Repräsentationen

Angewandte Computerlinguistik

Ziel: Erfolgreiche Modellierung sprachlichen Wissens auf einer Maschine, zur Lösung spezieller Anwendungsprobleme

- **Entwicklung von Formalismen zur adäquaten Modellierung linguistischer Teilaspekte**
  - Adäquate Ausdrucksstärke
  - Deklarative, sprachunabhängige Beschreibung
  - Allgemeine, modulare und effiziente Implementierung
- **Acquisition linguistischen Wissens** (‚Linguistische Ressourcen')
  - Lexika, grammatische Beschreibungen (= Grammatiken)
  - Sprachkorpora (gesprochene, geschriebene Sprache im Kontext)
- **Entwicklung von Algorithmen, Methoden**
  - Spracherkennung, Morphologie, Parsing, etc.
  - Informationssuche, -extraktion, Fragebeantwortung, Textzusammenfassung, ...
- **Evaluation**
  - Anhand real vorkommender Daten
  - Nach anerkannten Standards und Messverfahren

Beschreibungsebenen natürlicher Sprache

- Phonetik und Phonologie
  - Artikulatorische Merkmale und Lautstruktur
  - Wortsegmentierung, Aussprache, Prosodie
- Morphologie
  - Bildung und Struktur von Wörtern
  - Systematische Beziehungen zwischen Wörtern und Wortformen
    - Flexion, Derivation, Komposition *(schön – Schönheit – Schönheitskönigin)*
  - Prozesse/Regeln zur Erzeugung von Wortformen

Beschreibungsebenen natürlicher Sprache

- Syntax
  - Struktur von Sätzen
    - Konstituenz, Dependenz, Wortordnung
  - Grammatikalität (Wohlgeformtheit)
- Semantik
  - Bedeutung sprachlicher Einheiten (Wort, Satz, Text)
  - Komposition von Bedeutung aus Teilstrukturen
    - *Südliche Hemisphäre, kleiner Elefant*
- Pragmatik
  - Zweck, Wirkung, Intention sprachlicher Äußerungen („sprachliches Handeln“)
  - Präsuppositionen sprachlicher Äußerungen
    - *Fritz gelang es, das Auto wieder in Gang zu bekommen.*

Aspekte computerlinguistischer Modellierung

- *Repräsentation sprachlicher Strukturen* in einer Datenstruktur und Modellierung der *Prozesse (Verarbeitungsmechanismen, Algorithmen)*, die auf diesen Strukturen operieren
- Verarbeitungs“richtung“
  - *Analyse*: Sprachliche Eingabe → formale Repräsentation
  - *Generierung*: Formale Repräsentation → Erzeugung einer sprachlichen Oberflächenform

Aspekte computerlinguistischer Modellierung

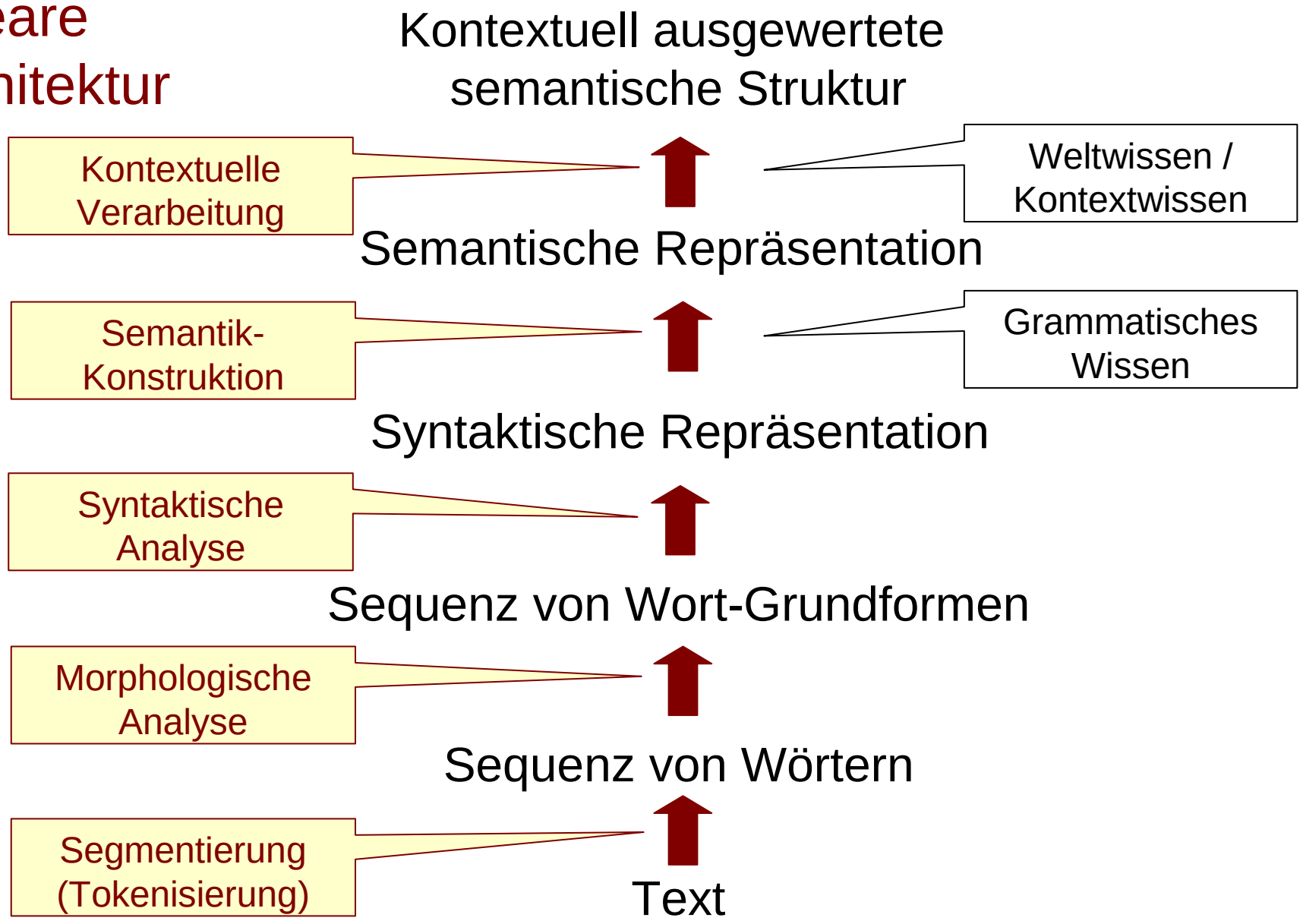
- Ebeneninteraktion: Wie stehen die Strukturen verschiedener Ebenen miteinander in Beziehung?
  - *Der Dieb stehlte den Diamanten.*
  - *Der Dieb gewitzte stahl Diamanten den.*
  - *Colourless green ideas sleep furiously.*
  - *Der Freak betrat die Kneipe. Sie bestellte ein Bier.*
- Verarbeitungsarchitektur
  - Lineares Modell: alle Ebenen vollständig und nacheinander
    - Fehler einer Ebene beeinflussen die Analyse „aufbauender“ Ebenen
  - Inkrementelle Verarbeitung: alle Ebenen partiell und parallel
    - Erwartungsgesteuert, prädiktiv

Holzwegsätze

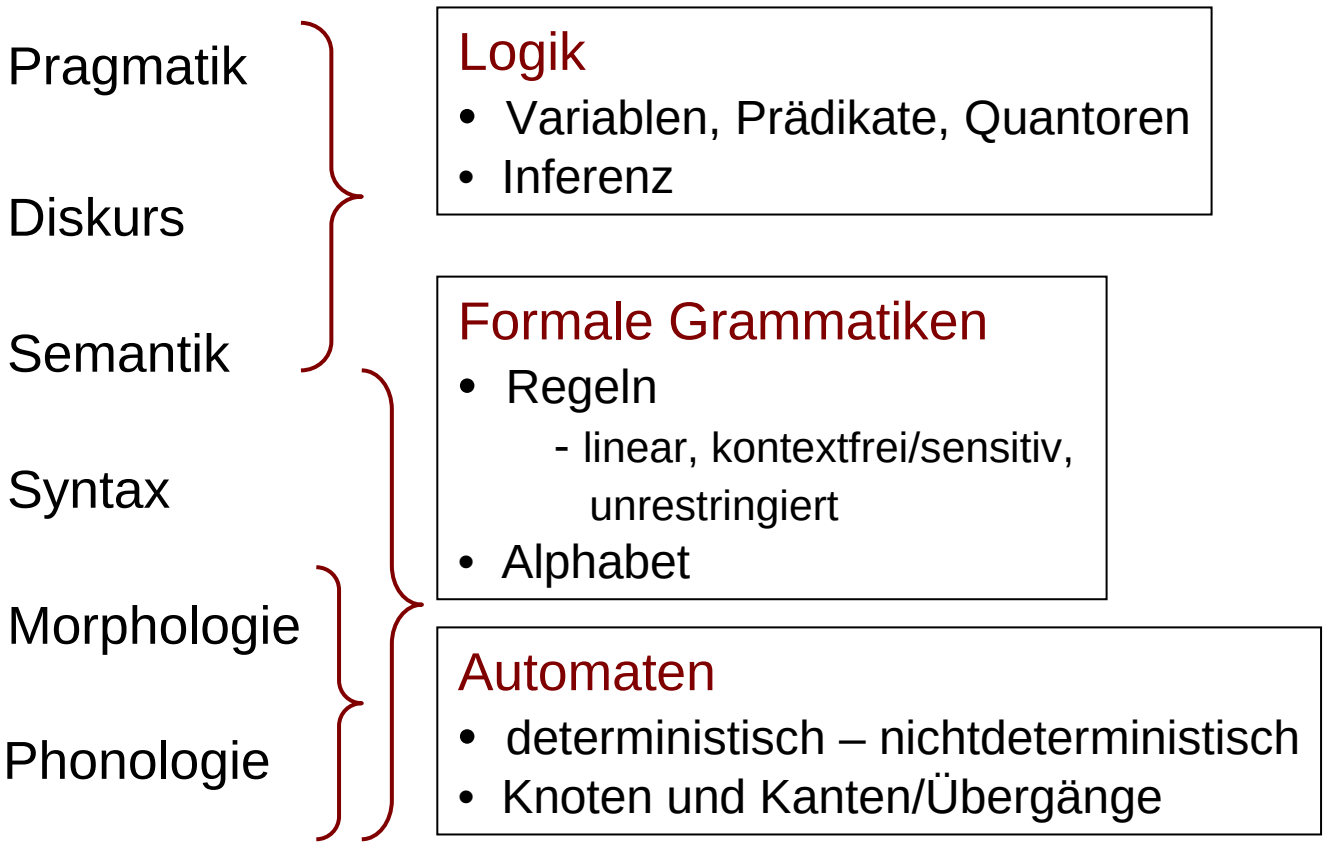
- *Er bezichtigte den Vater des Schreibens unkundiger Kinder.*
- *Peter beschuldigte sie der Geheimniskrämerei ähnlichen Verhaltens.*

Quelle: Hans Uszkoreit

Lineare Architektur



Formale Modelle und Algorithmen



Ambiguitäten auf allen Repräsentationsebenen

- Spracherkennung
  - *recognize speech – wreck a nice beach*
- Tokenisierung / Segmentierung
  - *Ich habe Zeit um fünf. Würde Ihnen das passen?*
    Ich habe Zeit. Um fünf würde Ihnen das passen.
- Morphologische Analyse
  - *Er kennt sich mit Kiefern (masc./fem.) aus.*
  - *Time flies(V/N) like(V/P) an arrow*
  - *Abteilungen: Ab-teil-ungen, Abt-ei-lungen, Staubecken, ..*
- *Syntaktische Analyse*
  - *Hans liebt Maria. – Aber sie/er ihn/sie nicht.*
  - *Fritz sieht die Frau mit dem Fernglas. -- Wer hat das Fernglas?*

Ambiguitäten auf allen Repräsentationsebenen

- Semantische Analyse
  - *Ein grüner Junge* (lexikalische Ambiguität)
  - *Jeder Holländer liebt eine Frau.* (strukturelle Ambiguität)
  - *Paul hat Maria nur zu seinem Fest eingeladen.*
    (a) .. Nicht ins Kino.
    (b) .. Er hat sie nicht geküßt.

- Kontextuelle Auswertung
  - *Paul kann Fritz nicht leiden. Er mag Maria.* (Anaphernauflösung)
  - *Fritz mag seine Mutter. Peter auch.* (Ellipsenauflösung)

Ambiguitäten auf allen Repräsentationsebenen

- Ambiguitäten potenzieren sich ... von Ebene zu Ebene
- Konsequenz: kombinatorische Explosion in Repräsentation und Verarbeitung!
- Wenige Alternativen werden ausgefiltert

Ambiguitäten

„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.“

Wieviele Lesarten besitzt dieser Satz?

Quelle: Hans Uszkoreit

Ambiguitäten

„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.“

Wieviele Lesarten besitzt dieser Satz? **–** 258.048

Quelle: Hans Uszkoreit

Ambiguitäten

„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.“

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

Quelle: Hans Uszkoreit

Ambiguitäten

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

Quelle: Hans Uszkoreit

Ambiguitäten

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

Wieviele Lesarten besitzt dieser Satz?

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

Quelle: Hans Uszkoreit

Ambiguitäten auf allen Repräsentationsebenen

- Strategien
  - Filtern/Disambiguierung: statistisch, präferenzgesteuert, first-best
    - Worthypothesen, PoS-Tagging, stochastische Grammatik
  - ‚Kompakte', nicht-redundante Repräsentationen für effiziente Verarbeitung ambiger Strukturen
    - Finite Automaten, Chartparsing, unterspezifizierte Strukturen

    Notwendig: nachfolgende Module müssen partielle/gepackte Strukturen verarbeiten können

Herausforderungen für die Computerlinguistik

- Ambiguität und Vagheit
  - Sprache ist mehrdeutig und vague (wie groß ist „*groß*“?)
- Komplexität
  - Feingranulare Analyse und Ambiguität → hohe Komplexität
- Vollständigkeit und Präzision
  - Sprache ist unaufzählbar und veränderlich (Neologismen,..)
  - Vollständigkeit *und* Präzision sind schwer zu erreichen
- Vielfalt der Sprachen
- Ambiguität und Disambiguierung
- Weltwissen

Empirische, datengetriebene, statistische Verfahren in der Computerlinguistik

- Erweiterung formaler Sprachmodelle durch *Wahrscheinlichkeiten*
- Beispiel: Wortartenambiguität
  - *book* – NN – 0.7 / *book* – VB – 0.3 / *book* – CD – 0
- Ermittlung der Wahrscheinlichkeiten durch statistische Analyse großer Sprachkorpora
  - *„take a body of English text (called a corpus) and learn the language by noting statistical regularities in that corpus“* (Charniak 1993)
- Grundlegende Annahme statistischer Modellierung
  - *Frequenzen* in einem gegebenen Korpus können zur Estimierung der *Wahrscheinlichkeit* von Lesarten benutzt werden
  - Je höher die empirisch ermittelte Wahrscheinlichkeit in einem Korpus, desto plausibler ist die entsprechende Lesart in einem neuen Kontext

➢ **Disambiguierung auf Basis empirisch ermittelter Wahrscheinlichkeiten**