

SPRACHTECHNOLOGIE

Unter diesem Begriff fasst man alle Bemühungen zusammen, die den Computer in die Lage versetzen sollen, mit menschlichen Sprachäußerungen *vernünftig* umzugehen.

- z.B.: • Gesprochene Sprache in geschriebene zu
 überführen, und umgekehrt
- Sprachliche Äußerungen von einer Sprache in eine andere zu übersetzen
 - Sprachliche Anfragen befriedigend beantworten zu können
 - Die immensen Mehrdeutigkeiten menschlicher Sprachen zu vereindeutigen
 - Texte zu ‚verstehen‘ und für Menschen ‚verständliche‘ Texte zu erzeugen
 - ...

Formen natürlicher Sprache

Gesprochene Äußerungen:

- Wörter sind in kontinuierlich gesprochener Sprache (meist) nicht durch Pausen getrennt.

„hasdumaneuro? kristmorgnwieda“

- Dieselbe Äußerung wird von verschiedenen Menschen auf ganz unterschiedliche Weise ausgesprochen.

(akustischer Fingerabdruck)

- Gesprochene Äußerungen sind oft unvollständig oder sogar grammatisch falsch.

Schriftliche Äußerungen:

- Orthographie und Wortstellung sind normiert (Duden)
- Wörter sind durch Leerzeichen und/oder Satzzeichen voneinander getrennt.
- Mimik und Gestik sind ausgeschlossen, da der Adressat eines Textes in der Regel nicht anwesend ist.
- Schriftliche Texte entstehen ‚überlegter‘ als gesprochene Äußerungen (kein Zeitdruck).

Wer hat sich bislang um solche Probleme gekümmert?

- **Linguistik / Sprachwissenschaft**
- **(Sprach-) Psychologie**
- **(Sprach-) Philosophie**
- **Computerlinguistik / Sprachverarbeitung**
- **Kognitionswissenschaft**

S P R A C H T E C H N O L O G I E

ist der Zweig der Computerlinguistik/bzw. Informatik, der weniger auf Ähnlichkeit zu menschlichen Prozessen der Sprachverarbeitung aus ist, sondern die Mensch-Rechner-Interaktion und deren Effizienz betrachtet.

Dies bedeutet allerdings nicht das Ausklammern kognitiver Kriterien.

Grundannahmen der modernen Linguistik

Geburtstunde: F. de Saussure: Cours de Linguistique Generale, 1916

Noam Chomsky: Aspects of the Theory of Syntax, 1965



- Sprachverstehen und Sprachproduktion sind vollständig formalisierbar.
- Die Grammatik beschreibt die Kompetenz eines idealen Sprecher/Hörers.
- Die linguistische Einheit der Grammatik ist der Satz.
Jeder Satz kann wohlgeformt oder nicht wohlgeformt sein.
- Mit einem endlichen Inventar von Regeln wird eine unendliche Menge von Sätzen erzeugt
(und die Struktur der wohlgeformten beschrieben).

Computerlinguistik hat die Erforschung, d.h. die Beschreibung und Erklärung des menschlichen Sprachverhaltens durch formale, auf dem Computer realisierbare Modelle zum Gegenstand und entwickelt, auf den Erkenntnissen dieser Forschung aufbauend, natürlichsprachliche Anwendungssysteme.

Ihre zentralen Aufgaben sind:

- die Entwicklung und Untersuchung von Formalismen, die die Repräsentation von grammatischem, semantischem und pragmatischem Sprachwissen ermöglichen und die Bereitstellung von Algorithmen zur Verarbeitung dieses Wissens,
- die Bereitstellung von Grammatikwerkzeugen und Arbeitsumgebungen für die Erforschung und Implementierung von natürlichsprachlichem Wissen,
- die Modellierung und Simulation der komplexen Zusammenhänge und Abläufe in Sprachverstehens- und Sprachproduktionsprozessen in geschriebener und gesprochener Sprache
- sowie die Entwicklung natürlichsprachlicher (NL-) Anwendungssysteme.

[Görz et al. 1990, 276]

Beschreibungsebenen für natürlichsprachliche Äußerungen

SYNTAX definiert die Struktur einer Äußerung (bzw. eines Satzes)

- Reihenfolge von Äußerungselementen (Satzstellung)
- Beziehungen zwischen Äußerungselementen
(z.B. Numeruskongruenz zwischen Subjekt und Verb)
- • •

** Den Stuhle Lehne zerbrochen worden warum.*

SEMANTIK definiert die Bedeutungsbeziehungen zwischen sprachlichen Einheiten und zu Elementen der Welt, die sie beschreiben.

- Kombinierbarkeit von Äußerungselementen zu ‚sinnvollen‘ Äußerungen
- Beschreibung konkreter Objekte, Handlungen
(Referenzsemantik)
- • •

** Drei eckige Bälle sangen geräuschlos ein Stück Käse.*

PRAGMATIK definiert den Verwendungszusammenhang sprachlicher Äußerungen

- Einfluss der Sprechsituation (Zeit, Ort, Sprecher, Hörer, ...)
- Intention von Sprachäußerungen
- • •

** Mach bitte die Tür zu! (Tür ist geschlossen)*

? Weißt Du, wie viel Uhr es ist? --- Ja.

Betreten der Baustelle verboten.

Eltern haften für ihre Kinder.

“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 (lexikalische, syntaktische und anaphorische Ambiguitäten) besitzt dieser Satz?

Tatsächlich sind es: 258.048!

Ohne Anaphernambiguität: 64.512!

Sprachtechnologische Anwendungen sind insbesondere für das sog. „Semantic Web“ interessant.

Tim Berners-Lee, James Hendler und Ora Lassila: /The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, May 17, 2001

http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

Der Zugriff auf Internet-Inhalte soll dadurch verbessert werden, dass man mit automatischen Verfahren die Dokumente um Annotationen anreichert, die ihre Semantik beschreiben.

Mit Hilfe solcher Annotationen können z.B. Suchmaschinen bessere Ergebnisse erzielen.