



Parsing Übersicht und Kurskonzept

Kursfolien

Karin Haenelt

01.04.2003

Karin Haenelt, Parsing – Stand 2003

1

Sprachtechnologie

Software

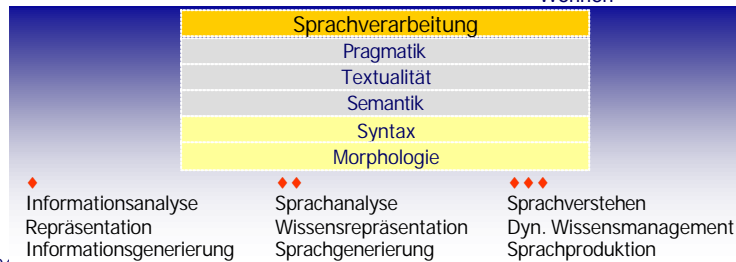
- Benutzerschnittstelle
- Strukturierung von semantischer Information

Kommunikation

- Diktiersysteme
- Dokumentenmanagement
- Rechtschreibung
- Suchmaschinen
- Übersetzung

Alltagstechnologie

- Consumer Electronic
- E-Commerce
- E-Learning
- Telekommunikation
- Telematik
- Verkehr, Auto
- Wohnen



01.04.2003

Karin Haenelt, Parsing – Stand 2003

Parsing - Stand

- Sprachverarbeitende Automaten sind defizitär im Vergleich zu sprachkompetenten Menschen
- Eine vollständige Lösung der Parsing-Aufgabe ist heute noch nicht bekannt
- Partielle Aspekte des Parsing sind heute gut bekannt
- Nicht für alle praktischen Aufgaben ist vollständiges Parsing erforderlich
- Für Massendaten wird robustes Parsing benötigt

01.04.2003

Karin Haenelt, Parsing – Stand 2003

3

Komplexe Phänomene

- Komplexe grammatikalische Phänomene
 - Morpho-syntaktische Merkmale
 - Kongruenz
 - Long distance dependencies
 - Freie Wortstellung
- Ambiguität sprachlicher Äußerungen
 - Morpho-syntaktische Ambiguität
 - Semantische Ambiguität
 - Pragmatische Ambiguität
- Kreativität des Sprachgebrauchs

01.04.2003

Karin Haenelt, Parsing – Stand 2003

4

Komplexe Aufgaben

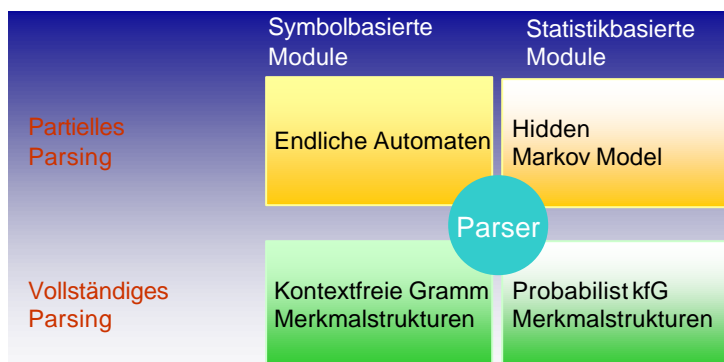
- Sprachtheoretische Beschreibung der Phänomene
- Definition der Repräsentation
 - Syntaktische Struktur
 - Semantische Struktur
 - Thematische Struktur
 - ...
- Reduktion der Komplexität und Modularisierung

01.04.2003

Karin Haenelt, Parsing – Stand 2003

5

Parsingmethoden



01.04.2003

Karin Haenelt, Parsing – Stand 2003

6

Parsingmethoden: vollständig / partiell

- Vollständiges Parsing
 - Ziel: vollständige und korrekte Analyse
 - Closed-World-Assumption
 - Grammatik ist vollständig
 - Gesucht wird die beste Lösung des gesamten Suchraumes
 - Probleme:
 - Nicht robust
 - keine Unterscheidung identifizierbarer und nicht identifizierbarer Phänomene
- Partielles Parsing
 - Ziel: effiziente, robuste und zuverlässige Erkennung identifizierbarer Information aus großen Textmengen
 - Open-World Assumption
 - Grammatik ist unvollständig
 - Gesucht wird nur nach bekannten Informationsbausteinen
 - Verzicht auf Vollständigkeit und Tiefe der Analyse

01.04.2003

Abney, 1996 7

Karin Haenelt, Parsing – Stand 2003

Probleme des vollständigen Parsing

- „All Grammars Leak!“ (Sapir)
- Keine klare Trennung zwischen syntaktischer und semantischer Information und keine adäquate Verarbeitung

01.04.2003

Karin Haenelt, Parsing – Stand 2003

8

„All Grammars Leak“*

*Edward Sapir, 1921

- Not possible to provide an exact and complete characterization
 - of all well-formed utterances
 - that cleanly divides them from all other sequences of words which are regarded as ill-formed utterances
- Rules are not completely ill-founded
- Somehow we need to make things looser, in accounting for the creativity of language use

01.04.2003

Karin Haenelt, Parsing – Stand 2003

Manning/Schütze, 1999, S. 3 9

Vermischung syntaktischer und semantischer Information

- Verarbeitungseinheit: Satz
 - Aufbau einer „syntaktischen Struktur“
 - Aufbau einer Konzeptstruktur
- Aber:
 - der Aufbau einer vollständigen Satzstruktur kann eine semantische Interpretation voraussetzen!
 - Vermischung von syntaktischer und semantischer Information in Syntax-Regeln führt zu kombinatorischer Explosion von Lesarten
- Lösungsansätze:
 - probabilistisches Parsing (Lesarten-Entscheidung)
 - Flaches Parsing (Unterspezifikation)

01.04.2003

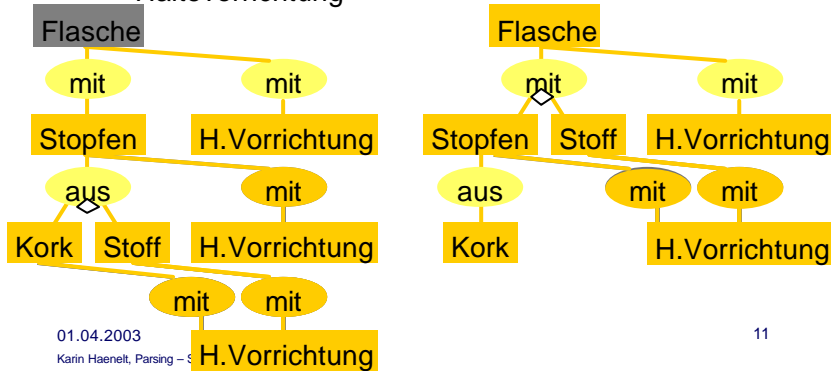
Karin Haenelt, Parsing – Stand 2003

10

Satzstruktur: syntaktische und semantische Information

Präpositionalphrase

Glasflaschen ... verschlossen mit einem pilzförmigen Stopfen aus Kork oder einem anderen Stoff mit Haltevorrichtung



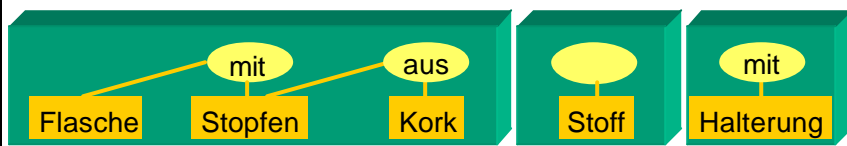
Semantisch offene Konstruktionen: satzintern

Flaschen mit Stopfen und besonderer Haltevorrichtung

Gesetz zur Besteuerung
Von Schaumwein und
Zwischenerzeugnissen



Flaschen mit Stopfen, der durch eine Haltevorrichtung befestigt ist



Vollständiges Parsing: Lesartenexplosion

<i>List the sales of products in 1973</i>	3 Lesarten
<i>List the sales of products produced in 1973</i>	10 Lesarten
<i>List the sales of products in 1973 with the products in 1972</i>	28 Lesarten
<i>List the sales of products produced in 1973 with the products produced in 1972</i>	455 Lesarten

Bod, 1998: 2

01.04.2003

13

Karin Haenelt, Parsing – Stand 2003

Lösung: flacheres syntaktisches Parsing

- Beschränkung des syntaktischen Parsing auf syntaktisch erkennbare Phänomene
- Flacheres Parsing
- Verbannung von Regeln der Art

$$VP \rightarrow VP PP$$
$$NP \rightarrow NP PP$$

01.04.2003

14

Karin Haenelt, Parsing – Stand 2003

Partielles Parsing: Operationen

Part-of-Speech-Tagging	Zuordnung einer Wortart
Chunking	<ul style="list-style-type: none"> - Chunks: nicht-rekursive Kerne der Hauptphrasen vom Anfang der Konstituente bis zum Kopf, ohne dem Kopf nachgestellte Attribute - Chunk-Kategorien: NX, VX, INF, VGX, VNX, AX, RX. - (Hauptphrasen: NP, VP, PP, AP, AdvP)
Clause Bracketing	Sätze und Teilsätze mit Zuordnung einer Kategorie
Chunk Attachment	<ul style="list-style-type: none"> - Identifikation weiterer Zuordnungsinformation (z.B. Köpfe von Präpositionalphrasen, postnominale Nominalattribute, Relativsätze), - Verbindung der Chunks mit weiteren Knoten und Kanten; - weitere Annäherung an Standardgraphen
Chunk Linking	<ul style="list-style-type: none"> - Identifikation von Satzfunktionen (z.B. Subjekt - Prädikat) und - Herstellung der Verbindung zwischen den Einheiten

01.04.2003

15

Karin Haenelt, Parsing – Stand 2003

Partielles Parsing: Beispiel

	Tagging	Chunking	Chunk Attachment	Chunk Linking	Clause Bracketing
Rechnungen	nomn	[NX]	[NP	[subje	
und	coor		"	"	
Messungen	nomn	[NX]	"]	"]	
haben	auxv	[VX	[VP	[praed	
ergeben	verb	"]	"]	"]	
,	,				
dass	hypo				
Handys	nomn	[NX]	[NP]	[subje]	
im	prpo		[PP	[pp	
Kopf	nomn	[NX]	"]	"]	
nur	advb				
eine	dete	[NX	[NP	[trans	
schwache	adje	"	"	"	
lokale	adje	"	"	"	
Erwärmung	nomn	"]	"]	"]	
von	prpo		[PP	"	
maximal	advb	[NX	"	"	
ca.	advb	"	"	"	
0.1	card	"	"	"	
°	masz	"]	"]	"]	
erzeugen	verb	[VX]	[NVP]	[praed]	
.	.				

01.04.2003

16

Karin Haenelt, Parsing – Stand 2003

Parsingmethoden: symbolisch / statistisch

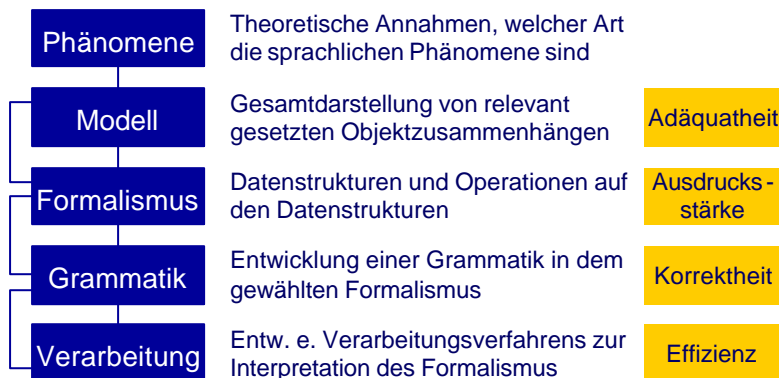
	symbolbasierte Systeme	statistikbasierte Systeme
mangelhafte Abdeckung	keine 100% Lösungen	
aufwändige Erstellung	Erstellung von Regeln	Annotation eines Referenzcorpus
	Entwicklung eines kategorialen Inventars	
	Kategorisierungsproblem ungelöst	
	Regeln extern und explizit kodiert	Regeln intern und implizit kodiert
einzel Sprachspezifisch	Regeln einzel Sprachspezifisch	Reduzierung der Kategorienmenge durch Bedarf an Menge statist. signifikanter Daten

01.04.2003

Karin Haenelt, Parsing – Stand 2003

Kiss, 2003 17

Grammatik-Entwicklung phänomen-orientiert



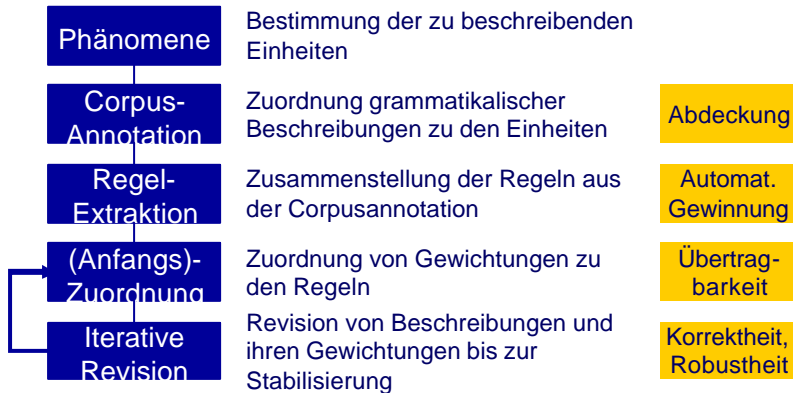
Aufgabe: Entwicklung einer geeigneten Kombination von Grammatik / Formalismus / Verarbeitungsverfahren¹⁸

01.04.2003

Karin Haenelt, Parsing – Stand 2003

(Hellwig, 1989, S. 349/350)

Grammatik-Entwicklung text-orientiert



Aufgabe: Entwicklung einer robusten Grammatik zur Verarbeitung großer Corpora

01.04.2003

Karin Haenelt, Parsing – Stand 2003

19

Projektmodelle

Phasenmodell

- Sukzessive Entwicklung
- Produkt entsteht in der letzten Projektphase

Zyklenmodell

- Inkrementelle Verfeinerung
- Erste Produktversion entsteht in der ersten Projektphase

01.04.2003

Karin Haenelt, Parsing – Stand 2003

20

Implementierungsmodelle

Zeichenkette : Ja/Nein	Zeichenkette : Interpretation	Zeichenkette : Interpret., Wahrscheinl.
Endliche Automaten - DEA - NEA - ϵ -NEA	Transducer - sequentiell - bidirektional	Hidden Markov Model
Erkennung	Parser - Earley - CKY - ...	Probabilist. Parser

01.04.2003

Karin Haenelt, Parsing – Stand 2003

21

Implementierungstechniken

Für alle Implementierungsmodelle

Agenda
zur Verwaltung
nicht-deterministischer
Information

Suchstrategien

Regeln

01.04.2003

Karin Haenelt, Parsing – Stand 2003

22

Organisation des Seminars

- Vorlesung
- Seminarprojekte

Seminar: Einführung

Einführung

(01) Parsing: Stand der Entwicklung

Seminar: Endliche Automaten

Endliche Automaten

- (02) Flaches Parsing mit endlichen Automaten
(Flex und JLex)
- (03) Endliche Automaten
deterministisch, nicht-deterministisch, mit
Epsilon-Transformationen
- (04) Transducer
sequentiell, bidirektional
- (05) Datenstrukturen und Algorithmen
- (06) Implementierung endlicher Automaten und
Transducer in Java

01.04.2003

Karin Haenelt, Parsing – Stand 2003

25

Seminar: Probabilistische Endliche Automaten

Wahrscheinlichkeitstheorie

- (07) Grundlagen der Wahrscheinlichkeitstheorie

Probabilistische Endliche Automaten

- (08) Hidden Markov Models
- (09) Forward-Algorithmus
- (10) Viterbi-Algorithmus

01.04.2003

Karin Haenelt, Parsing – Stand 2003

26

Seminar: Komplexität natürlicher Sprache

Komplexität

- (11) Komplexität natürlicher Sprache

01.04.2003

Karin Haenelt, Parsing – Stand 2003

27

Seminar: Partielles Parsing

Tagging

- (12) Ansätze des Tagging (Part-of-Speech und Konstituenten)
- (13) Grammatikstrukturen und Tag-Sets

Chunking, Chunk Attachment, Chunk Linking und Clause Bracketing

- (14) Chunking
- (15) Chunk Attachment und Chunk Linking
- (16) Clause Bracketing

01.04.2003

Karin Haenelt, Parsing – Stand 2003

28

Seminar: Klassisches Parsing

Standardalgorithmen

- (17) Der Earley-Algorithmus
- (18) Der Cocke-Kasami-Younger-Algorithmus
- (19) Robustes Parsing mit Standardalgorithmen

Probabilistische Standardalgorithmen

- (20) Probabilistische kontextfreie Grammatiken
- (21) Probabilistisches Parsing

Komplexität

- (22) Die O-Notation
- (23) Komplexität des Earley-Algorithmus

01.04.2003

Karin Haenelt, Parsing – Stand 2003

29

Seminarprojekte

- Teams übernehmen Projektaufgaben
- Stellen die Ergebnisse am Ende des Seminars vor
 - Elektronische Präsentation
 - Ggf. Systemvorführung
- Optional: Publikation der Ergebnisse
 - Kursseite im Internet
 - Kurs-CD
 - Publikation

01.04.2003

Karin Haenelt, Parsing – Stand 2003

30

Projekte: Endliche Automaten

Endliche Automaten

- (P01) Morphologie mit endlichen Automaten
- (P02) Flaches Parsing mit endlichen Automaten
(Flex und JLex)
- (P03) (Java)-Werkzeuge zur Entwicklung endlicher Automaten
- (P04) Implementierung und experimentelle Anwendung endlicher Automaten

Projekte: Probabilistische Endliche Automaten

Probabilistische Endliche Automaten

- (P05) Aufbau und Training von Hidden Markov Models
- (P06) Implementierung und Test des Viterbi-Algorithmus

Projekte: Partielles Parsing

Tagging

(P07) Ansätze des Tagging (Part-of-Speech und Konstituenten)

(P08) Fallstudie: Grammatikstrukturen und Tag-Sets. Anwendung auf ein Beispielcorpus.

Chunking, Chunk Attachment, Chunk Linking und Clause Bracketing

(P09) Chunking

(P10) Chunk Attachment und Chunk Linking

(P11) Clause Bracketing

01.04.2003

Karin Haenelt, Parsing – Stand 2003

33

Projekte: Klassisches Parsing

Standardalgorithmen

(P12) Robustes Parsing mit Standardalgorithmen

Probabilistische Standardalgorithmen

(P13) Probabilistische kontextfreie Grammatiken

(P14) Probabilistisches Parsing

01.04.2003

Karin Haenelt, Parsing – Stand 2003

34

Projekte

- Auch eigene Projektdefinitionen sind möglich

01.04.2003

Karin Haenelt, Parsing – Stand 2003

35

Literatur

- **Abney, Steven (1996)**: Part-of-Speech-Tagging and Partial Parsing. In: Ken Church, Steve Young und Gerrit Bloothoof (eds.): *Corpus-Based Methods in Language and Speech*. Dordrecht: Kluwer Academic Publishers.
- **Bod, Rens (1998)**: *Beyond Grammar. An Experience-Based Theory of Language*. CSLI Lecture Notes, 88, Stanford California: Center for the Study of Information and Language
- **Hellwig, Peter (1989)**: Parsing Natürlicher Sprachen. In: Bátor, István S.; Lenders, Winfried; Putschke, Wolfgang (Hrsg.): *Computational Linguistics = Computerlinguistik: An International Handbook on Computer Oriented Language Research and Applications = Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen*. Berlin: de Gruyter, 1989. pp. 348-432
- **Kiss, Tibor (2003)**: Anmerkungen zur scheinbaren Konkurrenz von numerischen und symbolischen Verfahren in der Computerlinguistik. In: Gerd Willée, Bernhard Schröder, Hans-Christian Schmitz (Eds.) *Computerlinguistik: Was geht, was kommt? Computational Linguistics: Achievements and Perspectives*. Festschrift für Winfried Lenders. Sankt Augustin: gardez! Verlag.
- **Manning, Christopher D.; Schütze, Hinrich (1999)**: *Foundations of Statistical Natural Language Processing*. Cambridge, Mass., London: The MIT Press. (vgl.: <http://www.sultry.arts.usyd.edu.au/fsnl/>)

01.04.2003

Karin Haenelt, Parsing – Stand 2003

36