

Einführung in die Computerlinguistik

Syntax II

WS 2010/2011

Manfred Pinkal

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Beispiel 1

CFG für einfache arithmetische Gleichungen:

$S \rightarrow \text{Term} = \text{Term}$ $\text{Term} \rightarrow x \mid y \mid z$
 $\text{Term} \rightarrow (\text{Term Op Term})$ $\text{Op} \rightarrow + \mid - \mid * \mid :$
 $\text{Term} \rightarrow - \text{Term}$

Konstituenten der Kategorie „Term“ sind zum Beispiel
 $x, y, -z, -(x*(y+z))$

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

CFG: Konstituentenstruktur

- Anders als endliche Automaten beschreibt eine CFG nicht nur die zulässigen Ausdrücke einer Sprache, sondern implizit auch deren Struktur.
- Sie ordnet den Sätzen der Sprache Ableitungsbäume zu (auch „Parse-Bäume“ genannt, Parsing = automatische syntaktische Analyse).
- Durch den Ableitungsbaum/ Parse-Baum werden Teilausdrücke (Teilketten) u des analysierten Wortes einer „Kategorie“ zugeordnet: dem nicht-terminalen Symbol A, aus dem u abgeleitet wurde. Wir nennen u eine „Konstituente“ von der Kategorie A, und sagen, dass A die Elemente von u „dominiert“.

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Beispiel 2

Die obige CFG für ein Fragment des Deutschen.

- *er* ist eine Konstituente der Kategorie Pro
- *er*, *der Student*, *der Student*, *der Informatik studiert* sind Konstituenten der Kategorie NP
- *der Informatik studiert* - *der arbeitet* sind Konstituenten der Kategorie SRel

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

CFG: Konstituentenstruktur

- CFGs drücken in eleganter Weise strukturelle Regularitäten aus:
 - „Eine Gleichung besteht aus zwei Termen, die durch ein Gleichheitszeichen verbunden sind“.
- Umgekehrt ausgedrückt: Ersetzungsregeln und Kategorien sollten so gewählt werden, dass die Grammatik in möglichst eleganter Weise strukturelle Regularitäten ausdrückt.
- Trivial für formale Sprachen: Die werden ja explizit mithilfe von CFGs definiert.
- Wie geht man aber bei natürlichen Sprachen vor?
- Unterschiedliche Tests („Heuristiken“), um zu bestimmen, welche Wortfolgen in Sätzen Konstituenten bilden.

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Tests für Konstituentenstruktur (1)

- *Peter hat der Dozentin [NP das neue Übungsblatt] heute ins Büro gebracht.*
- *[NP Das neue Übungsblatt] hat Peter der Dozentin heute ins Büro gebracht.*
- *Der Dozentin hat Peter heute [NP das neue Übungsblatt] ins Büro gebracht.*
- *Ins Büro hat heute Peter der Dozentin [NP das neue Übungsblatt] gebracht.*
- *Heute hat Peter [NP das neue Übungsblatt] der Dozentin ins Büro gebracht.*

Verschiebetest: Teilketten, die Konstituenten bilden, lassen sich (nur) gemeinsam im Satz verschieben.

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Tests für Konstituentenstruktur (2)

- *[NP Er] hat die Übungen gemacht.*
- *[NP Der Student] hat die Übungen gemacht.*
- *[NP Der interessierte Student] hat die Übungen gemacht.*
- *[NP Der an computerlinguistischen Fragestellungen interessierte Student] hat die Übungen gemacht.*
- *[NP Der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester] hat die Übungen gemacht.*
- *[NP Der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, der im Hauptfach Informatik studiert], hat die Übungen gemacht.*

Substitutionstest: Lassen sich Wortfolgen in einem gegebenen Kontext füreinander ersetzen, handelt es sich (vermutlich) um Konstituenten, und zwar Konstituenten der gleichen Kategorie.

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Tests für Konstituentenstruktur (3)

- *[NP Peter] hat der Dozentin das Übungsblatt heute ins Büro gebracht.*
- *[NP Das Übungsblatt] hat Peter der Dozentin heute ins Büro gebracht.*
- *[NP Der Dozentin] hat Peter heute das Übungsblatt ins Büro gebracht.*
- *[PP Ins Büro] hat heute Peter der Dozentin das Übungsblatt gebracht.*
- *[AdvP Heute] hat Peter das Übungsblatt der Dozentin ins Büro gebracht.*
- *[PP Ins Büro] hat das Übungsblatt der Dozentin Peter heute gebracht.*

"Verb-Zweit" bietet einen verlässlicher Konstituententest fürs Deutsche: Vor dem finiten Verb im Hauptsatz steht (im allgemeinen) genau eine Konstituente.

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Kriterien für Konstituenten

- Distributionelle Eigenschaften:
 - Verschiebbarkeit, Substituierbarkeit
 - Nominalausdrücke kommen in vielen Positionen im Satz vor (Subjekt, Objekt, Genitivattribut) und sind (modulo Flexionsform) füreinander einsetzbar. Dies ist eine Motivation für die Kategorie „Nominalphrase“.
 - Im deutschen Hauptsatz steht das Verb an zweiter Stelle: allerdings nicht in Wörtern gezählt, sondern in „Phrasen“. Diese und ähnliche Beobachtungen bilden ein Argument für die allgemeine Annahme von „**phrasalen Kategorien**“ als einer Grundebene der syntaktischen Beschreibung.
- Kriterien aus der internen Struktur:
 - Nominalausdrücke besitzen einheitlich als „Kopf“ ein Substantiv oder ein Pronomen.

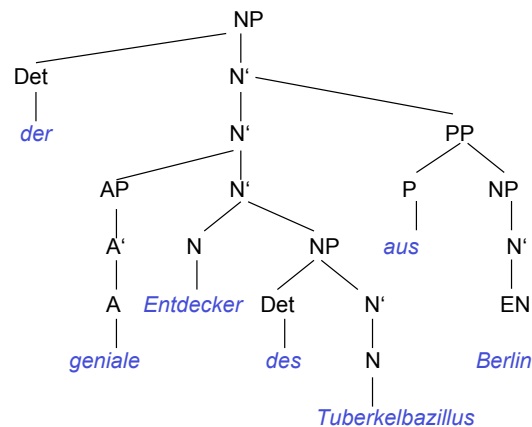
Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Phrasale Kategorien

- Für die drei „großen“ oder „offenen“ Wortarten Substantiv, Verb und Adjektiv und die Präpositionen werden üblicherweise vier **lexikalische Hauptkategorien** (S, V, A und P) angenommen.
- Entsprechend nimmt man vier **phrasale Hauptkategorien** (NP, VP, AP, PP) an, die Ausdrücke der jeweiligen lexikalischen Kategorie als Kopf besitzen:
 - **Nominalphrasen**: *der interessierte Student – die Übungen – computerlinguistische Fragestellungen*
 - **Präpositionalphrasen**: *an computerlinguistischen Fragestellungen – im ersten Semester – nach langer Überlegung*
 - **Adjektivphrasen**: *an computerlinguistischen Fragestellungen interessiert(e), sehr schön, viel größer als Peter*
 - **Verbphrasen**: *studiert Informatik – entscheidet sich für das Fach*

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

NP-Struktur: Ein Beispiel



Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Kategoriale Ebenen

- **Lexikalische Kategorien** („Präterminale Symbole“): Sie bilden die linke Seite von Regeln aufzutreten, deren rechte Seite aus einem Terminalsymbol (lexikalischen Ausdruck) besteht, z.B. N, A, V, Det, Pro, ...
- **Phrasale Kategorien** wie NP und PP, die „maximale Konstituenten“ bezeichnen, die im Satz eine relative Unabhängigkeit besitzen: kommen als „Satzteile“ innerhalb von anderen Phrasen vor, lassen sich verschieben, können nicht durch anderes Material unterbrochen werden.
- **Zwischenkategorien**: Hier nimmt man meist genau eine weitere Ebene an, die zwischen der phrasalen und der lexikalischen Ebene vermittelt. Sie werden üblicherweise als N', A', V' etc. notiert, alternativ mit einem Überstrich, daher als „N-Bar“, „V-Bar“ etc. ausgesprochen.

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Regelschemata

- NP-Struktur im Deutschen (vereinfacht)
 - $NP \rightarrow EN \mid Pro \mid Det N'$
 - $N' \rightarrow (AP) N'$
 - $N' \rightarrow N' (PP)$
 - $N' \rightarrow N (NP)$
- Allgemeines XP (X-Bar-)Schema:
 - $XP \rightarrow SpecXP X'$ "Specifier" + Kopf
 - $X' \rightarrow (YP) X'$ beliebig viele Prämodifikatoren + Kopf
 - $X' \rightarrow X' (YP)$ Kopf + beliebig viele Postmodifikatoren
 - $X' \rightarrow X YP_1 \dots YP_n$ lexikalischer Kopf +
n "subkategorisierte" Argumente

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Grammatiktheorie

- Die CFG als solche ist ein **Formalismus** zur syntaktischen Beschreibung. Die Frage, welche Ausdrücke als Konstituenten betrachtet werden sollen und welche Kategorien die Grammatik benutzen soll, ist eine Angelegenheit der **Grammatiktheorie**.
- Die Frage hat keine einfache Antwort. Unterschiedliche Auffassungen haben zu unterschiedlichen Grammatiktheorien geführt.
- Einvernehmen besteht darüber, dass es eine begrenzte Zahl von Ebenen für grammatische Kategorien und eine begrenzte Zahl von Hauptkategorien gibt, die sich an den Hauptwortarten ausrichten („X-Bar-Theorie“).

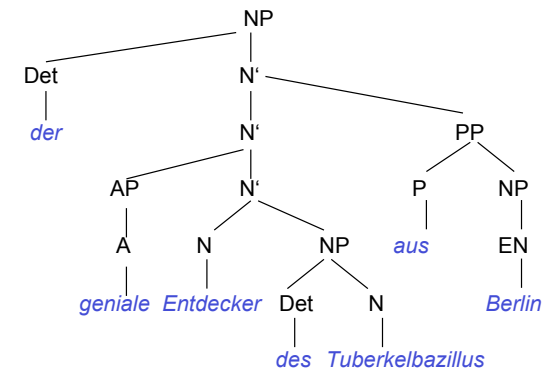
Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Kategorie und Funktion

- Syntaktische Kategorien** bezeichnen Klassen von Ausdrücken mit ähnlicher innerer Struktur und ähnlichem distributionellem Verhalten.
- Grammatische Funktionen** dagegen bezeichnen die Rolle, die eine Konstituente im größeren Ausdruck spielt. Eine NP kann, je nach Stellung im Satz unter anderem die Funktion von **Subjekt** oder (direktem oder indirektem) **Objekt** eines Satzes, (Genitiv-) **Attribut** einer anderen NP oder **Argument** einer Präpositionalphrase bilden. - Grammatische Funktionen sind relationale Konzepte!
- Unterschiedliche Kategorien können die gleiche Funktion ausüben: Subjekte können zum Beispiel Nominalphrasen oder Sätze sein:
 - *Dass es regnet, ist lästig*
 - *Der Regen ist lästig.*

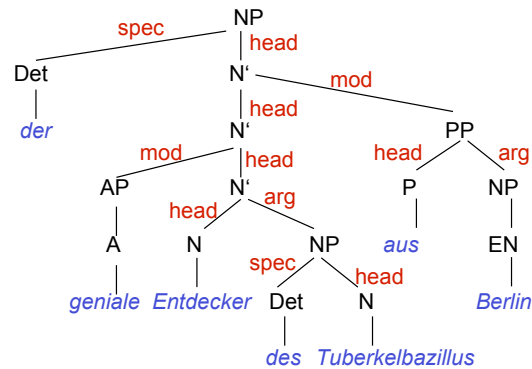
Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Ein Beispiel



Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Grammatische Funktionen



Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Grammatische Funktionen

- **Köpfe** sind die Kernbestandteile einer Konstituente, die für den syntaktischen „Charakter“ der Phrase verantwortlich sind. Die Merkmale des „lexikalischen Kopfes“ vererben sich über die „Kopflinie“ nach oben zur Phrase.
- **Argumente** werden durch lexikalische Köpfe „subkategorisiert“ oder „regiert“: Ein lexikalischer Ausdruck (V, N, A, P) kann ein oder mehrere Argumente mit bestimmten grammatischen Eigenschaften verlangen. Verbargumente sind Subjekt, direktes Objekt, Präpositionales Objekt etc.
- **Modifikatoren** sind freie Ergänzungen, die einen Ausdruck erweitern, ohne seine Kategorie zu verändern. Nominale Modifikatoren heißen **Attribute**, Satzmodifikatoren **Adjunkte** (auch „adverbiale Bestimmungen“).

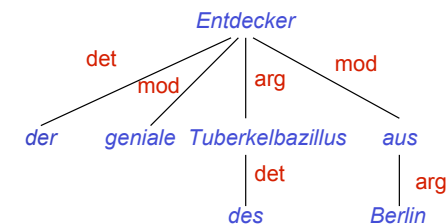
Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Ausblick: Dependenzgrammatik

- Kontextfreie Grammatiken beschreiben die Konstituentenstruktur und bestimmen in einem zweiten Schritt zusätzlich die grammatischen Funktionen.
- Dependenzgrammatik beschreibt die syntaktische Struktur primär durch die funktionalen Abhängigkeiten (Dependenzrelationen).
- Dependenzrelationen sind Relationen zwischen Wörtern, einem (lexikalischen) Kopf und einem abhängigen Wort (dem Dependenten).

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Dependenzgrammatik: Beispielstruktur



Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Kontextfreie Grammatik und Dependenzgrammatik

- Dependenzgrammatiken unterscheiden zwischen funktionaler Struktur (die im Dependenzbaum dargestellt wird) und Wortstellung, die durch separate Bedingungen geregelt wird: „Lineare Präzedenzregeln“.
- Dependenzgrammatiken eignen sich deshalb zur Modellierung von Sprachen mit variabler oder freier Wortstellung.
- Dependenzstrukturen sind „semantiknäher“ als Konstituentenstrukturen. Sie eignen sich deshalb besser als Grundlage für die semantische Interpretation.
- CFG-basierte Grammatiktheorien haben in der Linguistik und Computerlinguistik jahrzehntelang das Feld beherrscht. Seit einigen Jahren haben Dependenzgrammatiken stark an Bedeutung gewonnen.

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Syntaktische Struktur und semantische Interpretation

- Die syntaktische Struktur ist auch Grundlage für die semantische Interpretation.
- Beispiel Arithmetik:
 - Terme bezeichnen („bedeuten“) Zahlen
 - Operatoren bilden (Paare von) Termbedeutungen/Zahlen auf Termbedeutungen/Zahlen ab.
 - Bedeutung einer Gleichung ist ein Wahrheitswert.
- Die Bedeutung des Gesamtausdrucks wird „kompositionell“, entlang der syntaktischen Struktur berechnet.
- Strukturelle Mehrdeutigkeit, wenn Klammern und Klammerkonventionen fehlen. Beispiel: „3+4*5“: =23 oder =60?

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Syntaktische Struktur und semantische Interpretation

93.000€ in Gütersloh gefunden in der Handtasche einer Rentnerin, die auf einem Friedhof am Lenker eines Fahrrads baumelte.

Aus dem Spiegel, Rubrik „Hohlspiegel“

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Grammatische Mehrdeutigkeit, Beispiel

- Grammatiken für formale Sprachen werden so definiert, dass strukturelle Mehrdeutigkeit in jedem Fall vermieden wird.
- Natürliche Sprachen **sind** strukturell mehrdeutig.

Peter sah den Mann mit dem Teleskop

- Eine Grammatik, die die Mehrdeutigkeit modelliert:

$S \rightarrow NP VP$ $VP \rightarrow VT NP$ $VP \rightarrow VI$
 $NP \rightarrow ART N'$ $NP \rightarrow EN$ $NP \rightarrow NP PP$
 $PP \rightarrow P NP$ $VP \rightarrow VP PP$ $N' \rightarrow N (PP)$

- Die zwei Analysevarianten:

– $[S \textit{ Peter} [VP \textit{ sah} [NP \textit{ den} [N' \textit{ Mann} [PP \textit{ mit dem Teleskop}]]]]]]$
– $[S \textit{ Peter} [VP [VP \textit{ sah} [NP \textit{ den Mann}]]] [PP \textit{ mit dem Teleskop}]]]]$

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik