Einführung in die Computerlinguistik

Syntax II

WS 2009/2010

Manfred Pinkal

Kontextfreie Grammatik: Zur Notation

- Notationskonventionen:
  - Alternative Elemente werden durch „|"zusammengefasst (manchmal auch durch geschweifte Klammern)
  - Optionale Elemente werden durch runde Klammern notiert.

S → NP VI	S → NP VT NP
SRel → RPro VI	SRel → RPro NP VT
NP → Det N (SRel)	NP → Pro
VI → *schläft* | *arbeitet*	VT → *wählte* | *studiert*
N → *Student* | *Fach*	RPro → *der* | *das*
Det → *der* | *das*	Pro → *er* | *sie*

- Eine CFG wird meist nur durch die Angabe der Ersetzungsregeln spezifiziert: Das Startsymbol heißt per Konvention immer S; V und T sind aus den Regeln ablesbar.

Kontextfreie Sprachen und reguläre Sprachen

- Kontextfreie Sprachen sind eine echte Obermenge der regulären Sprachen: Jede reguläre Sprache kann von einer CFG erzeugt werden, und es gibt kontextfreie Sprachen, die nicht regulär sind.
- Endliche Automaten verwenden Iteration: Der Automat läuft beliebig oft durch Schleifen und arbeitet dabei Wiederholungen gleicher Symbolfolgen ab.
- Kontextfreie Grammatiken verwenden Rekursion. Produktionsregeln verwenden in der Definition eines Ausdruckstyps den Ausdruckstyp selbst: Nicht-terminale Symbole tauchen auf der linken und der rechten Seite von Regeln auf, z.B. in S → aSb.
- Rekursive Regeln erlauben die tiefe Schachtelung von Strukturen, und sie ermöglichen, dass eine Regel Elemente in Beziehung setzt, die in der Kette weit voneinander entfernt sind.

Kontextfreie Sprachen und natürliche Sprachen

- Kontextfreie Grammatiken bilden den Standard-Formalismus zur syntaktischen Beschreibung von formalen Sprachen (Logik, Arithmetik, Programmiersprachen).
- Ein alternatives, der CGF ähnliches Format zur Beschreibung kontextfreier Sprachen ist BNF (die „Backus-Naur-Form").
- Kontextfreie Grammatiken sind ein Standardformalismus zur Beschreibung der Grammatik natürlicher Sprachen.

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“.

Beispiel 1

CFG für einfache arithmetische Gleichungen:

S → Term = Term
Term → (Term Op Term)
Term → – Term

Term → x | y | z
Op → + | – | * | :

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

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

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?

Konstituentenstruktur: Verschiebbarkeit

- *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.*

Konstituentenstruktur: Ersetzbarkeit

- *[$_{NP}$ 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.*
- *[$_{NP}$ Der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, der im Hauptfach, für das er sich nach langer Überlegung entschieden hat, Informatik studiert], hat die Übungen gemacht.*

Konstituentenstruktur: „Verb-Zweit" im Deutschen

- *[$_{NP}$ 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.*

Kriterien für Konstituenten

- Distributionelle Eigenschaften:
  - Verschiebbarkeit, Ersetzbarkeit
  - 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:
  - Nominalaudrücke besitzen einheitlich als „Kopf" ein Substantiv oder ein Pronomen.

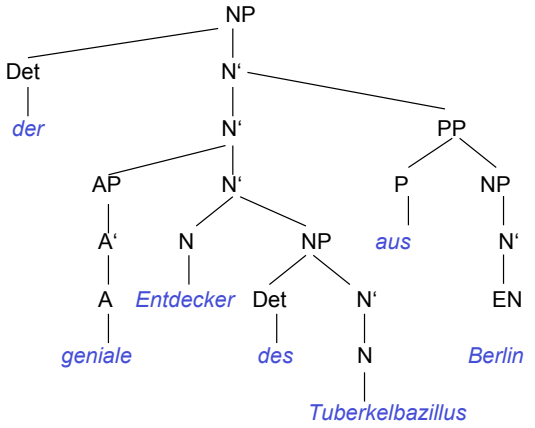
Phrasale Kategorien

- Entsprechend den lexikalischen Hauptkategorien Substantiv, Andjktiv, Verb und Präposition geht man üblicherweise von vier phrasalen Hauptkategorien aus:
  - Nominalphrasen: *er – der Student – der interessierte Student – die Übungen – computerlinguistischen Fragestellungen*
  - Präpositionalphrasen: *an computerlinguistischen Fragestellungen – im ersten Semester, – nach langer Überlegung*
  - Adjektivphrasen: *interessierte – an computerlinguistischen Fragestellungen interessierte*
  - Verbphrasen: *arbeitet - studiert Informatik - entscheidet sich für das Fach*

Kategoriale Ebenen

- Lexikalische Kategorien („Präterminale Symbole"): Sie bilden die linke Seite von Regeln auftauchen, 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" vor, lassen sich verschieben, können nicht durch anderes Material unterbrochen werden.
- Zwischenkategorien: Hier nimmt man meist genau eine Kategorie an, die zwischen der phrasalen und der lexikalischen Ebene vermittelt. Sie werden üblicherweise als N', A', V' etc. notiert.

Ein Beispiel



Grammatiktheorie

- Die CFG als solche ist ein Formalismus zur syntaktischen Beschreibung. Die Frage, welche 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.

Konstituentenstruktur und Semantik

- 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.

Grammatische Mehrdeutigkeit, Beipiel

- 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 → NP VP	VP → VT NP	VP → VI
NP → ART N‘	NP → EN	NP → NP PP
PP → P NP	VP → VP PP	N‘→ N (PP)

- Die zwei Analysevarianten:
  - [$_{S}$ *Peter* [$_{VP}$ *sah* [$_{NP}$ *den* [$_{N'}$ *Mann* [$_{PP}$ *mit dem Teleskop*]]]]]
  - [$_{S}$ *Peter* [$_{VP}$ [$_{VP}$ *sah* [$_{NP}$ *den Mann*]] [$_{PP}$ *mit dem Teleskop*]]]

Ein weiteres mehrdeutiges Beispiel

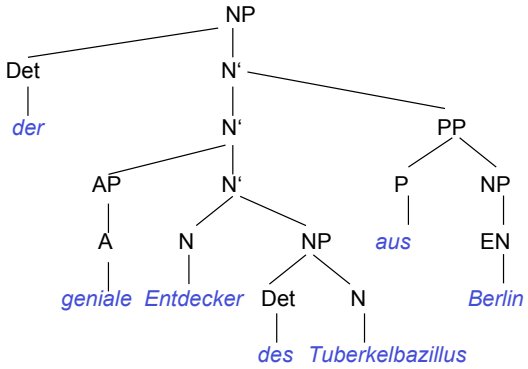
93.000€ in Gütersloh gefunden in der Handtasche einer Rentnerin, die auf einem Friedhof am Lenker eines Fahrrads baumelte.

Gefunden im Spiegel, Rubrik „Hohlspiegel“

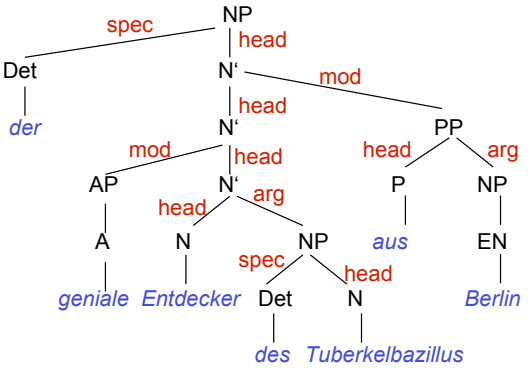
Kategorie und Funktion

- 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.*

Ein Beispiel



Grammatische Funktionen



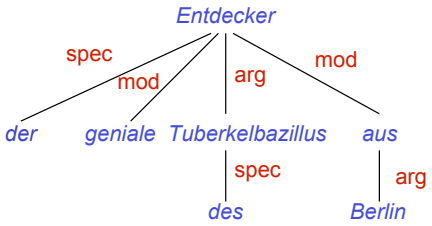
Grammatische Funktionen

- 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“).

Ausblick: Dependenzgrammatik

- CFG 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 (Dependent).

Dependenzgrammatik: Beispielstruktur



Kontextfreie Grammatik und Dependenzgrammatik

- Dependenzgrammatiken unterscheiden zwischen funktionaler Struktur (die im Dependenzbaum dargestellt wird) und Wortstellung (die durch separate Bedingungen geregelt wird).
- CFGs sind für Sprachen mit fester Wortstellung geeignet, Dependenzgrammatiken eignen sich für 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 holen Dependenzgramatiken stark auf.