Einführung in die Computerlinguistik

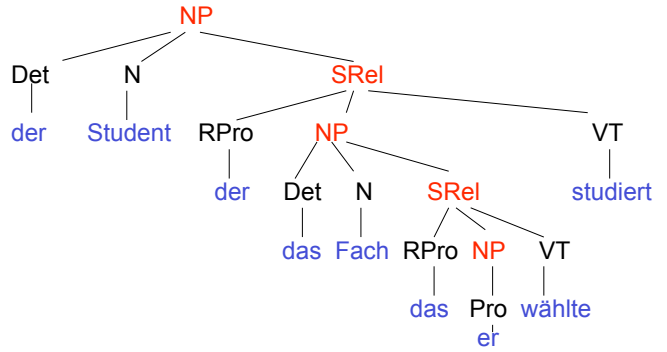
Syntax II

WS 2011/2012

Manfred Pinkal

Geschachtelte Strukturen in natürlicher Sprache

[$_{NP}$ *der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester,* [$_{SRel}$ *der* [$_{NP}$ *das Fach,* [$_{SRel}$ *das* [$_{NP}$ *er*] *nach langer Überlegung gewählt hat*]], *eifrig studiert*]]



Eine Kostprobe aus dem Rechtswesen

- "Der für die Werkstoffabholung auf der Annahme von drei An- und Abfahrten mit LKW, die Wertstoffe umfüllen, und zwei An- und Abfahrten eines LKW,der zuerst die volle Schrottmulde abholt und diese nach Leerung wiederabliefert, errechnete Beurteilungspegel..."

Eine kontextfreie Grammatik für deutsche Sätze

G1 = <V, Σ, P, S> mit

V = {S, SRel, NP, VI, VT, N, Det, RPro} ∪ Σ
Σ = {*schläft, arbeitet, studiert, wählte, Student, Fach, der, das, er*}

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*

CFG: Konstituentenstruktur

- Anders als endliche Automaten beschreibt eine CFG nicht nur die zulässigen Ausdrücke einer Sprache, sondern gibt ihnen implizit auch eine Struktur.
- Sie ordnet den Sätzen der Sprache Ableitungsbäume zu (auch „Parse-Bäume“ genannt, Parsing = automatische syntaktische Analyse).
- Durch den Ableitungsbaum werden Teilausdrücke (Teilketten) u von Wörtern (Terminalsymbolen) 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:

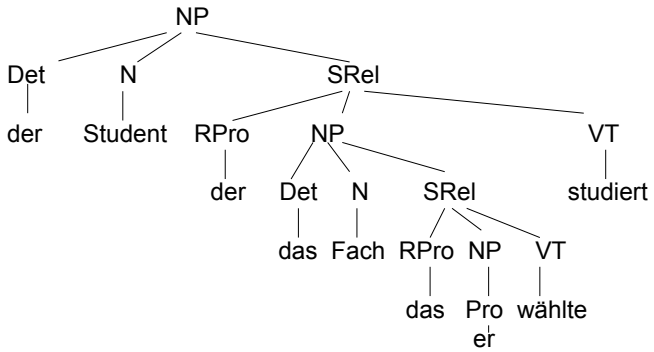
S → Term = Term
Term → (Term Op Term)
Term → – Term

Term → x | y | z
Op → + | – | * | :

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

Beispiel 2

[$_{NP}$ *der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester,* [$_{SRel}$ *der* [$_{NP}$ *das Fach,* [$_{SRel}$ *das* [$_{NP}$ *er*] *nach langer Überlegung gewählt hat*]], *eifrig studiert*]]



Beispiel 2

- *er* ist eine Konstituente der Kategorie Pro
- *er* , *das Fach*, und *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.
- Das ist trivial für formale Sprachen: Die werden ja explizit mithilfe von CFGs definiert.
- Wie geht man aber bei natürlichen Sprachen vor?

Konstituenten-Tests 1: Verschieben

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

Verschiebetest: Teilketten, die sich (nur) gemeinsam im Satz verschieben lassen, sind (tendenziell) Konstituenten

Konstituenten-Tests 2: Substituieren

- *$[_{NP}$ Er] hat die Übungen gemacht.*
- *$[_{NP}$ Peter] hat die Übungen gemacht.*
- *$[_{NP}$ Der Student] hat die Übungen gemacht.*
- *$[_{NP}$ Der an computerlinguistischen Fragestellungen interessierte Student] hat die Übungen gemacht.*

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

Konstituenten-Tests 3: Vorfeldbesetzung

- *$[_{NP}$ 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.*
- *$[_{Adv}$ Heute] hat Peter das Übungsblatt der Dozentin ins Büro gebracht.*

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

Weitere Kriterien für Konstituentenstruktur

- Distributionelle Eigenschaften:
  - Verschiebbarkeit, Substituierbarkeit
- Strukturelle Eigenschaften:
  - Komplexe Ausdrücke besitzen tendenziell einen „Kopf“, der ihren grammatischen Charakter bestimmt
  - Beispiel: Nominalausdrücke besitzen einheitlich als „Kopf“ ein Substantiv oder ein Pronomen
- Semantisches Kriterium:
  - Konstituenten beschreiben sinnvolle Bedeutungseinheiten
  - Beispiel: Nominalausdrücke bezeichnen/ "denotieren" Entitäten (Personen und Objekte)

Phrasale Kategorien

- Für die drei „großen“ oder „offenen“ Wortarten Substantiv, Verb und Adjektiv und die Präpositionen werden üblicherweise vier lexikalische Hauptkategorien (N, 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*

Kategoriale Ebenen

- Lexikalische Kategorien („Präterminale Symbole“): Sie bilden die linke Seite von Regeln, 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.

NP-Struktur

- NP-Struktur im Deutschen (vereinfacht)

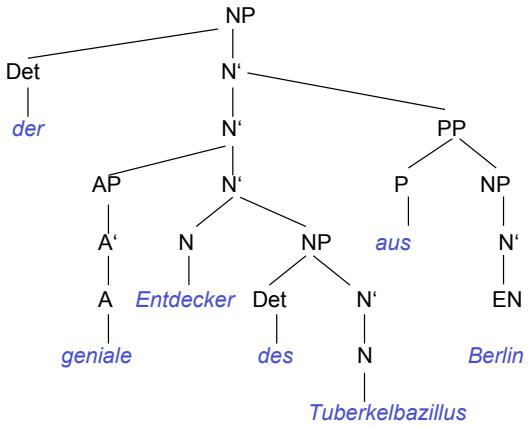
  NP → EN | Pro | Det N’

  N’ → AP N’

  N’ → N’ PP

  N’ → N (NP)

NP-Struktur: Ein Beispiel



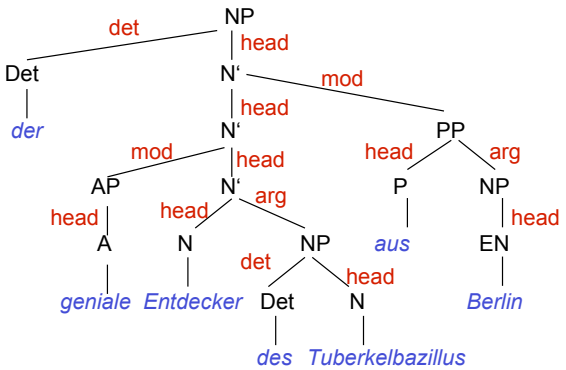
Kategorie und Funktion

- 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.*

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.; Substantive können Argumente als PP oder als Genitivattribut realisieren; die PP nimmt eine NP als Argument.
- Modifikatoren sind freie Ergänzungen, die einen Ausdruck erweitern, ohne seine Kategorie zu verändern. Nominale Modifikatoren heißen Attribute (pränominale AP, postnominale PP, Relativsatz), Satzmodifikatoren Adjunkte (auch „adverbiale Bestimmungen“).

Grammatische Funktionen



Regelschemata

- NP-Struktur im Deutschen (vereinfacht)

  NP → EN | Pro | Det N'

  N' → AP N'

  N' → N' PP

  N' → N (NP)

- Allgemeines XP (X-Bar-)Schema:

  XP → SpecXP X' "Specifier" + Kopf

  X' → (YP) X' beliebig viele Prämodifikatoren + Kopf

  X' → X' (YP) Kopf + beliebig viele Postmodifikatoren

  X' → X YP_1 ... Yp_n lexikalischer Kopf +
  n "subkategorisierte" Argumente

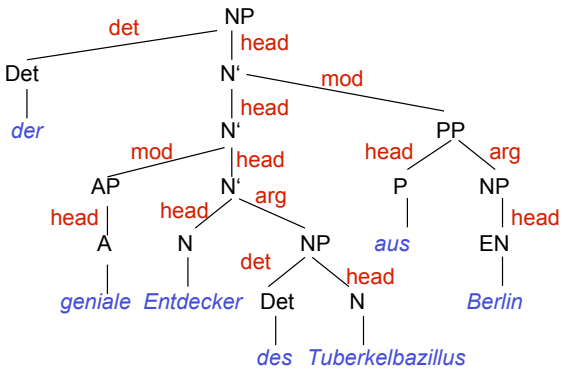
Grammatiktheorie

- Die CFG als solche ist ein Formalismus zur syntaktischen Beschreibung. Die Frage, welche Ausdrücke als Konstituenten betrachtet werden sollen und welche Kategorien und Funktionen die Grammatik annehmen soll, ist eine Angelegenheit der Grammatiktheorie.
- Die Frage hat keine einfache Antwort. Unterschiedliche Auffassungen haben zu unterschiedlichen Grammatiktheorien geführt.
- Einvernehmen besteht z.B. 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").

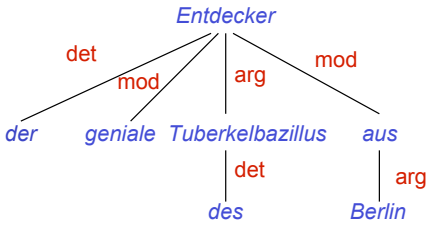
Ausblick: Dependenzgrammatik

- Kontextfreie Grammatiken beschreiben die Konstituenten-struktur 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).

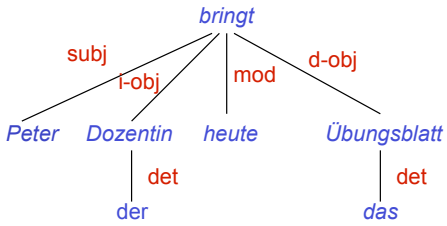
Grammatische Funktionen



Dependenzgrammatik: Beispielstruktur



Dependenzgrammatik: Beispielstruktur2



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.

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"

Syntaktische Struktur und semantische Interpretation

- Die syntaktische Struktur ist 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?

Grammatische Mehrdeutigkeit

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

- Die zwei Analysevarianten (Zwischenknoten z.T. weggelassen)
  - $[_S$ *Peter* $[_{VP}$ *sah* $[_{NP}$ *den* $[_{N'}$ *Mann* $[_{PP}$ *mit dem Teleskop*]]]]]
  - $[_S$ *Peter* $[_{VP}$ $[_{V'}$ *sah* $[_{NP}$ *den Mann*]] $[_{PP}$ *mit dem Teleskop*]]]