Einführung in die Computerlinguistik

Syntax II

WS 2013/14

Manfred Pinkal

Eigenschaften der syntaktischen Struktur [1]

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

Eigenschaften der syntaktischen Struktur [2]

Peter hat der Dozentin das Übungsblatt heute ins Büro gebracht.

Das Übungsblatt hat Peter der Dozentin heute ins Büro gebracht.

Der Dozentin hat Peter heute das Übungsblatt ins Büro gebracht.

Ins Büro hat heute Peter der Dozentin das Übungsblatt gebracht.

Heute hat Peter das Übungsblatt der Dozentin ins Büro gebracht.

?Ins Büro hat das Übungsblatt der Dozentin Peter heute gebracht.

** Ins Büro heute Peter das Übungsblatt hat gebracht der Dozentin.*

** Ins heute Büro der Peter Dozentin das hat Übungsblatt gebracht.*

Eigenschaften der syntaktischen Struktur [3]

- Wie finden Sie stattdessen die angehängten Bilder? Das sind Fotos, die im Rahmen des TALK-Projektes entstanden sind, uns gehören, und von BMW schon freigegeben waren. Außerdem vermitteln sie besser den Bezug zur Forschung.

Schachtelstruktur in natürlichen Sprachen

- "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

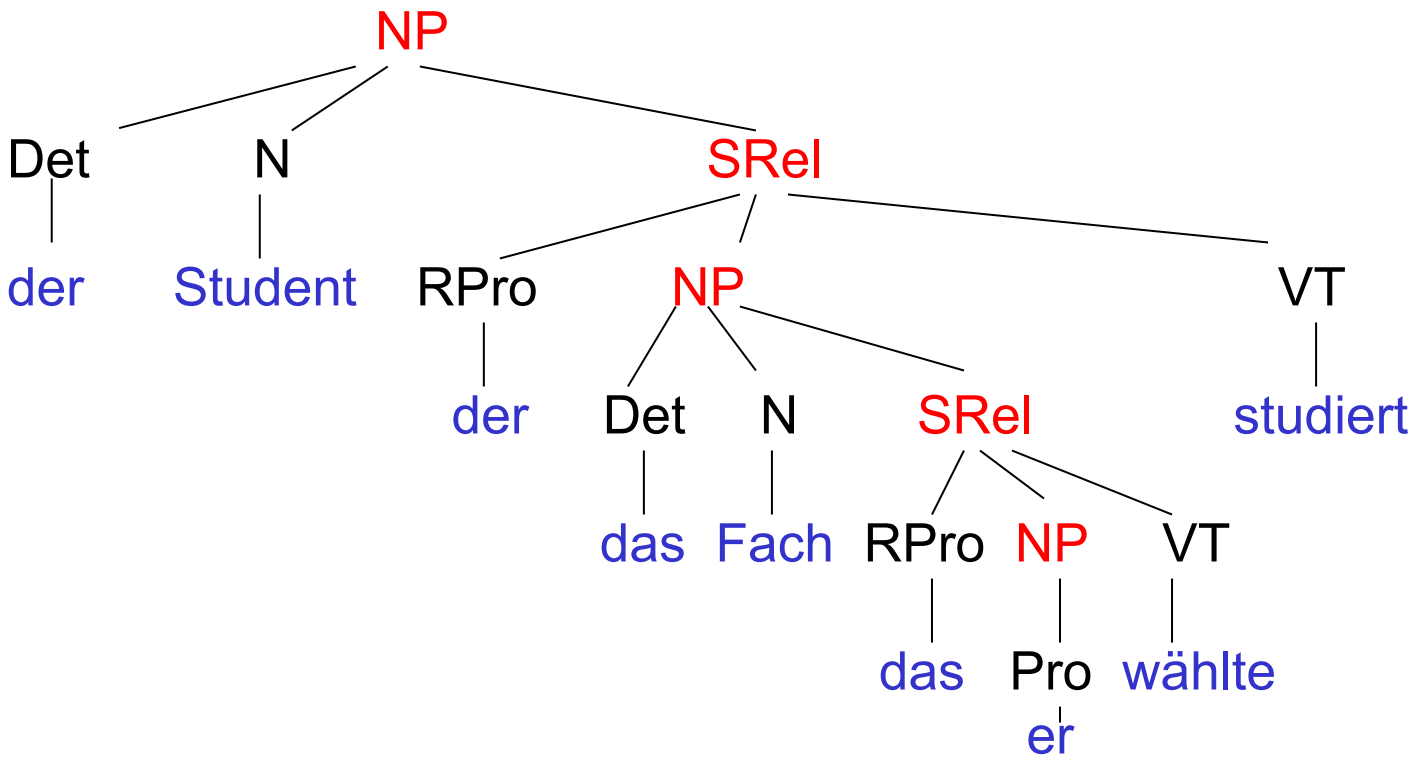
G1 = <V, Σ, P, S> mit

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* | *liest*	VT → *wählte* | *studiert*
N → *Student* | *Fach* | *Buch*	RPro → *der* | *das*
Det → *der* | *das* | *ein*	Pro → *er* | *sie*

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*]]



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) einem nicht-terminalen Symbol A zugeordnet, aus dem u abgeleitet werden kann. Wir nennen u eine „Konstituente“ von der „Kategorie“ A, und sagen, dass A die Elemente von u „dominiert“.

Kategorien und Konstituenten

- *er* ist eine Konstituente der Kategorie Pro
- *er – der Student – der Student, der Informatik studiert* sind Konstituenten der Kategorie NP
- *der das Fach, das er wählte, studiert* – *das er wählte* sind Konstituenten der Kategorie SRel

CFG: Konstituentenstruktur

- Ersetzungsregeln von CFGs erlauben nicht nur das Aufzählen von grammatisch korrekten Sätzen, sondern die Darstellung syntaktischer Struktur.
- Das heißt aber für den Grammatikschreiber, dass er nicht irgendwelche Ersetzungsregeln „erfindet“, sondern dass er Regeln und Kategorien so wählt, dass sie die syntaktische Struktur in geeigneter Weise repräsentieren.
- Das ist einfach für formale Sprachen. In der Arithmetik haben wir „Gleichung“, „Term“ und „Operator“ als Kategorien, und die syntaktische Struktur ist offensichtlich.
- Wie geht man aber bei natürlichen Sprachen vor? Was sind plausible Konstituenten, und wie findet man angemessene Kategorien?
- Diese Fragen stellt und beantwortet die Grammatiktheorie.

Kriterien für Konstituentenstruktur I

Verschiebetest:

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

Substitutionstest:

- *[$_{NP}$ Peter] hat [$_{NP}$ das neue Übungsblatt] [$_{NP}$ der Dozentin] heute ins Büro gebracht*
- *Er hat es ihr heute ins Büro gebracht*

„Vorfeld“-Test (für das Deutsche):

- *[$_{NP}$ Peter] hat der Dozentin das Übungsblatt heute ins Büro gebracht.*
- *[$_{NP}$ Das neue Übungsblatt] hat Peter der Dozentin heute ins Büro gebracht.*
- *[$_{PP}$ Ins Büro] hat heute Peter der Dozentin das Übungsblatt gebracht.*

Kriterien für Konstituentenstruktur II

- Distributionelle Eigenschaften:
  - Verschiebbarkeit, Substituierbarkeit
- Interne strukturelle Eigenschaften:
  - Ausdrücke besitzen tendenziell einen „Kopf“ eines bestimmten Typs, der ihren "grammatischen Charakter" bestimmt
  - Beispiel: Komplexe Nominalausdrücke besitzen einheitlich als „Kopf“ ein Substantiv, das Genus-, Numerus-, Kasus-Merkmale trägt, einen Artikel verlangt, durch Adjektive modifiziert werden kann, ...
- Semantische Eigenschaften:
  - Konstituenten beschreiben sinnvolle Bedeutungseinheiten; Konstituenten derselben Kategorie beschreiben tendenziell Bedeutungseinheiten desselben Typs.
  - Beispiel: Nominalausdrücke bezeichnen („referieren auf“) Entitäten (Personen und Objekte)

Lexikalische und phrasale Hauptkategorien

- 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*

Globale Satzstruktur

- In unserer Beispielgrammatik hatten wir die folgenden Regeln zur Satzstruktur angenommen:

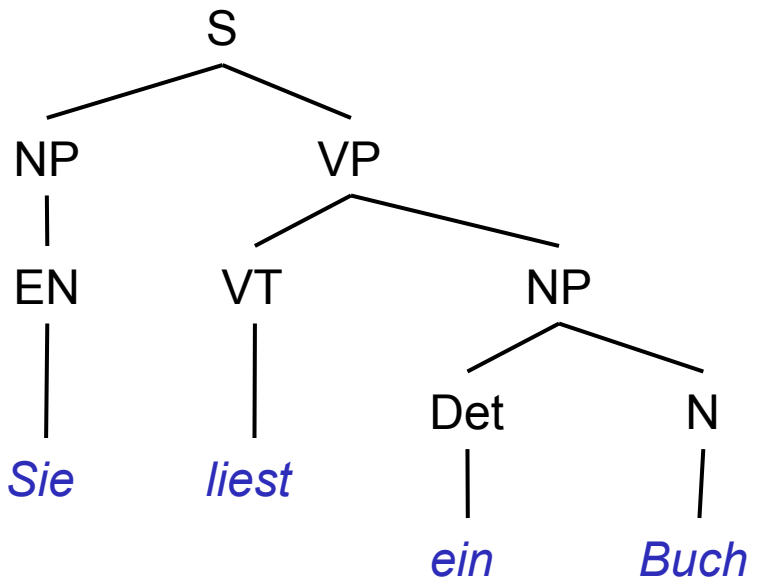
  S → NP VI S → NP VT NP

- Mit Verbphrasen als Hauptkategorien erhalten wir stattdessen:

S → NP VP

VP → VI

VP → VT NP



NP-Struktur

Beispiel:

der geniale Entdecker des Tuberkelbazillus aus Berlin

- NP-Struktur im Deutschen (vereinfacht)

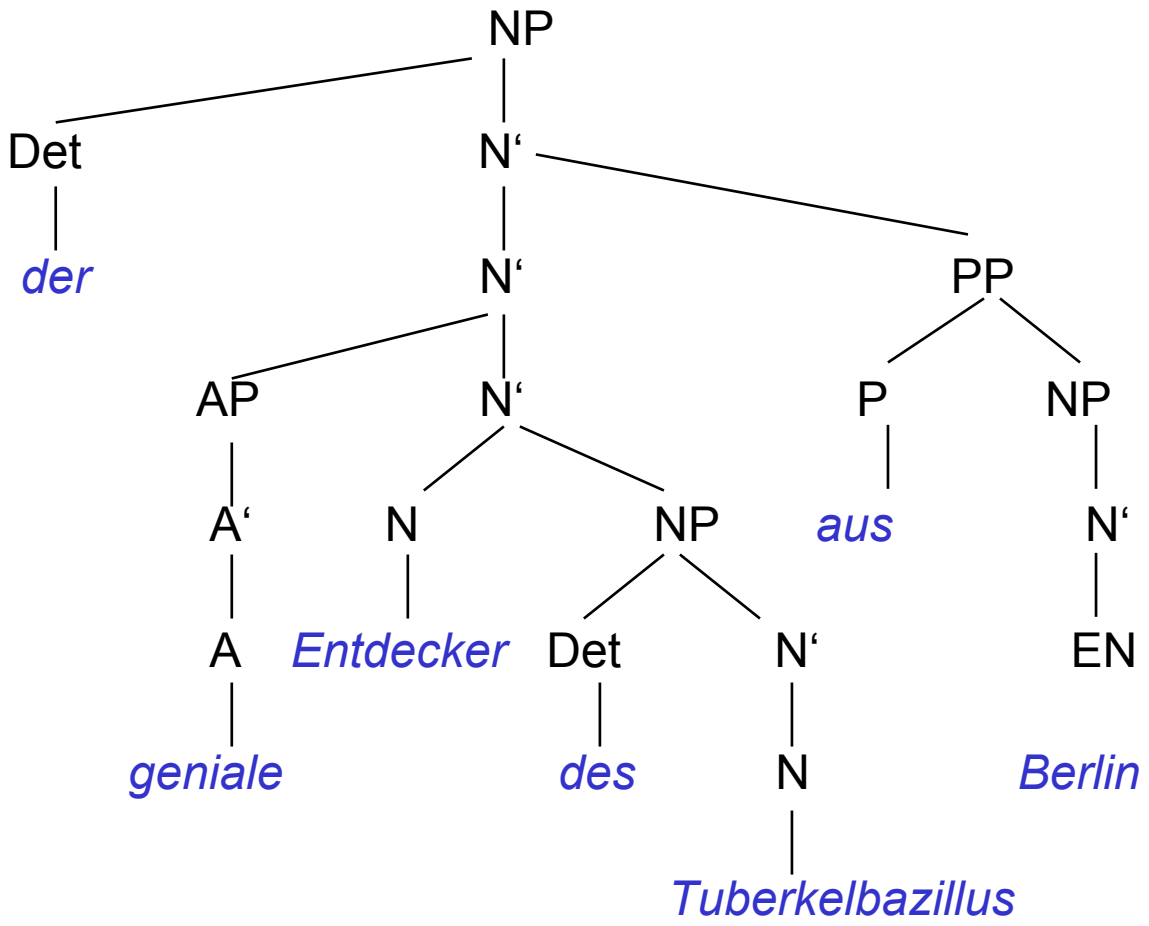
  NP → EN | Pro | Det N'

  N' → AP N'

  N' → N' PP

  N' → N (NP)

NP-Struktur: Ein Beispiel



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“ innerhalb von anderen Phrasen vor, lassen sich relativ leicht verschieben, können nur schwer durch anderes Material unterbrochen werden.

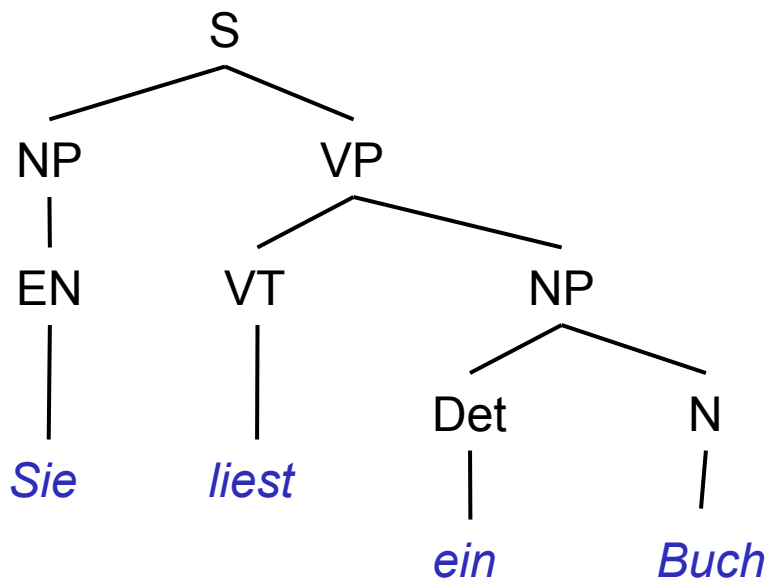
- Zwischenkategorien: Hier nimmt man meist genau eine weitere Ebene an, die zwischen der phrasalen und der lexikalischen Ebene vermittelt. Sie werden üblicherweise als N‘, A‘, V‘ etc. notiert, alternativ mit einem Überstrich, daher als „N-Bar“, „V-Bar“ etc. ausgesprochen.

Kategorie und Funktion

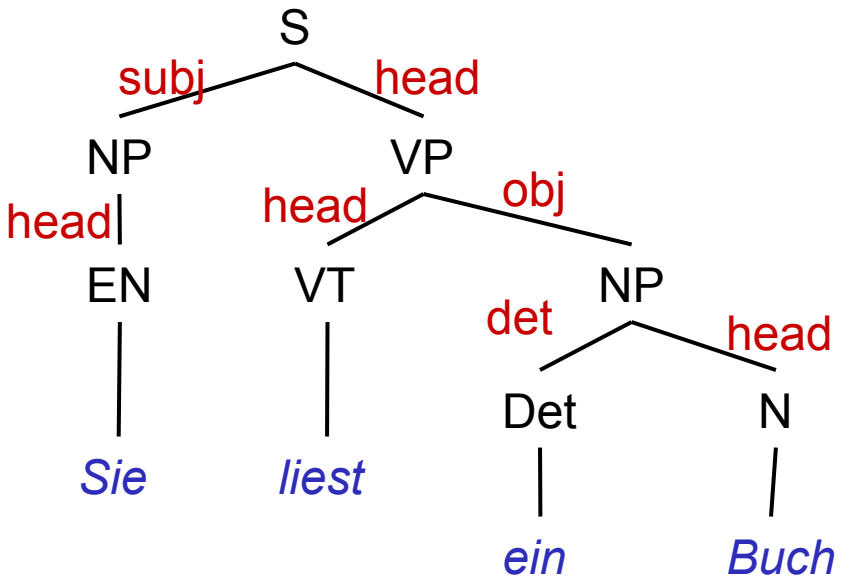
- Syntaktische Kategorien bezeichnen Klassen von Ausdrücken mit ähnlicher innerer Struktur und ähnlichem distributionellem Verhalten.

- Grammatische Funktionen dagegen bezeichnen die Rolle, die eine Konstituente im größeren Ausdruck spielt. Grammatische Funktionen sind relationale Konzepte! (Sie werden deshalb alternativ auch „grammatische Relationen“ genannt.)
- Eine Kategorie kann in unterschiedlichen Funktionen vorkommen: 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.
- Unterschiedliche Kategorien können die gleiche Funktion ausüben: Subjekte können zum Beispiel Nominalphrasen oder Sätze sein:
  - *Dass es regnet, ist lästig*
  - *Der Regen ist lästig.*

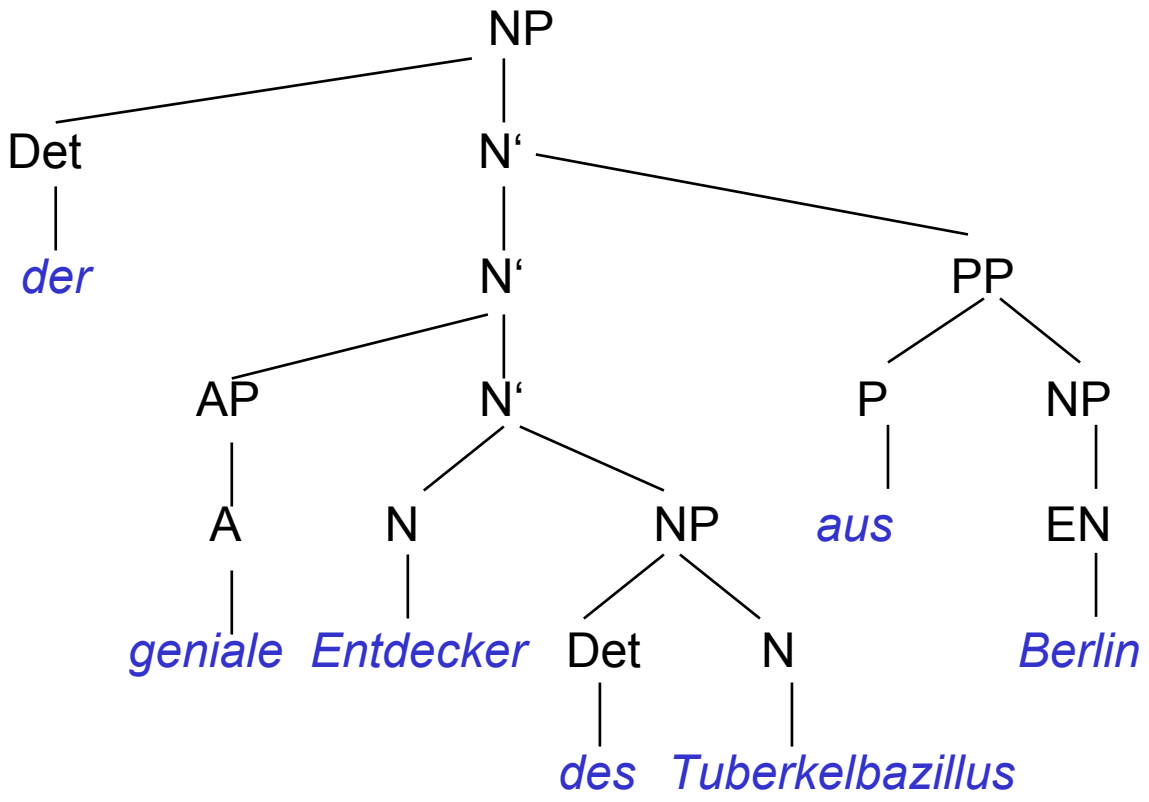
Ein Beispiel



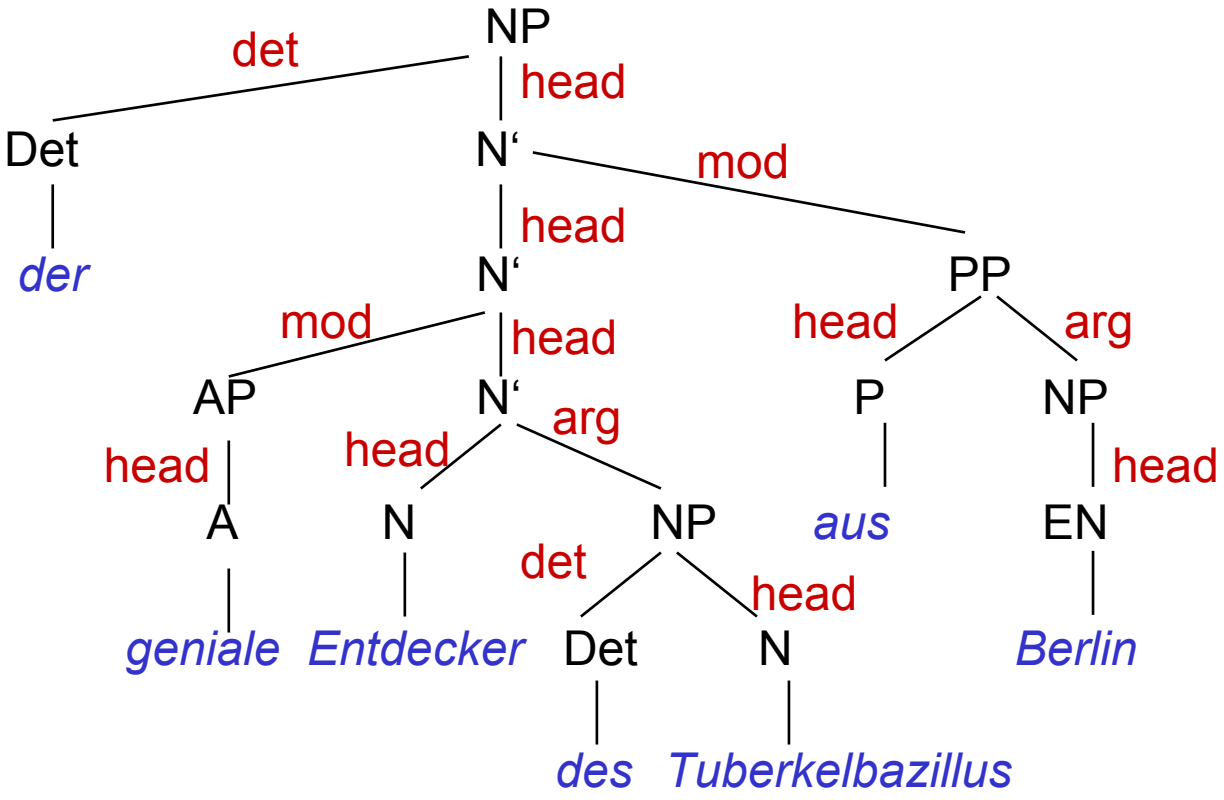
Ein Beispiel



Ein zweites Beispiel



Grammatische Funktionen



Haupttypen grammatischer 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 auch Attribute (pränominale AP, postnominale PP, Relativsatz), Satzmodifikatoren Adjunkte (auch „adverbiale Bestimmungen“).

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 → SpecX 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.

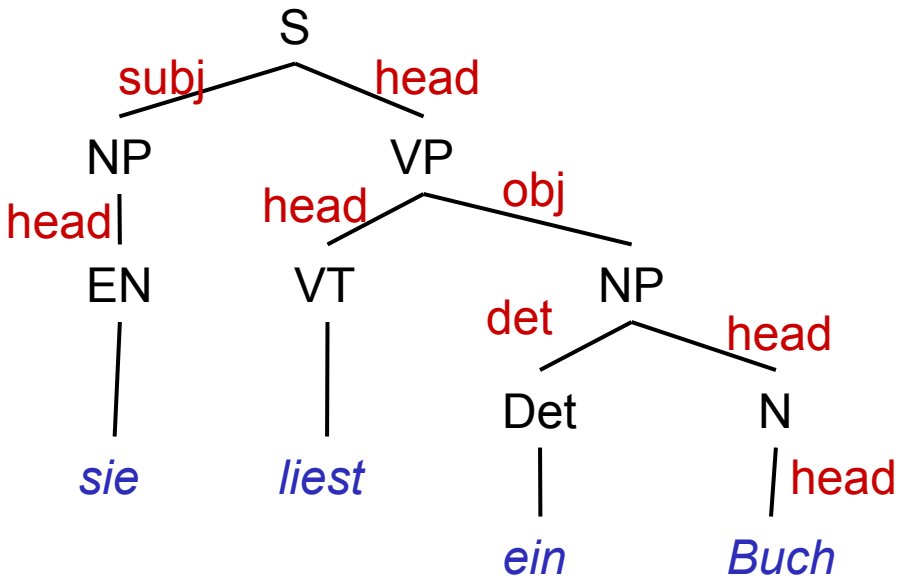
- Die Frage hat keine einfache Antwort. Unterschiedliche Auffassungen haben zu unterschiedlichen Grammatiktheorien geführt.

- 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“).

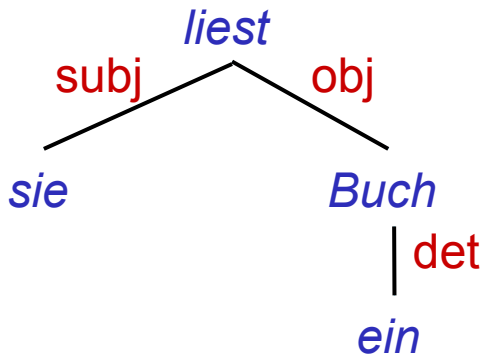
Ausblick: Dependenzgrammatik

- 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).

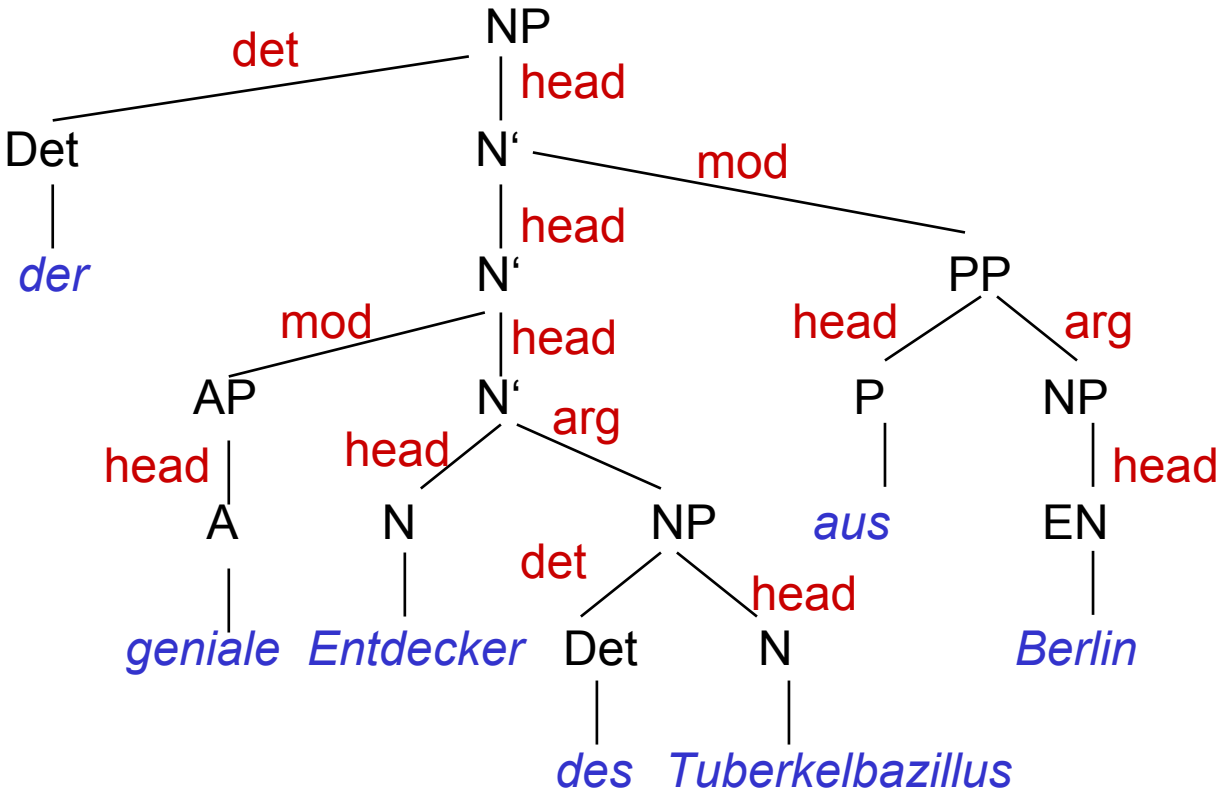
Beispiel: Konstituentenstruktur + grammatische Funktionen/Relationen



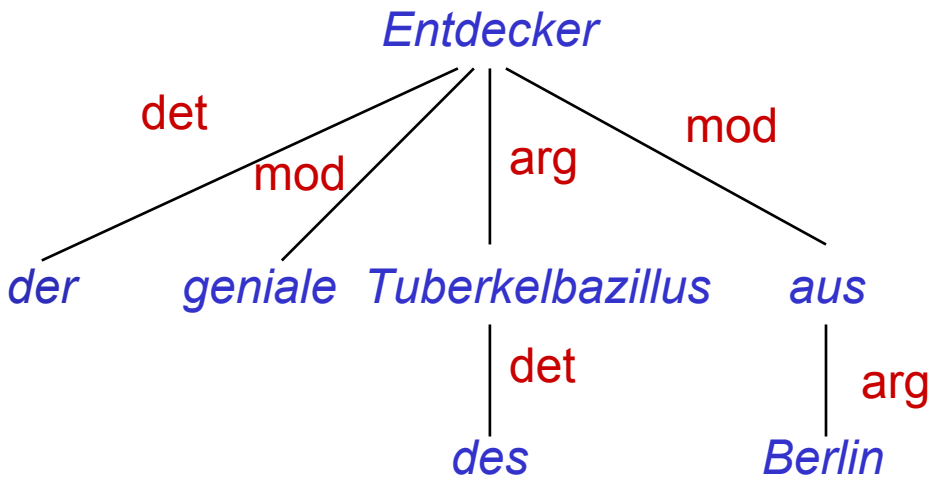
Ein Beispiel: Dependenzstruktur



Grammatische Funktionen



Dependenzgrammatik: Beispielstruktur



Kontextfreie Grammatik und Dependenzgrammatik

- 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

- 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* [$_{VP}$ [$_{V'}$ [$_{V‘}$ *sah* [$_{NP}$ *den Mann*]] [$_{PP}$ *mit dem Teleskop*]]]]

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“