

# Einführung in die Computerlinguistik

## Syntax II

WS 2012/2013

Manfred Pinkal

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## 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.*

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## 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.*

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## 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.

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## 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 wiederab liefert, errechnete Beurteilungspegel..."

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## Eine kontextfreie Grammatik für deutsche Sätze

$G_1 = \langle V, \Sigma, P, S \rangle$  mit

$V = \{S, SRel, NP, VI, VT, N, Det, RPro\} \cup \Sigma$

$\Sigma = \{schläft, arbeitet, studiert, wählte, Student, Fach, der, das, er\}$

$S \rightarrow NP VI$	$S \rightarrow NP VT NP$
$SRel \rightarrow RPro VI$	$SRel \rightarrow RPro NP VT$
$NP \rightarrow Det N (SRel)$	$NP \rightarrow Pro$
$VI \rightarrow schläft \mid arbeitet \mid liest$	$VT \rightarrow wählte \mid studiert$
$N \rightarrow Student \mid Fach \mid Buch$	$RPro \rightarrow der \mid das$
$Det \rightarrow der \mid das \mid ein$	$Pro \rightarrow er \mid sie$

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

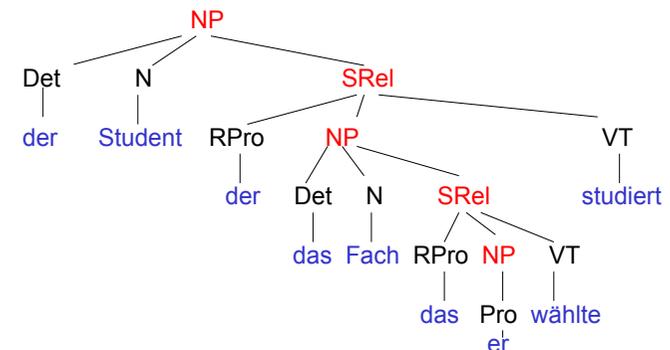
## CFG: Konstituentenstruktur

- 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“.

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## 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]



Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## 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

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## Konstituenten-Test 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

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## CFG: Konstituentenstruktur

- Ersetzungsregeln von CFGs erlauben die elegante Darstellung struktureller Regularitäten.
- Dafür müssen Ersetzungsregeln und Kategorien in geeigneter Weise spezifiziert werden.
- Das ist trivial für formale Sprachen: Die werden ja explizit mithilfe von CFGs definiert.
- Wie geht man aber bei natürlichen Sprachen vor? Wie bestimmt man Konstituenten, und wie ordnet man sie Kategorien zu?

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## Konstituenten-Test 2: Substituieren

- *[<sub>NP</sub> Er ] hat die Übungen gemacht.*
- *[<sub>NP</sub> Peter ] hat die Übungen gemacht.*
- *[<sub>NP</sub> Der Student ] hat die Übungen gemacht.*
- *[<sub>NP</sub> 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.

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## Konstituenten-Test 3 (für das Deutsche): Vorfeldbesetzung

- [<sub>NP</sub> Peter ] hat der Dozentin das Übungsblatt heute ins Büro gebracht.
- [<sub>NP</sub> Das Übungsblatt ] hat Peter der Dozentin heute ins Büro gebracht.
- [<sub>NP</sub> Der Dozentin ] hat Peter heute das Übungsblatt ins Büro gebracht.
- [<sub>PP</sub> Ins Büro ] hat heute Peter der Dozentin das Übungsblatt gebracht.
- [<sub>Adv</sub> Heute ] hat Peter das Übungsblatt der Dozentin ins Büro gebracht.

"Verb-Zweit" bietet einen verlässlicher Konstituententest fürs Deutsche: Was vor dem finiten Verb im Hauptsatz ("Vorfeld") steht, ist eine Konstituente.

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## Kriterien für Konstituentenstruktur

- Distributionelle Kriterien:
  - Verschiebbarkeit, Substituierbarkeit
- Lokale strukturelle Eigenschaften:
  - Komplexe Ausdrücke besitzen tendenziell einen „Kopf“ eines bestimmten Typs, der ihren "grammatischen Charakter" bestimmt
  - Beispiel: Nominalausdrücke besitzen einheitlich als „Kopf“ ein Substantiv oder ein Pronomen, das Genus-, Numerus- und Kasusmerkmale trägt, einen Artikel verlangt, durch Adjektive modifiziert werden kann, ...
- Semantisches Kriterium:
  - Konstituenten beschreiben sinnvolle Bedeutungseinheiten; Konstituenten einer Kategorie beschreiben Bedeutungseinheiten desselben oder ähnlichen Typs
  - Beispiel: Nominalausdrücke bezeichnen/ "denotieren" Entitäten (Personen und Objekte)

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## Etwas Grammatiktheorie: 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*

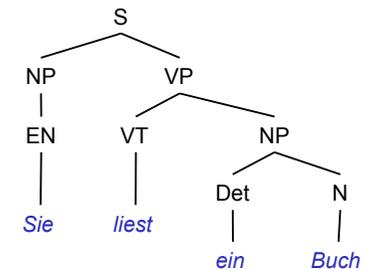
Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## Globale Satzstruktur

- In unserer Beispielgrammatik hatten wir die folgenden Regeln zur Satzstruktur angenommen:
 
$$S \rightarrow NP VI \quad S \rightarrow NP VT NP$$
- Wenn wir eine Verbphrase als Hauptkategorie annehmen, erhalten wir stattdessen:

$$S \rightarrow NP VP$$

$$VP \rightarrow VI$$

$$VP \rightarrow VT NP$$


Vorlesung "Einführung in die CL" 2012/2013 © 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 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.

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## NP-Struktur

- NP-Struktur im Deutschen (vereinfacht)

NP → EN | Pro | Det N'

N' → AP N'

N' → N' PP

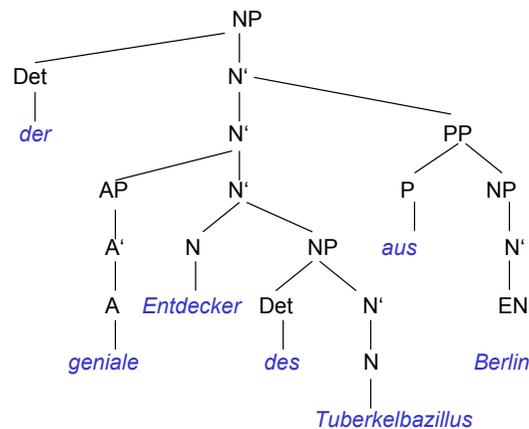
N' → N (NP)

Beispiel:

*der geniale Entdecker des Tuberkelbazillus aus Berlin*

Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## NP-Struktur: Ein Beispiel



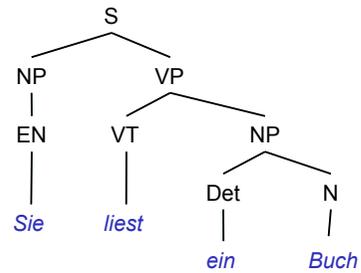
Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## 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.*

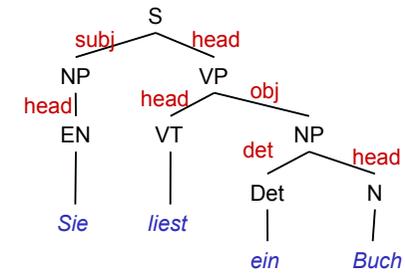
Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UdS Computerlinguistik

## Ein Beispiel



Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## Ein Beispiel



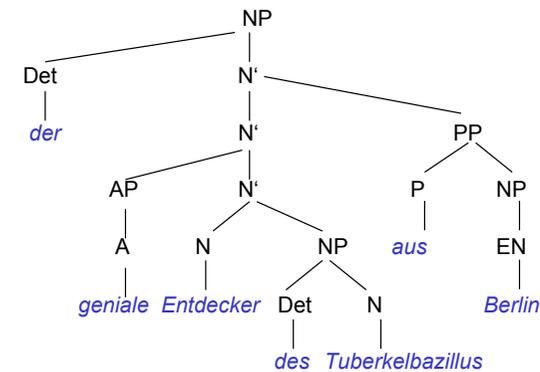
Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## 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. Verbarargumente 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“).

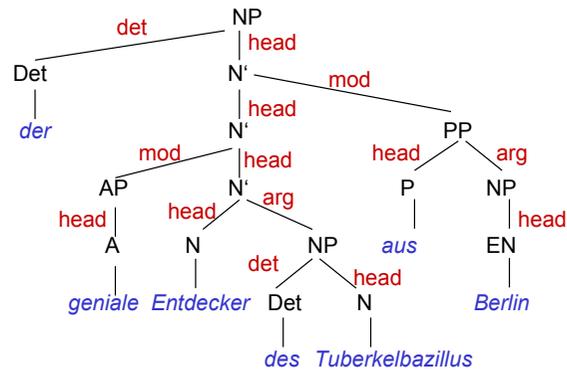
Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## Ein zweites Beispiel



Vorlesung "Einführung in die CL" 2012/2013 © M. Pinkal UoS Computerlinguistik

## Grammatische Funktionen



Vorlesung "Einführung in die CL" 2012/2013 © 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" 2012/2013 © 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 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“).

Vorlesung "Einführung in die CL" 2012/2013 © 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" 2012/2013 © M. Pinkal UdS Computerlinguistik

## 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:

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

- Die zwei Analysevarianten (Zwischenknoten z.T. weggelassen)
  - [<sub>S</sub> Peter [<sub>VP</sub> sah [<sub>NP</sub> den [<sub>N'</sub> Mann [<sub>PP</sub> mit dem Teleskop ] ] ] ] ]
  - [<sub>S</sub> Peter [<sub>VP</sub> [<sub>V'</sub> sah [<sub>NP</sub> den Mann ] ] [<sub>PP</sub> mit dem Teleskop ] ] ]