

# Einführung in die Computerlinguistik

## Syntax

WS 2021/2022  
Vera Demberg

## Erinnerungen / Infos

- Fragen zum Übungsblatt können Sie auch in den Wochen, in denen ein Übungsblatt rauskommt, im Freitagstutorium stellen.
- Sie können im Slot von 12-14 Uhr hier im Raum bleiben, um an einer anderen Veranstaltung, die nur remote angeboten wird, teilzunehmen.

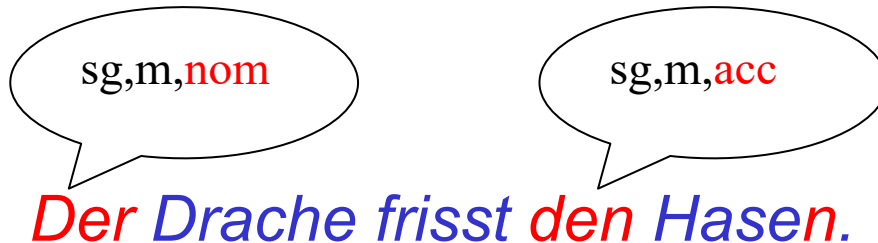
## Ziel heute

- Wie werden Sätze aufgebaut?
- Wie kann man Satzstruktur formal beschreiben?
- Wichtige Terminologie im Bereich Satzstruktur

Nächste Woche:

- Satzstrukturen mit dem Computer berechnen.

# Morphologie und Syntax



- Morphologie beschreibt die grammatischen Merkmale von Wörtern, die durch Wortform oder Flexionsmorpheme kodiert werden.
- Syntax beschreibt die Interaktion der grammatischen Merkmale unterschiedlicher Wörter im Satz.

# Morphologie und Syntax



- Morphologie beschreibt die grammatischen Merkmale von Wörtern, die durch Wortform oder Flexionsmorpheme kodiert werden.
- Syntax beschreibt die Interaktion der grammatischen Merkmale unterschiedlicher Wörter im Satz.

# Morphologie und Syntax

sg,m,nom

sg

*Der Drache frisst den Hasen.*

*Den Hasen frisst der Drache.*

*Die Drachen fressen den Hasen.*

pl,m,nom

pl

- Morphologie beschreibt die grammatischen Merkmale von Wörtern, die durch Wortform oder Flexionsmorpheme kodiert werden.
- Syntax beschreibt die Interaktion der grammatischen Merkmale unterschiedlicher Wörter im Satz.

# 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.

# Beispiele aus der juristischen Praxis

- "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 wiederabliefern, errechnete Beurteilungspegel..."
- "Bei der Umsetzung der Vorgaben der Gerichte für eine verfassungskonforme Regelung der Überführung von Ansprüchen und Anwartschaften aus den Zusatz- und Sonderversorgungssystemen der ehemaligen DDR..."

 **Beliebig lange und tiefe Schachtelung ist möglich.**



## 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.*

 variable Wortstellung (je nach Sprache)

## Eigenschaften der syntaktischen Struktur [3]

- Wie finden Sie stattdessen das angehängte Bild?  
Das ist ein Foto, das im Rahmen des TALK-Projektes entstanden ist, uns gehört, und von BMW schon freigegeben war. Außerdem vermittelt es besser den Bezug zur Forschung.

## 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.

➡ Abhängigkeiten von Wörtern zueinander in Wortform

# Eigenschaften der syntaktischen Struktur

- Sätze setzen sich aus Satzteilen (**Konstituenten**) zusammen, die wiederum aus einfachen oder ihrerseits komplexen Satzteilen bestehen können. Sätze können deshalb **beliebig lang und beliebig tief geschachtelt** sein.
- Die Syntax natürlicher Sprachen erlaubt **variable Wortstellung**: Wörter und Konstituenten mit der gleichen Funktion können an unterschiedlichen Stellen eines Satzes stehen. Unterschiedliche Sprachen erlauben sehr unterschiedliche Freiheitsgrade.
- Die **grammatischen Eigenschaften** unterschiedlicher Wörter und Konstituenten im Satz **hängen voneinander ab** – zum Teil auch in Fällen, in denen die Wörter und Konstituenten im Satz weit auseinander liegen.

# Fragen zur Repräsentation und Verarbeitung syntaktischer Strukturen

- Natürliche Sprachen sind Sprachen im Sinne der formalen Definition:
  - Wörter sind die Symbole
  - Das Lexikon ist das "Alphabet" ( $\Sigma$ )
  - Korrekte Sätze sind "Worte" über dem Alphabet
  - Die Menge der korrekten Sätze definiert die Sprache  $L \subseteq \Sigma^*$
- Kann man natürliche Sprachen mit endlichen Automaten beschreiben?  
Gibt es also für eine Sprache  $L$  einen Automaten  $A$  mit  $L(A) = L$ ?  
Anders gefragt: **Sind natürliche Sprachen durch einen regulären Ausdruck darstellbar, sind sie regulär?**
- Kann eventuell sogar jede denkbare Sprache mit einem endlichen Automaten beschrieben werden?

# Aufgabe

- Können Sie einen Automaten bauen, der die Sprache  $a^n b^n$  beschreibt?  
(also eine beliebige Anzahl von „a“ gefolgt von genauso vielen „b“)

Falls ja, wie?

Falls nein, warum nicht?

# Fragen zur Repräsentation und Verarbeitung syntaktischer Strukturen

Hier nochmal unsere Frage von vorhin:

*Kann eventuell sogar jede denkbare Sprache mit einem endlichen Automaten beschrieben werden?*

- Die Antwort ist **Nein**: Es gibt Sprachen, die sich nicht mit endlichen Automaten beschreiben lassen
- ... und zwar sehr einfache Sprachen wie  $a^n b^n$ .

# $a^n b^n$ und endliche Automaten

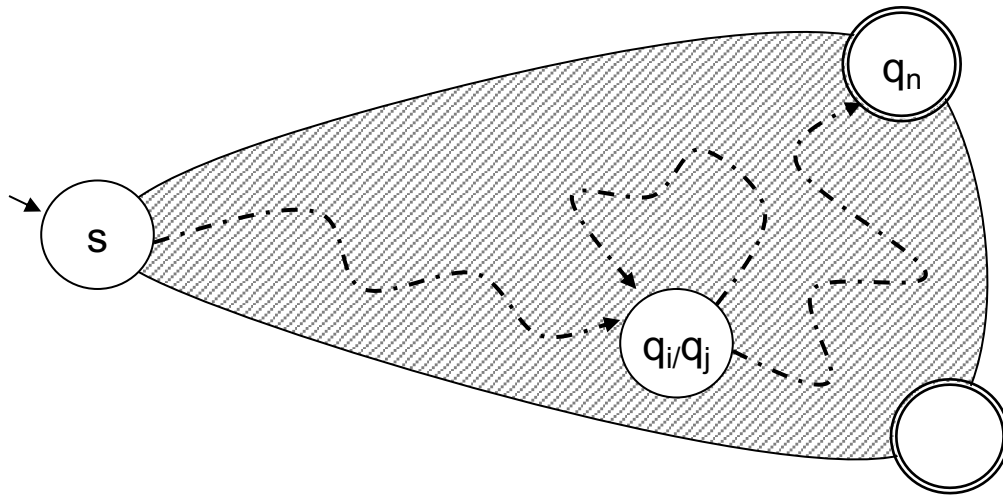
Um Zugehörigkeit zu  $a^n b^n$  zu erkennen, müsste sich der Automat beliebig lange Ketten von  $a$ 's merken können, weil er die Information anschließend beim Abarbeiten von  $b$ 's braucht.

Endliche Automaten haben eine fundamentale Einschränkung: Ihr „Gedächtnis“ ist **endlich**, durch die Anzahl ihrer Zustände beschränkt. Ein Automat mit  $k$  Zuständen kann sich nur an einen beschränkten Kontext „erinnern“, nämlich maximal die  $k$  vorausgegangenen Symbole. (Anders ausgedrückt: Er kann nur bis  $k$  zählen.)

Ein endlicher Automat kann deshalb nur solche Sprachen erkennen, bei denen die Zulässigkeit eines Symbols in einer Zeichenfolge auf der Grundlage eines Vorkontextes von begrenzter Länge entschieden werden kann.



# Beschränkungen endlicher Automaten



Ein endlicher Automat, der unendliche Sprachen erkennen kann, muss mind. eine Schleife besitzen.

Beispielsprachen:

$uv^n w$

$(ab)^n$

$a^n b^m$

“Pumping Lemma”  
(dt.: Schleifensatz)

# Was hat $a^n b^n$ mit natürlicher Sprache zu tun?

Gibt es etwas Ähnliches wie das „ $a^n b^n$ “ auch in natürlichen Sprachen?

# Kontextfreie Grammatik und natürliche Sprache

- 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.

# Kontextfreie Grammatik und natürliche Sprache

- "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 wiederabliefern, errechnete Beurteilungspegel..."

# Kontextfreie Grammatik: Ein neuer Formalismus

- Kontextfreie Grammatiken („KFG“, „CFG“) beschreiben Sprachen mithilfe von Ersetzungsregeln („rewrite rules“, **Produktionen**) der Form  **$A \rightarrow w$** 
  - Beispiel:  $S \rightarrow aSb$ ,  $S \rightarrow \varepsilon$  beschreibt  $L = a^n b^n$
- **$A \rightarrow u$**  ist zu lesen als: Ein Vorkommen von A in einer Symbolfolge/ einem Wort kann durch u ersetzt werden

# Kontextfreie Grammatik: Ein neuer Formalismus

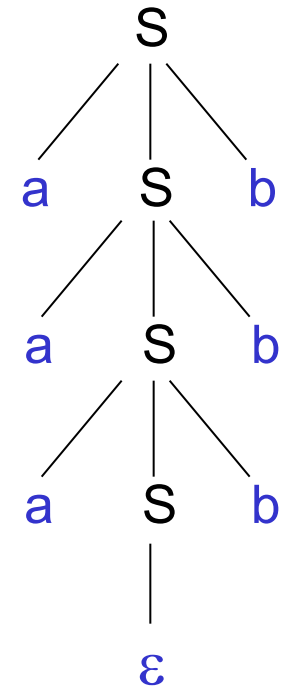
- Kontextfreie Grammatiken („KFG“, „CFG“) beschreiben Sprachen mithilfe von Ersetzungsregeln („rewrite rules“, **Produktionen**) der Form  $A \rightarrow w$ 
  - Beispiel:  $S \rightarrow aSb$ ,  $S \rightarrow \varepsilon$  beschreibt  $L = a^n b^n$
- $A \rightarrow u$  ist zu lesen als: Ein Vorkommen von A in einer Symbolfolge/ einem Wort kann durch u ersetzt werden
  - Beispiel: aaSbb wird zu aaaSbbb oder zu aa $\varepsilon$ bb = aabb
- Eine solche Ersetzung ist ein **zulässiger Ableitungsschritt**. Wir schreiben:  $aaSbb \Rightarrow aaSbbb$  bzw.  $aaSbb \Rightarrow aabb$ .
- Um ein Wort über der Sprache  $\{a, b\}$  abzuleiten, beginnen wir mit S (dem „Startsymbol“).
- Wir wenden Ersetzungsregeln an, bis ein Wort w entsteht, das nur noch a's und b's enthält („Terminalsymbole“).
  - Beispiel:  $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaSbbb \Rightarrow aaabbb$
- Wir zeigen damit, dass w durch die Regeln der Grammatik aus S ableitbar ist: w ein Wort der durch die Grammatik beschriebenen (erzeugten) Sprache L.

# Kontextfreie Grammatiken

- Die Ableitung  
 $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbbb$   
kann alternativ durch eine **Ableitungsbaum** dargestellt werden.
- Die **Wurzel** des Baumes ist das Startsymbol.
- Die **Blätter** des Baums ergeben, von links nach rechts gelesen und aneinandergehängt, das abgeleitete Wort.
- Alternative Schreibweise:  
 $[_s a[_s a[_s a[_s \varepsilon ] b] b] b]$

# Kontextfreie Grammatiken

- Die Ableitung  
 $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbbb$   
kann alternativ durch eine **Ableitungsbaum** dargestellt werden.
- Die **Wurzel** des Baumes ist das Startsymbol.
- Die **Blätter** des Baums ergeben, von links nach rechts gelesen und aneinandergehängt, das abgeleitete Wort.
- Alternative Schreibweise:  
 $[_s a[_s a[_s a[_s \varepsilon ] b] b] b]$





# Kontextfreie Grammatik: Definitionen

$G = \langle V, \Sigma, P, S \rangle$ , wobei

- $V$  nicht-leere Menge von Symbolen
- $\Sigma \subseteq V$  nicht-leere Menge von **Terminalsymbolen**
- $P \subseteq (V - \Sigma) \times V^*$  nicht-leere Menge von **Produktionsregeln**
- $S \in V - \Sigma$  das **Startsymbol**

Die Beispielgrammatik für  $L = a^n b^n$  in formaler Notation:

- $G_1 = \langle \{a, b, S\}, \{a, b\}, \{ \langle S, aSb \rangle, \langle S, \varepsilon \rangle \}, S \rangle$
- Für  $\langle A, \alpha \rangle \in P$  schreibt man üblicherweise  $A \rightarrow \alpha$ .

# Kontextfreie Grammatik: Definitionen

- Wenn  $A \rightarrow \alpha$  Produktion,  $w = uAv$  und  $w' = u\alpha v$ , so ist  $w'$  aus  $w$  in einem Schritt ableitbar:  $w \Rightarrow w'$
- $w'$  ist aus  $w$  ableitbar:  $w \Rightarrow^* w'$  gdw. es eine Folge von Ableitungsschritten gibt, die mit  $w$  beginnt und mit  $w'$  endet.
- Die durch  $G$  erzeugte Sprache  $L(G)$  ist die Menge aller Worte über  $\Sigma^*$ , die aus  $S$  ableitbar sind:  $L(G) = \{w \in \Sigma^* \mid S \Rightarrow^* w\}$
- Sprachen, die durch kontextfreie Grammatiken erzeugt werden, heißen kontextfreie Sprachen.

# Eine erste kontextfreie Grammatik für deutsche Sätze

$G1 = \langle V, \Sigma, P, S \rangle$  mit

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

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

$P =$	$S \rightarrow NP\ VI$	$NP \rightarrow Det\ N$
	$S \rightarrow NP\ VT\ NP$	$NP \rightarrow Det\ N\ SRel$
	$SRel \rightarrow RPro\ NP\ VT$	$NP \rightarrow Pro$
	$SRel \rightarrow RPro\ VI$	
	$VI \rightarrow schläft$	$N \rightarrow Student$
	$VI \rightarrow arbeitet$	$N \rightarrow Fach$
	$VT \rightarrow studiert$	$RPro \rightarrow der$
	$VT \rightarrow wählte$	$RPro \rightarrow das$
	$Det \rightarrow der$	$Det \rightarrow das$
	$Pro \rightarrow er$	$Pro \rightarrow sie$

# Eine kontextfreie Grammatik für deutsche Sätze

## Notationskonventionen:

- Alternative Elemente können durch „|“ zusammengefasst werden
- Optionale Elemente können durch runde Klammern notiert werden.

## Kompaktere Notation der gleichen Grammatik:

$S \rightarrow NP\ VI$

$S \rightarrow NP\ VT\ NP$

$SRel \rightarrow RPro\ VI$

$SRel \rightarrow RPro\ NP\ VT$

$NP \rightarrow Det\ N\ (SRel)$

$NP \rightarrow Pro$

$VI \rightarrow schl\ddot{a}ft\ |\ arbeitet$

$VT \rightarrow w\ddot{a}hlte\ |\ studiert$

$N \rightarrow Student\ |\ Fach$

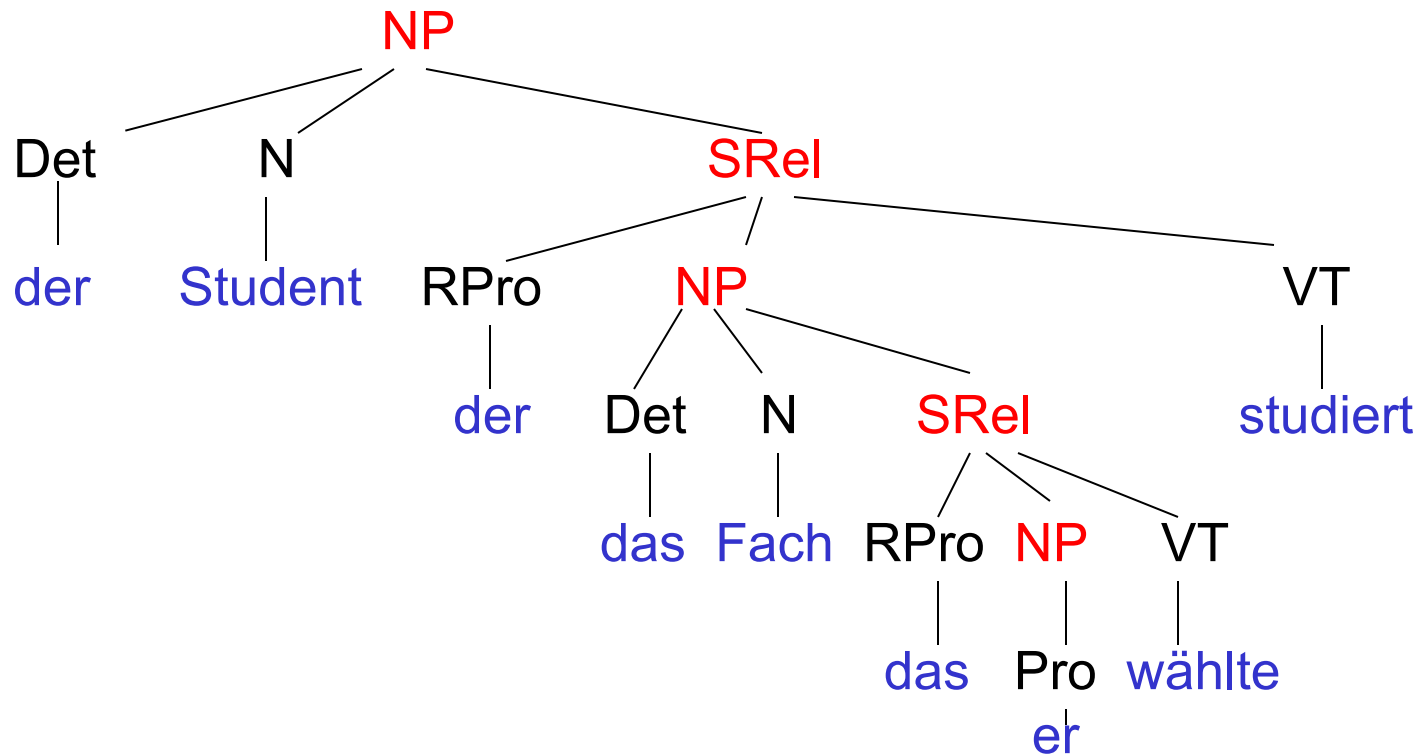
$RPro \rightarrow der\ |\ das$

$Det \rightarrow der\ |\ das$

$Pro \rightarrow er\ |\ sie$

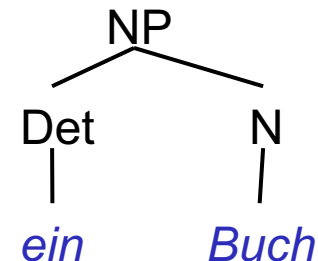
# Geschachtelte Strukturen in natürlicher Sprache

*[<sub>NP</sub> der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, [<sub>SRel</sub> der [<sub>NP</sub> das Fach, [<sub>SRel</sub> das [<sub>NP</sub> er ] nach langer Überlegung gewählt hat ]], eifrig studiert] ]*



# Kategorien und Konstituenten

Definitionen: 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**“.



- *er* ist eine Konstituente der Kategorie Pro
- *er – der Student – ein Buch – 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

# 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*

# 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)

# Globale Satzstruktur

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

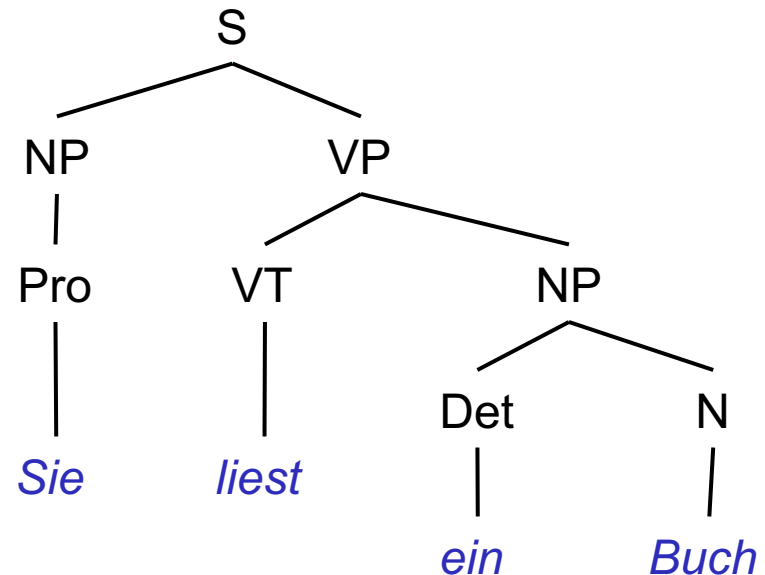
$S \rightarrow NP \text{ VI}$        $S \rightarrow NP \text{ VT NP}$

- Mit Verbphrasen als Hauptkategorien erhalten wir stattdessen:

$S \rightarrow NP \text{ VP}$

$VP \rightarrow \text{VI}$

$VP \rightarrow \text{VT NP}$



# NP-Struktur

Beispiel:

*der geniale Entdecker des Tuberkelbazillus aus Berlin*

- NP-Struktur im Deutschen (vereinfacht)

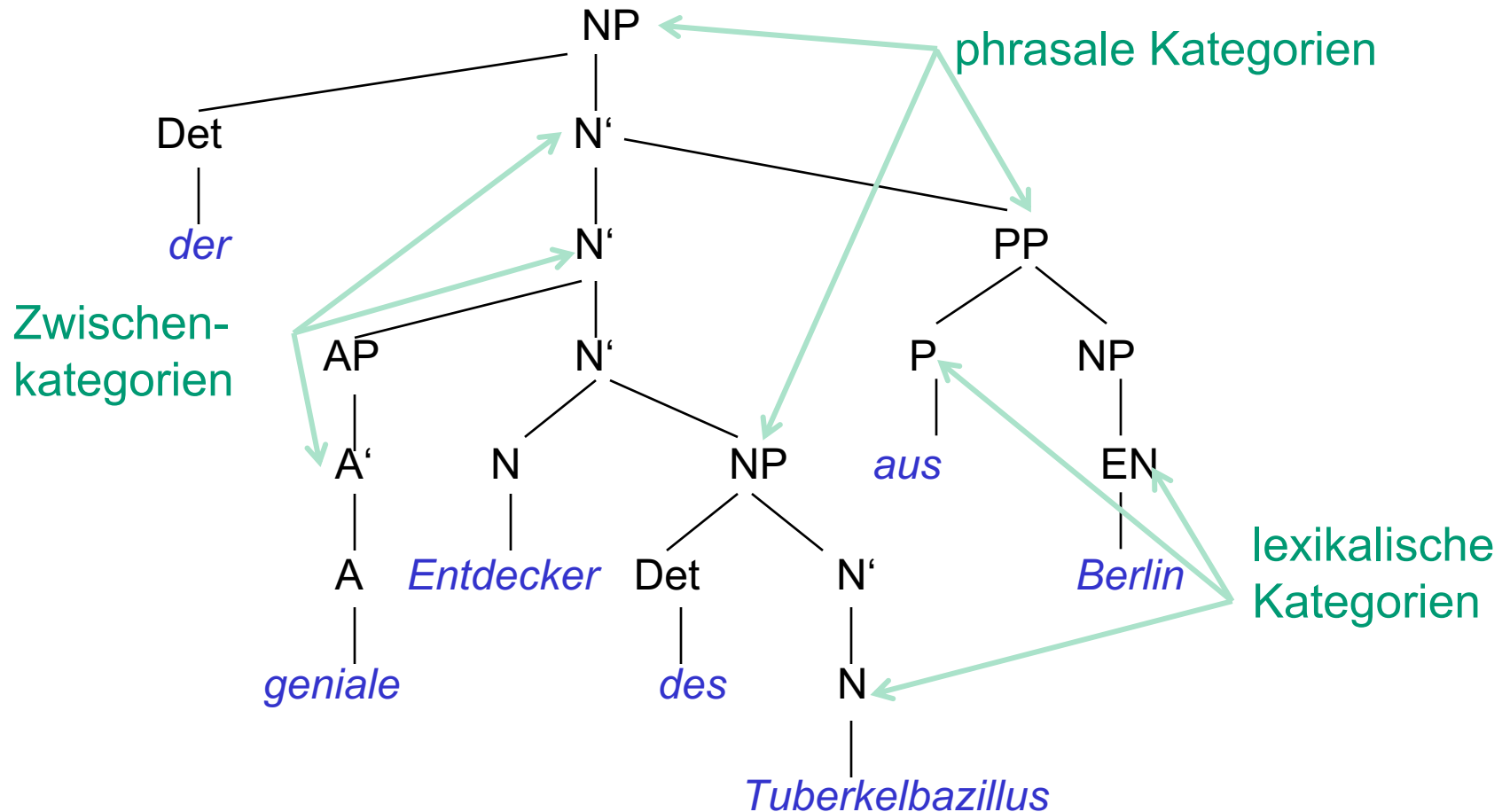
$NP \rightarrow EN \mid Pro \mid Det N'$

$N' \rightarrow AP N'$

$N' \rightarrow N' PP$

$N' \rightarrow N (NP)$

# 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

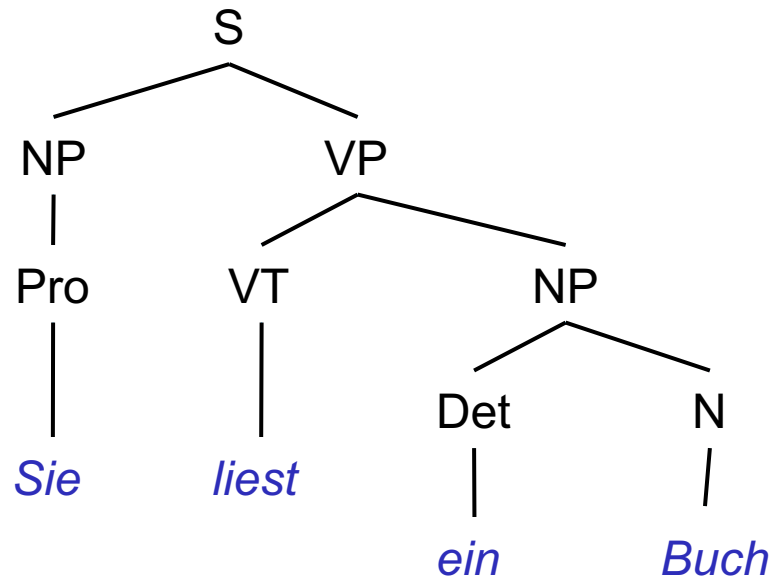
- **Syntaktische Kategorien** (z.B. NP, VP) bezeichnen Klassen von Ausdrücken mit ähnlicher innerer Struktur und ähnlichem distributionellem Verhalten.
- **Grammatische Funktionen** (z.B. subj, mod) dagegen bezeichnen die Rolle, die eine Konstituente im größeren Ausdruck spielt. Grammatische Funktionen sind relationale Konzepte! (Sie werden deshalb alternativ auch „grammatische Relationen“ genannt.)

# Kategorie und Funktion

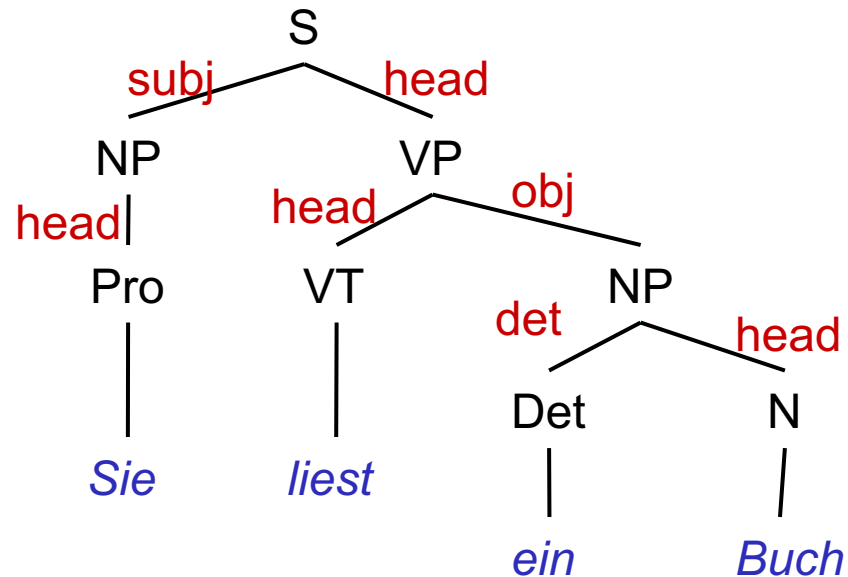
- Eine Kategorie kann in unterschiedlichen Funktionen vorkommen: Eine NP kann, je nach Stellung im Satz unter anderem die Funktion von **Subjekt** oder (direktem oder indirektem) **Objekt** eines Satzes, (Genitiv-) **Attribut** einer anderen NP oder **Argument** einer Präpositionalphrase bilden.
- Unterschiedliche Kategorien können die gleiche Funktion ausüben:  
Subjekte können zum Beispiel Nominalphrasen oder Sätze sein.
  - *Dass es regnet, ist lästig.*
  - *Der Regen ist lästig.*



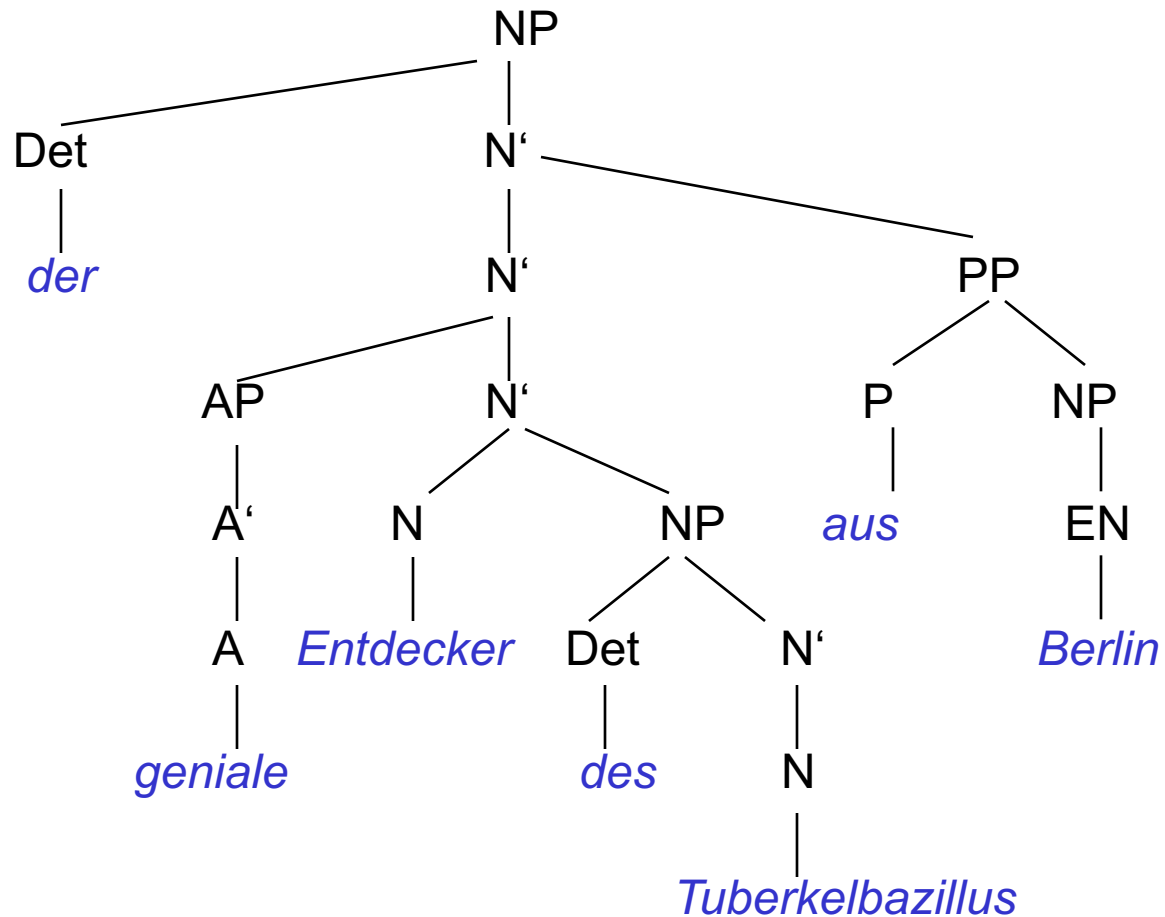
# Ein Beispiel



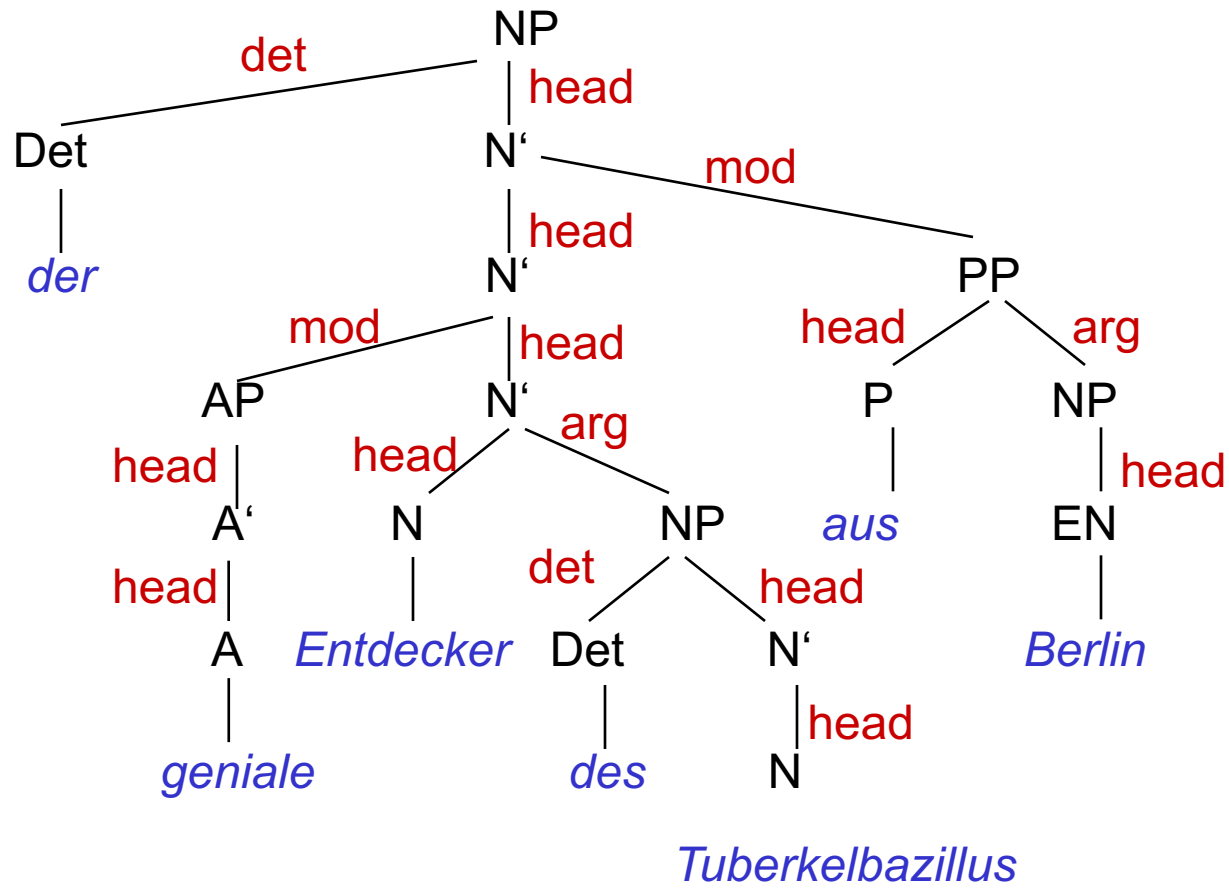
# Ein Beispiel



## 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. Verargumente 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“).

# 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“).

# Kontextfreie Sprachen und endliche Automaten

- 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. Die Regel  $S \rightarrow aSb$  besagt, dass ein Ausdruck, der mit einem  $a$  beginnt, mit einem  $b$  endet und dazwischen einen korrekten Ausdruck des Typs  $S$  enthält, ebenfalls ein korrekter Ausdruck vom Typ  $S$  ist.
- Rekursive Regeln erlauben die tiefe Schachtelung von Strukturen, und sie ermöglichen, dass eine Regel Elemente in Beziehung setzt, die in der Kette beliebig weit voneinander entfernt sind.

# Kontextfreie Sprachen und endliche Automaten

Kontextfreie Sprachen sind eine echte Obermenge der Sprachen, die von endlichen Automaten definiert werden („reguläre Sprachen“):

- Es gibt kontextfreie Sprachen, die nicht regulär sind.
- Jede reguläre Sprache kann von einer CFG erzeugt werden.



Chomsky Hierarchie

Kontextfreie Grammatiken sind ein Standardformalismus zur Beschreibung der Grammatik **natürlicher Sprachen**



## Sind alle natürlichen Sprachen kontextfrei?



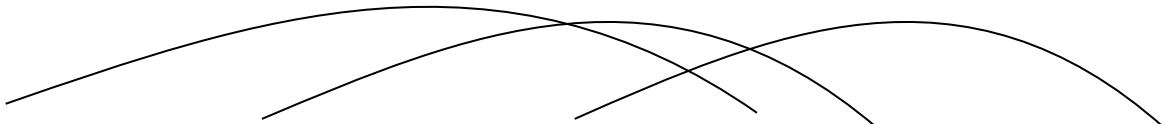
### Chomsky Hierarchie

Nein, es wurde gezeigt, dass einige natürliche Sprachen überkreuzende Abhängigkeiten haben und somit „mild kontextsensitiv“ sind (Schieber 1985).

Daher Verwendung auch von mächtigeren Grammatiken, e.g. „Baumadjunktionsgrammatik“ (Tree-Adjoining Grammar; Joshi 1985)

# Sprachen die nicht kontextfrei sind

- $L = \{a^n b^n c^n : n \geq 1\}$
- $L = \{a^m b^n x c^m d^n y\}$
- Schweizerdeutsch, Nebensatz



... mer d'chind em Hans es huus lönd hälfe aastriche.  
... *wir die Kinder dem Hans das Haus lassen helfen anstreichen.*  
... wir die Kinder dem Hans das Haus anstreichen helfen lassen.

# Was Sie nach dieser Vorlesung wissen sollten

- Eigenschaften syntaktischer Struktur:
  - Beliebig lange und tiefe Schachtelung
  - Variable Wortstellung
  - Abhängigkeit von Konstituenten voneinander
- Pumping Lemma
- Konstituentenstruktur vs. grammatische Funktionen
- Kontextfreie Grammatiken
  - Notation
  - X-bar Theorie
- Terminologie:
  - Konstituenten und Kategorien (NP, VP etc.)
  - Kontextfreie vs. kontextsensitive Sprachen