

# Einführung in die Computerlinguistik

## Syntax I

WS 2012/2013

Manfred Pinkal

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

## Anforderungen an Morphologiesysteme

- Korrektheit
- Abdeckung (engl. „coverage“)
- Effizienz

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

## Morphologiesysteme: Grundaufgaben

- Flexionsmorphologie: Lemmatisierung/Stemming
  - *veranstalt+et, Veranstaltung+en*
- Ableitungs-/Derivationsmorphologie
  - *Veranstalt+ung, un+glaubwürdig*
- Komposita-Zerlegung
  - *Fach+veranstaltung, glaub+würdig*

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

## Übergenerierung: Beispiele aus der Praxis

- Ein klassisches Beispiel aus der maschinellen Übersetzung (Systran, um 1980)
  - Barbarei
  - > nightclub nightclub egg
  - Bar|bar|ei
- Ein Beispiel aus der Rechtschreibkonversion (Corrigo, um 2000)
  - Hufeisenniere
  - > Hufeisennniere
  - Huf|ei|senn|niere

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

## Korrektheit

- Flexionsmorphologie: Typischerweise unproblematisch, Korrekt, wenn die Flexionsklassen im Lexikon korrekt angegeben sind.
- Kompositazerlegung:
  - **Übergenerierung** ist ein massives Problem
  - ... wenn sie nicht durch Zusatzmechanismen behoben wird (Blockierungslisten, statistische Gewichtung)
- Derivationsmorphologie:
  - Tendenziell Übergenerierung (Semiproduktivität)
  - Tendenziell semantisch irreführende Identifikation von Stämmen

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

## Abdeckung

- Aktuelle Morphologiesysteme haben eine gute bis sehr gute Abdeckung (s. z.B. Word-Rechtschreibung)
- Abdeckung und Korrektheit allein sind für sich genommen keine guten Bewertungskriterien:
  - Man kann Korrektheit billig auf Kosten der Abdeckung erreichen und umgekehrt.
  - Ziel: Zuverlässigkeit bei gleichzeitig großer Abdeckung

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

## Wortbildungsmorphologie: Ableitung/ Derivationsmorphologie

Derivationsmorphologie ist in verschiedener Hinsicht unsystematisch. Beispiele:

- die *Lesung* bezeichnet den Akt des Vorlesens,
- die *Singung* ist unmöglich
- die *Vorlesung* gibt es, bezeichnet aber nicht den Akt des Vorlesens,
- die *Schreibung* nicht den Akt des Schreibens

Viele Ableitungspräfixe und -suffixe sind semiproduktiv.

Viele Ableitungen sind semantisch "nicht transparent": Sie haben eine konventionelle, lexikalisierte Bedeutung, die mit der Bedeutung des Stammworts nicht in systematischer Beziehung steht.

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

## Effizienz

- Grundsätzlich: Morphologische Analyse benötigt lineare Zeit in Abhängigkeit von der Länge der Eingabe.
- Gute Morphologiesysteme liegen im mittleren  $\mu$ s-Bereich (pro Wortform).
- Durch Vorverarbeitung, Zwischenspeichern von Analysen, Indexierung etc. lässt sich für größere Dokumente die Zeit pro Textwort weiter erheblich drücken.
- Das ist
  - exzellent für Online- und Offline-Rechtschreibkorrektur
  - akzeptabel für begrenzte Datensammlungen (große Textkorpora, Firmen-Intranet etc.)
  - zu langsam fürs Web-Suchmaschinen (Alternative: Datenexpansion mit Wortformengenerierung)

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

## Morphologie und Syntax

- Gegenstand der **Morphologie** ist die **Struktur des Wortes**: der Aufbau von Wörtern aus Morphemen, den kleinsten funktionalen oder bedeutungstragenden Einheiten der Sprache.
- Gegenstand der **Syntax** ist die **Struktur des Satzes**: der Aufbau von Sätzen aus Wörtern.
- **Morphologie** beschreibt die **grammatischen Eigenschaften von Wörtern**, die durch Wortform oder Flexionsmorpheme kodiert werden.
- **Syntax** beschreibt die **Interaktion der grammatischen Eigenschaften** unterschiedlicher Wörter im Satz.

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

## Reale 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..."

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 [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 [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

## 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.

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

## 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 definieren die Sprache  $L \subseteq \Sigma^*$
- Können natürliche Sprachen mit endlichen Automaten beschrieben werden? 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 jede denkbare Sprache mit einem endlichen Automaten beschrieben werden?
- Die Antwort ist Nein. Es gibt sehr einfache Sprachen, die sich nicht mit endlichen Automaten beschreiben lassen.
- Ein Beispiel:  $a^n b^n$

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

## Beschränkungen endlicher Automaten

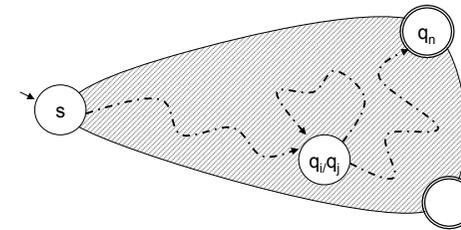
Um Zugehörigkeit zu  $a^n b^n$  zu erkennen, müsste sich der Automat beliebig lange Ketten von a's merken können, weil er die Information anschließend beim Abarbeiten von b's braucht.

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. Diese Eigenschaft heißt die „Markov-Eigenschaft“.

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

## Beschränkungen endlicher Automaten



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

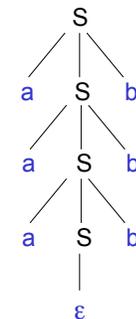
## Kontextfreie Grammatik: Ein neuer Formalismus

- Kontextfreie Grammatiken („KFG“, „CFG“) beschreiben Sprachen mithilfe von Ersetzungsregeln („rewrite rules“, Produktionen) der Form  $A \rightarrow w$ 
  - Beispiel:  $S \rightarrow aSb, S \rightarrow \epsilon$  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  $aa\epsilon bb = aabb$
- Eine solche Ersetzung ist ein zulässiger Ableitungsschritt. Wir schreiben:  $aaSbb \Rightarrow aaaSbbb$  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:  $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbb$
- Wir haben damit gezeigt, dass  $w$  durch die Regeln der Grammatik aus  $S$  ableitbar ist:  $w$  ein Wort der durch die Grammatik beschriebenen (erzeugten) Sprache  $L$ .

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

## Kontextfreie Grammatiken

- Die Ableitung  $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbb$  kann alternativ durch den **Ableitungsbaum** rechts dargestellt werden.



- Die **Wurzel** des Baumes ist das Startsymbol. Die **Blätter** des Baums ergeben, von links nach rechts gelesen und aneinandergelängt, das abgeleitete Wort.
- Alternative Schreibweise:  $[_s a[_s a[_s a[_s \epsilon ] b] b] b]$

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

## 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:

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

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

## $a^n b^n$ und die deutsche Satzstruktur

- 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 UdS Computerlinguistik

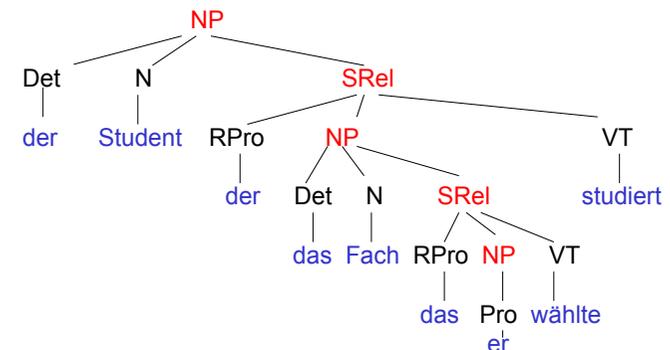
## Kontextfreie Grammatik: Definitionen

- 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**.

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

## Ein zweites Beispiel

- "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..."

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

## Eine erste kontextfreie Grammatik für deutsche Sätze

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

$V = \{S, SRel, NP, VI, VT, N, Det, RPro\} \cup \Sigma$   
 $\Sigma = \{schläft, arbeitet, studiert, wählte, Student, Fach, der, das, er\}$

$P =$

$S \rightarrow NP VI$	$NP \rightarrow Det N$
$S \rightarrow NP VT NP$	$NP \rightarrow Det N SRel$
$SRel \rightarrow RPro NP VT$	$NP \rightarrow Pro$
$SRel \rightarrow RPro VI$	

$VI \rightarrow schläft$	$N \rightarrow Student$
$VI \rightarrow arbeitet$	$N \rightarrow Fach$
$VT \rightarrow studiert$	$RPro \rightarrow der$
$VT \rightarrow wählte$	$RPro \rightarrow das$
$Det \rightarrow der$	$Det \rightarrow das$
$Pro \rightarrow er$	$Pro \rightarrow sie$

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

## Eine kontextfreie Grammatik für deutsche Sätze

Notationskonventionen:

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

Kompaktere Notation der Grammatik:

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

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

## 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. 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.

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

## Kontextfreie Sprachen und natürliche Sprachen

- Kontextfreie Grammatiken sind ein Standardformalismus zur Beschreibung der Grammatik **natürlicher 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“).

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

### Beispiel 1

CFG für einfache arithmetische Gleichungen:

$$\begin{aligned} S &\rightarrow \text{Term} = \text{Term} & \text{Term} &\rightarrow x \mid y \mid z \\ \text{Term} &\rightarrow ( \text{Term Op Term} ) & \text{Op} &\rightarrow + \mid - \mid * \mid : \\ \text{Term} &\rightarrow - \text{Term} \end{aligned}$$

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

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

## 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 generierten Satzes 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 UoS Computerlinguistik

### 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

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