Einführung in die Computerlinguistik

Syntax I

WS 2014/2015

Vera Demberg

Morphologiesysteme: Grundaufgaben

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

Morphologische Zerlegung: Ableitungen

Derivationsmorphologie ist in verschiedener Hinsicht unsystematisch. Beispiele:

- die *Lesung* bezeichnet den Akt des Vorlesens,
- die *Singung* ist 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.

Morphologische Zerlegung: Komposita

- Ein klassiches 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

Korrektheit und Abdeckung

- 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

- Flexionsmorphologie: Unproblematisch (wenn Lexikon mit Flexionsklassen verfügbar ist)
- Kompositazerlegung: Übergenerierung ist ein massives Problem
  - kann durch Zusatzmechanismen behoben werden (z.B. Blockierungslisten)
- Derivationsmorphologie: Neigung zur Übergenerierung (Semiproduktivität)
  - Korrekte Ableitungen werden normalerweise als Lemmata gelistet

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 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 wiederabliefert, 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..."

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 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.

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" (Σ)
  - Korrekte Sätze sind "Worte" über dem Alphabet
  - Die Menge der korrekten Sätze definiert 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 Sprachen, die sich nicht mit endlichen Automaten beschreiben lassen
- ... und zwar sehr einfache Sprachen.

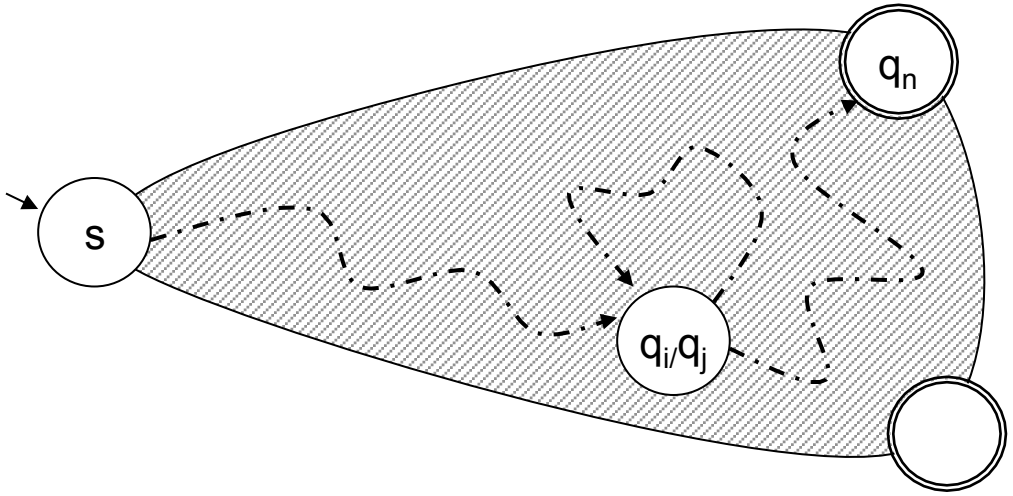
$a^n b^n$ und endliche Automaten

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 voraufgegangenen 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



"Pumping Lemma"

Kontextfreie Grammatik: Ein neuer Formalismus

- Kontextfreie Gramatiken („KFG“, „CFG“) beschreiben Sprachen mithilfe von Ersetzungsregeln („rewrite rules“, Produktionen) der Form A → w
  - Beispiel: S → aSb, S → ε beschreibt L= $a^n b^n$
- A → 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 aaεbb = aabb
- Eine solche Ersetzung ist ein zulässiger Ableitungsschritt. Wir schreiben: aaSbb ⇒ aaaSbbb bzw. aaSbb ⇒ 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 ⇒ aSb ⇒ aaSbb ⇒ aaaSbbb ⇒ 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.

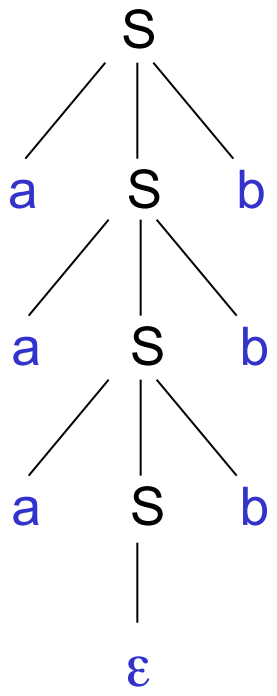
Kontextfreie Grammatiken

- Die Ableitung
  $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbb$
  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 = <V, Σ, P, S>, wobei

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

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

- G1 = < {a,b,S}, {a,b}, {<S, aSb>, <S, ε>}, S >

- Für <A, α> $\in$ P schreibt man üblicherweise $A \rightarrow \alpha$.

Kontextfreie Grammatik: Definitionen

- Wenn $A \rightarrow \alpha$ Produktion, w = uAv und w‘ = uαv, 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 Σ^*, die aus S ableitbar sind: $L(G) = \{w \in \Sigma^* | S \Rightarrow^* w\}$

- Sprachen, die durch kontextfreie Grammatiken erzeugt werden, heißen kontextfreie 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 wiederabliefert, errechnete Beurteilungspegel..."

Eine erste 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*}

P =

S → NP VI
S → NP VT NP
SRel → RPro NP VT
SRel → RPro VI

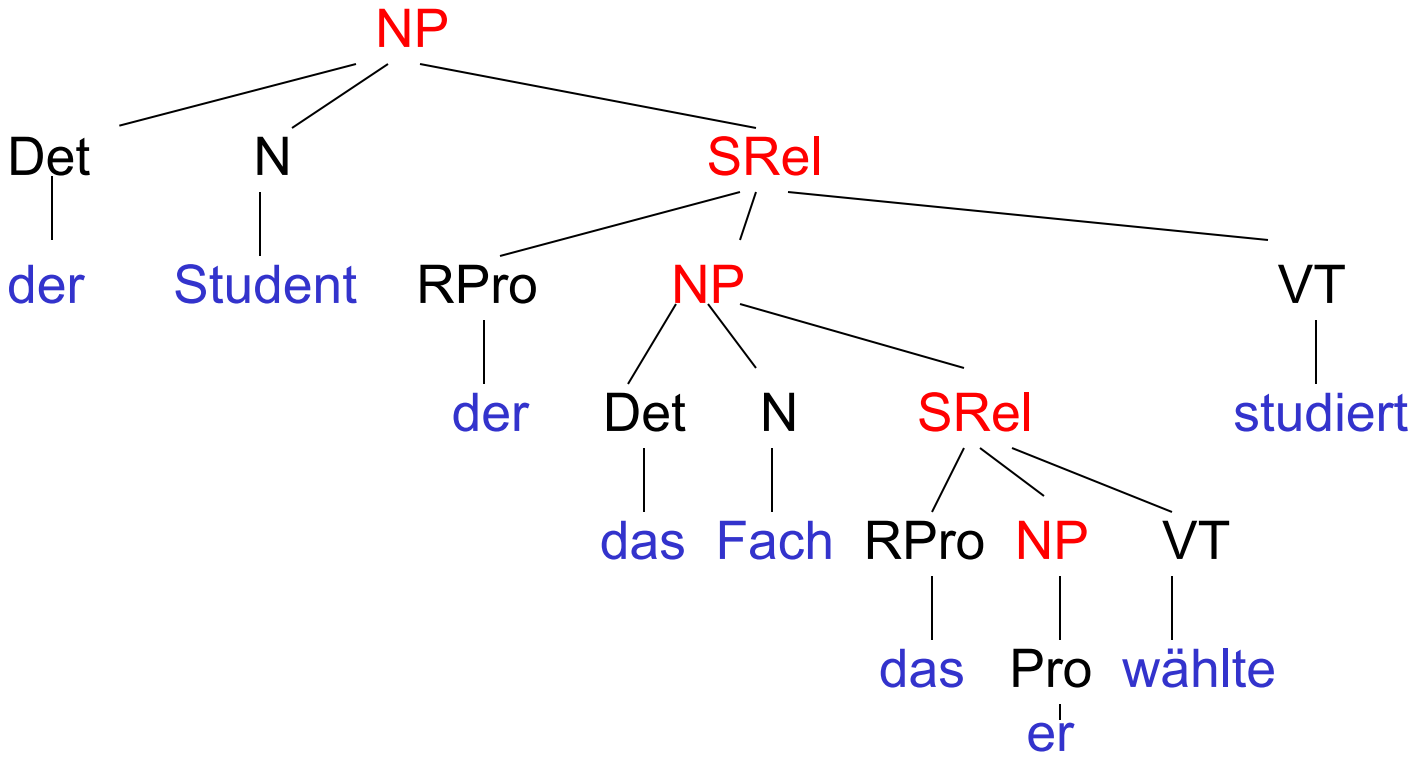
NP → Det N
NP → Det N SRel
NP → Pro

VI → *schläft*
VI → *arbeitet*
VT → *studiert*
VT → *wählte*
Det → *der*
Pro → *er*

N → *Student*
N → *Fach*
RPro → *der*
RPro → *das*
Det → *das*
Pro → *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*]]



Eine kontextfreie Grammatik für deutsche Sätze

Notationskonventionen:

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

Kompaktere Notation der Grammatik:

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*

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.

- 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 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“).