

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.

Eigenschaften der syntaktischen Struktur [1]

- 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.

Konstituenten

- 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.

Syntaktische Kategorien [1]

Konstituenten-Typen werden „syntaktische Kategorien“ genannt;
Beispiele:

- Nominalphrasen (Nominalausdrücke): *er – der Student – der interessierte Student – die Übungen – computerlinguistischen Fragestellungen*
- Präpositionalphrasen (Präpositionalausdrücke): *an computerlinguistischen Fragestellungen – im ersten Semester, – nach langer Überlegung*
- Adjektivphrasen: *interessierte – an computerlinguistischen Fragestellungen interessierte*
- Verben, Verbkomplex: *hat – gemacht – entschieden hat*
- Satz: Haupt- und Nebensätze unterschiedlicher Art

Syntaktische Kategorien [2]

Konstituenten /syntaktische Kategorien können beliebig ineinander verschachtelt sein:

- Der Nominalausdruck „*der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, der im Hauptfach, für das er sich nach langer Überlegung entschieden hat, Informatik studiert*“ enthält
- den (Relativ-)Satz „*der im Hauptfach, für das er sich nach langer Überlegung entschieden hat, Informatik studiert*“; der enthält
- den Nominalausdruck „*Hauptfach, für das er sich nach langer Überlegung entschieden hat*“; der enthält
- den (Relativ-)Satz „*für das er sich nach langer Überlegung entschieden hat*“; der enthält
- unter anderem den Nominalausdruck „*er*“.

Kategorie und Funktion

- Die syntaktische Kategorie ergibt sich aus dem internen Aufbau einer Konstituente, insbesondere aus der Wortart ihres „lexikalischen Kopfes“: Die Konstituenten
 - *der an computerlinguistischen Fragestellungen interessierte Student*
 - *an computerlinguistischen Fragestellungen interessiert*
 - *an computerlinguistischen Fragestellungen*sind Nominal-, Adjektiv- und Präpositionalphrase, weil der jeweilige Kopf Substantiv („Nomen“), Adjektiv, bzw. Präposition ist.
- Die grammatische Funktion dagegen bezeichnet die Rolle, die eine Konstituente im ganzen Satz spielt. Ein Nominalausdruck z.B. kann, je nach Stellung im Satz unter anderem die Funktion von Subjekt, (direktem oder indirektem) Objekt, (Genitiv-) Attribut oder Prädikatsnomen besitzen.

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.

Variable Wortstellung

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]

- 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.

Lässt sich syntaktische Struktur mit endlichen Automaten beschreiben?

- Überlegung: DEA, NEA und reguläre Ausdrücke sind äquivalente Formalismen, die komplizierte Sprachstrukturen beschreiben können und, wegen der garantierten Überführbarkeit in DEAs, in linearer Zeit analysieren. Es wäre wünschenswert, wenn wir das Instrumentarium auch auf syntaktische Strukturen anwenden könnten. Können wir nicht nur die Morphologie, sondern auch die Syntax natürlicher Sprachen durch endliche Automaten analysieren?
- Anders formuliert: Sind natürliche Sprachen reguläre Sprachen?
- Weitergehende Frage: Lassen sich alle Sprachen (über einem endlichen Alphabet) durch endliche Automaten beschreiben? Gibt es überhaupt Sprachen, die nicht regulär sind?

Es gibt nicht-reguläre Sprachen: Ein Existenzbeweis

Beweis:

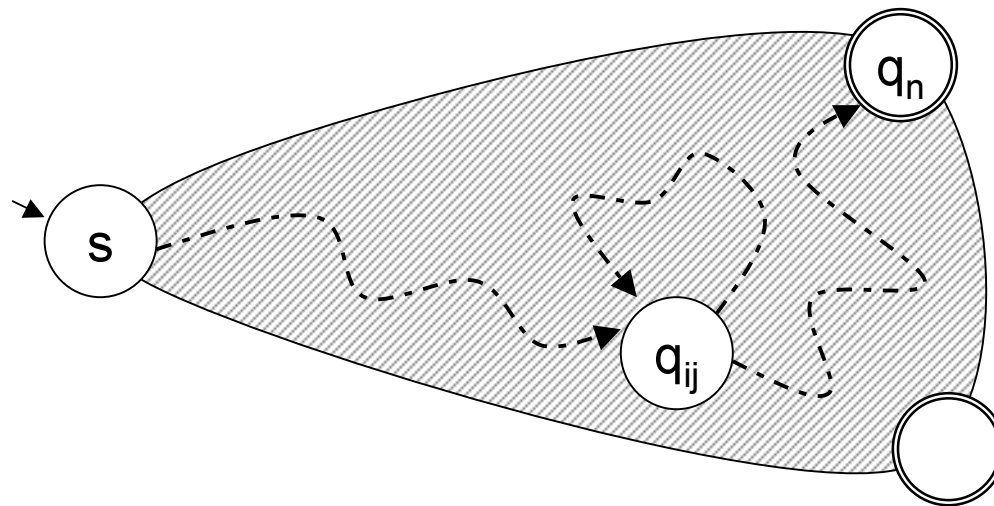
- Wenn Σ ein endliches Alphabet ist, ist Σ^* (die Menge aller Worte über Σ) abzählbar unendlich.
- Die Menge aller Sprachen über dem Alphabet ist $\wp(\Sigma^*)$. Die Potenzmenge einer abzählbar unendlichen Menge ist überabzählbar. Also gibt es zu einem endlichen Alphabet Σ eine überabzählbar unendliche Menge von Sprachen.
- Jede reguläre Sprache über Sprache über Σ wird durch einen regulären Ausdruck über Σ definiert. Reguläre Ausdrücke über Σ sind Worte über dem endlichen Alphabet $\Sigma \cup \{\emptyset, +, \cdot, *, (,)\}$. Es gibt nur abzählbar unendlich viele reguläre Ausdrücke über Σ , also nur abzählbar unendlich viele reguläre Sprachen.
- Also gibt es Sprachen über Σ , die nicht regulär sind.
- Aber was sind das für Sprachen?

Das "Pumping Lemma,, [1]

- Das Pumping Lemma (dt. „Pump-Lemma“ oder auch „uvw-Theorem“) gibt uns, anders als das vorher gezeigte Theorem, die Möglichkeit, konstruktiv zu beweisen, dass eine Sprache nicht regulär ist.
- Der Grundgedanke ist einfach: Jede reguläre Sprache L wird von einem endlichen Automaten akzeptiert. Ein endlicher Automat hat eine bestimmte, endliche Anzahl von Zuständen. Wenn ein Wort in L mehr Symbole hat als der Automat Zustände (genau genommen reichen schon mindestens so viele Symbole aus), dann muss beim Abarbeiten des Wortes ein Zustand mindestens zweimal vorkommen. Das bedeutet aber, dass beim Abarbeiten eine Schleife durchlaufen wird. Die könnte im Prinzip aber auch mehrfach durchlaufen (oder weggelassen) werden. Das heißt, für Wörter oberhalb einer bestimmten Länge durch beliebige Wiederholung eines Teilworts „aufgepumpt“ werden können und immer noch zur Sprache gehören.

Das "Pumping Lemma,, [2]

Sei $w = a_1 \dots a_n$ ein Wort, das vom DEA K mit k Zuständen erkannt wird, und $n = |w| \geq k$. Dann geht der Pfad vom Startzustand $q_0 = s$ zu einem Endzustand q_n , auf dem w gelesen wird, durch insgesamt $n+1$ Zustände. Das heißt, dass mindestens zwei Zustände q_i und q_j identisch sein müssen.



Das "Pumping Lemma,, [3]

Präzise Formulierung des Lemmas:

- Wenn L eine Sprache ist, die durch einen endlichen Automaten beschrieben werden kann ("reguläre Sprache"), dann gilt:
Wenn ein Wort $x \in L$ eine bestimmte Länge k erreicht oder überschreitet ($|x| \geq k$), dann läßt sich x so in drei Teile u , v und w zerlegen (mit $|v| \geq 1$), daß mit $x = uvw$ auch jedes $x' = uv^i w$ ($i=0$ oder $i>1$) Element von L ist.
- Um zu zeigen, dass eine Sprache L nicht regulär ist, genügt es, zu zeigen, dass L ausreichend lange Worte enthält, deren Teile nicht beliebig iterierbar sind.

Eine nicht reguläre Sprache [1]

Die Sprache $L = \{ a^n b^n \mid n \in \mathbb{N} \}$ (kurz: " $a^n b^n$ ") wird nicht von einem endlichen Automaten akzeptiert.

Beweis:

- Angenommen, die Sprache $L = \{ a^n b^n \mid n \in \mathbb{N} \}$ (kurz: " $a^n b^n$ ") wird von einem endlichen Automaten akzeptiert. Nach dem Pumping Lemma gibt es dann eine Zahl k , so daß für jedes Wort x mit $|x| \geq k$ eine Zerlegung in u , v und w möglich ist, so daß uw , $uvvw$, $uvvwv$, ... ebenfalls in L sind.
- Betrachten wir das Wort $a^k b^k$. Es gilt $|a^k b^k| \geq k$, das Wort muss also einen "duplizierbaren" Teil v besitzen. Um welchen Teil könnte es sich handeln?

Eine nicht reguläre Sprache [2]

- Drei Fälle sind denkbar:
 - Fall1: v liegt vollständig in der ersten Hälfte des Wortes, besteht also nur aus a 's. Dann müsste gelten, dass $uv^2w = a^{k+|v|}b^k \in L$: wegen $k+|v| \neq k$ unmöglich.
 - Fall2: v liegt vollständig in der zweiten Hälfte des Wortes, besteht also nur aus b 's. Dann müsste gelten, dass $uv^2w = a^kb^{k+|v|} \in L$: wegen $k+|v| \neq k$ unmöglich.
 - Fall 3: v überspannt die Mitte des Wortes, hat also die Form a^mb^m . Dann müsste gelten, dass $uv^2w = a^kb^ma^mb^k \in L$. Geht nicht, da a 's auf b 's folgen.
- Es gibt also für a^kb^k keine Zerlegung in uvw mit duplizierbarem Mittelteil. Also ist $L = a^n b^n$ nicht regulär.

Was steckt hinter dem Pumping Lemma?

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“: nach spätestens k gelesenen Symbolen ist er wieder in dem Zustand, den er vorher gehabt hat. Noch anders ausgedrückt: Der Automat kann nur bis k zählen. Um zum Beispiel $a^n b^n$ zu erkennen, müsste er sich aber beliebig lange Ketten von a 's merken können, weil er die Information anschließend beim Abarbeiten von b 's braucht.

Kontextfreie Grammatiken: Ein Beispiel

- $S \rightarrow aSb$, $S \rightarrow \varepsilon$ sind Produktionsregeln (Produktionen, Ersetzungsregeln)
- Beginne mit S (Startsymbol). Wenn in einer Kette das Symbol S vorkommt, ersetze es durch aSb (Regel 1) oder die leere Kette (Regel 2).
- Ersetze solange, bis nur noch a 's und b 's vorkommen („Terminalsymbole“)
- Eine gültige Ableitung der Beispielgrammatik:
 $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbbb$
- Die Beispielgrammatik erzeugt die Sprache $a^n b^n$

Kontextfreie Grammatik: Formale Definition

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

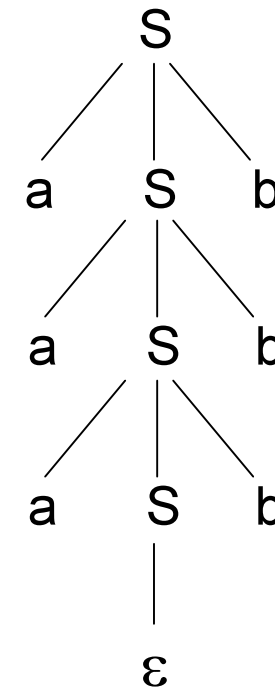
- V nicht-leere Menge von Symbolen
- $\Sigma \subseteq V$ nicht-leere Menge von Terminalsymbolen
- $P \subseteq (V - \Sigma) \times V^*$ nicht-leere Menge von Produktionsregeln
- $S \in V - \Sigma$ das Startsymbol

Die Beispielgrammatik in alternativer Notation:

- $G1 = \langle \{a,b,S\}, \{a,b\}, \{ \langle S, aSb \rangle, \langle S, \varepsilon \rangle \}, S \rangle$
- Für $\langle A, \alpha \rangle \in P$ schreibt man üblicherweise $A \rightarrow \alpha$.
- Ableitungsschritt: Wenn w ein nicht-terminales Symbol A enthält und $A \rightarrow \alpha$ Produktion ist, ersetze A in w durch α .
- Die durch G erzeugte Sprache ist die Menge aller Worte über Σ^* , die sich aus S ableiten lassen.

Kontextfreie Grammatiken: Ein Beispiel

- Das Wort aaabbb wird durch
 $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbb$
abgeleitet.
- Die Ableitung kann durch den folgenden
Ableitungsbaum dargestellt werden:
- Die Blätter des Baums ergeben, von links nach
rechts gelesen, das abgeleitete Wort.
- Eine weitere alternative Schreibweise:
 $[_S a[_S a[_S a[_S \varepsilon] b] b] b]$
Kontextfreie Grammatiken ermöglichen die
Darstellung beliebig tief geschachtelter
Strukturen.

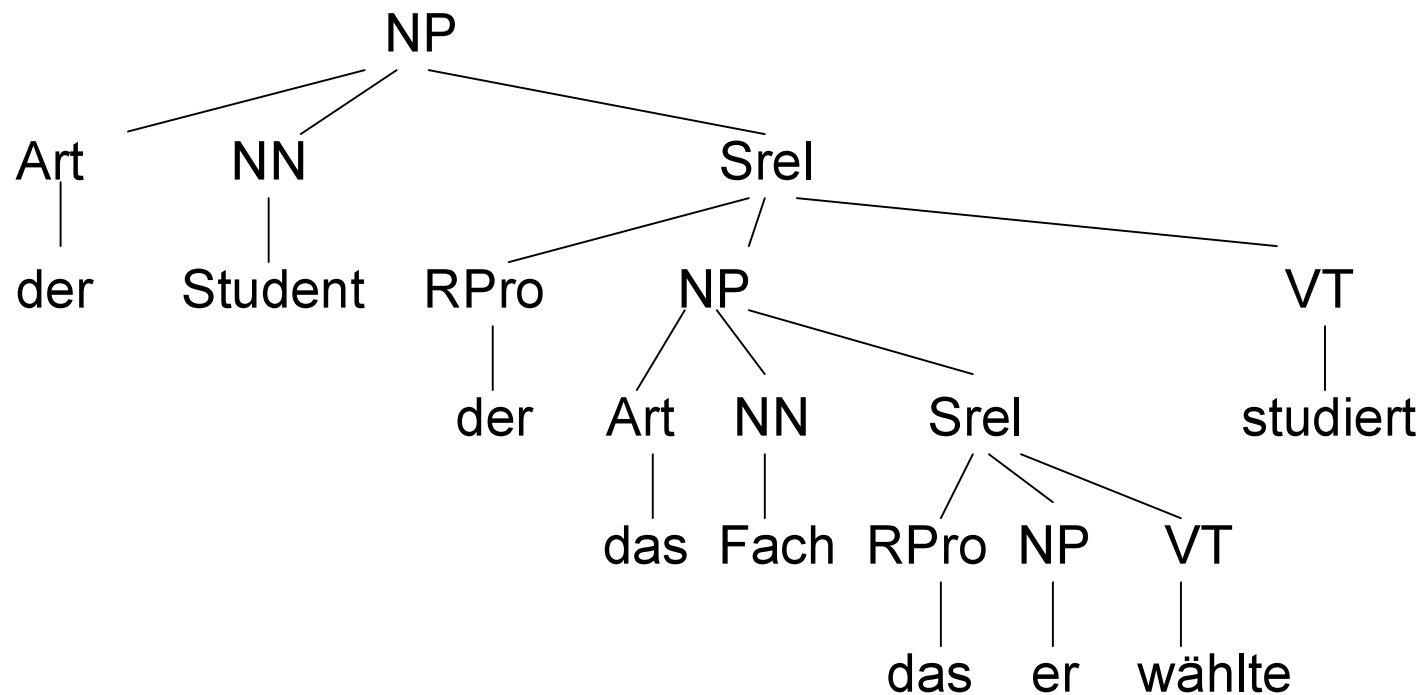


Geschachtelte Strukturen in natürlicher Sprache

- Der Nominalausdruck „*der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, der im Hauptfach, für das er sich nach langer Überlegung entschieden hat, Informatik studiert*“ enthält
- den (Relativ-)Satz „*der im Hauptfach, für das er sich nach langer Überlegung entschieden hat, Informatik studiert*“; der enthält
- den Nominalausdruck „*Hauptfach, für das er sich nach langer Überlegung entschieden hat*“; der enthält
- den (Relativ-)Satz „*für das er sich nach langer Überlegung entschieden hat*“; der enthält
- unter anderem den Nominalausdruck „*er*“.

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

$S \rightarrow NP VI$

$S \rightarrow NP VT NP$

$NP \rightarrow ART NN$

$VI \rightarrow \textit{schläft}$

$VI \rightarrow \textit{arbeitet}$

$VT \rightarrow \textit{studiert}$

$VT \rightarrow \textit{wählt}$

$ART \rightarrow \textit{der}$

$NP \rightarrow ART NN SREL$

$SREL \rightarrow RPRO VI$

$SREL \rightarrow RPRO NP VT$

$NN \rightarrow \textit{Student}$

$NN \rightarrow \textit{Fach}$

$RPro \rightarrow \textit{der}$

$RPro \rightarrow \textit{das}$

$ART \rightarrow \textit{das}$

- Kontextfreie Grammatiken heißen in der Linguistik auch Konstituentenstruktur-Grammatiken oder Phrasenstruktur-Grammatiken

Eine kontextfreie Grammatik für deutsche Sätze

Kompaktere Schreibweise:

- Optionale Konstituenten werden in Klammern geschrieben.
- Lexikalische Symbole werden mit Komma aneinandergehängt.

$S \rightarrow NP VI$ $NP \rightarrow ART NN (SREL)$

$SREL \rightarrow RPRO VI$ $S \rightarrow NP VT NP$

$SREL \rightarrow RPRO NP VT$

$VI \rightarrow \textit{schläft, arbeitet, wartet, ...}$

$VT \rightarrow \textit{wählt, studiert, liest, kennt, ...}$

$NN \rightarrow \textit{Student, Fach, Dozentin, Professor, Buch, ...}$

$RPro \rightarrow \textit{der, die, das}$

$ART \rightarrow \textit{der, die, das}$