

Grundlagen der Theoretischen Informatik

Till Mossakowski

Fakultät für Informatik
Otto-von-Guericke Universität
Magdeburg

Wintersemester 2014/15

Kontextfreie Grammatiken

Definition:

Eine Grammatik $G = (V, \Sigma, R, S)$ heißt **kontextfrei**, falls $R \subseteq V \times (V \cup \Sigma)^*$.

Auf der linken Seite einer Regel steht also stets genau ein Nichtterminal.

Definition:

Eine Sprache L heißt **kontextfrei**, falls es eine kontextfreie Grammatik G gibt, so dass $L = L(G)$.

Definition:

Ein **Kellerautomat** ist ein 6-Tupel $(K, \Sigma, \Gamma, \Delta, s, F)$, wobei gilt:

- K ist eine endliche Menge von **Zuständen**,
- Σ ist ein Alphabet, das **Eingabealphabet**,
- Γ ist ein Alphabet, das **Kelleralphabet**,
- $s \in K$ ist der **Startzustand**,
- $\Delta \subseteq (K \times (\Sigma \cup \{\epsilon\}) \times \Gamma^*) \times (K \times \Gamma^*)$ ist die **Übergangsrelation**,
- Δ ist endlich und
- $F \subseteq K$ ist die Menge der **Endzustände**.

Ein Kellerautomat ist also nach Definition nichtdeterministisch.

Kellerautomaten und kontextfreie Sprachen

Satz: [Kleene]

Die Klasse der kontextfreien Sprachen ist genau die Klasse der von Kellerautomaten akzeptierten Sprachen.

Abschlusseigenschaften

Satz:

Die Klasse der kontextfreien Sprachen ist abgeschlossen unter

- (a) Vereinigung,*
- (b) Konkatenation und*
- (c) Kleene Star.*

Es genügt, die Beweise für kontextfreie Grammatiken zu führen.
Wir skizzieren lediglich die Beweisideen.

(a) Vereinigung

Seien $G_1 = (V_1, \Sigma_1, R_1, S_1)$ und $G_2 = (V_2, \Sigma_2, R_2, S_2)$ kontextfreie Grammatiken und o.B.d.A. gelte $V_1 \cap V_2 = \emptyset$. Ferner sei $S \notin V_1 \cup V_2$.

Sei nun $G = (V_1 \cup V_2 \cup \{S\}, \Sigma_1 \cup \Sigma_2, R, S)$ mit

$$R = R_1 \cup R_2 \cup \{S \rightarrow S_1, S \rightarrow S_2\}$$

Dann gilt $L(G) = L(G_1) \cup L(G_2)$

(b) Konkatenation

Seien $G_1 = (V_1, \Sigma_1, R_1, S_1)$ und $G_2 = (V_2, \Sigma_2, R_2, S_2)$ kontextfreie Grammatiken und o.B.d.A. gelte $V_1 \cap V_2 = \emptyset$. Ferner sei $S \notin V_1 \cup V_2$.

Sei nun $G = (V_1 \cup V_2 \cup \{S\}, \Sigma_1 \cup \Sigma_2, R, S)$ mit

$$R = R_1 \cup R_2 \cup \{S \rightarrow S_1 S_2\}$$

Dann gilt

$$L(G) = L(G_1)L(G_2)$$

(c) Kleene Star

Sei $G_1 = (V_1, \Sigma_1, R_1, S_1)$ eine kontextfreie Grammatik. Ferner sei $S \notin V_1$.

Sei $G = (V_1 \cup \{S\}, \Sigma_1, R, S)$ mit

$$R = R_1 \cup \{S \rightarrow \varepsilon, S \rightarrow SS_1\}$$

Dann gilt $L(G) = L(G_1)^*$

Satz:

Der Schnitt einer kontextfreien Sprache mit einer regulären Sprache ist eine kontextfreie Sprache.

Beweisskizze: Sei $M_1 = (K_1, \Sigma, \Gamma, \Delta_1, s_1, F_1)$ ein Kellerautomat und sei $M_2 = (K_2, \Sigma, \delta, s_2, F_2)$ ein deterministischer endlicher Automat.

Wir simulieren M_1 und M_2 parallel mit dem Kellerautomat $M = (K_1 \times K_2, \Sigma, \Gamma, \Delta, (s_1, s_2), F_1 \times F_2)$ mit den Übergängen

$$(((q_1, q_2), a, \beta), ((p_1, \delta(q_2, a)), \gamma)) \in \Delta$$

für jedes $((q_1, a, \beta), (p_1, \gamma)) \in \Delta_1$ und jedes $q_2 \in K_2$ sowie

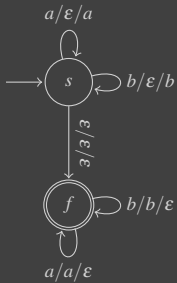
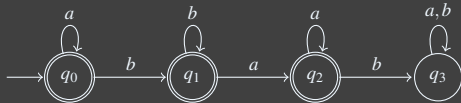
$$(((q_1, q_2), \varepsilon, \beta), ((p_1, q_2), \gamma)) \in \Delta$$

für jedes $((q_1, \varepsilon, \beta), (p_1, \gamma)) \in \Delta_1$ und jedes $q_2 \in K_2$.

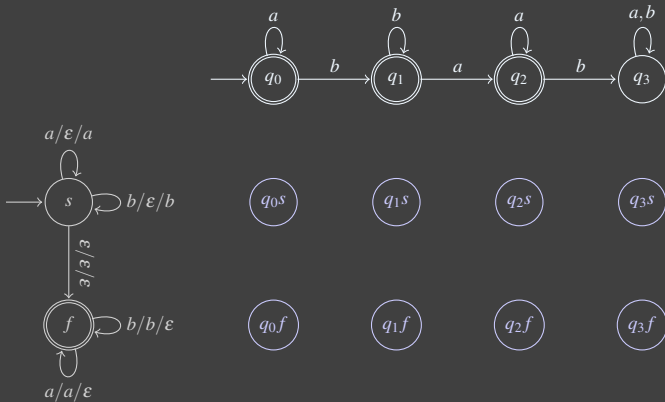
Dann gilt $L(M) = L(M_1) \cap L(M_2)$



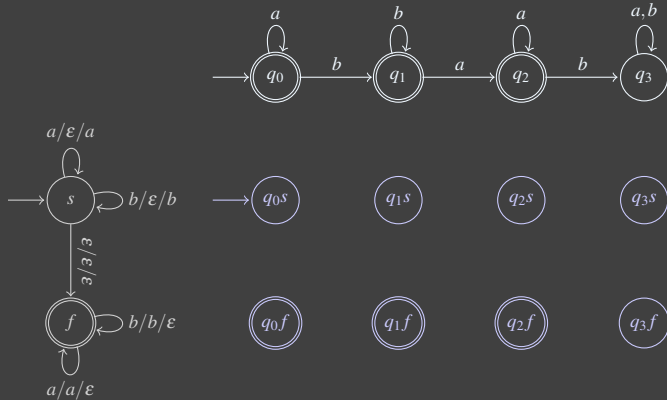
Beispiel:



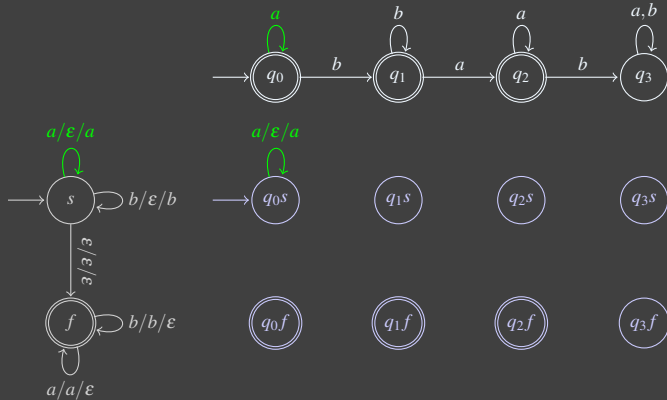
Beispiel:



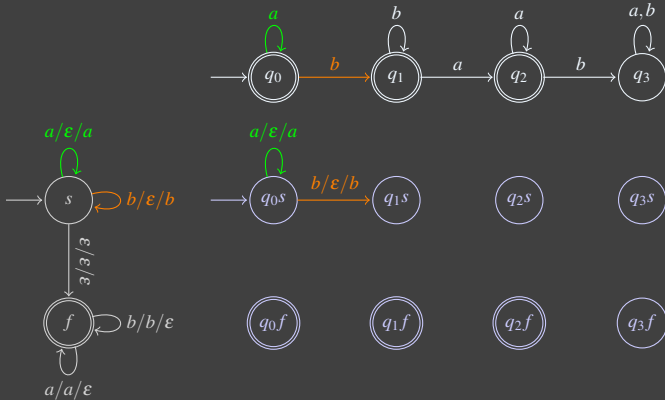
Beispiel:



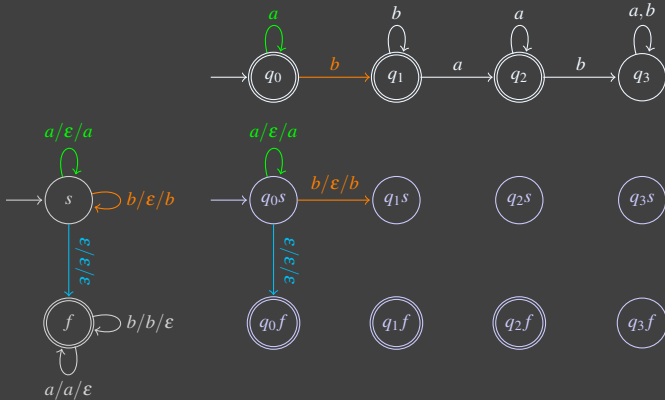
Beispiel:



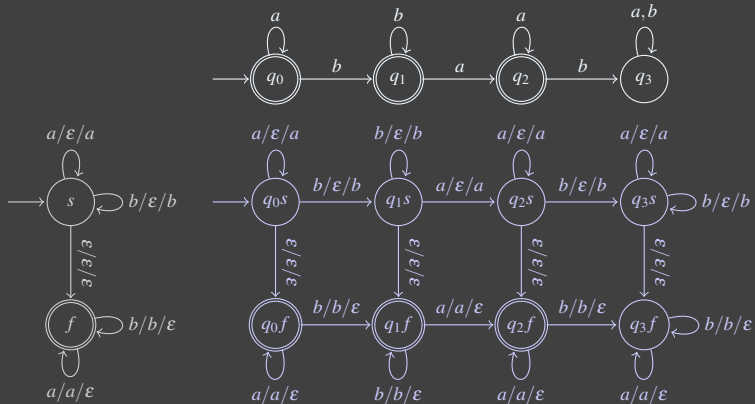
Beispiel:



Beispiel:



Beispiel:



$$L(M) = \{uu^R \mid u \in \{a,b\}^*\} \cap \{w \in \{a,b\}^* \mid bab \text{ ist kein Teilwort von } w\}$$

Kontextfreie und nichtkontextfreie Sprachen

Sei $G = (V, \Sigma, R, S)$ eine kontextfreie Grammatik.

Mit $\phi(G)$ bezeichnen wir die maximale Anzahl von Symbolen auf der rechten Seite einer Produktionsregel in R .

Die *Höhe eines Syntaxbaumes* ist die Länge des längsten Pfades von einem Blatt zur Wurzel des Baumes.

Satz:

Sei G eine kontextfreie Grammatik und T ein Syntaxbaum von G . Sei h die Höhe von T . Dann hat die Beschriftung von T Länge höchstens $\phi(G)^h$.

Pumping Lemma für kontextfreie Sprachen

uvwxy-Theorem:

Satz (Pumping Lemma):

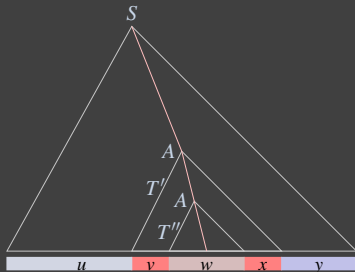
Sei L eine kontextfreie Sprache. Dann gibt es eine Zahl n , so dass sich alle Wörter $z \in L$ mit $|z| \geq n$ in $z = uvwxy$ zerlegen lassen, so dass

- (1) $vx \neq \varepsilon$
- (2) $|vwx| \leq n$
- (3) $uv^iwx^iy \in L$ für alle $i \geq 0$

Beweis: Sei $G = (V, \Sigma, R, S)$ eine kontextfreie Grammatik mit $L = L(G)$ und sei $n = \phi(G)^{|V|+1}$.

Sei nun $z \in L$ mit $|z| \geq n$ und T ein Syntaxbaum für z mit minimal vielen Blättern.

In T muss es einen Pfad der Länge mindestens $|V| + 1$ geben, auf dem nach dem Schubfachprinzip mindestens ein Nichtterminal mehrfach vorkommt. Sei A ein Nichtterminal, das unter den $|V| + 1$ untersten Nichtterminalen auf diesem Pfad mehrfach vorkommt.



Sei w die Beschriftung des Teilbaums T'' , dessen Wurzel mit dem untersten Vorkommen von A beschriftet ist, und sei vwx die Beschriftung des Teilbaums T' , dessen Wurzel mit dem zweituntersten Vorkommen von A beschriftet ist. Da wir einen Syntaxbaum für z mit minimal vielen Blättern gewählt haben, muss $vx \neq \varepsilon$ gelten.

Da wir A als ein Nichtterminal gewählt haben, das unter den $|V| + 1$ untersten Nichtterminalen auf dem langen Pfad mehrfach vorkommt, hat T' Höhe höchstens $|V| + 1$.

Somit hat vwx Länge höchstens $\phi(G)^{|V|+1} = n$. ■

Pumping Lemma:

für alle $L \in CF$
 gilt, es gibt $n \geq 1$
 so dass für alle $z \in L, |z| \geq n$
 gilt, es gibt $u, v, w, x, y, vx \neq \varepsilon, |vwx| \leq n, z = uvwxy$
 so dass für alle $i \geq 0$
 gilt

$$uv^iwx^iy \in L$$

Pumping Lemma:

$$\forall L \in CF$$

$$\exists n \geq 1$$

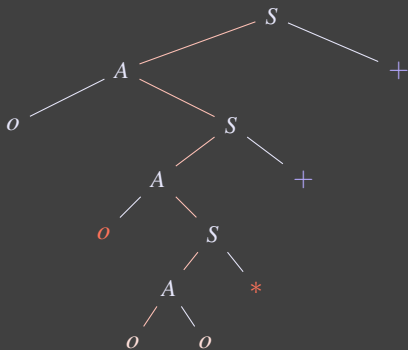
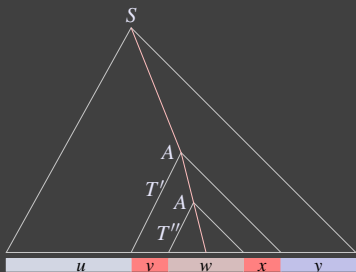
$$\forall z \in L, |z| \geq n$$

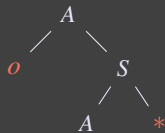
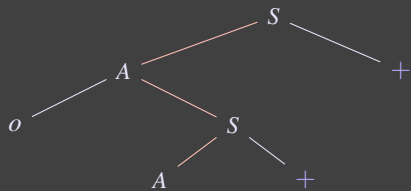
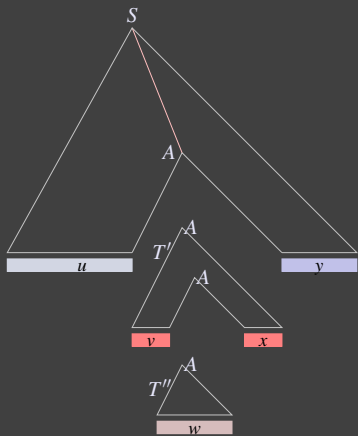
$$\exists u, v, w, x, y, vx \neq \varepsilon, |vwx| \leq n, z = uvwxy$$

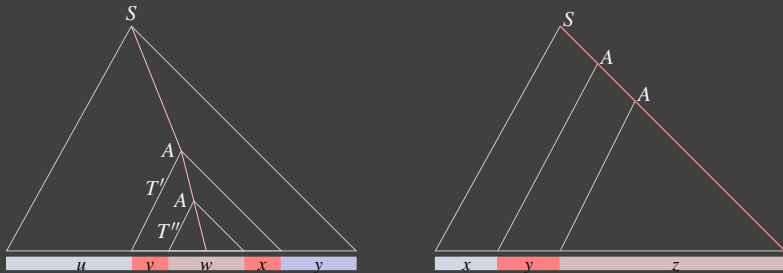
$$\forall i \geq 0$$

$$uv^iwx^iy \in L$$

Beispiel: $S \rightarrow A+ \mid A* \mid o$, $A \rightarrow SS \mid oS \mid oo$







Auch das Pumping Lemma für reguläre Sprachen lässt sich mit Hilfe von Syntaxbäumen unter Ausnutzung der speziellen Struktur der Syntaxbäume bei rechtslinearen Grammatiken beweisen.

Beispiele:

$L_1 = \{a^k b^k c^k \mid k \geq 0\}$ ist nicht kontextfrei:

Angenommen, L_1 wäre kontextfrei. Dann gäbe es ein n , so dass sich alle Wörter der Länge mindestens n wie im Pumping Lemma angeben zerlegen ließen.

Betrachte $z = a^n b^n c^n$. Dieses Wort müsste sich also in $uvwxy$ zerlegen lassen mit $vx \neq \varepsilon$.

Falls v mindestens zwei verschiedene Symbole enthielte, so enthielte uv^2wx^2y Symbole in falscher Reihenfolge. Analog für x .

Also wäre $v \in \mathcal{L}(\sigma_v^*)$ und $x \in \mathcal{L}(\sigma_x^*)$ für $\sigma_v, \sigma_x \in \{a, b, c\}$. Es gäbe also ein $\sigma \in \{a, b, c\} - \{\sigma_v, \sigma_x\}$. Das Wort uv^2wx^2y enthielte dann zuwenig Vorkommen von σ . **Widerspruch!**

$L_2 = \{w \in \{a,b,c\}^* \mid w \text{ enthält gleichviele } a, b \text{ und } c\}$ ist nicht kontextfrei:

Angenommen, L_2 wäre kontextfrei. Dann wäre auch der Schnitt von L_2 mit der regulären Sprache $\mathcal{L}(a^*b^*c^*)$ kontextfrei. Die Schnittmenge ist aber gerade die nicht kontextfreie Menge $L_1 = \{a^n b^n c^n \mid n \geq 0\}$. **Widerspruch!**

$L_3 = \{w cw \mid w \in \{a, b\}^*\}$ ist nicht kontextfrei:

Angenommen, L_3 wäre kontextfrei. Dann gäbe es ein n , so dass sich alle Wörter der Länge mindestens n wie im Pumping Lemma angegeben zerlegen ließen.

Betrachte $a^n b^n c a^n b^n$. Dieses Wort müsste sich also in $uvwxy$ zerlegen lassen mit $vx \neq \varepsilon$ und $|vwx| \leq n$.

vwx könnte weder vollständig links von c noch vollständig rechts von c enthalten sein, weil sonst der Teil links von c bzw. rechts von c beim Aufpumpen länger würde.

Also müsste vwx das c enthalten, genauer gesagt, w müsste c enthalten, weil sonst beim Aufpumpen Wörter entstünden, die mehr als ein c enthielten.

Da $|vwx| \leq n$ sein müsste, könnte vwx weder den Block von a Symbolen links von c noch den Block von b Symbolen rechts von c überlappen. vwx würde also links von c nur bs enthalten und rechts von c nur as . Nach Aufpumpen würden also die Wortteile links von c und rechts von c jeweils verschieden viele a und b enthalten.

Widerspruch!

Abschlusseigenschaften


Satz:

Die Klasse der kontextfreien Sprachen ist unter Schnitt nicht abgeschlossen.

Beweis: Die Sprachen

$$L_3 = \{a^n b^n c^m \mid n, m \geq 0\}$$

$$L_4 = \{a^m b^n c^n \mid n, m \geq 0\}$$

sind beide kontextfrei, aber $L_3 \cap L_4 = \{a^n b^n c^n \mid n \geq 0\}$ ist nicht kontextfrei. 

Satz:

Die Klasse der kontextfreien Sprachen ist nicht abgeschlossen unter Komplementbildung.

Beweis:

Seien L und L' kontextfreie Sprachen über den Alphabeten Γ_1 bzw. Γ_2 . Dann sind L und L' Sprachen über $\Sigma = \Gamma_1 \cup \Gamma_2$. Wegen

$$L \cap L' = \overline{\overline{L} \cup \overline{L'}} = \Sigma^* - ((\Sigma^* - L) \cup (\Sigma^* - L'))$$

ist die Klasse der kontextfreien Sprachen auch unter Komplementbildung nicht abgeschlossen. ■

Satz:

Sei $h : \Sigma^* \rightarrow \Gamma^*$ ein Homomorphismus. Falls $L \subseteq \Sigma^*$ eine kontextfreie Sprache ist, dann ist auch $h(L) = \{h(w) \mid w \in L\}$ eine kontextfreie Sprache.

Satz:

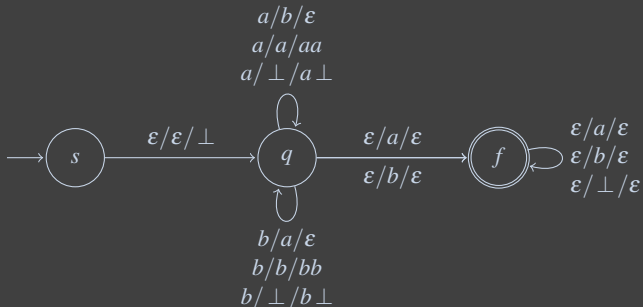
Sei $h : \Sigma^* \rightarrow \Gamma^*$ ein Homomorphismus. Falls $L \subseteq \Gamma^*$ eine kontextfreie Sprache ist, dann ist auch $h^{-1}(L) = \{w \mid h(w) \in L\}$ eine kontextfreie Sprache.

Satz:

Falls L eine kontextfreie Sprache ist, dann ist auch $L^R = \{w \mid w^R \in L\}$ eine kontextfreie Sprache.

$L_4 = \{w \in \{a,b,c\}^* \mid w \text{ enthält nicht gleichviele } a, b \text{ und } c\}$ ist kontextfrei:

$L_{a \neq b} = \{w \in \{a,b,c\}^* \mid |w|_a \neq |w|_b\}$ ist kontextfrei:



Schließlich gilt $L_4 = L_{a \neq b} \cup L_{a \neq c} \cup L_{b \neq c}$.

Satz von Parikh: *Jede kontextfreie Sprache über einem einelementigen Alphabet ist regulär.*

$L_5 = \{w \in \{a,b\}^* \mid |w| \text{ ist keine Primzahl}\}$ ist nicht kontextfrei:

Angenommen, L_5 wäre kontextfrei. Dann wäre auch $h(L_5)$ kontextfrei für den Homomorphismus h mit $h(a) = a$ und $h(b) = a$. Nun wäre $h(L_5)$ nach dem Satz von Parikh sogar regulär.

Da reguläre Sprachen unter Komplement abgeschlossen sind, wäre auch $\overline{h(L_5)}$ regulär. Es gilt jedoch

$$\overline{h(L_5)} = \{a^p \mid p \text{ ist eine Primzahl}\}$$

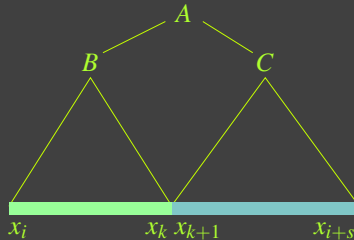
Also wäre $\{a^p \mid p \text{ ist eine Primzahl}\}$ regulär. **Widerspruch!**

CYK-Algorithmus

Wortproblem für kontextfreie Sprachen: Sei $L \subseteq \Sigma^*$ eine kontextfreie Sprache und sei $w = x_1 \dots x_n \in \Sigma^*$, gilt $w \in L$?

Der folgende von Cocke, Kasami und Younger unabhängig voneinander entworfene Algorithmus löst das Wortproblem für kontextfreie Sprachen, die durch eine kontextfreie Grammatik in Chomsky Normalform gegeben sind.

Sei $G = (V, \Sigma, R, S)$ eine kontextfreie Grammatik in Chomsky Normalform und sei $w = x_1 \dots x_n \in \Sigma^*$.



Für $1 \leq i \leq j \leq n$ sei $N[i, j]$ die Menge aller Symbole aus V , aus denen das Teilwort $x_i \dots x_j$ abgeleitet werden kann.

CYK($G = (V, \Sigma, R, S), w = x_1 \dots x_n$)

```

1  for  $i \leftarrow 1$  to  $n$ 
2      do  $N[i, i] \leftarrow \{A \in V \mid \text{es gibt } A \rightarrow x_i \text{ in } R\}$ 
3      for  $j \leftarrow i+1$  to  $n$ 
4          do  $N[i, j] \leftarrow \emptyset$ 
5  for  $s \leftarrow 1$  to  $n-1$ 
6      do for  $i \leftarrow 1$  to  $n-s$ 
7          do for  $k \leftarrow i$  to  $i+s-1$ 
8              do if ( es gibt  $A \rightarrow BC$  in  $R$  mit
9                       $B \in N[i, k]$  und  $C \in N[k+1, i+s]$  )
10                 then füge  $A$  zu  $N[i, i+s]$  hinzu
11  if (  $S \in N[1, n]$  )
12      then return " $w \in L(G)$ "
13      else return " $w \notin L(G)$ "

```

Lemma:

Nach s Iterationen, $0 \leq s \leq n$, von $\text{CYK}((V, \Sigma, R, S), x_1 \dots x_n)$ gilt für alle $i = 1, \dots, n - s$,

$$N[i, i + s] = \{A \in V \mid A \Rightarrow_G^* x_i \dots x_{i+s}\}$$

Beweisskizze: Induktion über s . ■

Aus dem Lemma folgt, dass $x \in L(G)$ genau dann wenn $S \in N[1, n]$.

Kontextsensitive Sprachen

Definition:

Eine Grammatik $G = (V, \Sigma, R, S)$ heißt **monoton**, falls für alle Regeln $w_1 \rightarrow w_2$ in R gilt $|w_1| \leq |w_2|$.

Definition:

Eine Grammatik $G = (V, \Sigma, R, S)$ heißt **kontextsensitiv**, falls alle Regeln in R von der Form $\alpha A \beta \rightarrow \alpha \gamma \beta$ mit $A \in V$, $\alpha, \beta \in (V \cup \Sigma)^*$ und $\gamma \in (V \cup \Sigma)^+$ sind.

Kontextsensitive Grammatiken sind also monotone Grammatiken.

Definition:

Eine Sprache L heißt **kontextsensitiv**, falls es eine kontextsensitive Grammatik G gibt, so dass $L - \{\varepsilon\} = L(G)$.

Satz:

Zu jeder monotonen Grammatik gibt es eine äquivalente kontextsensitive Grammatik.

Beispiel:

Die monotone Grammatik $G = (\{S, B\}, \Sigma, R, S)$ mit

$$S \rightarrow aSBc$$

$$S \rightarrow aBC$$

$$cB \rightarrow Bc$$

$$aB \rightarrow ab$$

$$bB \rightarrow bb$$

erzeugt $L(G) = \{a^n b^n c^n \mid n \geq 1\}$.

Beispiel:

Die kontextsensitive Grammatik $G = (\{S, B, C, H\}, \Sigma, R, S)$ mit

$$\begin{array}{ll}
 S & \rightarrow aSBC & S & \rightarrow aBC \\
 CB & \rightarrow HB & HB & \rightarrow HC \\
 HC & \rightarrow BC & aB & \rightarrow ab \\
 bB & \rightarrow bb & bC & \rightarrow bc \\
 cC & \rightarrow cc & &
 \end{array}$$

erzeugt ebenfalls $L(G) = \{a^n b^n c^n \mid n \geq 1\}$.