Automaten und Formale Sprachen
13. Vorlesung

Prof. Dr. Dietrich Kuske

FG Automaten und Logik, TU Ilmenau

Wintersemester 2023/24

Abschlußeigenschaften

Erinnerung

Die Klasse der regulären Sprachen ist abgeschlossen unter:

- Vereinigung (L_1, L_2 regulär $\Rightarrow L_1 \cup L_2$ regulär)
- Schnitt (L_1, L_2 regulär $\Rightarrow L_1 \cap L_2$ regulär)
- Komplement (L regulär $\Rightarrow \Sigma^* \setminus L$ regulär)
- Produkt/Konkatenation (L_1, L_2 regulär $\Rightarrow L_1 L_2$ regulär)
- Stern-Operation (L regulär $\Rightarrow L^*$ regulär)

Frage

Welche dieser Eigenschaften gelten für die Klasse der kontextfreien Sprachen?

Satz

Die Klasse der kontextfreien Sprachen ist abgeschlossen unter:

1. Vereinigung (L_1, L_2 kontextfrei $\Rightarrow L_1 \cup L_2$ kontextfrei)
2. Produkt/Konkatenation (L_1, L_2 kontextfrei $\Rightarrow L_1 L_2$ kontextfrei)
3. Stern-Operation (L kontextfrei $\Rightarrow L^*$ kontextfrei)

Die Klasse der kontextfreien Sprachen ist nicht abgeschlossen unter:

4. Schnitt (es gibt L_1, L_2 kontextfrei, so daß $L_1 \cap L_2$ nicht kontextfrei ist)
5. Komplement (es gibt L kontextfrei, so daß $\Sigma^* \setminus L$ nicht kontextfrei ist)

Beweis:
(1)-(3): siehe Übungsaufgabe 6(3)

Bisher haben wir keine Methode, um zu zeigen, daß eine Sprache <u>nicht</u> kontextfrei ist. - Der Beweis wird auf Folie 13.28 vervollständigt werden.

Das Pumping-Lemma für kontextfreie Sprachen

Idee: Man versucht auszunutzen, daß eine kontextfreie Sprache von einer Grammatik mit endlich vielen Nichtterminalen erzeugt werden muß. Das bedeutet auch: Wenn ein Ableitungsbaum ausreichend tief ist, so gibt es einen Ast, der ein Nichtterminal mehrfach enthält. Die durch diese zwei Vorkommen bestimmten Teilbäume werden wir „pumpen“.

Pumping Lemma (Bar-Hillel, Perles, Shamir ’61)

Wenn L eine **kontextfreie** Sprache ist,
dann gibt es $n \geq 1$ derart,
daß für alle $z \in L$ mit $|z| \geq n$ gilt:
es gibt Wörter $u, v, w, x, y \in \Sigma^*$ mit

(i) $z = uvwxy$,

(ii) $|vwx| \leq n$,

(iii) $|vx| \geq 1$ und

(iv) $uv^iwx^iy \in L$ für alle $i \geq 0$.

Dieses Lemma spricht nicht über kontextfreie Grammatiken, sondern nur über die Eigenschaften der Sprache. Daher ist es dazu geeignet, Aussagen über Nicht-Kontextfreiheit zu machen. Wir zeigen zunächst an einem Beispiel, wie dies funktioniert:

Beispiel (Fortsetzung von Folie 2.8 bzw. 2.11)

Betrachte die kontextsensitive Grammatik G mit den Produktionen

$$
\begin{array}{rclrclrcl}
S & \to & aSBC \mid aBC & CB & \to & HB & HB & \to & HC \\
HC & \to & BC & aB & \to & ab & bB & \to & bb \\
bC & \to & bc & cC & \to & cc\,. & & &
\end{array}
$$

Sie erzeugt z.B. das Wort $aabbcc$:

$$
\begin{aligned}
\underline{S} &\Rightarrow a\underline{S}BC \Rightarrow aaB\underline{CB}C \Rightarrow aaB\underline{HB}C \\
&\Rightarrow aaB\underline{HC}C \Rightarrow a\underline{aB}BCC \Rightarrow aa\underline{bB}CC \\
&\Rightarrow aab\underline{bC}C \Rightarrow aabb\underline{cC} \Rightarrow aabbcc
\end{aligned}
$$

Die erzeugte Sprache ist $L_1 = \{a^m b^m c^m \mid m \geq 1\}$.

Diese kontextsensitive Sprache ist nicht kontextfrei. Also gilt $\mathcal{L}_2 \subsetneq \mathcal{L}_1$.

Beweis: indirekt
Angenommen, L_1 wäre kontextfrei.

Nach dem Pumping-Lemma gibt es ein $n \geq 1$, so daß die folgende Aussage gilt:

> Für jedes $z \in L_1$, $|z| \geq n$, gibt es $u, v, w, x, y \in \Sigma^*$ mit (i), (ii), (iii) und (iv). (*)

Wir wählen nun $z = a^n b^n c^n$.
Dann ist $z \in L_1$ und $|z| = 3n > n$.
Nach der Aussage (*) gibt es also $u, v, w, x, y \in \Sigma^*$ mit

(i) $z = uvwxy$,
(ii) $|vwx| \leq n$,
(iii) $|vx| \geq 1$ und
(iv) $uv^i wx^i y \in L_1$ für alle $i \geq 0$.

$$uvwxy \overset{(i)}{=} z = \underbrace{aaaaa \cdots aaaa}_{n-\text{mal}}\underbrace{bbbbb \cdots bbbb}_{n-\text{mal}}\underbrace{ccccc \cdots cccc}_{n-\text{mal}}$$

Wegen $|vwx| \leq n$ nach (ii) gilt $vwx \in L(a^* b^*) \cup L(b^* c^*)$ und damit $vx \in L(a^* b^*) \cup L(b^* c^*)$.

Mit anderen Worten: es ist nicht möglich, daß sowohl ein a als auch ein c in vx vorkommen.
Wir unterscheiden zwei Fälle:

- c kommt in vx nicht vor: dann enthält uv^0wx^0y (immer noch) n Vorkommen von c, aber

$$|uv^0wx^0y|_a + |uv^0wx^0y|_b = 2n - |vx| < 2n$$

  wegen (iii) und damit $uv^0wx^0y \notin L_1$, im Widerspruch zu (iv).
- a kommt in vx nicht vor: man erhält analog einen Widerspruch zu (iv).

Also ist L_1 tatsächlich nicht kontextfrei. $\square$

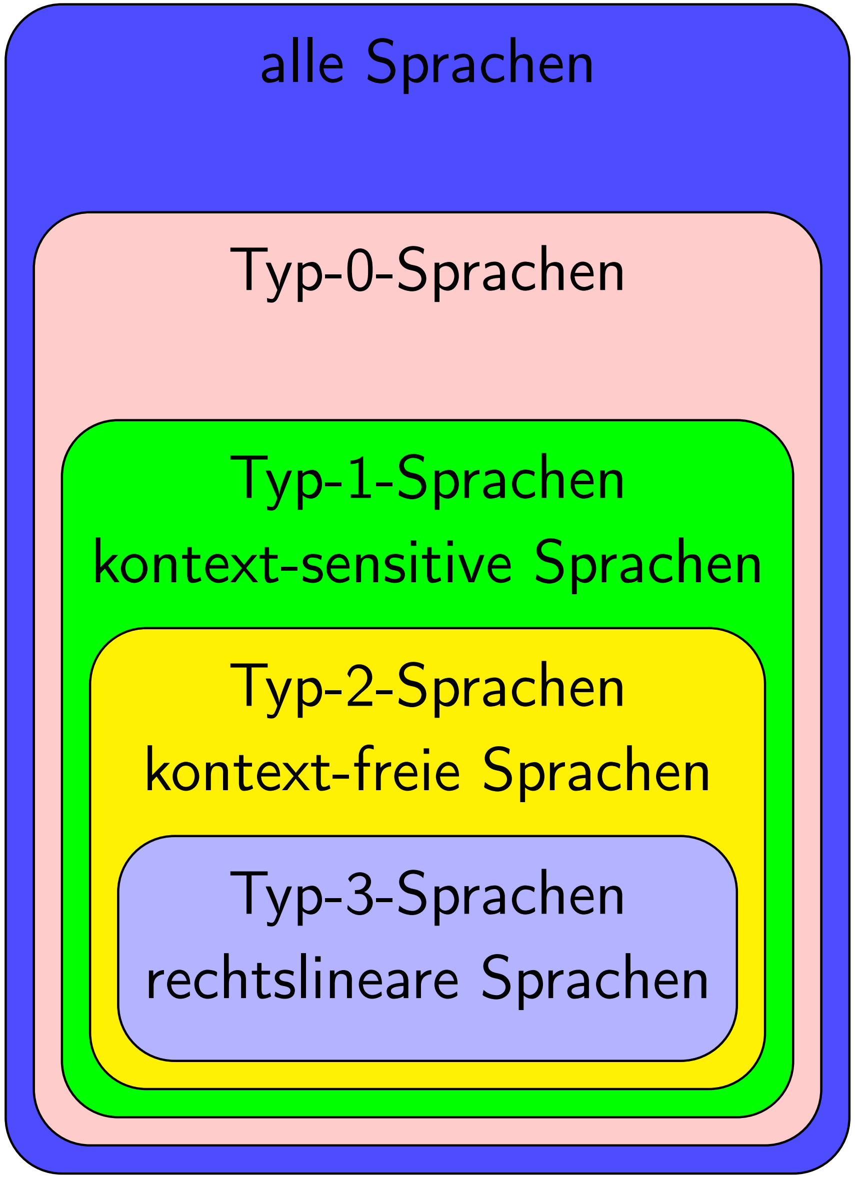
Auf Folien 2.23 und 9.29 wurde
$\mathcal{L}_3 \subseteq \mathcal{L}_2 \subseteq \mathcal{L}_1 \subseteq \mathcal{L}_0$ gezeigt.

Auf Folie 6.8 wurde $\mathcal{L}_0 \subsetneq$ Klasse aller Sprachen und auf Folie 6.13 auch $\mathcal{L}_3 \subsetneq \mathcal{L}_2$ gezeigt.

Jetzt wissen wir, daß auch
$\mathcal{L}_2 \subsetneq \mathcal{L}_1$ gilt.

Wir wissen noch nicht, ob
$\mathcal{L}_1 = \mathcal{L}_0$ oder $\mathcal{L}_1 \subsetneq \mathcal{L}_0$ gilt.

Wir haben also
$\mathcal{L}_3 \subsetneq \mathcal{L}_2 \subsetneq \mathcal{L}_1 \subseteq \mathcal{L}_0 \subsetneq$ Klasse aller Sprachen.



Im Beweis des Pumping-Lemmas verwenden wir die folgende Aussage.

Lemma

Sei B ein Binärbaum (d.h., jeder Knoten in B hat entweder null oder zwei Kinder) mit mindestens 2^k Blättern.
Dann hat B einen von der Wurzel ausgehenden Ast, der aus mindestens k Kanten und $k+1$ Knoten besteht.

Beweis: Induktion über k.

IA: $k = 0$.

Sei B ein Binärbaum mit mindestens 2^0 Blättern, d.h. mit mindestens einem Blatt.

Dann hat B einen Ast, der aus mindestens einem Knoten besteht.

IS: $k \geq 0$.

Sei B ein Binärbaum mit mindestens $2^{k+1} = 2^k + 2^k > 1$ Blättern.

Seien v_1 und v_2 die beiden Kinder der Wurzel und seien B_1 und B_2 die Binärbäume mit Wurzel v_1 bzw. v_2.

Dann existiert $i \in \{1, 2\}$, so daß B_i mindestens 2^k Blätter hat.

IV $\Rightarrow$ in B_i gibt es einen Ast mit mindestens k Kanten und $k+1$ Knoten.

$\Rightarrow$ in B gibt es einen Ast mit mindestens $k+1$ Kanten und $k+2$ Knoten.

$\square$

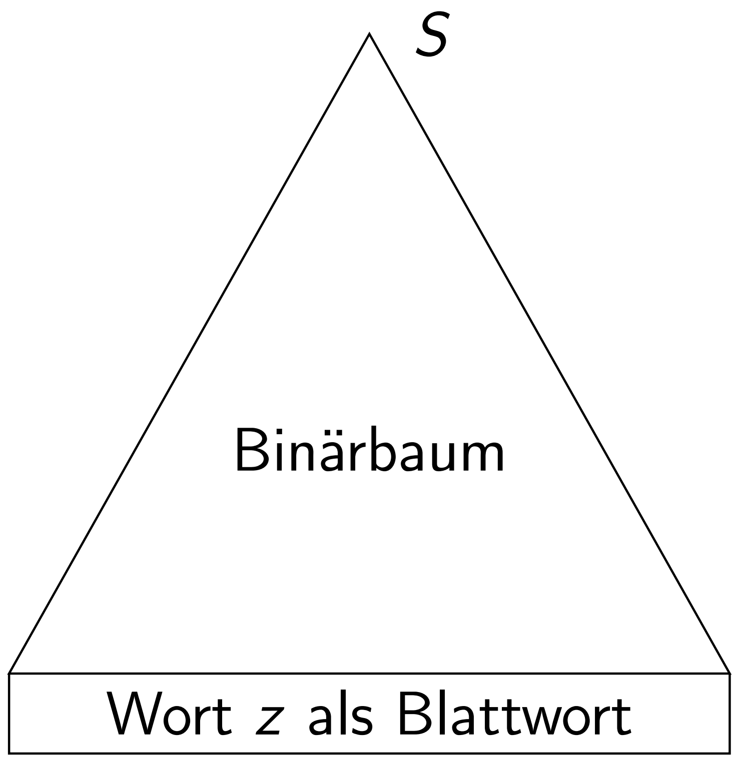
Beweis des Pumping-Lemmas:

Sei nun $G = (V, \Sigma, P, S)$ eine kontextfreie Grammatik mit $L = L(G)$ in Chomsky-Normalform. Setze $k := |V|$ und $n := 2^k$.

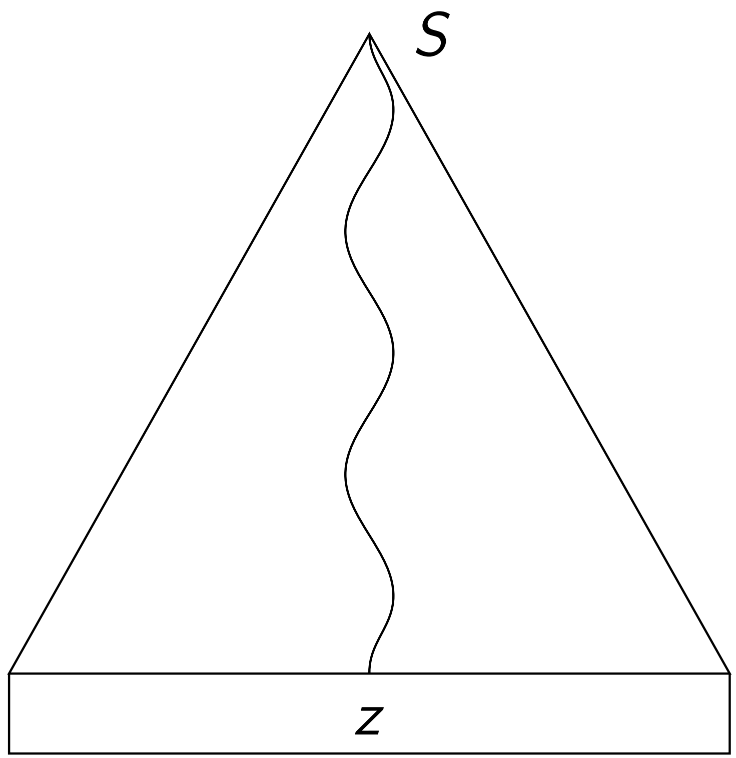
Sei weiter $z \in L(G)$ mit $|z| \geq n = 2^k$.

Wegen $z \in L(G)$ existiert ein Ableitungsbaum T mit $\alpha(T) = z$.

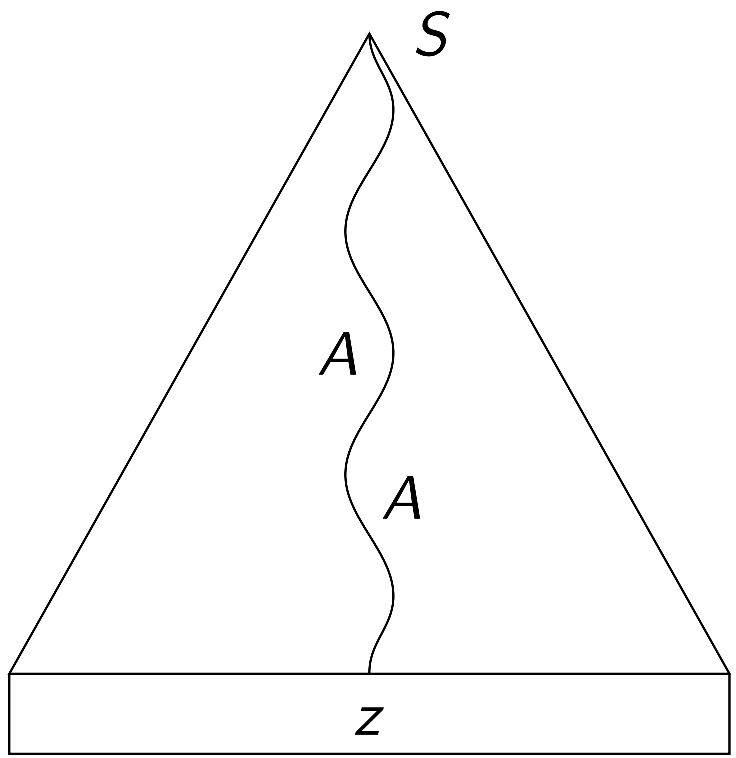
Der Ableitungsbaum T mit $\alpha(T) = z$ und $|z| \geq n = 2^k$.



Nach dem vorherigen Lemma existiert ein Ast in T mit mindestens $k+1$ Knoten.

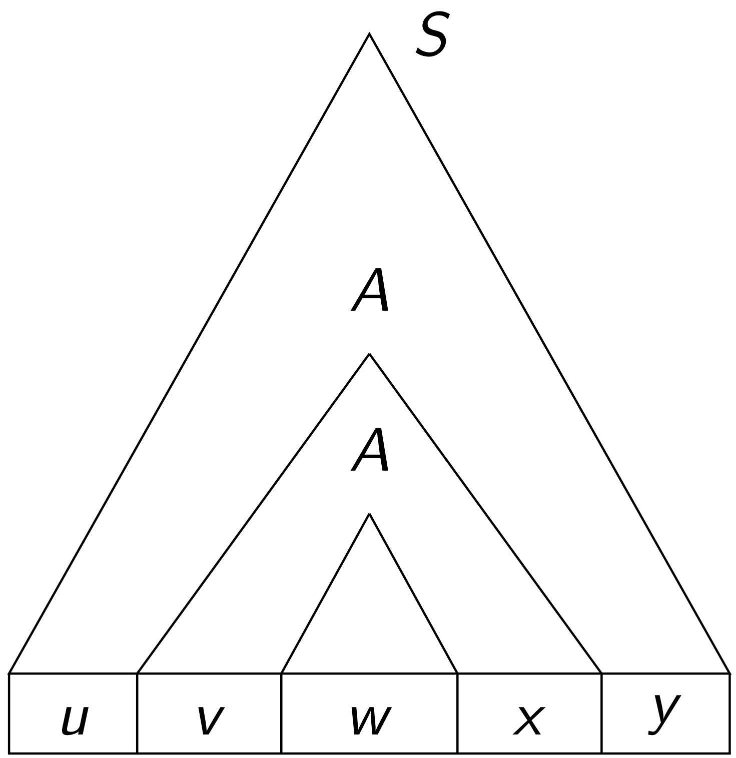


Also gibt es ein Nichtterminal (etwa A), das auf dem Ast zweimal auftaucht.
Wir können annehmen, daß auf jedem Ast, der vom oberen A ausgeht, höchstens das A mehrfach und dieses nicht dreimal vorkommt.

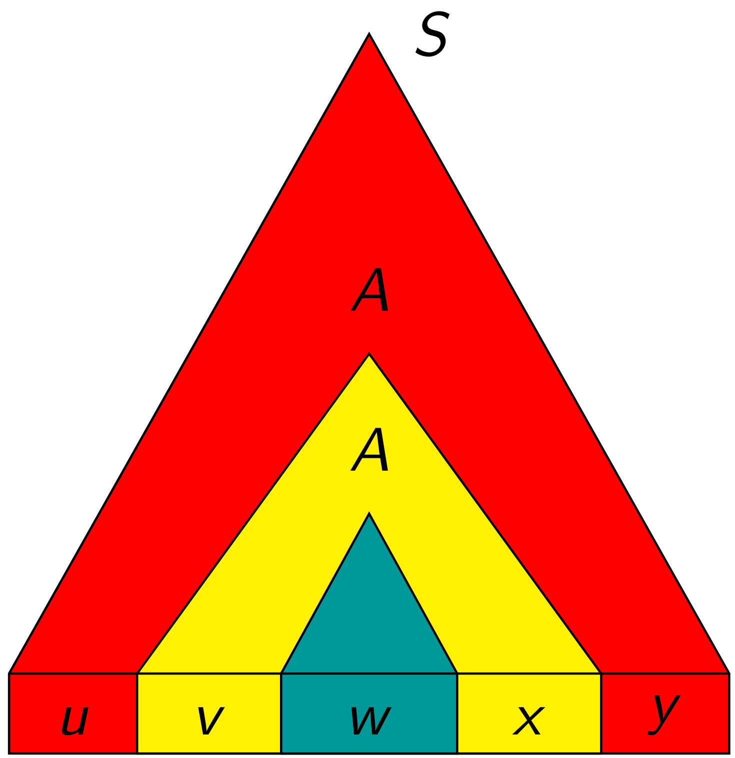


Das Wort z wird nun in fünf Teilwörter u, v, w, x, y aufgespalten (womit (i) gesichert ist):

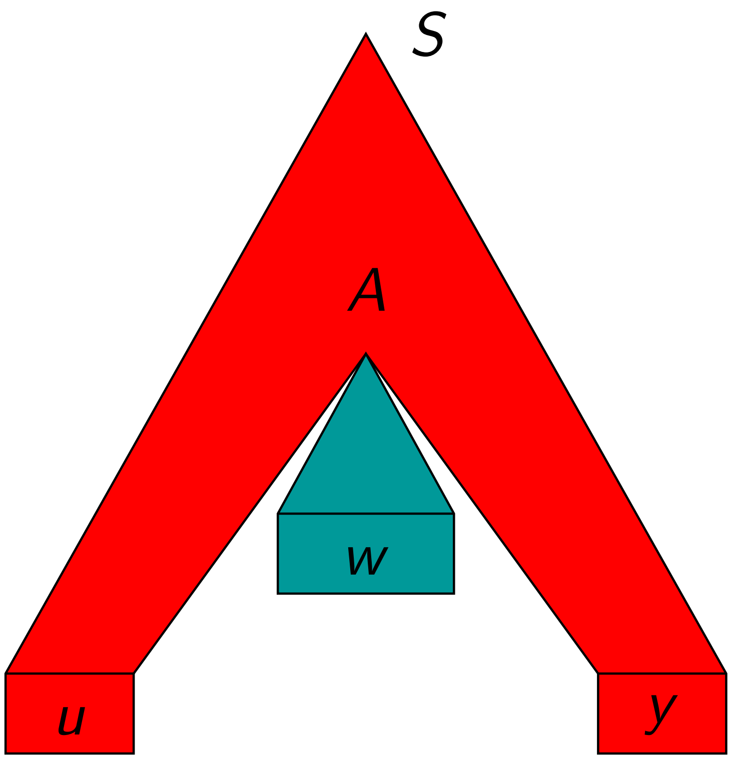
- w wird aus dem unteren A abgeleitet: $A \Rightarrow^* w$
- vwx wird aus dem oberen A abgeleitet: $A \Rightarrow^* vAx \Rightarrow^* vwx$



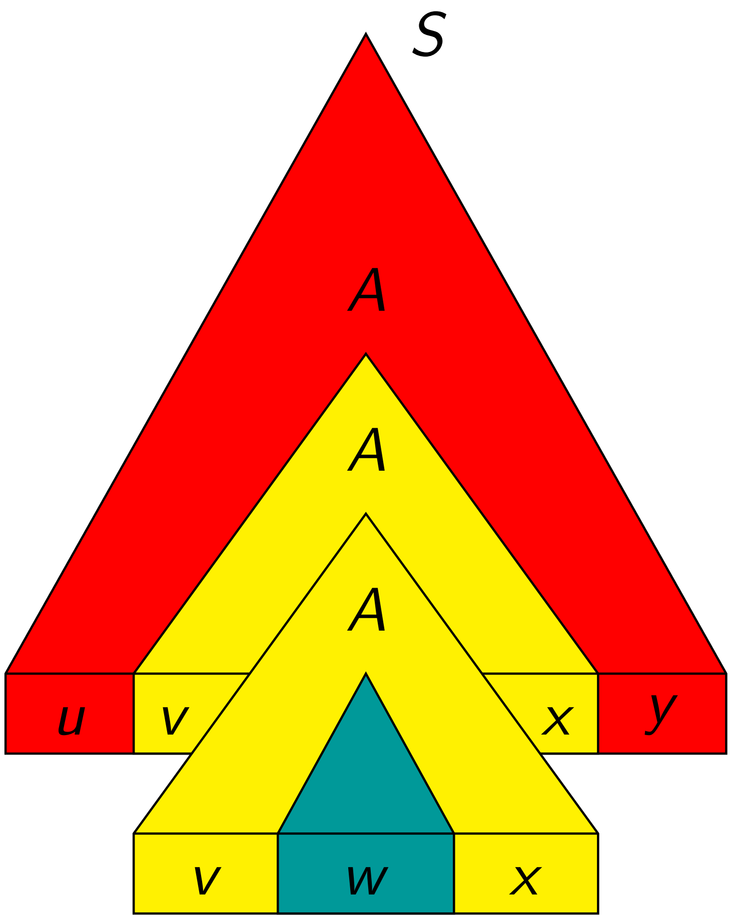
Damit erhält man drei ineinander enthaltene Teil-Ableitungsbäume, die man neu zusammenstecken kann.



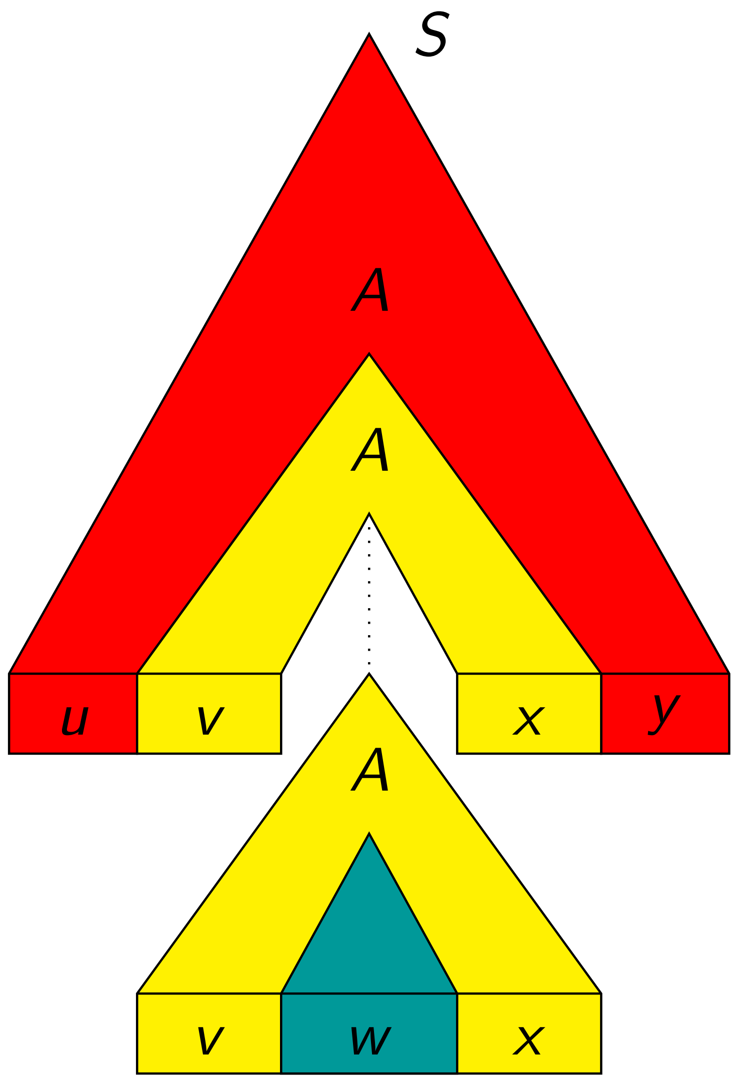
Durch Weglassen des mittleren Teilbaums erhält man einen Ableitungsbaum für uwy. Damit gilt: $uwy \in L(G)$.



Durch Verdoppeln des mittleren Teilbaums erhält man einen Ableitungsbaum für uv^2wx^2y. Damit gilt: $uv^2wx^2y \in L(G)$.



Durch Ver-i-fachen des mittleren Teilbaums erhält man einen Ableitungsbaum für uv^iwx^iy. Damit gilt: $uv^iwx^iy \in L(G)$, d.h. wir haben (iv) gezeigt.



Wir zeigen als nächstes (ii), d.h. $|vwx| \leq n$:

Sei T_1 der Teilbaum von T, der am oberen A beginnt (d.h. es gilt $\alpha(T_1) = vwx$).

Da wir das obere A so gewählt haben, daß davon ausgehende Äste höchstens das A mehrfach (und dieses nicht dreimal) enthalten, kann kein Ast in $T_1 > |V| + 1$ Knoten enthalten.

Das vorhergehende Lemma sichert $|vwx| =$ Blattanzahl in $T_1 \ \leq 2^{|V|} = n$.

Es bleibt noch (iii) zu zeigen, also $|vx| \geq 1$:

Sei T_2 der Teilbaum von T, der am unteren A beginnt (d.h. es gilt $\alpha(T_2) = w$).
Da T nur binäre Verzweigungen enthält und T_2 echter Teilbaum von T_1 ist, hat T_1 wenigstens ein Blatt, das nicht zu T_2 gehört. Da es keine ε-Produktionen gibt, ist dieses Blatt mit $a \in \Sigma$ beschriftet. Also gilt $vx \neq \varepsilon$. $\square$

Beispiel

Die Sprache $L_2 = \{0^{m^2} \mid m \in \mathbb{N}\}$ ist nicht kontextfrei.

Beweis: indirekt
Angenommen, L_2 wäre kontextfrei.

Nach dem Pumping-Lemma gibt es ein $n \geq 1$, so daß die folgende Aussage gilt:

> Für jedes $z \in L_2$, $|z| \geq n$, gibt es $u, v, w, x, y \in \Sigma^*$ mit (i), (ii), (iii) und (iv). (*)

Wir wählen nun $z = 0^{n^2}$.
Dann ist $z \in L_2$ und $|z| = n^2 \geq n$.
Nach der Aussage (*) gibt es also $u, v, w, x, y \in \Sigma^*$ mit

(i) $z = uvwxy$, (ii) $|vwx| \leq n$,

(iii) $|vx| \geq 1$ und (iv) $uv^iwx^iy \in L_2$ für alle $i \geq 0$.

Nach (ii) und (iii) gilt $1 \leq |vx| \leq |vwx| \leq n$.

Aus $|uv^2wx^2y| \overset{(i)}{=} |z| + |vx| = n^2 + |vx|$ folgt somit
$n^2 < |uv^2wx^2y| \leq n^2 + n < (n+1)^2$

und damit $uv^2wx^2y \notin L_2$, im Widerspruch zu (iv).

Also ist L_2 tatsächlich nicht kontextfrei. $\square$

Unsere Beweise, daß L_1 und L_2 nicht kontextfrei sind, folgten dem folgenden Schema:

Behauptung: Die Sprache L ist nicht kontextfrei.

[0] (wörtlich) **Beweis:** indirekt. Angenommen, L wäre kontextfrei. Nach dem Pumping-Lemma gibt es ein $n \geq 1$, so daß die folgende Aussage gilt:

Für jedes $z \in L$, $|z| \geq n$, gibt es $u, v, w, x, y \in \Sigma^*$ mit (i), (ii), (iii) und (iv). (*)

[1] (problemspezifisch) Wir wählen ein **geeignetes** $z \in L$ mit $|z| \geq n$, so daß Schritt [3] ausführbar ist.

[2] (wörtlich) Nach der Aussage (*) gibt es $u, v, w, x, y \in \Sigma^*$ mit (i)-(iv).

[3] (problemspezifisch) Wir wählen zu u, v, w, x, y ein **passendes** $i \geq 0$ und zeigen, daß uv^iwx^iy nicht in L sein kann.

[4] (wörtlich) Widerspruch zu (iv). □

Dieses Beweisschema ist die Umsetzung der folgenden Formulierung des Pumping-Lemmas mit logischen Operatoren:

L kontextfrei
$\rightarrow \exists n \;\; \forall z \in L$ mit $|z| \geq n \;\; \exists u, v, w, x, y$ mit (i–iii) $\;\forall i : uv^i wx^i y \in L$

Das Pumping-Lemma ist also logisch äquivalent zu:

$\forall n \quad \exists z \in L$ mit $|z| \geq n \quad \forall u, v, w, x, y$ mit (i-iii) $\quad \exists i : uv^i wx^i y \notin L$
$\rightarrow L$ ist nicht kontextfrei

Um zu zeigen, daß eine Sprache L nicht kontextfrei ist, reicht es also zu zeigen, daß es für alle n ein $x \in L$ gibt ...

Diese Umformulierung können wir auch in dem folgenden **Spielschema** fassen:

Wir (die Beweiser oder Braven) wollen zeigen, daß die Sprache L nicht kontextfrei ist. Dazu müssen wir das folgende Spiel (gegen den Gegner oder den Gemeinen) gewinnen:

Runde 1 G wählt eine Zahl $n \geq 1$.

Runde 2 B wählt ein $z \in L$ mit $|z| \geq n$

Runde 3 G wählt u, v, w, x und y mit
(i) $z = uvwxy$, (ii) $|vwx| \leq n$ und (iii) $|vx| \geq 1$.

Runde 4 B wählt ein i und zeigt, daß $uv^iwx^iy \notin L$.

Die Sprache L ist **nicht** kontextfrei, falls B unabhängig von den Wahlen von G in Runden 1 und 3 immer so wählen kann (in Runden 2 und 4), daß schließlich $uv^iwx^iy \notin L$ gilt.

Beispiel

$L_3 = \{w2w \mid w \in \{0,1\}^*\}$ ist nicht kontextfrei.

Beweis: wir zeigen, daß B im Spielschema immer so wählen kann, daß $uv^iwx^iy \notin L_3$ gilt:

Runde 1 G wählt eine Zahl $n \geq 1$.

Runde 2 B wählt $z = 0^n1^n20^n1^n$ (natürlich gelten $z \in L_3$ und $|z| \geq n$)

Runde 3 G wählt u, v, w, x und y mit
(i) $z = uvwxy$, (ii) $|vwx| \leq n$ und (iii) $|vx| \geq 1$.

Runde 4 B wählt $i = 0$ und zeigt, daß $uv^iwx^iy \notin L_3$:

1. Fall: v oder x enthält die 2.
   Dann enthält uv^0wx^0y keine 2, gehört also nicht zu L_3.
2. Fall: u enthält 2.
   Dann stehen in uv^0wx^0y links von der 2 mehr Buchstaben als rechts, also $uv^0wx^0y \notin L_3$.
3. Fall: y enthält 2.
   analog
4. Fall: w enthält 2.
   Wegen (ii) ist vwx Faktor von 1^n20^n, der 2 enthält. Also gilt $v \in L(1^*)$ und $x \in L(0^*)$ und damit $uv^0wx^0y = 0^n1^{n-|v|}20^{n-|x|}1^n \notin L_3$, da $|v| > 0$ oder $|x| > 0$ nach (iii).

Also kann B so wählen, daß $uv^iwx^iy \notin L_3$, d.h., L_3 ist tatsächlich nicht kontextfrei. $\square$

Das Lemma von Ogden

Lemma (William Ogden ’68)

Wenn L eine **kontextfreie** Sprache ist,
dann gibt es $n \geq 1$ derart,
daß für alle $z \in L$ ~~mit $|z| \geq n$~~, in denen n Positionen markiert sind,
gilt:
es gibt Wörter $u, v, w, x, y \in \Sigma^*$ mit

(i) $z = uvwxy$,

~~(ii)~~ $\cancel{|vwx| \leq n}$,

(iii) $\cancel{|vx| \geq 1}$ v oder x enthält wenigstens eine der Markierungen und

(iv) $uv^iwx^iy \in L$ für alle $i \geq 0$.

Beweis: ähnlich zum Beweis des Pumping-Lemmas. □

Beispiel

$L_4 = \{a^k b^\ell c^m \mid 0 \leq k \leq \ell \leq m\}$ ist nicht kontextfrei.

Beweis: Nach dem Spielschema, das man aus dem Lemma von Ogden gewinnen kann. Wir zeigen also, daß B immer so wählen kann, daß $uv^i wx^i y \notin L_4$ gilt:

Runde 1 G wählt eine Zahl $n \geq 1$.

Runde 2 B wählt $z = a^n b^n c^n$ (natürlich gilt $z \in L_4$) und markiert die Vorkommen von a

Runde 3 G wählt u, v, w, x und y mit
(i) $z = uvwxy$ und (iii) v oder x enthält eine Markierung.

Runde 4 B wählt $i = 2$ und zeigt, daß $uv^i wx^i y \notin L_4$:

1. Fall: v oder x enthält zwei verschiedene Sorten von Buchstaben. Dann gilt $uv^2wx^2y \notin L(a^*b^*c^*)$ und damit $uv^2wx^2y \notin L_4$.
2. Fall: $x \in L(a^*)$.
   Dann gilt $|uv^2wx^2y|_a = n + |vx|_a > n = |uv^2wx^2y|_b$, also $uv^2wx^2y \notin L_4$.
3. Fall: $x \in L(b^*)$.
   Dann enthält x keine Markierungen, aus (iii) folgt also, daß v wenigstens ein a enthält. Dann gilt $|uv^2wx^2y|_a \geq n + 1 > n = |uv^2wx^2y|_c$, also $uv^2wx^2y \notin L_4$.
4. Fall: $x \in L(c^*)$.
   Wir können $v \in L(a^* + b^* + c^*)$ annehmen, woraus nach (iii) $v \in L(a^+)$ folgt. Analog zum 3. Fall folgt nun $|uv^2wx^2y|_a \geq n + 1 > n = |uv^2wx^2y|_b$, also $uv^2wx^2y \notin L_4$.

B kann also so wählen, daß $uv^iwx^iy \notin L_4$, d.h., ist L_4 nicht kontextfrei. □

Beispiele

weitere Sprachen, die nicht kontextfrei sind:

$$
\begin{array}{ll}
\{a^k b^m a^k b^m \mid k, m \geq 0\} & \{a^k b^m a^{k \cdot m} \mid k, m \geq 0\} \\
\{w\, w \mid w \in \{0,1\}^*\} & \{0^m\, 1\, 0^m\, 1\, 0^m \mid m \geq 0\} \\
\{0^m\, 1^{2m}\, 0^{4m} \mid m \geq 0\} & \{0^p \mid p \text{ Primzahl } \}
\end{array}
$$

Beweis jeweils mit dem Pumping-Lemma oder dem Lemma von Ogden.

Beweis der Aussagen (4) und (5) des Satzes auf Folie 13.3

4.
$$\begin{aligned} L_5 &= \{a^k b^\ell c^m \mid 0 \le k \le \ell, 0 \le m\} \\ L_6 &= \{a^k b^\ell c^m \mid 0 \le k, 0 \le \ell \le m\} \end{aligned}$$

sind kontextfreie Sprachen (einfach).
Ihr Schnitt

$$L_5 \cap L_6 = \{a^k b^\ell c^m \mid 0 \le k \le \ell \le m\} = L_4$$

ist jedoch nicht kontextfrei (siehe Folie 13.25).

5. $L_7 = \{a^k b^\ell c^m \mid k > \ell \text{ oder } \ell > m\} \cup \{a, b, c\}^+ \setminus L(a^* b^* c^*)$
ist eine kontextfreie Sprache (einfach).

Ihr Komplement

$$\{a, b, c\}^* \setminus L_7 = \{a^k b^\ell c^m \mid 0 \le k \le \ell \le m\} = L_4$$

ist jedoch nicht kontextfrei (siehe Folie 13.25). □

Zusammenfassung 13. Vorlesung

in dieser Vorlesung neu

- die Klasse der kontextfreien Sprachen ist abgeschlossen unter Vereinigung, Verkettung und Iteration
- Pumping-Lemma für kontextfreie Sprachen
- die Klasse der kontextfreien Sprachen ist nicht abgeschlossen unter Schnitt und auch nicht unter Komplement

kommende Vorlesung

- der Schnitt einer kontextfreien und einer regulären Sprache ist kontextfrei
- Wird L von einem deterministischen PDA akzeptiert, so auch das Komplement