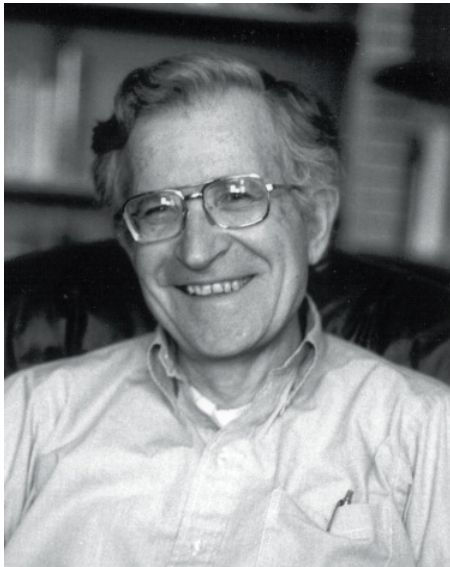


Noam CHOMSKY, Sheila GREIBACH

Noam CHOMSKY (*1928)



Sheila GREIBACH (*1939)



Normalformen für kontextfreie Grammatiken

Grammatik $G = (N, T, P, S)$

GREIBACH Normalform:

$$A \rightarrow aw, \quad w \in N^*$$

Erweiterte GREIBACH Normalform:

$$A \rightarrow aw, \quad w \in (N \cup T)^*$$

CHOMSKY Normalform:

$$A \rightarrow BC, \quad A \rightarrow a, \quad A, B, C \in N, \quad a \in T$$

$S \rightarrow \varepsilon$ ist nur dann erlaubt, wenn S nicht auf der rechten Seite einer Produktion vorkommt.

Homomorphismen auf kontextfreien Sprachen

Satz. Die Familie der kontextfreien Sprachen ist gegenüber beliebigen Homomorphismen abgeschlossen.

Beweis:

Sei Grammatik $G = (N, T, P, S)$ eine Typ-2-Grammatik, ohne Beschränkung der Allgemeinheit in CHOMSKY-Normalform, und $h: T^* \rightarrow W^*$ ein Homomorphismus mit $N \cap W = \{ \}$ (sonst Variablen in N umbenennen !).
Konstruiere nun eine kontextfreie Grammatik G' ,
 $G' = (N, W, P', S')$ mit

$$P' = P - \{A \rightarrow a \mid A \rightarrow a \in P, A \in N, a \in T\} \\ \cup \{A \rightarrow h(a) \mid A \rightarrow a \in P, A \in N, a \in T\}.$$

Klarerweise gilt auf Grund der Konstruktion
 $L(G') = h(L(G))$.



Ableitungen in kontextfreien Grammatiken

„normale“ Ableitung \Rightarrow :

ein beliebiges Nonterminal wird ersetzt.

Links-Ableitung \Rightarrow_L :

das in der Satzform am weitesten links vorkommende Nonterminal wird ersetzt

Rechts-Ableitung \Rightarrow_R :

das in der Satzform am weitesten rechts vorkommende Nonterminal wird ersetzt

Parallel-Ableitung $\Rightarrow_{||}$:

alle in der Satzform vorkommenden Nonterminale werden gleichzeitig ersetzt

Alle Ableitungsvarianten ergeben dieselbe Sprache!

Bäume

Ein (geordneter, markierter gerichteter) Baum über V und W ist ein Graph $g = (K, E, L)$ über V und W , der folgende Bedingungen erfüllt:

- 1) $W = [1..n]$ für ein $n \in \mathbb{N}_1$.
- 2) Es gibt genau einen ausgezeichneten Knoten p_0 (**Wurzel**), der keinen Vorgänger hat. Außerdem gibt es von der Wurzel aus zu jedem anderen Knoten q von g einen Pfad $(p_0, e_1, p_1, \dots, p_{k-1}, e_k, q)$ der Länge $k \geq 1$, $(p_i, e_{i+1}, p_{i+1}) \in E$, $0 \leq i < k$, $q = p_k$.
- 3) Jeder von der Wurzel verschiedene Knoten hat genau einen Vorgänger.
- 4) Jeder Knoten ohne Nachfolger heißt **Blatt**.
- 5) Ist p kein Blatt, so sind die Nachfolger von p geordnet (die Kanten tragen die Bezeichnungen 1 bis k für ein k).

Ableitungsbäume für kontextfreie Grammatiken

Sei $G = (N, T, P, S)$ eine kontextfreie Grammatik. Ein Baum $g = (K, E, L)$ über $V = N \cup T \cup \{\varepsilon\}$ und $W = [1..n]$ heißt **Ableitungsbaum für G** , wenn Folgendes gilt:

- 1) Ist p_0 die **Wurzel** von g , so gilt $L(p_0) = S$.
- 2) Ist p kein Blatt, so muss $L(p) \in N$ gelten.
- 3) Ist p ein **Blatt** mit $L(p) = \varepsilon$, so ist p der einzige Nachfolger seines Vorgängers.
- 4) Ist $\langle (p, i, q_i) \rangle_{i \in [1..k]}$ die geordnete Menge der von p mit $L(p) = A$ wegführenden Kanten, so ist $A \rightarrow L(q_1) \dots L(q_k)$ eine Produktion in P .

A-Baum für G : für Wurzel gilt $L(p_0) = A$.

Ableitungsbaum ist ein S-Baum.

Front eines Ableitungsbaumes

Sei $G = (N, T, P, S)$ eine kontextfreie Grammatik und $g = (K, E, L)$ ein A-Baum für G .

Wir definieren nun eine Ordnungsrelation auf den Pfaden von g : Seien $P(j) = (p(j,0), e(j,1), p(j,1), \dots, e(j,k_j), p(j,k_j))$ für $j \in \{1,2\}$ zwei voneinander verschiedene Pfade in g , die in der Wurzel beginnen (i.e., $p(1,0) = p(2,0) = p_0$) und zu einem Blatt von g führen, dann definieren wir $P(1) < P(2)$ genau dann, wenn es ein $m \geq 1$ so gibt, dass $e(1,i) = e(2,i)$ für alle $1 \leq i < m$ und $e(1,m) \neq e(2,m)$.

Betrachten wir nun alle derartigen Pfade in g , so sind diese wohlgeordnet und sind $p(1), \dots, p(k)$ die Blätter dieser Pfade, so ist die **Front** von g durch $L(p(1)) \dots L(p(k))$ definiert.

Ableitungen und Ableitungsbäume

Sei $G = (N, T, P, S)$ eine kontextfreie Grammatik und $g = (K, E, L)$ ein A-Baum für G sowie $w \in (N \cup T)^*$.

Dann gilt $A \Rightarrow_G w$ genau dann, wenn es einen A-Baum für G mit Front w gibt.

Jeder Linksableitung in G kann man eindeutig einen Ableitungsbaum zuordnen. Gibt es zwei verschiedene Linksableitungen in G für ein Wort w , so sind die entsprechenden Ableitungsbäume nicht äquivalent.

Eindeutigkeit, (inhärente) Mehrdeutigkeit

Sei $G = (N, T, P, S)$ eine kontextfreie Grammatik.

G heißt **eindeutig**, wenn es zu jedem in G ableitbaren Terminalwort genau eine Linksableitung in G gibt.

Ansonsten heißt G **mehrdeutig**.

Eine kontextfreie Sprache heißt **inhärent mehrdeutig**, wenn jede Grammatik, die L erzeugt, mehrdeutig ist.

Beispiel: Die kontextfreie Sprache $L = L(1) \cup L(2)$ mit
 $L(1) = \{a^n b^n c^m \mid n, m \in \mathbf{N}\}$ und
 $L(2) = \{a^n b^m c^m \mid n, m \in \mathbf{N}\}$
ist inhärent mehrdeutig.

Bemerkung: $L(1) \cap L(2) = \{a^n b^n c^n \mid 1 \leq n\}$.

Pumping Lemma für kontextfreie Sprachen

Sei L eine unendliche kontextfreie Sprache. Dann gibt es eine (nur von L abhängige) Schranke $m > 0$ so, dass für jedes Wort z in L mit $|z| \geq m$ Wörter $u, x, v, y, w \in \Sigma^*$ so existieren, dass

$$z = uxvyw \text{ mit}$$

$$|xvy| \leq m \text{ und}$$

$$|xy| > 0$$

sowie

$$z(i) = ux^i v y^i w \quad \text{für alle } i \geq 0 \text{ ebenfalls in } L \text{ liegt.}$$

Pumping Lemma für kontextfreie Sprachen – Beweis 1

Sei $G = (N, T, P, S)$ eine kontextfreie Grammatik, die L erzeugt, $k = \text{card}(N)$ und $m := 2^k$.

Wegen $|z| \geq m (= 2^k)$ muss jeder Ableitungsbaum für z einen Pfad mit einer Länge von mindestens $k+1$ haben.

So ein Pfad hat aber mindestens $k+2$ Knoten, wobei alle bis auf den letzten mit einem Nonterminal markiert sind.

Also muss es mindestens ein Nonterminal A aus N geben, das in diesem Pfad mindestens zweimal als Markierung eines Knotens vorkommt (Schubfachprinzip!).

Pumping Lemma für kontextfreie Sprachen – Beweis 2

Sei nun (p_0, e_1, \dots, p_l) so ein Pfad maximaler Länge in einem Ableitungsbaum von z , d.h., $l \geq k$.

Dann kann man in diesem Pfad zwei Knoten p_{v_1} und p_{v_2} so auswählen, dass Folgendes gilt:

- 1) $0 < v_2 - v_1 \leq k$ und $v_1 \geq l - k - 1$;
- 2) $L(p_{v_1}) = L(p_{v_2}) = A$ für ein $A \in N$ und
 $L(p_j) \neq A$ für alle $j \in \{i \mid v_1 < i \leq l\} - \{v_2\}$.

Der A-Baum T' mit der Wurzel p_{v_1} repräsentiert die Ableitung eines Teilwortes z' von z mit einer Länge von höchstens 2^k , da es lt. Voraussetzung nur Pfade mit einer maximalen Länge von $k+1$ geben kann; z' ist somit die Front von T' ; bezeichnet man die Front des vom Knoten p_{v_2} ausgehenden A-Baumes mit v , so kann man $z' = xvy$ für gewisse Wörter x, v, y schreiben.

Pumping Lemma für kontextfreie Sprachen – Beweis 3

Da G (außer eventuell $S \rightarrow \varepsilon$) keine ε -Produktionen enthält und $v_2 \neq v_1$ gilt, muss $|xy| \geq 1$ gelten, d.h.:

$A \Rightarrow^* xAy$ und $A \Rightarrow^* v$,
wobei $|xvy| \leq 2^k = m$ und $|xy| \geq 1$.

$A \Rightarrow^* x^i A y^i \Rightarrow^* x^i v y^i$ für alle $i \geq 0$.

Offensichtlich gibt es nun noch Wörter u und w so, dass man $z = uxvyw$ schreiben kann, d.h., man erhält
 $S \Rightarrow^* ux^i v y^i w$ für alle $i \geq 0$. □

Korollare zum Pumping Lemma für kontextfreie Sprachen

Korollar A.

Sei $L = \{a^{f(m)} \mid m \in \mathbf{N}\}$ eine formale Sprache und $f: \mathbf{N} \rightarrow \mathbf{N}$ eine monoton wachsende Funktion über den natürlichen Zahlen derart, dass für jedes $c \in \mathbf{N}$ ein $k \in \mathbf{N}$ mit $f(k+1) > f(k)+c$ existiert. Dann ist L nicht kontextfrei.

Korollar B.

Sei $L = \{a^{f(m)} \mid m \in \mathbf{N}\}$ eine formale Sprache und $f: \mathbf{N} \rightarrow \mathbf{N}$ eine monoton wachsende Funktion über den natürlichen Zahlen derart, dass für ein $d > 0$ $f(k+1) > f(k) + dk$ für alle $k \in \mathbf{N}$ gilt. Dann ist L nicht kontextfrei.

Beispiele zum Pumping Lemma für kontextfreie Sprachen

Bemerkung: Korollar B folgt direkt aus Korollar A.

Beispiel. $L = \{a^p \mid p \text{ prim}\}$

ist nach Korollar A nicht kontextfrei, da es beliebig große Primzahlücken gibt.

Aufgabe PL2A. $L = \{a^{f(m)} \mid m \in \mathbf{N}\}$ ist nicht kontextfrei für:

1) $f(m) = m^d$, $d \geq 2$,

2) $f(m) = k^m$.

Aufgabe PL2B. L ist nicht kontextfrei für:

1) $L = \{a^n b^n c^n \mid n \geq 1\}$,

2) $L = \{ww \mid w \in \{0,1\}^*\}$.

Abschlusseigenschaften kontextfreier Sprachen

Die Familie der kontextfreien Sprachen ist weder gegenüber Durchschnitt noch gegenüber Komplement abgeschlossen.

Die Sprachen $L(1)$ und $L(2)$ mit
 $L(1) = \{a^n b^n c^m \mid n, m \in \mathbf{N}\}$ und
 $L(2) = \{a^n b^m c^m \mid n, m \in \mathbf{N}\}$
sind kontextfreie Sprachen.

$L(1) \cap L(2) = \{a^n b^n c^n \mid 1 \leq n\}$ ist aber nicht kontextfrei.

Außerdem gilt

$$L(1) \cap L(2) = \{a, b, c\}^* - ((\{a, b, c\}^* - L(1)) \cup (\{a, b, c\}^* - L(2))).$$

Wäre also L_2 gegenüber Komplement abgeschlossen,
dann wäre L_2 , da gegenüber Vereinigung abgeschlossen,
auch gegenüber Durchschnitt abgeschlossen; Widerspruch!

Kontextfreie Sprachen aus $\{a\}^*$

Jede kontextfreie Sprache über einem einelementigen Alphabet ist regulär.

Die Korollare A und B zum Pumping Lemma für kontextfreie Sprachen würden somit schon aus dem Pumping Lemma für reguläre Sprachen ableitbar sein.

Charakterisierung regulärer Sprachen aus $\{a\}^*$:

Für jede reguläre Sprache L aus $\{a\}^*$ gibt es natürliche Zahlen $d, m \geq 0$ sowie $c(k)$, $1 \leq k \leq m$, derart, dass

$$L = \bigcup_{1 \leq k \leq m} \{a^{dx+c(k)} \mid x \in \mathbf{N}\}.$$

Aufgabe: Wie schauen die entsprechenden Minimalautomaten aus?

Normalformen für kontextfreie Grammatiken

Grammatik $G = (N, T, P, S)$

GREIBACH Normalform:

$$A \rightarrow aw, \quad w \in N^*$$

Erweiterte GREIBACH Normalform:

$$A \rightarrow aw, \quad w \in (N \cup T)^*$$

CHOMSKY Normalform:

$$A \rightarrow BC, \quad A \rightarrow a, \quad A, B, C \in N, \quad a \in T$$

$S \rightarrow \varepsilon$ ist nur dann erlaubt, wenn S nicht auf der rechten Seite einer Produktion vorkommt.

Normalformen für Grammatiken

Grammatik $G = (N, T, P, S)$

Normalform für monotone Grammatiken:

$A \rightarrow BC, AD \rightarrow BC, A \rightarrow a, \quad A, B, C, D \in N, a \in T$

$S \rightarrow \varepsilon$ ist nur dann erlaubt, wenn S nicht auf der rechten Seite einer Produktion vorkommt.

Normalform für unbeschränkte Grammatiken:

$A \rightarrow BC, AD \rightarrow BC, A \rightarrow a, \quad A, B, C, D \in N, a \in T \cup \{\varepsilon\}$

Varianten von regulären Grammatiken

Grammatik $G = (N, T, P, S)$

Normalform für reguläre Grammatiken:

$A \rightarrow aB, A \rightarrow a, A, B \in N, a \in T$

$S \rightarrow \varepsilon$ ist nur dann erlaubt, wenn S nicht auf der rechten Seite einer Produktion vorkommt.

„Maximalvariante“ für reguläre Grammatiken:

$A \rightarrow wB, A \rightarrow w, A, B \in N, w \in T^*$

CHOMSKY - Hierarchie

Grammatik $G = (N, T, P, S)$; betrachte Normalformen:

Normalform für unbeschränkte Grammatiken:

$A \rightarrow BC, AD \rightarrow BC, A \rightarrow a, A, B, C, D \in N, a \in T \cup \{\varepsilon\}$

Normalform für monotone Grammatiken:

$A \rightarrow BC, AD \rightarrow BC, A \rightarrow a, A, B, C, D \in N, a \in T$

CHOMSKY Normalform:

$A \rightarrow BC, A \rightarrow a, A, B, C \in N, a \in T$

Normalform für reguläre Grammatiken:

$A \rightarrow aB, A \rightarrow a, A, B \in N, a \in T$

CHOMSKY-Hierarchie: $L_3 \subset L_2 \subset L_1 \subset L_0$

$L_3 \subset L_2 \subset L_1 \subset L_{\text{rek}} \subset L_0$

Bedeutung der ε -Produktionen

Grammatik $G = (N, T, P, S)$:

Normalform für unbeschränkte Grammatiken:

$A \rightarrow BC, AD \rightarrow BC, A \rightarrow a, A, B, C, D \in N, a \in T \cup \{\varepsilon\}$

Normalform für monotone Grammatiken:

$A \rightarrow BC, AD \rightarrow BC, A \rightarrow a, A, B, C, D \in N, a \in T$

Unterschied: ε -Produktionen der Gestalt $A \rightarrow \varepsilon$

Eine ε -Produktion der Gestalt $A \rightarrow \varepsilon$ reicht bereits:

Zu jeder Typ-0-Sprache $L \subseteq T^*$ gibt es eine monotone Sprache $L' \subseteq (T \cup \{e\})^*$ derart, dass gilt:

- Für jedes Wort $w \in L$ existiert ein Wort $we^n \in L'$ für ein $n \in \mathbb{N}$.
- Jedes Wort $v \in L'$ ist von der Gestalt we^n für ein $w \in L$ und ein $n \in \mathbb{N}$.

Homomorphismen auf monotonen Sprachen

Beweis:

Sei Grammatik $G = (N, T, P, S)$ eine Typ-0-Grammatik, ohne Beschränkung der Allgemeinheit in Normalform.

Konstruiere dazu eine monotone Grammatik G' ,
 $G' = (N \cup \{S', E\}, T \cup \{e\}, P', S')$ mit den folgenden Produktionen in P' :

$S' \rightarrow Se$, $ED \rightarrow DE$ für alle $D \in N$, $Ee \rightarrow ee$,

$A \rightarrow E$ für alle Produktionen $A \rightarrow \varepsilon$ in P ,

$A \rightarrow BC$, $AD \rightarrow BC$, $A \rightarrow a$, $A, B, C, D \in N$, $a \in T$,
für alle derartigen Produktionen in P . □

$h(L(G')) = L$ für den Homomorphismus

$h: (T \cup \{e\})^* \rightarrow T^*$ mit $h(a) = a$, $a \in T$, und $h(e) = \varepsilon$.

Analog dazu gilt $L(G'') = L$ für

$G'' = (N \cup \{S', E, e\}, T, P' \cup \{e \rightarrow \varepsilon\}, S')$.

Abgeschlossenheit gegenüber Homomorphismen

Aufgabe HOMA:

Zeigen Sie, dass alle Sprachfamilien der CHOMSKY-Hierarchie gegenüber ε -freien Homomorphismen abgeschlossen sind.

Aufgabe HOMB:

Zeigen Sie, dass die Sprachfamilien der CHOMSKY-Hierarchie L_3 , L_2 und L_0 gegenüber beliebigen Homomorphismen abgeschlossen sind, L_1 und L_{rek} hingegen nicht.

Sprachfamilien – (volle) Trios

Sprachfamilie: nichttriviale Menge formaler Sprachen
(enthält zumindest eine nichtleere Sprache)

TRIO:

Abgeschlossen gegenüber $\cap R, h^{-1}, h_{-\varepsilon}$
(ε -freier Homomorphismus)

full (volles) TRIO:

Abgeschlossen gegenüber $\cap R, h^{-1}, h$

(volle) Abstrakte Sprachfamilien

(full) abstract family of languages: (f)AFL

AFL:

TRIO und abgeschlossen gegenüber $\cup, \bullet, ^+$

full AFL:

volles TRIO und abgeschlossen gegenüber $\cup, \bullet, ^*$

\cup oder \bullet folgt bereits aus den jeweils 5 anderen Eigenschaften!

Abschlusseigenschaften von Sprachfamilien

	TRIO	fTRIO	AFL	fAFL	$L_3 \subset$	$L_2 \subset$	$L_1 \subset$	L_0
\cup			+	+	+	+	+	+
\bullet			+	+	+	+	+	+
+			+	+	+	+	+	+
*				+	+	+	+	+
$\cap R$	+	+	+	+	+	+	+	+
$h_{-\varepsilon}$	+	+	+	+	+	+	+	+
h		+	+	+	+	+	-	+
h^{-1}	+	+	+	+	+	+	+	+
$gsm_{-\varepsilon}$	+	+	+	+	+	+	+	+
gsm		+		+	+	+	-	+
gsm^{-1}	+	+	+	+	+	+	+	+
Kompl.					+	-		-

Quotient von Sprachen

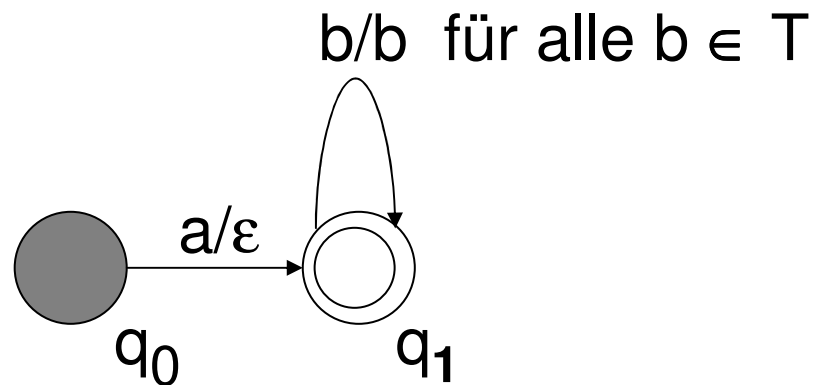
Quotient von Sprachen L/M

$L/M = \{ w \mid \text{es gibt ein } u \in M \text{ sodass } wu \in L \}$

$L \setminus M = \{ w \mid \text{es gibt ein } u \in M \text{ sodass } uw \in L \}$

Ist F eine Sprachfamilie, die gegenüber gsm-Abbildungen abgeschlossen ist, so ist für jede Sprache $L \subseteq T^*$ aus F auch $L \setminus \{a\}$ für jedes $a \in T$ aus F .

Beweis:



Weitere Operationen auf Sprachen

$$\text{INIT}(L) = \{ w \mid wu \in L \}$$

$$\text{FIN}(L) = \{ w \mid uw \in L \}$$

$$\text{SUB}(L) = \{ w \mid xwu \in L \}$$

Aufgabe AFLA: Zeigen Sie, dass jede Sprachfamilie, die gegenüber gsm-Abbildungen abgeschlossen ist, auch gegenüber den Operationen INIT, FIN und SUB abgeschlossen ist.