Einführung in die Computerlinguistik

Pumping-Lemma für kontextfreie Sprachen, Abschlußeigenschaften kontextfreier Sprachen

und die Komplexität natürlicher Sprachen

Dozentin: Wiebke Petersen

WS 2004/2005

Nachtrag: deterministische und nichtdeterministische Kellerautomaten

Genau wie bei den endlichen Automaten unterscheidet man zwischen deterministischen und nichtdeterministischen Kellerautomaten.

So ist jeder Kellerautomat, der die Sprache der Palindrome über dem Alphabet $\{a, b\}$ akzeptiert notwendigerweise nichtdeterministisch, da ein Automat, der die Eingabekette strikt von links nach rechts abarbeitet, nicht erkennen kann, wann die Mitte der Eingabekette erreicht ist.

Vorsicht, nichtdeterministische Kellerautomaten bedeuten nicht, daß die akzeptierte Sprache ambig ist: die Grammatik, die die Sprache der Palindrome über $\{a, b\}$ generiert ist nicht ambig!

$$S \rightarrow aSa \quad S \rightarrow bSb \quad S \rightarrow \epsilon$$

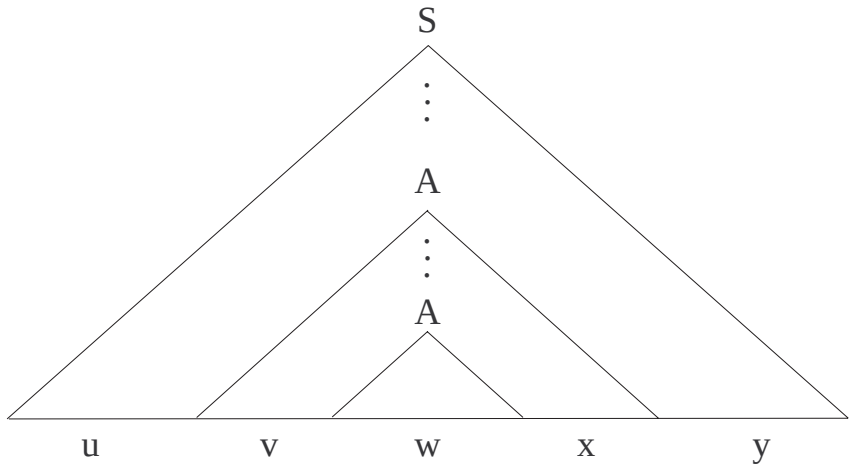
Die Menge der Sprachen, die von deterministischen Kellerautomaten akzeptiert wird, ist eine echte Teilmenge der kontextfreien Sprachen.

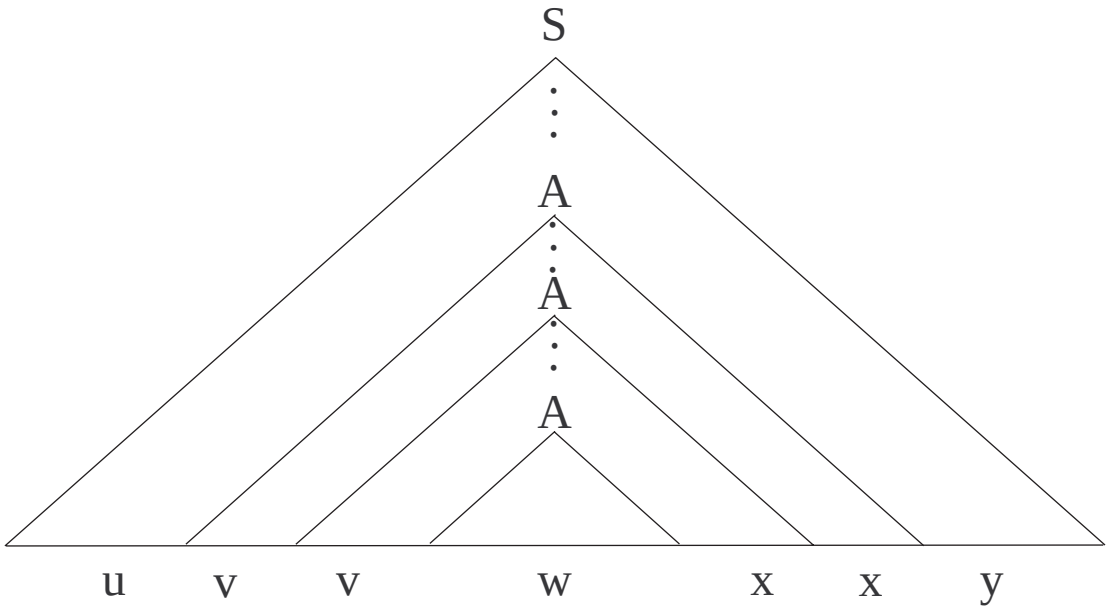
Pumpinglemma für kontextfreie Sprachen

Lemma 1. [Pumpinglemma für kontextfreie Sprachen] *Für jede kontextfreie Sprache L gibt es $p \in \mathbb{N}$, so daß für jedes $z \in L$ gilt, wenn $|z| > p$, dann gibt es eine Zerlegung $z = uvwxy$, mit*

- $u, v, w, x, y \in T^*$,
- $|vwx| \leq p$,
- $vx \neq \epsilon$ *und*
- $uv^iwx^iy \in L$, *für jedes* $i \geq 0$.

Beweisskizze zum Pumpinglemma





$|vwx| \leq p$, $vx \neq \epsilon$ und $uv^iwx^iy \in L$, für jedes $i \geq 0$.

Existenz von nicht-kontextfreien Sprachen

- $L_1 = \{a^n b^n c^n\}$
- $L_2 = \{a^n b^m c^n d^m\}$
- $L_1 = \{ww : w \in \{a, b\}^*\}$

Abschlußeigenschaften kontextfreier Sprachen

	Typ3	Typ2	Typ1	Typ0
Vereinigung	+	+	+	+
Schnittmenge	+	-	+	+
Komplementmenge	+	-	+	-
Konkatenation	+	+	+	+
Kleene Stern	+	+	+	+
Schnittmenge mit einer regulären Sprache	+	+	+	+

Vereinigung: $G = (N_1 \cup N_2 \cup \{S\}, T_1 \cup T_2, S, P)$ mit $P = P_1 \cup P_2 \cup \{S \rightarrow S_1, S \rightarrow S_2\}$

Schnittmenge: $L_1 = \{a^i b^i a^j\}$, $L_2 = \{a^i b^j a^j\}$, aber $L_1 \cap L_2 = \{a^i b^i a^i\}$

Komplementmenge: *de Morgan*

Konkatenation: $G = (N_1 \cup N_2 \cup \{S\}, T_1 \cup T_2, S, P)$ mit $P = P_1 \cup P_2 \cup \{S \rightarrow S_1 S_2\}$

Kleene Stern: $G = (N_1 \cup \{S\}, T_1, S, P)$ mit $P = P_1 \cup P_2 \cup \{S \rightarrow S_1 S, S \rightarrow \epsilon\}$

Zusammenfassung der bisherigen Ergebnisse

- Wir haben gesehen, daß die Menge der regulären Sprachen eine echte Teilmenge der Menge der kontextfreien Sprachen ist.

- Wir haben gesehen, daß es Sprachen gibt, die nicht kontextfrei sind.

- Auch für die nicht-kontextfreien Sprachen lassen sich allgemeine Regelgrammatiken angeben:
  Beispiel $a^n b^n c^n$:

$$\begin{array}{rclrclrcl} S & \rightarrow & abc & S & \rightarrow & aAbc & & & \\ Ab & \rightarrow & bA & Ac & \rightarrow & Bbcc & & & \\ bB & \rightarrow & Bb & aB & \rightarrow & aaA & aB & \rightarrow & aa \end{array}$$

$$S \vdash aAbc \vdash abAc \vdash abBbcc \vdash aBbbcc \vdash aabbcc$$

- Es gibt Hinweise darauf, daß eine Hierarchie der Sprachklassen in Abhängigkeit von den zulässigen Regelformen existiert.

Chomsky-Hierarchie

- Wenn man die Form der Regeln einschränkt erhält man Teilmengen der Menge aller durch eine Grammatik erzeugten Sprachen.
- Die Chomsky-Hierarchie ist eine Hierarchie über die Regelbedingungen (den verschiedenen Sprachklassen entsprechen Einschränkungen über die rechten und linken Regelseiten).
- Die Chomsky Hierarchie reflektiert eine spezielle Form der Komplexität, andere Kriterien sind denkbar und führen zu anderen Hierarchien.
- Die Sprachklassen der Chomsky Hierarchie sind in der Informatik intensiv untersucht worden (Berechnungskomplexität, effektive Parser).
- Für Linguisten ist die Chomsky Hierarchie besonders interessant, da sie die Form der Regeln zentral stellt, und somit Aussagen über Grammatikformalismen zuläßt.

Noam Chomsky

Noam Chomsky
(∗ 7.12.1928, Philadelphia)
Noam Chomsky, *Three Models for the Description of Language*, IRE Transactions on Information Theory (1956).

Chomsky Hierarchie (1956)

$A \in N$, $a, b \in T$, $\alpha \in (N \cup T)^* \setminus T^*$, $\beta \in (N \cup T)^*$

- **reguläre Sprachen** (Typ 3, REG): $A \rightarrow bA$
  - Beispiel: a^*b^*
  - endliche Automaten (Wortproblem in linearer Zeit lösbar)
- **kontextfreie Sprachen** (Typ 2, CFL): $A \rightarrow \beta$
  - Beispiele: $a^n b^n$, ww^{-1}
  - nichtdeterministische Kellerautomaten (Wortproblem in kubischer Zeit lösbar)
- **kontextsensitive Sprachen** (Typ 1, CSL): $\alpha \rightarrow \beta$ (mit $|\alpha| \leq |\beta|$ und evtl. $S \rightarrow \epsilon$)
  - Beispiele: $a^n b^n c^n$, ww
  - linear beschränkte Automaten (Wortproblem in exponentieller Zeit lösbar)
- **rekursiv aufzählbare Sprachen** (Typ 1, RE): $\alpha \rightarrow \beta$
  - Turingmaschinen (Wortproblem nicht entscheidbar)

$$REG \subset CFL \subset CSL \subset RE$$

Zeitkomplexität

Angenommen ein Rechner kann einen Rechenschritt in einer Mikrosekunde (10^-6 s) durchführen, dann ergeben sich (abhängig von der Länge der Eingabe n) folgende unterschiedliche Berechnungszeiten für quadratische (n^2) und exponentielle (2^n) Probleme.

n	n^2	2^n
2	0.000004 Sekunden	0.000004 Sekunden
10	0.0001 Sekunden	0.001 Sekunden
20	0.0004 Sekunden	1.05 Sekunden
30	0.0009 Sekunden	17.9 Minuten
40	0.0016 Sekunden	12.7 Tage
50	0.0025 Sekunden	35.7 Jahre

Vokabular der Theorie der Entscheidbarkeit

Algorithmus: Eine aus endlich vielen Schritten bestehende Verarbeitungsvorschrift, die, mechanisch angewandt zur Lösung eines Problems führt.

Entscheidbarkeit: Ein Problem ist entscheidbar, wenn ein Algorithmus existiert, der bei Eingabe einer Instantiierung des Problems nach endlich vielen Schritten angibt, ob dieses lösbar ist oder nicht.

Entscheidbarkeitsprobleme

Gegeben: Grammatiken $G = (N, \Sigma, S, P)$, $G' = (N', \Sigma, S', P')$, Wort $w \in \Sigma^*$

Wortproblem Ist w in G ableitbar?

Leerheitsproblem Erzeugt G eine nichtleere Sprache?

Äquivalenzproblem Erzeugen G und G' die gleichen Sprachen $(L(G) = L(G'))$?

Ergebnisse zu Entscheidbarkeitsproblemen

	Typ3	Typ2	Typ1	Typ0
Wortproblem	E	E	E	U
Leerheitsproblem	E	E	U	U
Äquivalenzproblem	E	U	U	U

E steht für entscheidbar
U steht für unentscheidbar

Wortproblem: Argumentation über Wortlänge

Leerheitsproblem: Markiere die Symbole der Regeln aus denen ein Terminalwort ableitbar ist (wenn Startsymbol markiert, dann ist die Sprache nicht leer).

Äquivalenzproblem: Zurückführbar auf das Postsche Korrespondenz-Problem

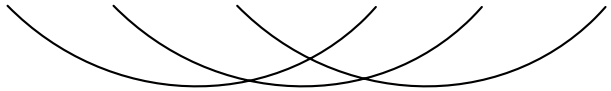
Sind natürliche Sprachen kontextfrei?

Nebensatzeinbettung im Schweizerdeutschen

mer d'chind em Hans es huus lönd hälfe aastriiche

wir die Kinder-AKK Hans-DAT das Haus-AKK ließen helfen anstreichen

$NP_1\ NP_2\ NP_3\ VP_1\ VP_2\ VP_3$ "cross serial dependencies"



*mer d'chind de Hans es huus lönd hälfe aastriiche

wir die Kinder-AKK Hans-AKK das Haus-AKK ließen helfen anstreichen

Nebensatzeinbettung im Deutschen

weil er die Kinder dem Hans das Haus streichen helfen ließ

$NP_1\ NP_2\ NP_3\ VP_3\ VP_2\ VP_1$ "nested dependencies"

NL $\not\subset$ CF: Beweis Shieber 1985

Homomorphismus:	f("laa") = c	f("es huus haend wele") = x
f("d'chind") = a	f("hälfe") = d	f("Jan säit das mer") = w
f("em Hans") = b	f("aastriiche") = y	f(s) = z otherwise

- f(Schweizerdeutsch) $\cap$ wa*b*xc*d*y = $wa^mb^nxc^md^ny$
- $wa^mb^nxc^md^ny$ ist nicht kontextfrei ($\rightarrow$Pumping Lemma)
- wa*b*xc*d*y ist regulär
- kontextfreie Sprachen sind abgeschlossen unter
  - Homomorphismen
  - Schnitt mit regulären Sprachen

➢ Das Schweizerdeutsch ist nicht kontextfrei

potentielle Angriffspunkte des Beweis

falsche Daten

- Grammatikalitätsurteile
- andere Konstituentenstrukturen sind auch möglich

Kasus ist nicht syntaktisch

- dann wäre Kasus bestimmt durch Semantik

Die Länge der Sätze ist beschränkt

- Shieber: "Down this path lies tyranny. Acceptance of this argument opens the way to proofs of natural languages as regular, nay, finite. The linguist proposing this counterargument to salvage the context-freeness of natural language may have won the battle, but has certainly lost the war.