Formale Methoden 1

Gerhard Jäger
Gerhard.Jaeger@uni-bielefeld.de

Uni Bielefeld, WS 2007/2008

30. Januar 2008

Kontextfreie Sprachen und Kellerautomaten

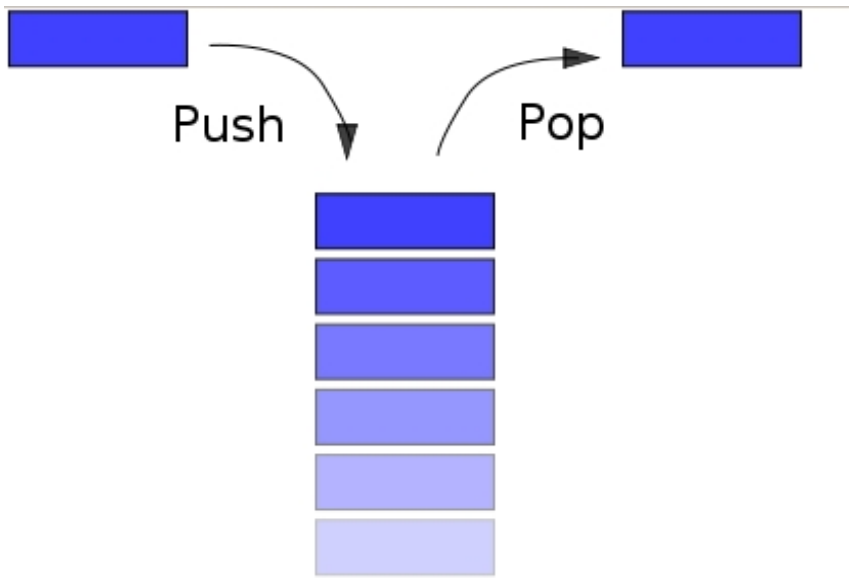
- **Kontextfreie Grammatiken (Typ-2-Grammatiken):** Alle Regeln haben die Form

$$A \to \gamma$$

  wobei A ein Nicht-Terminalsymbol ist und γ eine Kette von Terminalsymbolen
- **Kontextfreie Sprache:** Sprache, die von einer Typ-2-Grammatik erkannt wird
- Jede reguläre Sprache ist kontextfrei.
- Beispiele für kontextfreie Sprachen (die nicht regulär sind):
  - $a^n b^n$
  - $a^n b^{2n}$
  - ww^R (Palindrom-Sprache)

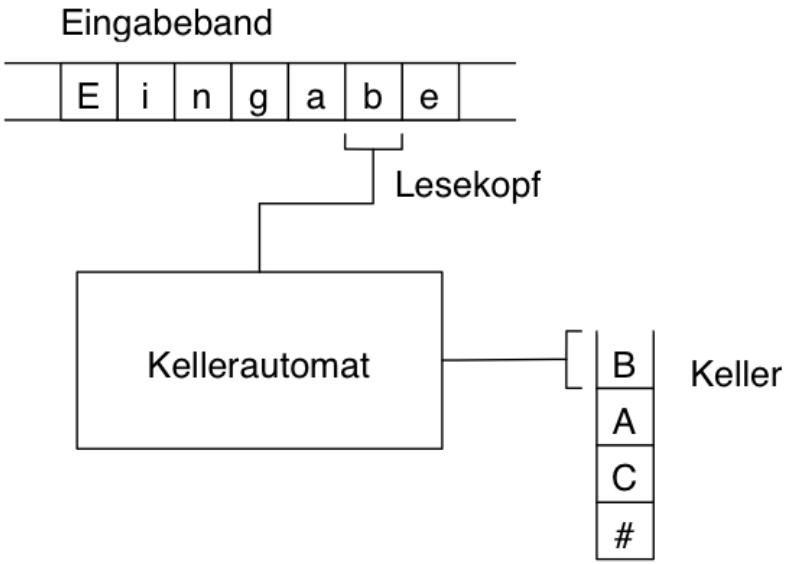
Kontextfreie Sprachen und Kellerautomaten

- **Kellerautomat:** endlicher Automat mit einem *Kellerspeicher*
- **Kellerspeicher:**
  - Stapelspeicher (engl. *stack*)
  - ordnet Symbole in linearer Sequenz an
  - Manipulation nach dem Prinzip *last in—first out*



Kontextfreie Sprachen und Kellerautomaten

- Im Startzustand ist Kellerspeicher leer.
- pro Zustandsübergang: entferne maximal ein Element vom Kellerspeicher und füge eine endliche Anzahl von Elementen hinzu
- Eine Input-Kette ist akzeptiert, wenn
  - der Automat nach Abarbeitung der Kette in einem Endzustand ist, und
  - der Kellerspeicher dann leer ist.



Kontextfreie Sprachen und Kellerautomaten

- Beispiel für einen Kellerspeicherautomaten, der $a^n b^n$ akzeptiert:

$$
\begin{array}{lll}
\text{Zustände:} & K & = \{Z_0, Z_1\} \\
\text{Input-Alphabet:} & \Sigma & = \{a, b\} \\
\text{Keller-Alphabet:} & \Gamma & = \{A\} \\
\text{Anfangszustand:} & Z_0 & \\
\text{Endzustände:} & F & = \{Z_0, Z_1\} \\
\text{Zustandsübergänge:} & \Delta & = \left\{ \begin{array}{l} (Z_0, a, \epsilon) \rightarrow (Z_0, A) \\ (Z_0, b, A) \rightarrow (Z_1, \epsilon) \\ (Z_1, b, A) \rightarrow (Z_1, \epsilon) \end{array} \right\}
\end{array}
$$

Kontextfreie Sprachen und Kellerautomaten

Theorem

Jeder Kellerautomat akzeptiert eine kontextfreie Sprache, und jede kontextfreie Sprache wird von einen Kellerautomaten erkannt.

Pumping-Lemma für kontextfreie Sprachen

- Wenn eine Kette x von einer kf Grammatik G generiert wird, gibt es eine Syntax-Baum für x, der nur Regeln aus G benutzt.
- Es gibt endliche viele Regeln in G. Sei r die Anzahl der Regeln in G.
- Jede Regel aus G hat ein bestimmte Zahl von Symbolen auf der rechten Seite. Sei s die maximale Anzahl von Symbolen auf der rechten Seite einer Regel.

Pumping-Lemma für kontextfreie Sprachen

- Angenommen,
  - x wird von G generiert,
  - T ist der Syntax-Baum für x.
  - Es gibt kein Nichtterminal-Symbol, was sich in T selbst dominiert.

  Dann gilt:
  - Es gibt maximal s^r Äste in T.
  - Also gibt es nicht mehr als $r \times s^r$ viele Regelanwendungen in der Ableitung von x.
  - Bei jeder Regelanwendungen werden höchstens s Terminal-Symbole generiert.
  - Also ist die Länge von x höchtsens $s \times r \times s^r$.

Pumping-Lemma für kontextfreie Sprachen

Wenn $L(G)$ unendlich ist, enthält sie Ketten, die länger sind als $s \times r \times s^r$. Der zugehörige Syntaxbaum enthält dann mindestens ein Nichtterminal-Symbol, was sich selbst dominiert. Präziser gesagt: es gibt zwei Knoten α und β, die mit dem selben Nichtterminal-Symbol etikettiert sind, so dass α β dominiert. Daraus ergibt sich das folgende Resultat:

Theorem (Pumping-Lemma für kontextfreie Sprachen)

Sei L eine unendliche kontextfreie Sprache. Dann gibt es eine Zahl n, so dass sich alle Wörter $x \in L$ zerlegen lassen in $x = u \frown v \frown w \frown y \frown z$, so dass

- $l(v) + l(y) > 0$,
- $l(v) + l(w) + l(y) \leq n$, *und*
- *für alle* $i \in \mathbb{N} : u \frown v^i \frown w \frown y^i \frown z \in L$.

NL and the Chomsky Hierarchy

The respectively argument

- Bar-Hillel and Shamir (1960):
  - English contains copy-language
  - cannot be context-free
- Consider the sentence

  John, Mary, David, ... are a widower, a widow, a widower, ..., respectively.
- Claim: the sentence is only grammatical under the condition that if the nth name is male (female) then the nth phrase after the copula is *a widower (a widow)*

NL and the Chomsky Hierarchy

- suppose the claim is true
- intersect English with regular language

$$L_1 = (\textit{Paul}|\textit{Paula})^+ \textit{are}[(\textit{a widower}|\textit{a widow})^+ \textit{respectively}$$

English $\cap L_1 = L_2$

- homomorphism $L_2 \rightsquigarrow L_3$:

$$\begin{aligned} \textit{John, David, Paul, ...} &\mapsto a \\ \textit{Mary, Paula, Betty, ...} &\mapsto b \\ \textit{a widower} &\mapsto a \\ \textit{a widow} &\mapsto b \\ \textit{are, respectively} &\mapsto \epsilon \end{aligned}$$

NL and the Chomsky Hierarchy

- result: copy language L_3

$$\{ww|w \in (a|b)^+\}$$

- copy language is not cf due to pumping lemma (exercise: why is this so?)
- hence L_2 is not cf
- hence English is not cf

NL and the Chomsky Hierarchy

Counterargument

- crossing dependencies triggered by *respectively* are semantic rather than syntactic
- compare above example to

  (Here are John, Mary and David.) They are a widower, a widow and a widower, respectively.

Cross-serial dependencies in Dutch

- Huybregt (1976):
  - Dutch has copy-language like structures
  - thus Dutch is not context-free

(1) dat Jan Marie Pieter Arabisch laat zien schrijven
THAT JAN MARIE PIETER ARABIC LET SEE WRITE
'that Jan let Marie see Pieter write Arabic'

NL and the Chomsky Hierarchy

Counterargument

- crossing dependencies only concern argument linking, i.e. semantics
- Dutch has no case distinctions
- as far as plain string are concerned, the relevant fragment of Dutch has the structure

$$NP^n V^n$$

which is context-free

Sind natürliche Sprachen kontextfrei?

- definitives Argument (Huybregts 1985, Shieber 1987): Schweizerdeutsch ist nicht kontextfrei
- grundsätzliche Einsichten:
  - kontextfreie Grammatiken können beliebig tief **geschachtelte** Abhängigkeiten beschreiben
  - kontextfreie Grammatiken können keine beliebig lange **überkreuzende** Abhängigkeiten beschreiben
  - in natürlichen Sprachen kommen, wenn auch mariginal, überkreuzende Abhängigkeiten vor

Sind natürliche Sprachen kontextfrei?

- Typ-1-Grammatiken („kontext-sensitive Grammatiken“) sind im Allg. zu „mächtig“ für linguistische Zwecke
- **Mild kontextsensitive Grammatiken:** Familie von Grammatikformalismen, die nur unwesentlich mächtiger sind als Typ-2-Grammatiken, aber überkreuzende Abhängigkeiten erfassen
- wichtigste Vertreter:
  - Baum-Adjunktions-Grammatiken (*Tree Adjoining Grammars/TAG*)
  - Kombinatorische Kategorialgrammatik (*Combinatory Categorial Grammar/CCG*