Grundlagen der theoretischen Informatik

Kurt Sieber

Fakultät IV, Department ETI
Universität Siegen

SS 2013

Vorlesung vom 07.05.2013 (Teil 2)

Kontextfreie Sprachen

- Reguläre Sprachen haben eine sehr einfache Struktur.

  Deshalb kann man mit ihnen nur die kleinsten (d.h. einfachsten) syntaktischen Einheiten einer Programmiersprache erfassen,

  z.B. Schlüsselwörter, Zahldarstellungen, Bezeichner, Kommentare.

- Wie beschreibt man die komplizierteren syntaktischen Strukturen,

  z.B. arithmetische und boolesche Ausdrücke, Anweisungen, Deklarationen?

  Am besten durch eine *Konstruktionsvorschrift*,

  d.h. aus mathematischer Sicht: durch eine *induktive Definition*.

Beispiel: Vollständig geklammerte arithmetische Ausdrücke

Konstruktionsvorschrift

1. Jede Dezimalzahl ist ein arithmetischer Ausdruck.
2. Jeder Bezeichner ist ein arithmetischer Ausdruck.
3. Wenn e ein arithmetischer Ausdruck ist, dann ist auch $(-e)$ ein arithmetischer Ausdruck.
4. Wenn e_1, e_2 arithmetische Ausdrücke sind, dann ist auch $(e_1 + e_2)$ ein arithmetischer Ausdruck.
5. Wenn e_1, e_2 arithmetische Ausdrücke sind, dann ist auch $(e_1 - e_2)$ ein arithmetischer Ausdruck.
6. Wenn e_1, e_2 arithmetische Ausdrücke sind, dann ist auch $(e_1 * e_2)$ ein arithmetischer Ausdruck.

Mathematische Formulierung

Die Sprache L der vollständig geklammerten arithmetischen Ausdrücke ist *induktiv definiert* durch:

1. Jede Dezimalzahl liegt in L.
2. Jeder Bezeichner liegt in L.
3. Wenn $e \in L$, dann ist auch $(-e) \in L$.
4. Wenn $e_1, e_2 \in L$, dann ist auch $(e_1 + e_2) \in L$.
5. Wenn $e_1, e_2 \in L$, dann ist auch $(e_1 - e_2) \in L$.
6. Wenn $e_1, e_2 \in L$, dann ist auch $(e_1 * e_2) \in L$.

Was bedeutet '*induktiv definiert*'? Es bedeutet, dass L die *kleinste* Menge mit den Eigenschaften 1. bis 6. ist (eine andere Menge mit diesen Eigenschaften ist z.B. Σ^*).

Kontextfreie Sprachen

Die mathematische Formulierung genügt uns nicht (so wie uns die Mengenausdrücke nicht genügten). Wir brauchen eine *formale Schreibweise* (wie die regulären Ausdrücke).

Definition 2.38 *Eine kontextfreie Grammatik (kurz: KFG) ist ein 4-Tupel $G = (\Sigma, N, S, P)$, wobei gilt:*

- *Σ und N sind Alphabete mit $\Sigma \cap N = \emptyset$. Die Zeichen aus Σ heißen Terminalzeichen, die aus N heißen Nichtterminalzeichen von G (engl.: Nonterminals).*
- *$S \in N$. S heißt Startzeichen von G.*
- *$P \subseteq N \times (\Sigma \cup N)^*$. Die Elemente von P heißen Regeln oder Produktionen von G.*

Eine Produktion ist also ein *Paar* (A, u), wobei A ein Nichtterminalzeichen ist und u ein Wort, das sowohl Terminal- als auch Nichtterminalzeichen enthalten darf. Statt (A, u) schreibt man $A \to u$.

Konvention:
Als Nichtterminalzeichen benutzen wir Großbuchstaben.

Beispiel:

$G = (\Sigma, N, S, P)$ mit

- $\Sigma = \{0, 1, x, y, -, +, *, (,)\}$
- $N = \{E\}$
- $S = E$
- $P = \{E \to 0,\ E \to 1,\ E \to x,\ E \to y,\ E \to (-E),$
  $E \to (E + E),\ E \to (E - E),\ E \to (E * E)\}$

ist eine KFG für vollständig geklammerte arithmetische Ausdrücke (in denen nur die Zahlen $0, 1$ und nur die Bezeichner x, y vorkommen).

Kurzschreibweise:

Man schreibt

$$A \to u_1 \mid \ldots \mid u_n$$

als Abkürzung für eine *Menge* von Produktionen

$$A \to u_1, \ldots, A \to u_n$$

mit dem gleichen Nichtterminalzeichen auf der linken Seite.

Im Beispiel kann man also *alle* Produktionen zusammenfassen zu

$$E \to 0 \mid 1 \mid x \mid y \mid (-E) \mid (E + E) \mid (E - E) \mid (E * E)$$

wobei man "|" als "oder" liest.

Kontextfreie Sprachen

Wie ist eine KFG zu verstehen, d.h. welche Sprache beschreibt sie?

- Entweder als *Konstruktionsvorschrift* für eine Sprache:

  Dann dienen die Produktionen dazu, die Wörter der Sprache *abzuleiten* oder zu *erzeugen*.

- Oder als *induktive Definition* einer Sprache:

  Dann liest man die Produktionen als *Regeln*, die die Sprache erfüllen muss.

Beide Auffassungen sind äquivalent zueinander, üblich ist aber die Formulierung als Konstruktionsvorschrift.

Definition 2.39 *Sei $G = (\Sigma, N, S, P)$ eine KFG.*

Auf der Menge $(\Sigma \cup N)^*$ definieren wir eine Relation $\Rightarrow_G$ (oder kürzer: $\Rightarrow$) durch:

$$u \Rightarrow_G v \quad \Leftrightarrow \quad \begin{array}{l} \text{es gibt eine Produktion } A \to y \text{ in } P \\ \text{und Wörter } x, z \in (\Sigma \cup N)^* \\ \text{so dass } u = xAz \text{ und } v = xyz \end{array}$$

Man bezeichnet $u \Rightarrow_G v$ als einen Ableitungsschritt (in G).

In Worten: Ein Ableitungsschritt $u \Rightarrow_G v$ besteht darin, ein Vorkommen eines Nichtterminalzeichens A im Wort u durch die rechte Seite y einer Produktion $A \to y$ zu ersetzen.

Mit $\overset{n}{\Rightarrow}_G$ $(n \geq 0)$, $\overset{+}{\Rightarrow}_G$ und $\overset{*}{\Rightarrow}_G$ bezeichnen wir wieder die n-te Potenz, den transitiven Abschluss und den reflexiven, transitiven Abschluss der Relation $\Rightarrow_G$.

Definition 2.40 *Sei G eine KFG.*

1. *v heißt ableitbar aus u (in G), wenn $u \overset{*}{\Rightarrow}_G v$, d.h. wenn eine Folge $u = w_0 \Rightarrow_G \ldots \Rightarrow_G w_n = v$ $(n \geq 0)$ von Ableitungsschritten in G existiert.*

   Eine solche Folge bezeichnet man als Ableitung von v aus u.

2. *Die von G erzeugte Sprache $L(G)$ ist definiert durch*

$$L(G) \;=\; \{w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w\}$$

   $L(G)$ besteht also aus allen Wörtern über dem Terminalalphabet Σ, die aus dem Startzeichen S ableitbar sind.

Definition 2.41 *Eine Sprache $L \subseteq \Sigma^*$ heißt* kontextfrei, *wenn es eine kontextfreie Grammatik G gibt mit $L = L(G)$.*

Beispiel:

Sei G die oben definierte KFG für arithmetische Ausdrücke.

Dann gilt z.B.
$E \Rightarrow (E+E) \Rightarrow ((E*E)+E) \Rightarrow ((x*E)+E) \Rightarrow ((x*y)+E) \Rightarrow ((x*y)+1)$
also $((x * y) + 1) \in L(G)$.

Die Sprache $L(G)$ ist (per Definition) kontextfrei.

Aber ist $L(G)$ die gewünschte Sprache L der vollständig geklammerten arithmetischen Ausdrücke?
Um diese Frage zu beantworten, bräuchte man eine Definition für L, die unabhängig von der Grammatik ist. Die ist schwer zu finden!
Also stellen wir uns auf den Standpunkt, dass L durch die Grammatik G *definiert* ist. Dann ist nichts mehr zu beweisen.

Kontextfreie Sprachen

Sei G wie oben.

Neben der bereits genannten Ableitung
$E \Rightarrow (E+E) \Rightarrow ((E*E)+E) \Rightarrow ((x*E)+E) \Rightarrow ((x*y)+E) \Rightarrow ((x*y)+1)$
gibt es andere Ableitungen für das gleiche Wort, z.B.
$E \Rightarrow (E+E) \Rightarrow (E+1) \Rightarrow ((E*E)+1) \Rightarrow ((E*y)+1) \Rightarrow ((x*y)+1)$

In beiden Ableitungen werden die 'gleichen' Ableitungsschritte in unterschiedlicher Reihenfolge benutzt.

Die Reihenfolge der Ableitungsschritte spielt keine Rolle, weil die Ableitungsschritte in einer kontextfreien Grammatik sehr einfach aussehen: Es wird stets ein Nichtterminalzeichen A durch ein Wort y ersetzt. Der *Kontext* des Zeichens A, d.h. der Text vor und hinter A, spielt dabei keine Rolle. Er beeinflusst den Ableitungsschritt nicht, und er wird durch den Ableitungsschritt nicht verändert (daher die Bezeichnung '*kontextfrei*'). Also ist es unerheblich, ob man zuerst A oder zuerst ein Nichtterminalzeichen im Kontext von A ersetzt.

Diese Überlegungen kann man präzisieren,

- entweder indem man eine spezielle Reihenfolge für die Ableitungsschritte *vorschreibt*, und dann zeigt, dass jede Ableitung so umgeformt werden kann, dass sie der Vorschrift genügt,
- oder indem man eine Schreibweise benutzt, die von der Reihenfolge der Ableitungsschritte *abstrahiert*.

Definition 2.42 *Ein Ableitungsschritt $u \Rightarrow v$ heißt* Linksableitungsschritt, *wenn es Wörter $x \in \Sigma^*$, $z \in (N \cup \Sigma)^*$ und eine Produktion $A \to y$ in G gibt mit $u = xAz$ und $v = xyz$. Eine* Linksableitung *ist eine Ableitung, die nur aus Linksableitungsschritten besteht.*

Eine Linksableitung ist also eine Ableitung, bei der stets das erste, d.h. das am weitesten links stehende Nichtterminalzeichen ersetzt wird.

Ein Beispiel für eine Linksableitung haben wir schon gesehen:

$E \Rightarrow (E+E) \Rightarrow ((E*E)+E) \Rightarrow ((x*E)+E) \Rightarrow ((x*y)+E) \Rightarrow ((x*y)+1)$

Das folgende Lemma besagt, dass Linksableitungen ausreichen.

Lemma 2.43 *$u \stackrel{*}{\Rightarrow}_G v$ gilt genau dann, wenn es eine Linksableitung von v aus u in G gibt.*

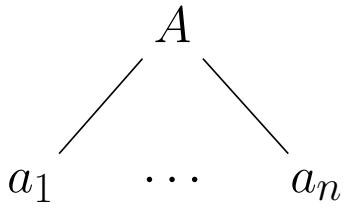
Beweisidee:

Man kann die Schritte einer beliebigen Ableitung so umordnen, dass eine Linksableitung entsteht. Die Details sind mühsam! □

Man kann Lemma 2.43 so interpretieren:
Wenn wir beliebige Ableitungen betrachten, so gibt es (fast in jeder Grammatik) viele Ableitungen, die sich nur 'unwesentlich' unterscheiden, nämlich nur in der Reihenfolge der Ersetzungen. Betrachten wir nur Linksableitungen, so entfallen diese unwesentlichen Unterschiede, weil die Reihenfolge der Ableitungsschritte fest vorgeschrieben ist.

Ein alternativer Ansatz besteht darin, von der Reihenfolge der Ableitungsschritte zu abstrahieren, indem man *Ableitungsbäume* betrachtet.

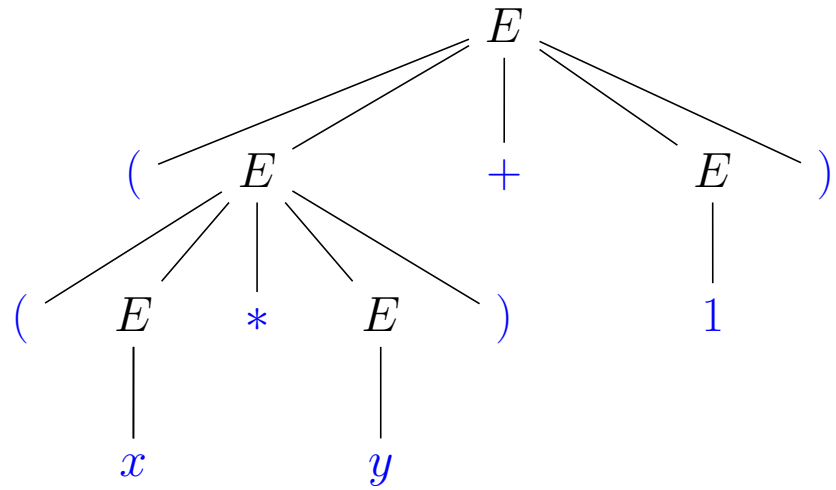
Definition 2.44 *Ein* Ableitungsbaum *(oder* Syntaxbaum*) für die Grammatik* $G = (\Sigma, N, S, P)$ *ist ein Baum, dessen Knoten mit Zeichen aus* $\Sigma \cup N$ *markiert sind, und zwar so, dass jeder innere Knoten die Form*



hat, wobei $A \to a_1 \dots a_n$ *eine Produktion in* P *ist.*

Beispiel:

Ein Ableitungsbaum für unsere Grammatik G ist



An diesem Baum sieht man, dass $((x * y) + 1)$ aus E ableitbar ist (ohne dass die Reihenfolge der Ableitungsschritte festgelegt wird).