Grammatik

Chr. Wagenknecht & M. Hielscher

15. April 2008

Sprache

Rückblick

Sprache

- Eine Sprache L ist eine endliche oder im Allg. eine abzählbar unendliche Teilmenge der Wortmenge A^* über einem bestimmten Alphabet A. Die Elemente jeder Sprache sind Wörter aus A^*.

Auswahlprozess

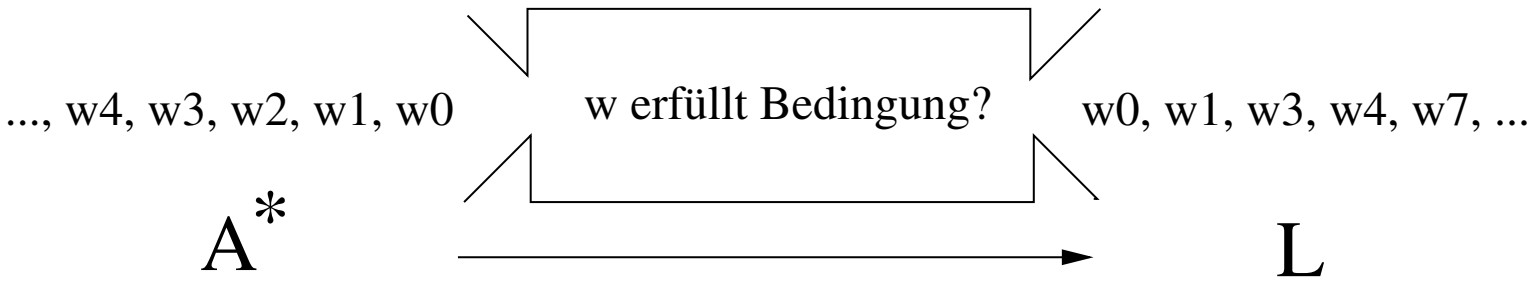
- Zur Definition einer bestimmten Sprache findet ein *Auswahlprozess* statt.
- Für endliche Sprachen werden konkrete Wörter aus A^* einzeln herauszugreifen.
- Zur Bestimmung unendlicher Sprachen braucht man so etwas wie einen „Auswahlprozessor“.

Sprache

Auswahlprozess

Auswahlprozess

- Auf jedes Wort aus A^* wird dieser angewandt $\Rightarrow$ er nimmt genau die Wörter in L auf, die die entsprechende Auswahlbedingung erfüllen.



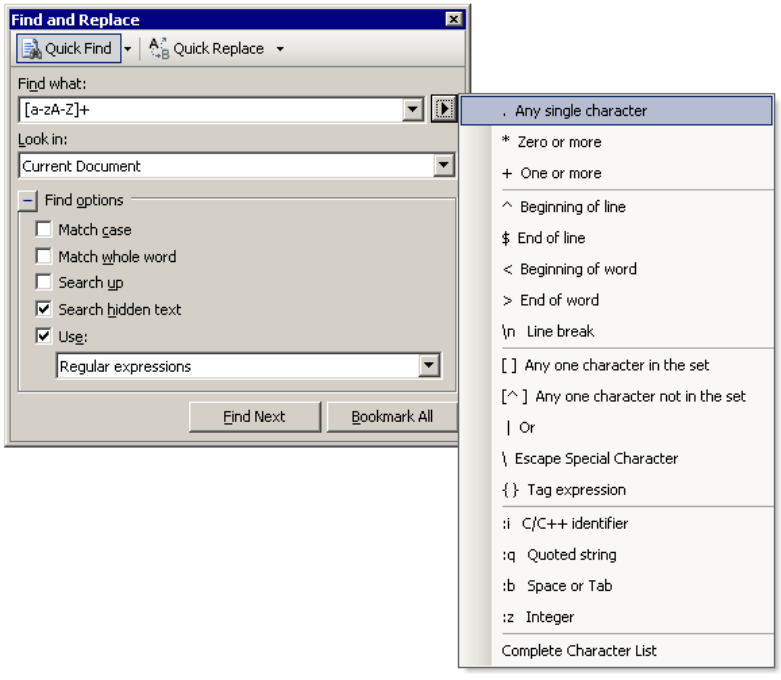
- Da A^* abzählbar ist, könnten wir alle Wörter aus A^* durch diesen Auswahlprozess schicken, um so L aufzuzählen.

Sprache

Auswahlprozess

Auswahlprozess

- Die *Auswahlbedingung* charakterisiert die Sprache L über A vollständig.
- Wir können uns einen Filter vorstellen der wie eine Schablone oder Muster funktioniert. (wie in Text/Entwicklungs-Editoren)



Sprache

Auswahlprozess

Auswahlprozess

- Der Auswahlprozess wird deshalb als *Pattern matching* (Mustervergleich) bezeichnet.
- Das zu beurteilende Wort w aus A^* wird mit dem Muster verglichen und wenn es passt in L aufgenommen ansonsten verworfen.
- Das Verfahren ist bequem, aber wird nicht zur Beschreibung beliebiger Sprachen ausreichen!

Grammatik

- Die Beschreibung einer Sprache durch *Regeln* (Woche 1 - Konkrete Syntax) kann auch als Auswahlprozessor verwendet werden.

Formale Grammatik

Definition

Grammatik

- Die intuitive Anwendung dieser Regeln haben wir uns in Ableitungsbäumen angesehen.

Definition

Eine (formale) *Grammatik* G ist ein 4-Tupel
$G = (N, T, P, s)$ mit folgenden Eigenschaften:

N ... Menge der *Nichtterminale* (grammatikalische Variablen)

T ... Menge der *Terminale* (Alphabetzeichen)

N, T sind nichtleere, endliche und disjunkte Mengen, d.h. $N \cap T = \emptyset$.

P ... endliche Menge von *Regeln* oder *Produktionen*

$P = \{\alpha \to \beta \mid \alpha \in (N \cup T)^* \setminus T^* \text{ und } \beta \in (N \cup T)^*\}$

s ... *Start-* oder *Satz-* oder *Spitzensymbol*, wobei $s \in N$.

Formale Grammatik

Bestandteile einer Grammatik

Bestandteile

- In der Grammatik ist das Alphabet die sogenannte Terminalmenge T.
- Die Nichtterminale N sind Platzhalter wie etwa Artikel für die Terminale `der`, `die` und `das`.
- Die Menge $N \cup T$ wird häufig als *Vokabular* V bezeichnet.
- $(N \cup T)^*$ bezeichnet die Menge aller *Zeichenketten über dem Vokabular*.
- Dies sind „gemischte Zeichenketten“, die aus beliebig vielen Nichtterminalen und Terminalen in beliebiger Reihenfolge bestehen. Man nennt sie *Satzformen*.

Formale Grammatik

Gestalt der Produktionsregeln

Produktionsregeln

- In der Grammatik-Definition wird eine sehr freizügige Regelgestalt $\alpha \to \beta$ erlaubt.
- Auf der rechten Seite von Produktionen (β) dürfen beliebige Satzformen stehen. Links (α) sind lediglich Folgen von Terminalen ausgeschlossen.

Beispiel

- Beispiel zeigt eine Grammatik mit uneingeschränkter Regelgestalt. In der Notation orientieren wir uns an der BNF.
- Nicht konktextfreie Grammatik (kfG) $G = (N, T, P, s)$, mit

$$\begin{aligned} N &= \{S,\ A,\ B\}\ T = \{\mathtt{a},\ \mathtt{b},\ \mathtt{c}\} \\ P &= \{S \to AS \mid \mathtt{cc}S\mathtt{b};\ \mathtt{c}S \to \mathtt{a};\ AS \to S\mathtt{bb};\ \mathtt{c}S\mathtt{b} \to \mathtt{c}\} \\ s &= S \end{aligned}$$

Formale Grammatik

Gestalt der Produktionsregeln

Beispiel

- Wir betrachten die folgende sehr einfache Grammatik:
  $G = (N, T, P, s)$, mit

$$
\begin{aligned}
N &= \{\mathit{Number},\ \mathit{Digit}\} \\
T &= \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} \\
P &= \{\mathit{Number} \longrightarrow \mathit{Digit} \mid \mathit{DigitNumber} \\
&\qquad \mathit{Digit} \longrightarrow 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9\} \\
s &= \mathit{Number}
\end{aligned}
$$

- Diese Grammatik ist eine kfG: Sämtliche linken Regelseiten bestehen aus genau einem Nichtterminal.
- Anhand des Regelaufbaus kann man verschiedene *Sprachklassen* unterscheiden.
- Wir konzentrieren uns zunächst auf kfG (die im Compilerbau wichtigste Sprachklasse).

Formale Grammatik

Übung

Übung 1

- Entwickeln Sie eine kfG für die Sprache der Uhrzeit im Format: `SS:MM`
- Beginnen Sie mit der Definition von T.

Formale Grammatik

Übung

Übung 1

- Entwickeln Sie eine kfG für die Sprache der Uhrzeit im Format: `SS:MM`
- Beginnen Sie mit der Definition von T.

- Mögliche Lösung:

  Uhrzeit $\longrightarrow$ Stunden : Minuten

  Stunden $\longrightarrow$ 0 Z0bis9 | 1 Z0bis9 | 2 Z0bis3

  Minuten $\longrightarrow$ Z0bis5 Z0bis9

  Z0bis3 $\longrightarrow$ 0 | 1 | 2 | 3

  Z0bis5 $\longrightarrow$ 0 | 1 | 2 | 3 | 4 | 5

  Z0bis9 $\longrightarrow$ 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

Ableitung

Definition

Relationen $\Rightarrow_G$ und $\overset{*}{\Rightarrow}_G$

- Mit $u \in (N \cup T)^* \setminus T^*$ und $v \in (N \cup T)^*$ stehen u und v in Relation $\Rightarrow_G$, d.h. $u \Rightarrow_G v$, und man sagt: „u geht unter G unmittelbar über in v“, falls u und v die folgenden Formen besitzen:
  $u = xyz, \quad v = xy'z, \quad x, z \in (N \cup T)^*$ und $(y \to y') \in P$.
- Für $d \Rightarrow_G e \Rightarrow_G \ldots \Rightarrow_G f$ schreibt man verkürzend $d \overset{*}{\Rightarrow}_G f$. Es ist üblich, das an den Doppelpfeil angehängte G wegzulassen und kurz: „u geht unmittelbar über in v“ zu sprechen.
- α steht in Relation $\Rightarrow$ mit β, wenn es in α (mindestens) eine Teilzeichenkette gibt, die mit der linken Seite (mindestens) einer Regel aus P übereinstimmt.

Ableitung

Definition

Ableitung

- Unter Verwendung von $\Rightarrow$ bzw. $\overset{*}{\Rightarrow}$ kann nun der Auswahlprozess für jedes Wort w aus A^* gemäß G als *Ableitung* exakt beschrieben werden.
- Eine Folge von Satzformen $(\alpha_0, \alpha_1, \ldots, \alpha_n)$, mit $\alpha_0 = s$, $\alpha_n = w$, $w \in T^*$ und $s \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \ldots \Rightarrow w$, heißt *Ableitung von w*.
- Wörter, die zu der von G definierten Sprache $L(G)$ gehören, müssen vom Startsymbol aus ableitbar sein.

Zusammenhang G und Sprache L

- Die durch G definierte Sprache $L(G)$ ist $L(G) = \{w \mid w \in T^* \text{ und } s \overset{*}{\Rightarrow}_G w\}$.

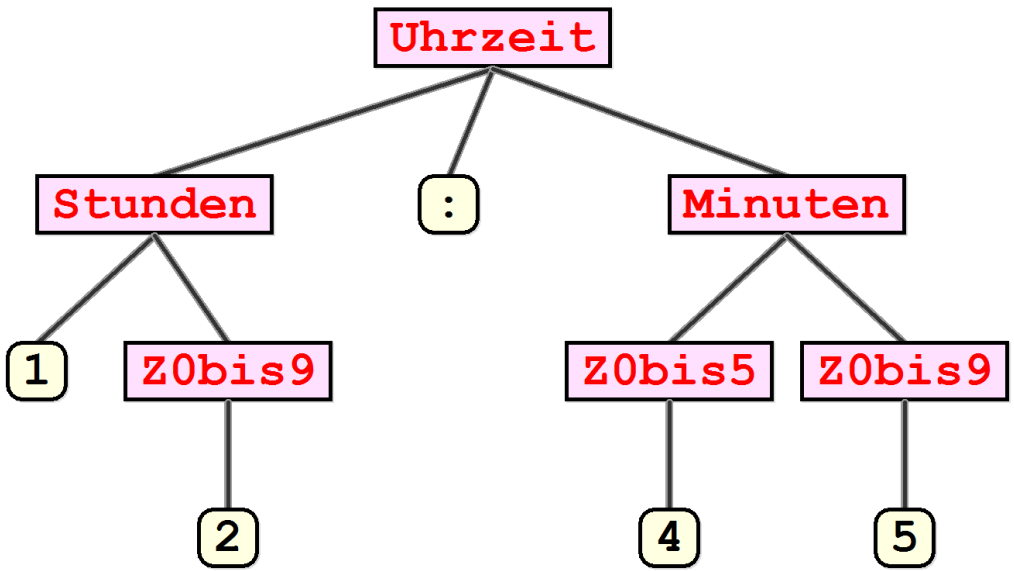
Ableitung

Definition

Am Beispiel Uhrzeit

Zu prüfendes Wort w sei 12:45. Die Ableitung beginnt am Spitzensymbol Uhrzeit:

Uhrzeit $\Rightarrow$ Stunden : Minuten$\Rightarrow$ 1 Z0bis9 : Minuten$\Rightarrow$
1 2 : Minuten$\Rightarrow$ 1 2 : Z0bis5 Z0bis9$\Rightarrow$ 1 2 : 4 Z0bis9$\Rightarrow$ 1 2 : 4 5



Formale Grammatik

Übung

Übung 2

- Entwickeln Sie eine kfG für die Sprache der arithmetischen Ausdrücke für die Grundrechenarten: Addition, Subtraktion, Multiplikation und Division.
- Beginnen Sie wieder mit der Definition von T.

Formale Grammatik

Übung

Übung 2

- Entwickeln Sie eine kfG für die Sprache der arithmetischen Ausdrücke für die Grundrechenarten: Addition, Subtraktion, Multiplikation und Division.
- Beginnen Sie wieder mit der Definition von T.
- Mögliche Lösung:

  Ausdruck $\longrightarrow$ Ausdruck $+$ Ausdruck
  Ausdruck $\longrightarrow$ Ausdruck - Ausdruck
  Ausdruck $\longrightarrow$ Ausdruck * Ausdruck
  Ausdruck $\longrightarrow$ Ausdruck / Ausdruck
  Ausdruck $\longrightarrow$ Zahl
  Zahl $\longrightarrow$ Ziffer $\mid$ Ziffer Zahl
  Ziffer $\longrightarrow$ 0 $\mid$ 1 $\mid$ 2 $\mid$ 3 $\mid$ 4 $\mid$ 5 $\mid$ 6 $\mid$ 7 $\mid$ 8 $\mid$ 9
- Welchen Nachteil hat diese Grammatik?

Formale Grammatik

Übung

Übung 3

- Ausdruck → Ausdruck + Ausdruck
  Ausdruck → Ausdruck - Ausdruck
  Ausdruck → Ausdruck * Ausdruck
  Ausdruck → Ausdruck / Ausdruck
  Ausdruck → Zahl
  Zahl → Ziffer | Ziffer Zahl
  Ziffer → 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
- Leiten Sie das Wort 23+1-8 ab (notieren Sie alle Satzformen die Wahrend der Ableitung entstehen).

Ableitungsprozesses

Nichtdeterminismus des Ableitungsprozesses

Linksableitung

- Wir haben bislang: In jeder „abzuleitenden“ Satzform das am weitesten links stehende Nichtterminal im nächsten Schritt (durch eine entsprechende rechte Regelseite) ersetzt.
- Man bezeichnet diese Verabredung deshalb als *Linksableitung*.

Rechtsableitung

- Wir können aber auch vereinbaren, dass stets das am weitesten rechts stehende Nichtterminal ersetzt wird.
- Zu welchem Schluss kommen wir, wenn wir den Ableitungsbaum von beiden Varianten vergleichen?
- Könnte ein anderes Ergebnis durch beliebige Wahl des zu ersetzenden Nichterminals entstehen?

Ableitungsprozesses

Nichtdeterminismus des Ableitungsprozesses

Übung

- Gegeben sei die Grammatik $G = (N, T, P, s)$, mit $s = S$, $T = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{-}\}$, $N = \{A, S\}$ und $P = \{S \xrightarrow{1} A,\ S \xrightarrow{2} S\mathtt{-}A,\ A \xrightarrow{3} \mathtt{a},\ A \xrightarrow{4} \mathtt{b},\ A \xrightarrow{5} \mathtt{c}\}$.
- Mit Hilfe der Ziffern über den Pfeilen der fünf Produktionen werden wir uns auf die jeweilige Regel beziehen.
- Prüfen Sie ob das Wort $\mathtt{a-b-b}$ zur Sprache $L(G)$ gehört und geben Sie die angewendeten Regeln als Ziffernfolge an.

Ableitungsprozesses

Nichtdeterminismus des Ableitungsprozesses

Übung

- Gegeben sei die Grammatik $G = (N, T, P, s)$, mit $s = S$, $T = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{-}\}$, $N = \{A, S\}$ und $P = \{S \xrightarrow{1} A,\ S \xrightarrow{2} S\mathtt{-}A,\ A \xrightarrow{3} \mathtt{a},\ A \xrightarrow{4} \mathtt{b},\ A \xrightarrow{5} \mathtt{c}\}$.
- Mit Hilfe der Ziffern über den Pfeilen der fünf Produktionen werden wir uns auf die jeweilige Regel beziehen.
- Prüfen Sie ob das Wort $\mathtt{a-b-b}$ zur Sprache $L(G)$ gehört und geben Sie die angewendeten Regeln als Ziffernfolge an.
- $S \Rightarrow S\mathtt{-}A \Rightarrow S\mathtt{-}A\mathtt{-}A \Rightarrow A\mathtt{-}A\mathtt{-}A \Rightarrow \mathtt{a}\mathtt{-}A\mathtt{-}A \Rightarrow \mathtt{a}\mathtt{-}\mathtt{b}\mathtt{-}A \Rightarrow \mathtt{a}\mathtt{-}\mathtt{b}\mathtt{-}\mathtt{b}$ daraus folgt: $\mathtt{a}\mathtt{-}\mathtt{b}\mathtt{-}\mathtt{b} \in L(G)$.
  Die gewählte Ableitung ist $(2, 2, 1, 3, 4, 4)$.
  Die Rechtsableitung ergibt $(2, 4, 2, 4, 1, 3)$.
- Ebenso hätte man $(2, 4, 2, 1, 4, 3)$ nehmen können.

Ableitungsprozesses

Nichtdeterminismus des Ableitungsprozesses

Ersetzungsreihenfolge

- Durch Änderung der Ersetzungsreihenfolge der betreffenden Nichtterminale ergeben sich keine neuen Substitutionsmöglichkeiten.
- Es ergibt sich bei kfG stets das gleiche Ergebnis.
- Aber: Bei einer nicht konktextfreien Grammatik wie
  $G = (N, T, P, s)$, mit
  $N = \{S,\ A,\ B\}\ T = \{\mathtt{a},\ \mathtt{b},\ \mathtt{c}\}$
  $P = \{S \to AS \mid \mathtt{cc}S\mathtt{b};\ \mathtt{c}S \to \mathtt{a};\ AS \to S\mathtt{bb};\ \mathtt{c}S\mathtt{b} \to \mathtt{c}\}$
  $s = S$
  Hier können neuen Konstellationen für Satzformen entstehen, die bei einem veränderten Ableitungsprozess nicht anwendbar wären.

Ableitungsprozesses

Nichtdeterminismus des Ableitungsprozesses

Angewendete Regelen

- Die optisch identischen Bäume sind *dynamisch* verschieden, d.h. sie haben unterschiedliche „Entstehungsgeschichten“ .
- Die angewendeten Regeln sind die gleichen, jedoch in einer anderen Reihenfolge.
- Es ist sinnvoll sich bei kfG auf *Linksableitungen* festzulegen.

Definition Linksableitung

Eine Ableitung für ein Wort w gemäß einer kfG heißt *Linksableitung* genau dann, wenn es für jedes $i = 0, 1, \ldots, n-1$ eine Produktion $Y_i \to Q_i$ derart gibt, dass $\alpha_i = v_i Y_i w_i$, $\alpha_{i+1} = v_i Q_i w_i$, wobei $s = \alpha_0 \Rightarrow \alpha_1 \Rightarrow \ldots \Rightarrow \alpha_n = w$.
Es gilt $n \geq 1$, $Y_i \in N$, Q_i und $w_i \in (N \cup T)^*$ und $v_i \in T^*$.

Ableitungsprozesses

Nichtdeterminismus des Ableitungsprozesses

Schlussfolgerung

Bei „kfS“ gibt es zu jedem Ableitungsbaum *genau eine* Linksableitung.

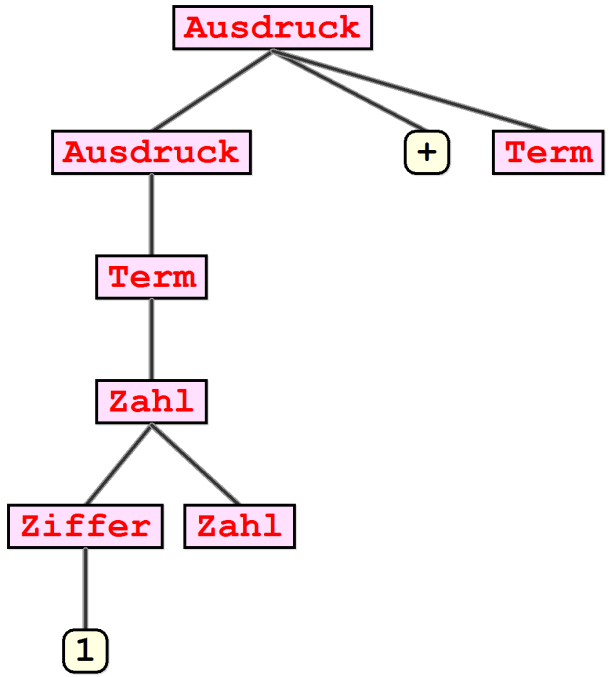
Nichtdeterminismus

- Wir haben uns (meist) im ersten Versuch für die jeweils passende Regel entschieden, auch wenn es mehrere zur Auswahl gab.
- Wie kann dies der Rechner entscheiden? Welche Regel sollte zur Substitution des in einer Satzform ausgewählten Nichtterminals verwendet werden, wenn es mehrere Regeln für dieses Nichtterminal gibt?

Ableitungsprozesses

Nichtdeterminismus des Ableitungsprozesses

Sackgassen



Die Abbildung zeigt den Einstieg in eine *Sackgasse* für das Eingabewort 1+3: Der Ableitungsprozess befindet sich am Ende einer Sackgasse, wenn auf eine aktuelle Satzform keine Regel anwendbar ist.

Ableitungsprozesses

Nichtdeterminismus des Ableitungsprozesses

Sackgassen

- Das Vorliegen einer Sackgasse bedeutet im Allg. nicht, dass das Eingabewort nicht akzeptiert wird.
- Wir müssen die jeweils zuletzt vorgenommene Ersetzung revidieren und – falls möglich – eine alternative Ersetzung durchführen.
- Dieses *Zurückkehren an den letzten Entscheidungspunkt* nennt man *Backtracking*.
- Erst wenn alle Möglichkeiten ausgeschöpft sind und sich darunter keine Ableitung für das zu analysierende Wort w ergab, ist nachgewiesen, dass w nicht zur Sprache gehört.
- Teilweise entstehen unendlich viele Möglichkeiten und wir müssen Abbruchbedingungen definieren.

Mehrdeutigkeit
in kontextfreier Grammatiken

Mehrdeutigkeit

- Jedem Ableitungsbaum kann genau eine Linksableitung zugeordnet werde, aber nicht jede Linksableitung für ein bestimmtes Wort besitzt genau einen Ableitungsbaum (Mehrdeutigkeit).
- Es gibt kfG mit der Eigenschaft, dass es für manche Wörter der zugehörigen Sprache mehrere *strukturell verschiedene* Ableitungsbäume gibt.

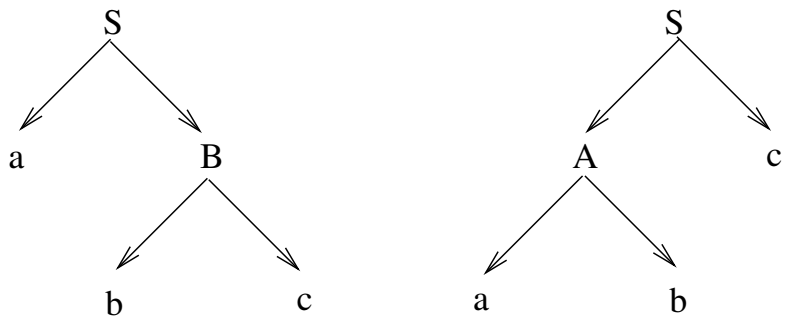
Definition

Eine kfG ist *mehrdeutig*, wenn es ein Wort in $L(G)$ gibt, das zwei Linksableitungen besitzt. Andernfalls ist die Grammatik eindeutig. Eine kfS L ist *eindeutig*, wenn es (mindestens) eine eindeutige kfG G, mit $L = L(G)$, gibt. Ansonsten ist L (inhärent) *mehrdeutig*.

Mehrdeutigkeit
in kontextfreier Grammatiken

Beispiel

Gegeben sei die Grammatik $G = (N, T, P, s)$, mit $s = S$,
$T = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, $N = \{A, B, S\}$ und
$P = \{S \xrightarrow{1} \mathtt{a}B,\ S \xrightarrow{2} A\mathtt{c},\ A \xrightarrow{3} \mathtt{ab},\ B \xrightarrow{4} \mathtt{bc}\}$.
Für das Wort abc gibt es zwei Linksableitungen:
$S \overset{1}{\Rightarrow} \mathtt{a}B \overset{4}{\Rightarrow} \mathtt{abc}$ und $S \overset{2}{\Rightarrow} A\mathtt{c} \overset{3}{\Rightarrow} \mathtt{abc}$



Dieses Beispiel sollte uns an die enge Verbindung von AST (bei prinzipiell erfolgreicher Syntaxanalyse) und Semantik erinnern. Mehrdeutige Grammatiken sind daher äußerst unerwünscht bzw. unbrauchbar.

Mehrdeutigkeit
in kontextfreier Grammatiken

Mehrdeutigkeit

- Leider gibt es *keinen* Universalalgorithmus, der für eine beliebige kfS entscheidet, ob sie mehrdeutig ist. Im Einzelfall gelingt es aber, solche Mehrdeutigkeiten durch Angabe einer *äquivalenten Grammatik*zu eliminieren.
- Allerdings kann nicht zu jeder mehrdeutigen kfG eine äquivalente eindeutige angegeben werden. Gelingt dies nicht, spricht man von *inhärenter Mehrdeutigkeit*.

Definition *äquivalente Grammatik*

Zwei Grammatiken G_1 und G_2 heißen *äquivalent*, geschrieben: $G_1 \cong G_2$, wenn die zugehörigen erzeugbaren Sprachen übereinstimmen, d.h. wenn $L(G_1) = L(G_2)$.

Mehrdeutigkeit
in kontextfreier Grammatiken

Übung

Ausdruck → Ausdruck + Ausdruck
Ausdruck → Ausdruck - Ausdruck
Ausdruck → Ausdruck * Ausdruck
Ausdruck → Ausdruck / Ausdruck
Ausdruck → Zahl
Zahl → Ziffer | Ziffer Zahl
Ziffer → 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

Formen Sie die Grammatik so um, dass keine Mehrdeutigkeiten mehr entstehen.

Mehrdeutigkeit
in kontextfreier Grammatiken

Übung

Eine mögliche Lösung:

```
Ausdruck -> Ausdruck + Term
          | Ausdruck - Term
          | Term
Term     -> Term * Zahl
          | Term / Zahl
          | Zahl
Zahl     -> Ziffer | Ziffer Zahl
Ziffer   -> 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
```