8 Syntax von Programmiersprachen und Syntaxanalyse

Mit Hilfe des Pumping Lemmas wurde im vorigen Abschnitt gezeigt, dass schon einfachste Klammerstrukturen mit n öffnenden Klammern (repräsentiert durch a) gefolgt von n schließenden Klammern (repräsentiert durch b) für beliebige $n \in \mathbb{N}$ nicht von endlichen Automaten erkannt werden können. Das ist auch anschaulich klar. Denn beim Lesen von links nach rechts kann sich ein endlicher Automat mit seiner beschränkten Zahl von Zuständen nur beschränkt viele öffnende Klammern merken, während das Eingabewort beliebig viele enthalten kann, so dass später kein Vergleich mehr mit der Zahl der schließenden Klammern möglich ist. Da Klammerstrukturen in praktisch allen Programmiersprachen vielfältig vorkommen, für die dann eine analoge Beweisführung möglich ist wie für die einfachen Klammerstrukturen, stellt sich heraus, dass Programme von Programmiersprachen in der Regel nicht durch endliche Automaten erkannt werden können.

Wer ein Programm in einer Programmiersprache X schreiben möchte, wählt sich häufig einen Texteditor Y aus und erstellt mit dessen Hilfe eine bestimmte Zeichenkette, die dann als Eingabe für den Compiler von X dient. Dies ist in Abbildung 5 skizziert.

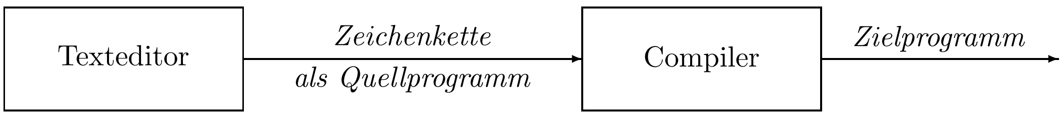

Abbildung 5: Erstellung eines Programms

Man kann nicht erwarten, dass jeder Text, der mit dem Editor Y entstehen kann, bereits ein Programm in X ist. Denn mit einem Texteditor lassen sich in der Regel beliebige Zeichenketten aufbauen, während ein Programm eine bestimmte Form haben muss. Zeichenketten jedoch, die keine Programme sind, wird der Compiler als unübersetzbar zurückweisen. Woher weiß aber eine Programmiererin oder ein Programmierer, wie die Zeichenkette aussehen muss, um ein Programm zu sein? Wie findet der Compiler heraus, ob irgendeine Zeichenkette ein übersetzbares Programm darstellt?

Die Form von Programmen einer Programmiersprache wird durch ihre Syntax festgelegt. Die Syntaxdefinition macht meist äußerst restriktive Vorschriften über die Anordnung und Plazierung von Zeichen, damit eine Zeichenkette ein syntaktisch richtiges Programm ist. Eine Person, die ein Programm mit Hilfe eines Texteditors erstellen will, sollte die Syntax recht gut kennen, weil andernfalls sicherlich häufig Syntaxfehler auftreten. Der Compiler dagegen übersetzt die eingegebene Zeichenkette in ein Zielprogamm, falls die Eingabe ein syntaktisch korrekt gebildetes Programm ist. Um das festzustellen, besitzen Compiler eine Syntaxanalyse-Komponente, die diese Aufgabe übernimmt.

Bei der Syntaxanalyse wird für eingegebene Zeichenketten untersucht, ob sie Programme sind oder nicht. Im positiven Fall wird außerdem der syntaktische Aufbau ermittelt, weil diese Information bei der weiteren Übersetzung maßgeblich genutzt wird. Im negativen Fall werden meist noch Hinweise auf Syntaxfehler gegeben. Die Trennung in "richtige" und "falsche" Zeichenketten ist in der Regel das entscheidende algorithmische Problem, weil bei der Lösung die zusätzlichen Informationen ohne allzu große Mühe nebenbei gewonnen werden können.

Die syntaktisch richtigen Programme einer Programmiersprache bilden als Menge von Zeichenketten eine "formale Sprache". Solche Zeichenkettenmengen sind Gegenstand der Untersuchung in der Theorie formaler Sprachen, die Konzepte für die syntaktische Definition formaler Sprachen bereitstellt und Methoden liefert, um die Eigenschaften formaler Sprachen analysieren zu können. Zu den wichtigsten Anwendungsfeldern der Theorie formaler Sprachen gehören die Syntaxdefinition von Programmiersprachen und die Syntaxanalyse. Die Schlüsselfrage der Syntaxanalyse, ob eine Zeichenkette ein Programm ist oder nicht, wird auch *Wortproblem* genannt. Betrachtet man die Menge aller richtigen Programme als formale Sprache, dann besteht das Problem darin, ob eine Zeichenkette in der Sprache liegt oder nicht. Da die Lösung des Wortproblems eine zentrale Rolle bei der Implementierung von Programmiersprachen spielt, aber keineswegs immer auf der Hand liegt, zieht sich die Behandlung des Wortproblems wie ein roter Faden durch die Theorie formaler Sprachen.

Aber erst einmal zurück zur Syntaxdefinition. Eine grundlegende Weise, formale Sprachen, einschließlich Programmiersprachen, zu spezifizieren, besteht darin, die übliche Art der Begriffsbildung (unter *dem und dem* versteht man *das und das*) in formalisierter Form zu nutzen. Einem zu definierenden, also noch undefinierten, syntaktischen Konstrukt wird ein definierender Ausdruck zugeordnet. Beides zusammen bildet eine Syntaxregel, das (noch) Undefinierte wird linke, das Definierende rechte Regelseite genannt. Der definierende Ausdruck der rechten Seite ist eine Zeichenkette, die rekursiv auch wieder undefinierte Konstrukte enthalten darf. Ist das noch Undefinierte der linken Seite durch ein einziges Zeichen dargestellt, spricht man von einer kontextfreien Regel. Die Syntaxdefinition einer Programmiersprache besteht in einem ersten Anlauf meist in der Angabe von kontextfreien Regeln, die dann später um Syntaxteile ergänzt werden, die sich nicht durch kontextfreie Regeln ausdrücken lassen.

Der kontextfreie Anteil der Syntax von Programmiersprachen wird häufig in der sogenannten Backus-Naur-Form geschrieben, wobei die kontextfreien Regeln als linke Seiten nichtterminale Zeichen, die zu definierende syntaktische Konstrukte der Sprache benennen, und als rechte Seiten Zeichenketten aus terminalen und nichtterminalen Zeichen besitzen. Die rechten Seiten zur selben linken Seite werden als Alternativen nebeneinandergestellt, durch einen senkrechten Strich voneinander getrennt. Linke und rechte Seiten werden durch das Trennzeichen "::=" auseinandergehalten. Nichtterminale Zeichen sind in spitze Klammern eingeschlossen.

Für die Beschreibung der Form von Programmen – oder wie man auch sagt: ihrer Syntax – werden also andere Mittel gebraucht. Sehr häufig werden dafür kontextfreie Grammatiken verwendet, die in ähnlicher Form auch für die grammatikalische Beschreibung natürlicher

Sprachen eingesetzt werden. In diesem Abschnitt werden solche Grammatiken motiviert und informell eingeführt. Die formale Behandlung folgt dann in den folgenden Abschnitten.

Um das Prinzip zu illustrieren, sollen Boolesche Ausdrücke einfacher Art beschrieben werden. Ein solcher Ausdruck kann eine der Booleschen Konstanten $true$ oder $false$ sein, eine Boolesche Variable, ein negierter Boolescher Ausdruck oder die Komposition zweier Boolescher Ausdrücke durch einen Booleschen Operator. In den beiden letzten Fällen werden jeweils Klammern verwendet, damit Anfang und Ende der Ausdrücke eindeutig festgelegt sind. Neben den zu definierenden Ausdrücken sind auch die Variablen ein syntaktisches Konstrukt, das definiert werden muss. Der Einfachheit halber geschieht das durch jeweils ein $''b''$ gefolgt von einer Ziffernfolge. Desweiteren sind Boolesche Operatoren und Ziffernfolge sowie Ziffern als syntaktische Konstrukte eingeführt und definiert. Die folgenden Syntaxregeln reflektieren die verbale Beschreibung:

$$
\begin{array}{l}
< boolexp >::= true \mid false \mid < var > \mid \\
(\neg < boolexpr >) \mid (< boolexpr >< boolop >< boolexpr >) \\
< boolop >::= \wedge \mid \vee \mid \Rightarrow \mid \Leftrightarrow \\
< var >::= b < cipherseq > \\
< cipherseq >::=< cipher > \mid < cipher >< cipherseq > \\
< cipher >::= 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9
\end{array}
$$

Die Regeln können zum Aufbau syntaktisch korrekter Boolescher Ausdrücke benutzt werden, indem mit dem nichtterminalen Zeichen $< boolexpr >$ begonnen und dann nach und nach in den aktuellen Zeichenketten je ein nichtterminales Zeichen durch eine zugehörige rechte Regelseite ersetzt wird, bis keine nichtterminalen Zeichen mehr vorkommen. Ein Beispiel dafür ist:

$$
\begin{aligned}
< boolexp > &\rightarrow (< boolexpr >< boloop >< boolexpr >) \\
&\rightarrow (< boolexpr > \Leftrightarrow < boolexpr >) \\
&\rightarrow (< boolexpr > \Leftrightarrow true) \\
&\rightarrow ((< boolexpr >< boloop >< boolexpr >) \ \Leftrightarrow true) \\
&\rightarrow ((< boolexpr > \vee < boolexpr >) \Leftrightarrow true) \\
&\rightarrow ((< boolexpr > \vee (\neg < boolexpr >)) \ \Leftrightarrow true) \\
&\rightarrow ((< var > \vee (\neg < boolexpr >)) \ \Leftrightarrow true) \\
&\rightarrow ((< var > \vee (\neg < var >)) \Leftrightarrow true) \\
&\rightarrow ((b < cipherseq > \vee (\neg < var >)) \ \Leftrightarrow true) \\
&\rightarrow ((b < cipherseq > \vee (\neg\ b < cipherseq >)) \ \Leftrightarrow true) \\
&\rightarrow ((b < cipher > \vee (\neg\ b < cipherseq >)) \ \Leftrightarrow true) \\
&\rightarrow ((b < cipher > \vee (\neg\ b < cipher >)) \ \Leftrightarrow true) \\
&\rightarrow ((b < cipher > \vee (\neg\ b\ 0)) \ \Leftrightarrow true) \\
&\rightarrow ((b\ 0 \vee \ (\neg\ b\ 0)) \ \Leftrightarrow true)
\end{aligned}
$$

Wie dieses Gesetz vom ausgeschlossenen Dritten lassen sich alle Booleschen Ausdrücke mit Hilfe der Regeln herleiten.

9 Kontextfreie Grammatiken

Die im vorigen Abschnitt erläuterte Art, Zeichenketten einer bestimmten Form durch Syntaxregeln festzulegen, wird hier mit dem Konzept kontextfreier Grammatiken und ihrer erzeugten Sprachen formal definiert.

9.1 Definition kontextfreier Grammatiken

Eine kontextfreie Grammatik besteht aus endlich vielen Regeln, die in der Literatur oft auch Produktionen genannt werden. Eine kontextfreie Regel hat die Form $A ::= u$, wobei A ein nichtterminales Zeichen ist und u ein Wort, das aus nichtterminalen und terminalen Zeichen zusammengesetzt sein kann. Außerdem gibt es ein nichtterminales Startsymbol.

1. Eine *kontextfreie Grammatik* ist ein System $G = (N, T, P, S)$, wobei N eine Menge *nichtterminaler Zeichen*, T eine Menge *terminaler Zeichen*, P eine endliche Menge kontextfreier Produktionen und $S \in N$ ein *Startsymbol* ist. Dabei ist eine kontextfreie Produktion (Regel) ein Paar $(A, u) \in N \times (N \cup T)^*$, das meist als $A ::= u$ geschrieben wird.
2. Das Zeichen A wird *linke Seite*, die Zeichenkette u *rechte Seite* von p genannt.
   Zur Abkürzung können mehrere Produktionen $A ::= u_1, \ldots, A ::= u_k$ $(k \geq 2)$ mit derselben linken Seite zu $A ::= u_1 \mid \cdots \mid u_k$ zusammengefasst werden.
   Soweit nichts anderes gesagt wird, nimmt man an, dass kein nichtterminales Zeichen gleichzeitig terminal ist, d.h. $N \cap T = \emptyset$.

9.2 Ableitungsprozess

Produktionen werden auf Zeichenketten angewendet indem man in einer Zeichenkette ein Zeichen sucht, das die linke Seite einer Produktion ist, und es durch die rechte Seite ersetzt.

1. Seien $w, w', x, y, u, v \in (N \cup T)^*$. Dann wird w' aus w *direkt* durch Anwendung der Produktion $p = (A ::= u)$ *abgeleitet*, falls $w = xAy$ und $w' = xuy$. In diesem Falle wird $w \xrightarrow[p]{} w'$ geschrieben.
   Die Anwendung einer Produktion wird *direkte Ableitung* genannt. Ist P eine Menge von Produktionen und $p \in P$, so kann man statt $w \xrightarrow[p]{} w'$ auch $w \xrightarrow[P]{} w'$ schreiben.
2. Die Iteration direkter Ableitungen ergibt das Konzept der *Ableitung*:
   $$w_0 \xrightarrow[p_1]{} w_1 \xrightarrow[p_2]{} \cdots \xrightarrow[p_n]{} w_n$$
   für $w_0, \ldots, w_n \in (N \cup T)^*$ und Produktionen $p_1, \ldots, p_n$ $(n \geq 1)$. Stammen alle angewendeten Produktionen aus P, so kann man die obige Ableitung auch schreiben als $w_0 \xrightarrow[P]{} \cdots \xrightarrow[P]{} w_n$ oder $w_0 \xrightarrow[P]{n} w_n$. Für manche Zwecke ist es sinnvoll, auch *Nullableitungen* zuzulassen: $w \xrightarrow[P]{0} w$ für alle $w \in (N \cup T)^*$. Statt $w \xrightarrow[P]{n} w'$ für $n \in \mathbb{N}$

darf auch $w \xrightarrow[P]{*} w'$ geschrieben werden. Außerdem kann man bei Ableitungen und direkten Ableitungen das Subskript P weglassen, wenn die Produktionsmenge aus dem Kontext klar ist.

9.3 Erzeugte Sprache

Der Ableitungsprozess bildet die operationelle Semantik, die durch eine Produktionsmenge syntaktisch beschrieben ist. Betrachtet man diejenigen Zeichenketten, die aus dem Startsymbol einer kontextfreien Grammatik $G = (N, T, P, S)$ ableitbar sind und nur aus terminalen Zeichen bestehen, so erhält man auf der Basis des Ableitungsprozesses eine erzeugte Sprache.

Sei $G = (N, T, P, S)$ eine kontextfreie Grammatik. Dann enthält die von G *erzeugte Sprache* alle mit Produktionen in P aus dem Startsymbol S ableitbaren terminalen Zeichenketten:

$$L(G) = \{w \in T^* \mid S \xrightarrow[P]{*} w\}.$$

Auf diese Weise stellen kontextfreie Grammatiken ein syntaktisches Instrument dar, um formale Sprachen zu spezifizieren, die dann entsprechend *kontextfreie Sprachen* genannnt werden. Ausführliche Darstellungen von kontextfreien Grammatiken findet man in praktisch jedem Buch über formale Sprachen (siehe z.B. Hopcroft, Ullman und Motwani [HUM01] mit der deutschen Übersetzung [HUM02], Hopcroft und Ullman [HU79] mit der deutschen Übersetzung [HU90], Moll, Arbib und Kfoury [MAK88] und Salomaa [Sal73]). Die Verbindung von kontextfreien Sprachen und der syntaktischen Behandlung von Programmiersprachen wird umfassend in Aho und Ullman [AU72] abgehandelt.

9.4 Beispiele

1. Mit der Produktion $S ::= aS$ lässt sich das Zeichen a hochzählen:

   $$S \to aS \to a^2S \to \cdots \to a^nS.$$

   Entsprechend kann man mit $S ::= a^kS$ ($k \in \mathbb{N}$) ein Vielfaches von k hochzählen. Terminieren lässt sich dieser Vorgang mit $S ::= \lambda$, so dass

   $$L((\{S\}, \{a\}, \{S ::= a^kS \mid \lambda\}, S)) = \{a^{n \cdot k} \mid n \in \mathbb{N}\}.$$

2. Auch das getrennte Zählen zweier (bzw. mehrerer) Größen ist kein Problem. Sei $T_l = \{a_1, \ldots, a_l\}$ ($l \geq 1$), $N_l = \{S\} \cup \{A_1, \ldots, A_l\}$ und $P_l = \{S ::= A_1 \cdots A_l\} \cup \{A_i ::= a_iA_i \mid i = 1, \ldots, l\} \cup \{A_i ::= \lambda \mid i = 1, \ldots, l\}$.
   Dann gilt:

   $$L((N_l, T_l, P_l, S)) = \{a_1^{n_1} \cdots a_l^{n_l} \mid n_i \geq 0,\ i = 1, \ldots, l\}.$$

3. Fast genauso einfach ist es, zwei Größen gleichzeitig hochzuzählen:

$$L((\{S\}, \{a, b\}, \{S ::= aSb \mid \lambda\}, S)) = \{a^n b^n \mid n \in \mathbb{N}\}.$$

   Es erweist sich also als ausgesprochen einfach, Klammerausdrücke, an denen endliche Automaten scheitern, durch eine kontextfreie Grammatik zu spezifizieren.
4. Auch kompliziertere Klammerausdrücke mit mehreren Klammern, die nicht nur ineinander, sondern auch nebeneinander gesetzt werden können, wie das in Programmiersprachen üblich ist, lassen sich kontextfrei spezifizieren.
   Seien $(a_1, b_1), \ldots, (a_k, b_k)$ k Klammerpaare und $T_{bracket} = \{a_1, \ldots a_k\} \cup \{b_1, \ldots, b_k\}$. Dann erzeugt die kontextfreie Grammatik $G_{bracket} = (\{S\}, T_{bracket}, P_{bracket}, S)$ mit den Produktionen

$$S ::= SS \mid a_i b_i \mid a_i S b_i \text{ für } i = 1, \ldots, k$$

   alle Klammerstrukturen mit den gegebenen Klammerpaaren.
   Für die Klammerpaare [,] und $<, >$ lässt sich beispielsweise folgende Struktur ableiten:

$$S \rightarrow SS \rightarrow S[S] \rightarrow < S > [S] \rightarrow < SS > [S] \rightarrow < []S > [S] \rightarrow$$
$$< []S > [<>] \rightarrow < [][] > [<>]$$

Literatur

[AU72] Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation, and Compiling, Vol. I: Parsing.* Prentice-Hall, Englewood-Cliffs, New Jersey, 1972.

[HU79] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Compuation.* Addison-Wesley, Reading, Mass., 1979.

[HU90] John E. Hopcroft and Jeffrey D. Ullman. *Einführung in die Automatentheorie, Formale Sprachen und Komplexitätstheorie.* Addison-Wesley (Deutschland), Bonn, 1990.

[HUM01] John E. Hopcroft, Rajeev Motwani, Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, 2001.

[HUM02] John E. Hopcroft, Rajeev Motwani, Jeffrey D. Ullman. *Einführung in die Automatentheorie, Formale Sprachen und Komplexitätstheorie.* Addison-Wesley, 2002.

[MAK88] Robert N. Moll, Michael A. Arbib, and A.J. Kfoury. *An Introduction to Formal Language Theory.* Springer-Verlag, New York, 1988.

[Sal73] Arto K. Salomaa. *Formal Languages.* Academic Press, New York, 1973.

10 Übersetzung endlicher Automaten in rechtslineare Grammatiken

Kontextfreie Grammatiken können nicht nur Sprachen erzeugen, die endliche Automaten nicht erkennen, sondern stellen selbst eine Verallgemeinerung endlicher Automaten dar. Genauer gesagt, lässt sich ein Übersetzer von endlichen Automaten in kontextfreie Grammatiken angegeben, bei dem im Prinzip nur die Zustandsüberführung in Regelform umgewandelt wird. Es wird gezeigt, dass der Übersetzer korrekt ist, d.h. dass der eingegebene Automat die Sprache erkennt, die die ausgegebene Grammatik erzeugt.

Die bei der Übersetzung entstehenden Regeln haben eine spezielle Form, die rechtslinear genannt wird.

Theorem 4
Sei $A = (Z, I, d, s_0, F)$ ein endlicher Automat. Dann wird durch $GRA(A) = (Z, I, P_A, s_0)$ mit

$$P_A = \{s ::= xs' \mid s' \in d(s, x)\} \cup \{s'' ::= \lambda \mid s'' \in F\}$$

eine rechtslineare Grammatik konstruiert, so dass gilt: $L(A) = L(GRA(A))$.

Diese Übersetzung (einschließlich der semantischen Verträglichkeit) lässt sich mit dem Diagramm in Abbildung 6 illustrieren.

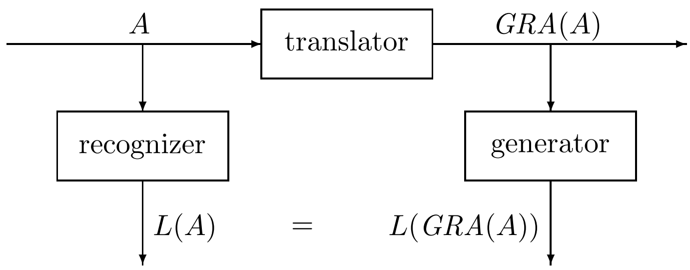

Abbildung 6: Korrekte Übersetzung endlicher Automaten in rechtslineare Grammatiken

Beweis.
Da $GRA(A)$ offensichtlich eine rechtslineare Grammatik ist, muss nur die Sprachgleichheit gezeigt werden. Dazu wird ein Zusammenhang zwischen der fortgesetzten Zustandsüberführung von A und den Ableitungen von $GRA(A)$ im anschließenden Lemma 5 hergestellt. Die behauptete Sprachgleichheit folgt dann vergleichsweise einfach:

$w \in L(A)$ bedeutet $d^*(s_0, w) \cap F \neq \emptyset$, d.h. es gibt einen Zustand $s' \in F$ mit $s' \in d^*(s_0, w)$. Nach Definition der Regeln und Lemma 5 ist das gleichbedeutend zu $s' ::= \lambda \in P_A$ und $s_0 \xrightarrow[P_A]{*} ws'$. Daraus lässt sich die Ableitung $s_0 \xrightarrow[P_A]{*} ws' \xrightarrow[s' ::= \lambda]{} w\lambda = w$ zusammensetzen. Die Ableitung $s_0 \xrightarrow[P_A]{*} w$ liefert aber gerade $w \in L(GRA(A))$.

Umgekehrt zerfällt eine Ableitung der Form $s_0 \xrightarrow[P_A]{*} w$ immer in Ableitungen $s_0 \xrightarrow[P_A]{*} ws'$ und $ws' \xrightarrow[s' ::= \lambda]{} w$, weil das die einzige Möglichkeit zum Terminieren ist. Damit lässt sich die gesamte Überlegung auch umdrehen. $\square$

Lemma 5
Seien A und $GRA(A)$ wie in Theorem 4.
Dann gilt für alle $s, s' \in Z$ und $w \in I^*$: $s \xrightarrow[P_A]{*} ws'$ gdw. $s' \in d^*(s, w)$.

Beweis (mit Induktion über die Länge von w)**.**

IA (für $w = \lambda$):

$s \xrightarrow[P_A]{*} \lambda s' = s'$ gdw. $s = s'$ gdw. $s' \in \{s\} = d^*(s, \lambda)$,
wobei in der zweiten Äquivalenz die Definition von d^* ausgenutzt wird und bei der ersten die Tatsache, dass jede Regelanwendung in $GRA(A)$ ein terminales Zeichen erzeugt oder das nichtterminale löscht.

IV: Die Behauptung gelte für $w \in I^*$.

IS (für wx mit $w \in I^*$ und $x \in I$):

Eine Ableitung der Form $s \xrightarrow[P_A]{*} wxs'$ zerfällt immer in zwei Ableitungen der Form $s \xrightarrow[P_A]{*} w\bar{s}$ und $w\bar{s} \xrightarrow[\bar{s} ::= xs']{} wxs'$, weil nur Regeln der Form $\bar{s} ::= xs'$ Wörter verlängern und weil das nur am rechten Ende geschehen kann. Nach der Induktionsvoraussetzung korrespondiert die Ableitung $s \xrightarrow[P_A]{*} w\bar{s}$ zu $\bar{s} \in d^*(s, w)$. Die im Ableitungsschritt angewendete Regel korrespondiert zu $s' \in d(\bar{s}, x)$. Beides zusammen bedeutet nach Definition von d^* gerade $s' \in \bigcup_{t \in d^*(s,w)} d(t, x) = d^*(s, wx)$. $\square$