

Softwareparadigmen

Dieses Skriptum basiert auf der Softwareparadigmen Übung im
Sommersemester 2011 und dem Vorlesungsskriptum 2007.
Vorlesung von Alexander Felfernig

Übungsskriptum verfasst von Daniel Gruß. Version 10. Februar 2012.
Fehlerfunde bitte melden an gruss@student.tugraz.at.

Inhaltsverzeichnis

1	Syntax	1
1.1	Grundlegende Definitionen	1
1.2	Grammatiken und Sprachen	1
1.3	Chomsky-Sprachhierarchie	3
1.4	Parser	4
1.6	Lexikalische Analyse	4
1.7	Grammatikalische Analyse	5
1.8	LL(1)-Grammatiken	5
1.9	LL(1)-Tabellen	6
2	Semantik von Programmiersprachen	9
2.1	Sprache \mathcal{A} - einfache arithmetische Ausdrücke	9
2.2	Sprache \mathcal{VA} - arithmetische Ausdrücke mit Variablen	10
2.3	Datentypen	11
2.4	Sprache der Terme \mathcal{T}	12

Kapitel 1

Syntax

1.1 Grundlegende Definitionen

Definition 1.1 (Alphabet): Ein Alphabet Σ ist eine endliche Menge von Symbolen.

Binärzahlen haben das Alphabet $\Sigma = \{0, 1\}$. Eine einfache Variante der Markup-Sprache HTML hat das Alphabet $\Sigma = \{<, /, >, \dots, a, b, c, \dots\}$.

Definition 1.2: Σ^* ist die Menge aller beliebigen Konkatenationen von Symbolen aus Σ . Ein Element aus Σ^* nennen wir Wort.

Für $\Sigma = \{0, 1\}$ ist $\Sigma^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, \dots\}$.

Definition 1.3 (Sprache): Eine Sprache \mathcal{L} ist eine Teilmenge (\subseteq) von Σ^* .

Sei $\Sigma = \{0, 1\}$. Definieren wir die Sprache der Binärzahlen \mathcal{B} , müssen wir zu jedem Wort aus Σ^* entscheiden ob dieses Wort in \mathcal{B} enthalten ist. Im Fall der Binärzahlen könnten wir z.B. ε herausnehmen, dann ist die Sprache $\mathcal{B} = \Sigma^* \setminus \{\varepsilon\}$.

Würden wir die Programmiersprache \mathcal{C} als formale Sprache definieren, so wäre ein gesamtes (gültiges) \mathcal{C} -Programm ein Wort der Sprache, also ein Element der Menge \mathcal{C} .

Definition 1.4 (Compiler): Seien \mathcal{A} und \mathcal{B} Programmiersprachen. Ein Compiler ist ein Programm, welches Programme von \mathcal{A} nach \mathcal{B} übersetzt.

Ein einfacher Compiler würde beispielsweise Binärzahlen zu Dezimalzahlen übersetzen.

1.2 Grammatiken und Sprachen

Definition 1.5 (Grammatik): Eine Grammatik ist ein 4-Tupel (V_N, V_T, S, Φ) .

- V_N ist eine endliche Menge von Nonterminalen (entsprechen Zuständen der FSM),
- V_T ist eine endliche Menge von Terminalen (entsprechen Alphabet der Sprache),
- $S \in V_N$ das Startsymbol (wie Initialzustand der FSM),
- $\Phi = \{\alpha \rightarrow \beta\}$ eine endliche Menge von Produktionsregeln (Zustandsübergänge).

α ist hierbei eine beliebige Aneinanderreihung von Terminalen und Nonterminalen, die zumindest ein Nonterminal enthält, also $(V_N \cup V_T)^* V_N (V_N \cup V_T)^*$.

β ist eine beliebige Aneinanderreihung von Terminalen und Nonterminalen, einschließlich der leeren Menge, also $(V_N \cup V_T)^*$.

Um zu verifizieren wann ein Wort von einer Grammatik erzeugt werden kann müssen wir nun zuerst definieren was eine Ableitung ist.

Definition 1.6: Sei (V_N, V_T, S, Φ) eine Grammatik und $\alpha, \beta \in (V_N \cup V_T)^*$. Wenn es 2 Zeichenfolgen τ_1, τ_2 gibt, so dass $\alpha = \tau_1 A \tau_2$, $\beta = \tau_1 B \tau_2$ und $A \rightarrow B \in \Phi$, dann kann β direkt (in einem Schritt) von α abgeleitet werden ($\alpha \rightarrow \beta$).

Diese Definition ist nur geeignet um eine Aussage darüber zu treffen was in genau einem Schritt abgeleitet werden kann. Wir definieren daher die reflexive Hülle dieses Operators.

Definition 1.7: Sei (V_N, V_T, S, Φ) eine Grammatik und $\alpha, \beta \in (V_N \cup V_T)^*$. Wenn es $n \in \mathbb{N}$ Zeichenfolgen τ_1, τ_n gibt, so dass $\alpha \rightarrow \tau_1, \tau_1 \rightarrow \tau_2, \dots, \tau_{n-1} \rightarrow \tau_n, \tau_n \rightarrow \beta$, dann kann β von α (in n Schritten) abgeleitet werden ($\alpha \xrightarrow{+} \beta$).

Diese Definition ist für $n \geq 1$ geeignet. Wenn $\alpha = \beta$ ist, wäre $n = 0$. Wir definieren die reflexive, transitive Hülle durch eine Verknüpfung dieser beiden Fälle.

Definition 1.8: Es gilt $\alpha \xrightarrow{*} \beta$, genau dann wenn $\alpha \xrightarrow{+} \beta$ oder $\alpha = \beta$ (reflexive, transitive Hülle).

Mit dieser Definition können wir alle Wörter ableiten die diese Grammatik produziert.

Definition 1.9: Sei (V_N, V_T, S, Φ) eine Grammatik G . G akzeptiert die Sprache $L(G) = \{w \mid S \xrightarrow{*} w, w \in V_T^*\}$, d.h. die Menge aller Wörter w die in beliebig vielen Schritten aus dem Startsymbol S ableitbar sind und in der Menge aller beliebigen Konkatenationen von Terminalsymbolen V_T enthalten sind.

Die Äquivalenz von zwei Sprachen zu zeigen ist im Allgemeinen nicht trivial. Wenn wir versuchen eine Sprache \mathcal{L} durch eine Grammatik G zu beschreiben ist es im Allgemeinen nicht trivial zu zeigen dass $L(G) = \mathcal{L}$.

An dieser Stelle möchten wir festzuhalten: Um die Gleichheit zweier Mengen (Sprachen) M, N zu zeigen muss gezeigt werden, dass jedes Element (Wort) aus M in N enthalten ist und jedes Element (Wort) aus N in M enthalten ist. Ungleichheit ist daher viel leicht-

ter zu zeigen, da es genügt ein Element (Wort) zu finden welches nicht in beiden Mengen (Sprachen) enthalten ist. Der geneigte Leser kann probieren die Gleichheit oder Ungleichheit der Sprache unserer oben definierten Grammatik und der Sprache der Binärzahlen zu zeigen.

Definition 1.10: Ein Programm $P_{\mathcal{L}}$ welches für ein Wort w entscheidet ob es in der Sprache \mathcal{L} enthalten ist (d.h. `true` dann und nur dann zurückliefert wenn es enthalten ist), nennen wir **Parser**.

1.3 Chomsky-Sprachhierarchie

Wir haben Produktionsregeln definiert durch $\alpha \rightarrow \beta$, wobei α zumindest ein Nonterminal enthält.

Definition 1.11: Eine Grammatik ist nach der Chomsky-Sprachhierarchie:

- allgemein/uneingeschränkt (unrestricted)

Keine Restriktionen

- Kontext-sensitiv (context sensitive): $|\alpha| \leq |\beta|$

Es werden nicht mehr Symbole gelöscht als produziert.

- Kontext-frei (context free): $|\alpha| \leq |\beta|$, $\alpha \in V_N$

Wie Kontext-sensitiv; außerdem muss α genau **ein** Non-Terminal sein

- regulär (regular): $|\alpha| \leq |\beta|$, $\alpha \in V_N$, $\beta = aA$, $a \in V_T \cup \{\varepsilon\}$, $A \in V_N \cup \{\varepsilon\}$

Wie Kontext-frei; außerdem ist $\beta = aA$ wobei a ein Terminal oder ε ist und A ein Nonterminal oder ε . (Anmerkung: $\varepsilon\varepsilon = \varepsilon$)

Es gilt $\mathbb{L}_{\text{regular}} \subset \mathbb{L}_{\text{context free}} \subset \mathbb{L}_{\text{context sensitive}} \subset \mathbb{L}_{\text{unrestricted}}$ (\mathbb{L}_x Menge aller Sprachen der Stufe x).

Nicht alle Grammatiken können in eine äquivalente Grammatik einer stärker eingeschränkten Stufe umgewandelt werden. Um zu zeigen dass es sich um echte Teilmengen ($A \subset B$) handelt müssen wir zeigen dass alle Elemente aus A in B enthalten sind und mindestens ein Element aus B nicht in A enthalten ist.

Die Chomsky-Sprachhierarchie unterscheidet Sprachen anhand der Komplexität der produzierten Sprache.

1.4 Parser

Wir überspringen an dieser Stelle den BPARSE-Algorithmus (siehe Vorlesungsskriptum) und betrachten stattdessen einen Recursive Descent Parser (RDP).

Bei einem RDP werden alle Nonterminale in Funktionen übersetzt und diese Funktionen behandeln die verschiedenen Produktionsregeln. Die Eingabe wird in die Terminale unterteilt (auch Tokens genannt). Wir verwenden im Pseudo-Code die Variable `token`, die immer das aktuelle Token enthält, sowie die Funktion `nextToken()`, die `token` auf das nächste Token setzt. Der Parser ruft die Startfunktion auf und gibt `true` zurück wenn `token` leer ist (d.h. das Ende der Eingabe erreicht wurde). `ERROR` führt dazu dass der Parser die Eingabe (das Wort) nicht akzeptiert.

Definition 1.12 (Mehrdeutig, Eindeutig): Sei $G = (V_N, V_T, S, \Phi)$ eine Grammatik. Wenn es für ein Wort $w \in L(G)$ mehrere unterschiedliche Ableitungssequenzen ω, ψ , d.h. $S \rightarrow \omega_1, \omega_1 \rightarrow \dots, \dots \rightarrow \omega_n, \omega_n \rightarrow w$, $S \rightarrow \psi_1, \psi_1 \rightarrow \dots, \dots \rightarrow \psi_k, \omega_k \rightarrow w$ wobei $\exists \psi_i : \psi_i \neq \omega_i$, ist es mehrdeutig. Wenn eine Grammatik ist genau dann eindeutig, wenn sie nicht mehrdeutig ist.

Definition 1.13 (Linksrekursiv): Eine Grammatik ist direkt linksrekursiv wenn sie eine Produktion der Form $A\alpha \rightarrow A\beta$ enthält, wobei A ein Nonterminal ist. Eine Grammatik ist indirekt linksrekursiv wenn sie Produktionen der Form $A\alpha \rightarrow A_1\beta_1, A_1\alpha_1 \rightarrow A_2\beta_2, \dots, A_n\alpha_n \rightarrow A\beta_n$ enthält, wobei A, A_i Nonterminale sind. Eine Grammatik ist linksrekursiv wenn sie direkt oder indirekt linksrekursiv ist.

Nun können wir mit Sprachen und Grammatiken umgehen und diese nach ihrer Komplexität einstufen.

1.6 Lexikalische Analyse

Definition 1.14: Ein regulärer Ausdruck A ist wie folgt rekursiv definiert (mit regulären Ausdrücken Q und R):

$$A = \begin{cases} \varepsilon & \text{Leerstring} \\ t & \text{ein Terminal, d.h. } t \in V_T \\ Q|R & \text{entweder } A = Q \text{ oder } A = R \\ QR & \text{Konkatenation zweier regulärer Ausdrücke} \\ Q? & \text{entspricht } Q|\varepsilon, \text{ d.h. } Q \text{ ist optional} \\ Q* & \text{beliebige Konkatenation von } Q \text{ mit sich selbst, d.h. } A \in \{\varepsilon, Q, QQ, \dots\} \\ Q+ & \text{entspricht } QQ*, \text{ d.h. mindestens ein } Q \\ (Q) & \text{Klammerung des Ausdrucks } Q \end{cases}$$

Zur Vereinfachung erlauben wir Variablen (Bezeichner) in regulären Ausdrücken:

Definition 1.15: Sei B die Menge aller Variablen (Bezeichnungen). Ein regulärer Ausdruck der eine Variable b aus B enthält ist ein erweiterter regulärer Ausdruck. Wenn E ein erweiterter regulärer Ausdruck ist und c eine Variable aus B , dann ist $c := E$ eine reguläre Definition.

Übersetzen wir dazu den erweiterten regulären Ausdruck zu einer Funktion so beobachten wir:

- Terminale werden zu `if`-Abfragen,
- Q^* wird zu einem Schleifenkonstrukt,
- $Q|R$ wird zu einer `if, else if, else`-Abfrage wobei der `else` Fall zu einem Ablehnen der Eingabe führt,
- QR bedeutet dass zuerst Q überprüft wird, danach R .

1.7 Grammatikalische Analyse

Bei kontext-freien oder regulären Grammatiken werden wir fortan nur noch die Produktionsregeln mit unterstrichenen Nonterminalen anschreiben, da die Grammatik dadurch vollständig definiert wird:

- V_N ist die Vereinigung über alle Symbole auf der linken Seite der Produktionsregeln (d.h. alle Nonterminale),
- V_T ist die Vereinigung aller unterstrichenen Zeichenfolgen (d.h. alle Terminale),
- S , das Startsymbol ist (soweit nicht anders festgehalten) das erste aufgeführte Nonterminal,
- Φ wird explizit angegeben.

1.8 LL(1)-Grammatiken

Definition 1.16 (LL(1)-Grammatik): Eine kontextfreie Grammatik G ist eine LL(1)-Grammatik (ist in LL(1)-Form) wenn sie

- keine Linksrekursionen enthält
- keine Produktionen mit gleichen Prefixen für die selbe linke Seite enthält
- ermöglicht immer in einem Schritt (d.h. nur mit Kenntnis des nächsten Tokens) zu entscheiden, welche Produktionsregel zur Ableitung verwendet werden muss.

LL(1) steht für “Left to right”, “Leftmost derivation”, “1 Token Look-ahead”.

Wir definieren einfache Regeln zwecks Auflösung von:

- **Indirekten Linksrekursionen**

Gibt es Produktionen der Form $A \rightarrow \alpha B \beta$ sowie $B \rightarrow \gamma$, dann kann man die Produktion $A \rightarrow \alpha \gamma \beta$ einfügen. Wenn $\alpha = \varepsilon$ ist können so indirekte Linksrekursionen gefunden werden.

- **Direkten Linksrekursionen**

Gibt es Produktionen der Form $A \rightarrow A\alpha$ und $A \rightarrow \beta$, dann kann man diese in Produktionen der Form $A \rightarrow \beta R$ sowie $R \rightarrow \alpha R | \varepsilon$ umwandeln.

- **Linksfaktorisierungen**

Gibt es Produktionen der Form $A \rightarrow \alpha B$ sowie $A \rightarrow \alpha C$, wobei α der größte gemeinsame Prefix von αB und αC ist, so können wir diese Produktion durch $A \rightarrow \alpha R$ und $R \rightarrow B | C$ ersetzen. Der größte gemeinsame Prefix einer Sequenz von Terminalen und Nonterminalen ist die größte gemeinsame Zeichenkette dieser Sequenzen. Der größte gemeinsame Prefix von if E then E else E und if E then E ist if E then E.

Mit diesen Regeln können wir oftmals Grammatiken in äquivalente LL(1)-Grammatiken umformen. Es gibt natürlich Grammatiken bei denen dies nicht möglich ist. In diesem Fall bleibt nur die Möglichkeit eine andere Parsing-Methode zu verwenden oder die Sprache zu verändern.

1.9 LL(1)-Tabellen

LL(1)-Tabellen (auch: LL(1)-Parser-Tabellen) ermöglichen es uns einen generischen LL(1)-Parser zu schreiben, der mit einer beliebigen Tabelle (und damit Sprache) arbeiten kann. Wir werden uns nun erarbeiten wie eine solche Tabelle berechnet werden kann. Für die Definition der FIRST- und Follow-Mengen stellen wir uns Nonterminale wieder als Zustände in einem Graph oder eine Maschine vor.

Definition 1.17 (FIRST-Menge): Die FIRST-Menge eines Nonterminals X ist die Menge aller Terminalsymbole die im Zustand X **als erstes** geparst werden können. Die FIRST-Menge eines Terminals x ist immer das Terminalsymbol selbst.

Formal bedeutet dies:

1. Wenn x ein Terminal ist: $\text{FIRST}(x) = \{x\}$
2. Wenn die Grammatik Produktionsregeln enthält so dass $X \rightarrow \dots \rightarrow \varepsilon$, dann ist: $\varepsilon \in \text{FIRST}(X)$

3. Für jede Produktionsregel $X \rightarrow Y_1 Y_2 \dots Y_n$, ist $x \in \text{FIRST}(X)$ wenn $x \in \text{FIRST}(Y_i)$ und für alle Y_j mit $j < i$ gilt, dass $\varepsilon \in \text{FIRST}(Y_j)$.

Für aufeinanderfolgende Terminal- bzw. Nonterminalsymbole $X_1 X_2 \dots X_n$:

1. Wenn $x \in \text{FIRST}(X_i)$ und für alle X_j mit $j < i$ gilt, dass $\varepsilon \in \text{FIRST}(X_j)$, dann ist $x \in \text{FIRST}(X_1 X_2 \dots X_n)$.
2. Wenn für alle X_i $\varepsilon \in \text{FIRST}(X_i)$ ist, dann ist auch $\varepsilon \in \text{FIRST}(X_1 X_2 \dots X_n)$.

Aus dieser Definition folgt beispielsweise für einfache Regeln $A \rightarrow B$, dass $\text{FIRST}(A) = (\text{FIRST}(B) \setminus \{\varepsilon\}) \cup \dots$ ist.

Definition 1.18: Zwecks Übersichtlichkeit definieren wir die FIRST^* -Menge als FIRST -Menge ohne ε :

$$\text{FIRST}^*(X) = \text{FIRST}(X) \setminus \{\varepsilon\}.$$

Definition 1.19: Jede Eingabe des Parsers endet mit dem “End of Input” Symbol \$.

Definition 1.20 (Follow-Menge): Die Follow-Menge eines Nonterminals X ist die Menge aller Terminalsymbole die direkt auf Zustand X **folgen** können. Das heißt, alle Terminalsymbole die nach Abarbeitung des Zustands X als erstes geparkt werden können.

Formal bedeutet dies:

1. Wenn X das Startsymbol ist, dann ist $\$ \in \text{FOLLOW}(X)$
2. Für alle Regeln der Form $A \rightarrow \alpha B \beta$ ist $\text{FIRST}^*(\beta) \subseteq \text{FOLLOW}(X)$
3. Für alle Regeln der Form $A \rightarrow \alpha B$ bzw. $A \rightarrow \alpha B \beta$ mit $\varepsilon \in \text{FIRST}(\beta)$ ist $\text{FOLLOW}(A) \subseteq \text{FOLLOW}(B)$

Algorithmus (Parse-Table): Der Parse-Table Algorithmus berechnet aus einer Grammatik eine Parse-Table (auch LL(1)-Tabelle) M .

1. Für jede Produktion $X \rightarrow \alpha$:
 - (a) Für jedes Element $y \in \text{FIRST}^*(\alpha)$ bzw. wenn $\alpha = y$:
 - i. Füge $X \rightarrow \alpha$ in $M(X, y)$ ein.
 - (b) Wenn $\varepsilon \in \text{FIRST}(\alpha)$:
 - i. Für alle $y \in \text{FOLLOW}(X)$:
 - A. Füge $X \rightarrow \alpha$ in $M(X, y)$ ein.

Leere Einträge in der Tabelle sind Fehlerfälle. Diese können auch explizit mit ERROR beschriftet werden.

Definition 1.21: Eine Grammatik ist eine LL(1)-Grammatik, wenn die berechnete Parse-Tabelle keine Mehrfacheinträge hat.

Definition 1.22: Eine LL(1)-Parsing Tabelle (oder auch Parsing-Tabelle) stellt einen vollständigen Parse-Vorgang dar. Jede Zeile entspricht einem Bearbeitungsschritt im Parse-Vorgang. In Spalten werden Stack, Eingabe und die angewandte Produktionsregel aufgetragen.

Algorithmus (LL(1)-Parsing mit Tabelle): Sei X das oberste Stack-Element, t das aktuelle Token der Eingabe w und \mathcal{L} die von der Grammatik akzeptierte Sprache.

1. Wenn X ein Non-Terminal ist:
 - (a) Nimm den Wert von $M(X, t)$
 - (b) Ist der Eintrag leer oder ein Fehlereintrag: Abbruch ($w \notin \mathcal{L}$).
 - (c) Sonst: Ersetze das oberste Stack-Element X durch Produktion in umgekehrter Reihenfolge (WVU wenn $M(X, t) = X \rightarrow UVW$).
2. Andernfalls (X ist ein Terminal):
 - (a) Wenn $X = t = \$$: Parsing erfolgreich ($w \in \mathcal{L}$).
 - (b) Sonst, wenn $X = t \neq \$$, dann nimm X vom Stack und gehe zum nächsten Token im Input.
 - (c) Sonst: Abbruch ($w \notin \mathcal{L}$).

Diese Definition kann für andere Parser leicht angepasst werden.

Kapitel 2

Semantik von Programmiersprachen

In funktionalen Programmiersprachen besteht jedes Programm aus einer oder mehreren Funktionen.

Definition 2.1 (Funktion): Eine Funktion ist eine Relation zwischen einer Menge A und einer Menge B . Jedem Element aus der Menge A wird genau ein Element der Menge B zugeordnet. Das heißt: Für jeden möglichen Eingabewert gibt es genau einen Ausgabewert.

2.1 Sprache \mathcal{A} - einfache arithmetische Ausdrücke

Arithmetische Ausdrücke sind Funktionen. Wir können beispielsweise die Funktionen Addition, Subtraktion und Multiplikation von zwei Zahlen in \mathbb{R} definieren mit einem Eingabewert in $\mathbb{R} \times \mathbb{R}$ und einen Ausgabewert in \mathbb{R} . Auch die Division können wir als Funktion definieren von $\mathbb{R} \times \mathbb{R} \setminus \{0\}$ (Division durch 0 schließen wir damit aus, da die Division in diesem Fall nicht als Funktion definiert ist) auf Ausgabewerte in \mathbb{R} .

Definition 2.2: Die Sprache \mathcal{A} definieren wir mit Alphabet $\Sigma = \{\underline{0}, \dots, \underline{9}, (,), \pm\}$. Zwecks Einfachheit definieren wir Ziffern (\mathcal{D} , digits) und Zahlen:

- $\mathcal{A}_D = \{\underline{0}, \dots, \underline{9}\}$
- $\text{ZAHL} = (\mathcal{A}_D \setminus \{\underline{0}\} \mathcal{A}_D^*) \cup \{\underline{0}\}$

Wir definieren die Sprache \mathcal{A} nun nicht mehr über eine Grammatik sondern durch eine induktive Beschreibung (Basisfall und allgemeine Fälle):

1. $\text{ZAHL} \subset \mathcal{A}$

2. Wenn $x, y \in \mathcal{A}$, dann ist auch $\underline{(x)} \pm \underline{(y)} \in \mathcal{A}$.

An dieser Stelle sei noch einmal darauf hingewiesen dass wir x, y nicht unterstreichen dürfen, da sie keine Sprachelemente sind sondern Platzhalter, mathematisch würde man sie auch als Variablen bezeichnen.

Definition 2.3: Die Semantik der Sprache \mathcal{A} definieren wir durch:

1. $I_{\mathcal{A}}(x) = \langle x \rangle$ wenn $x \in \mathcal{A}_N$. x ist dabei eine Zeichenkette im Programm, $\langle x \rangle$ die entsprechende Repräsentation in \mathbb{N}_0 .
2. $I_{\mathcal{A}}(\underline{(x)} + \underline{(y)}) = I_{\mathcal{A}}(x) + I_{\mathcal{A}}(y)$ wenn $x, y \in \mathcal{A}$.

2.2 Sprache \mathcal{VA} - arithmetische Ausdrücke mit Variablen

Wir erweitern die Sprache \mathcal{A} durch Variablen und schaffen so eine mächtigere Sprache \mathcal{VA} . Um mit Variablen umgehen zu können brauchen wir nun einerseits eine Menge zulässiger Variablennamen und andererseits eine Funktion die von Variablennamen auf eine Wertemenge der semantischen Ebene (z.B. \mathbb{N}_0) abbildet. Die Menge der zulässigen Variablennamen nennen wir IVS (Individuenvariablensymbole).

Definition 2.4: Zwecks Einfachheit erlauben wir nur wenige Variablennamen und definieren daher

$$\text{IVS} = \{\underline{a}, \underline{b}, \dots, \underline{z}\} \cup \{\underline{x1}, \underline{x2}, \dots\}.$$

Die Funktion die von Variablennamen auf eine Wertemenge abbildet nennen wir ω -Environment, (Variablen-)Umgebung. Man kann sich diese Funktion auch als Tabelle vorstellen bzw. in einem Interpreter als Tabelle implementieren.

Definition 2.5: Die Menge aller Environments sei

$$\text{ENV} = \bigcup_{x \in \text{IVS}, y \in \Lambda} \{(x, y)\},$$

das heißt, die Vereinigung über alle Tupel Variablenname $x \in \text{IVS}$ und Wert auf semantischer Ebene $y \in \Lambda$.

Für die Sprache \mathcal{VA} ist $\Lambda = \mathbb{N}_0$.

Definition 2.6: Die Syntax der Sprache \mathcal{VA} ist definiert durch:

1. $\text{ZAHL} \subset \mathcal{VA}$
2. $\text{IVS} \subset \mathcal{VA}$
3. $\underline{(x)} \pm \underline{(y)} \in \mathcal{C}$, wenn $x, y \in \mathcal{VA}$.

Die Interpretation eines Programms hängt nun nicht mehr allein vom Programm selbst ab, sondern auch von den Werten der Variablen im ω -Environment.

Definition 2.7: Die Interpretationsfunktion $I_{\mathcal{VA}} : \text{ENV} \times \mathcal{VA} \rightarrow \Lambda$ weist jedem Tupel aus Environment und Programm einen Wert in Λ zu.

1. $I_{\mathcal{VA}}(\omega, k) = \langle k \rangle$ wenn $k \in \text{ZÄHL}$, $\omega \in \text{ENV}$.
2. $I_{\mathcal{VA}}(\omega, v) = \omega(v)$ wenn $vk \in \text{IVS}$, $\omega \in \text{ENV}$.
3. $I_{\mathcal{VA}}(\omega, \underline{(x)} + \underline{(y)}) = I_{\mathcal{VA}}(\omega, x) + I_{\mathcal{VA}}(\omega, y)$ wenn $x, y \in \mathcal{VA}$, $\omega \in \text{ENV}$.

2.3 Datentypen

Bisher haben wir eine Sprache nur für einen Datentypen definiert. Dieser war implizit in der Definition der Sprache drin (beispielsweise die natürlichen Zahlen). Derartige Definitionen erlauben kein Ersetzen des Datentyps ohne die Definition der Sprache wesentlich zu überarbeiten. Da wir dies aber häufig wollen werden wir nun zuerst Datentypen auf der semantischen Ebene und anschließend die Repräsentation von Datentypen auf der syntaktischen Ebene definieren.

Definition 2.8 (Datentyp): Ein Datentyp ist ein Tupel $\Psi = (A, F, P, C)$ mit

- A : Grundmenge (Wertebereich)
- F : Menge von Funktionen $f_i : A^{k_i} \rightarrow A^{l_i}$.
 f_i ist die i -te Funktion in der Menge, k_i die Dimension vom Urbild (Dimension des Inputs, Anzahl der Funktionsargumente) und l_i die Dimension vom Bild der i -ten Funktion (Dimension des Outputs, Anzahl der Funktionsrückgabewerte).
- P : Menge von Prädikaten $p_i : A^{k_i} \rightarrow \{T, F\}$.
 p_i ist die i -te Funktion in der Menge und k_i die Dimension vom Urbild (Dimension des Inputs, Anzahl der Funktionsargumente) der i -ten Funktion.
- C : Menge von Konstanten c_i wobei $c \subseteq A$.

Die Mengen F^Σ , P^Σ und C^Σ enthalten die entsprechenden Symbole für die syntaktische Repräsentation:

- Funktionssymbole F^Σ : je ein Symbol f_i^Σ (z.B. Name der Funktion) für jede Funktion f_i
- Prädikatensymbole P^Σ : je ein Symbol p_i^Σ (z.B. Name des Prädikats) für jedes Prädikat p_i
- Konstantensymbol C^Σ : je ein Symbol c_i^Σ (z.B. ausgeschriebene Form der Konstante) für jede Konstante c_i

Konstanten sind eigentlich nur spezielle Funktionen (0 Argumente) und wir unterscheiden nur zwecks Übersicht.

Beispiel 2.1: Definieren Sie den Datentyp Integer mit Funktionen für Addition, Subtraktion und Multiplikation sowie Prädikaten für “kleiner” und Gleichheit.

Lösung: Im Fall der Integer ist die Definition der Funktionen und Prädikate trivial, da alle Funktionen und Prädikate durch die entsprechenden Operationen auf den ganzen Zahlen \mathbb{Z} definiert sind. Daher genügt es zu definieren welche Funktion welcher Operation entspricht.

- Grundmenge $A = \mathbb{Z}$
- Funktionen $f_1 : +$, $f_2 : -$, $f_3 : *$

Die Funktionen $+$, $-$, $*$ sind auf \mathbb{Z} definiert. Auf der syntaktischen Ebene definieren wir: $f_1^\Sigma : \underline{\text{plus}}$, $f_2^\Sigma : \underline{\text{minus}}$, $f_3^\Sigma : \underline{\text{mult}}$.

- Prädikate $p_1 : <$, $p_2 : =$

Die Prädikate $<$, $=$ sind auf \mathbb{Z} definiert. Syntaktische Ebene: $p_1^\Sigma : \underline{\text{lt?}}$, $p_2^\Sigma : \underline{\text{eq?}}$.

- Konstanten $c_1 : 0$, $c_2 : 1$

Syntaktische Ebene: $c_1^\Sigma : \underline{\text{null}}$, $c_2^\Sigma : \underline{\text{eins}}$.

Hinzunahme von Division ist problematisch / da das Ergebnis nicht unbedingt $\in A$ ist.

Wir können bei der Definition eines Datentyps oft (z.B. bei den verschiedenen Datentypen für Zahlen) auf bekannte algebraische Strukturen (Halbgruppen, etc.) zurückgreifen.

2.4 Sprache der Terme \mathcal{T}

Wir müssen nun um den Datentyp verwenden zu können eine grundlegende Sprache definieren die diesen Datentyp verwendet. Auf dieser Sprache können dann weitere Sprachen aufgebaut werden.

Definition 2.9 (Sprache der Terme \mathcal{T}): Sei $\Psi = (A, F, P, C)$ ein Datentyp. Das Alphabet Σ ist dann eine Vereinigung aus den Mengen der

- IVS (Individuenvariablensymbole)
- Funktionssymbole F^Σ
- Prädikatensymbole P^Σ
- Konstantensymbol C^Σ
- $(,)$ und $_{\perp}$
- Sondersymbole (Keywords): if, then, else, begin, end, ...

Die Syntax von $\mathcal{T} \subseteq \Sigma$ über einem beliebigen Datentypen ist dann definiert durch:

1. $C^\Sigma \subseteq \mathcal{T}$, d.h. Konstantensymbole sind Terme
2. $\text{IVS} \subseteq \mathcal{T}$, d.h. Individuenvariablensymbole sind Terme
3. Wenn f_i^Σ ein n -stelliges Funktionssymbol ist und t_1, \dots, t_n Terme, dann ist auch $f_i^\Sigma(t_1, \dots, t_n)$ ein Term (Unterstreichungen beachten!).

Die Semantik von \mathcal{T} definieren wir durch die Interpretationsfunktion $I_{\mathcal{T}} : \text{ENV} \times \mathcal{T} \rightarrow A$.

1. $I_{\mathcal{T}}(\omega, c'_i) = c_i$ mit $c_i \in C$ (Semantik-Ebene), $c'_i \in C^\Sigma$ (Syntax-Ebene) und $\omega \in \text{ENV}$.
2. $I_{\mathcal{T}}(\omega, v) = \omega(v)$ mit $v \in \text{IVS}$ und $\omega \in \text{ENV}$.
3. $I_{\mathcal{T}}(\omega, f'_i(t_1, \dots, t_n)) = f_i(I_{\mathcal{T}}(\omega, t_1), \dots, I_{\mathcal{T}}(\omega, t_n))$ mit $f_i \in F$ (Semantik-Ebene), $f'_i \in F^\Sigma$ (Syntax-Ebene) und $\omega \in \text{ENV}$.

Beispiel 2.2: Führen Sie das Programm plus(plus(x,y),plus(eins,z)) mit dem Environment $\omega(\underline{x}) = 0$, $\omega(\underline{y}) = 1$, $\omega(\underline{z}) = 2$ aus.

Lösung:

$$\begin{aligned}
 I_{\mathcal{T}} \left(\omega, \underline{\text{plus}(\text{plus}(x, y), \text{plus}(\text{eins}, z))} \right) &= + \left(I_{\mathcal{T}} \left(\omega, \underline{\text{plus}(x, y)} \right), I_{\mathcal{T}} \left(\omega, \underline{\text{plus}(\text{eins}, z)} \right) \right) \\
 &= + \left(+ \left(I_{\mathcal{T}}(\omega, \underline{x}), I_{\mathcal{T}}(\omega, \underline{y}) \right), + \left(I_{\mathcal{T}}(\omega, \underline{\text{eins}}), I_{\mathcal{T}}(\omega, \underline{z}) \right) \right) \\
 &= + \left(+ \left(\omega(\underline{x}), \omega(\underline{y}) \right), + \left(1, \omega(\underline{z}) \right) \right) \\
 &= + \left(+ \left(0, 1 \right), + \left(1, 2 \right) \right) = + \left(1, 3 \right) = 4
 \end{aligned}$$

Laut Definition von \mathcal{T} gibt es nur die Konstanten 0 und 1. Variablen können natürlich jeden beliebigen Wert in \mathbb{Z} annehmen. Es kann aber auch gezeigt werden dass alle ganzen Zahlen durch einen variablenfreien Term dargestellt werden können. Terme sind rekursiv definiert. Die **vollständige Induktion** ist die übliche Beweistechnik für Beweise über rekursive bzw. rekursiv definierte Ausdrücke.

Beispiel 2.3: Wir wollen zeigen, dass alle ganzen Zahlen durch einen variablenfreien Term dargestellt werden können. Die **vollständige Induktion** besteht aus 3 Einzelschritten: In der **Induktionsbasis** werden ein oder mehrere Basisfälle direkt bewiesen. In der **Induktionshypothese** wird versucht eine allgemeine Aussage (eine Hypothese) zu treffen von der angenommen wird dass sie bis zum n -ten Fall gilt. Im **Induktionsschritt** gehen wir einen Schritt weiter, also in den Fall $n + 1$ und versuchen diesen zu beweisen. Hier muss unbedingt auf die Induktionshypothese zurückgegriffen werden, sonst wurde keine vollständige Induktion durchgeführt.

Um eine Induktion durchführen zu können müssen die Elemente unbedingt aufgezählt werden können (d.h. man muss eine eindeutige Reihenfolge/Sortierung für die Elemente angeben können).

Bevor wir die Induktion durchführen versuchen wir ein Muster zu erkennen.

$$\begin{aligned}
 I_{\mathcal{T}}(\omega, \underline{\text{null}}) &= 0 && \text{(trivial)} \\
 I_{\mathcal{T}}(\omega, \underline{\text{eins}}) &= 1 && \text{(trivial)} \\
 I_{\mathcal{T}}(\omega, \underline{\text{plus}(\text{eins}, \text{eins})}) &= 2 && \text{(eine Addition)} \\
 I_{\mathcal{T}}(\omega, \underline{\text{plus}(\text{eins}, \text{plus}(\text{eins}, \text{eins}))}) &= 3 && \text{(verschachtelte Addition)}
 \end{aligned}$$

Wir erkennen das Muster: Alle Zahlen $\in \mathbb{N}$ können durch rekursive Addition von 1 dargestellt werden. Diese Rekursion kann beliebig tief werden, hat allerdings immer die gleiche Form. Wir definieren uns einen Platzhalter t_k um beliebig lange solcher Ausdrücke einfach darzustellen:

$$\begin{aligned}
 t_2 &= \underline{\text{plus}(\text{eins}, \text{eins})} \\
 t_3 &= \underline{\text{plus}(\text{eins}, \text{plus}(\text{eins}, \text{eins}))} \\
 &\vdots \\
 t_{k+1} &= \underline{\text{plus}(\text{eins}, t_k)}
 \end{aligned}$$

k entspricht der Anzahl der eins im Ausdruck.

- **Induktionsbasis:**

In der Induktionsbasis beweisen wir einen oder mehrere Basisfälle. Das sind in unserem Fall die beiden Konstanten sowie der Fall t_2 :

$$\begin{aligned}
 I_{\mathcal{T}}(\omega, \underline{\text{null}}) &= 0 \\
 I_{\mathcal{T}}(\omega, \underline{\text{eins}}) &= 1 \\
 I_{\mathcal{T}}(\omega, t_2) &= I_{\mathcal{T}}(\omega, \underline{\text{plus}(\text{eins}, \text{eins})}) = 2
 \end{aligned}$$

- **Induktionshypothese:**

In der Induktionshypothese treffen wir eine Aussage über ein Element n oder mehrere (evtl. alle) Elemente bis zu einem gewissen Element n . In unserem Fall sagen wir dass

$$I_{\mathcal{T}}(\omega, t_n) = n$$

- **Induktionsschritt:**

Im Induktionsschritt erfolgt zeigen wir nun dass unter Annahme der Korrektheit der Induktionshypothese der Beweis auch für das Element $n + 1$ (also das erste Element über das wir keine Annahme getroffen haben) erbracht werden kann.

(Häufiger Fehler: Wird im Induktionsschritt die Induktionshypothese nicht verwendet so handelt es sich nicht um eine vollständige Induktion!) Wir müssen also zeigen dass:

$$\begin{aligned}
 I_{\mathcal{T}}(\omega, t_{n+1}) &= n + 1 && \text{(Einsetzen von } t_{n+1}) \\
 I_{\mathcal{T}}\left(\omega, \underline{\text{plus}(\text{eins}, t_n)}\right) &= n + 1 && \text{(Interpretationsfunktion durchführen)} \\
 + (I_{\mathcal{T}}(\omega, \underline{\text{eins}}), I_{\mathcal{T}}(\omega, t_n)) &= n + 1 \\
 + (1, I_{\mathcal{T}}(\omega, t_n)) &= n + 1
 \end{aligned}$$

Unter Verwendung der Induktionshypothese (d.h. wir setzen die Induktionshypothese $I_{\mathcal{T}}(\omega, t_n) = n$ ein):

$$+ (1, n) = n + 1$$

Damit ist der Beweis erbracht.