

PicoC-Compiler

Übersetzung einer Untermenge von C in den Befehlssatz der RETI-CPU

BACHELORARBEIT

Abgabedatum: 13. September 2022

Autor:
Jürgen Mattheis

Gutachter:
Prof. Dr. Scholl

Betreuung:
M.Sc. Seufert

ERKLÄRUNG

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen/Hilfsmittel verwendet habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe. Darüber hinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, bereits für eine andere Prüfung angefertigt wurde.

Danksagungen

Bevor der Inhalt dieser Schriftlichen Ausarbeitung der Bachelorarbeit anfängt, will ich einigen Personen noch meinen Dank aussprechen.

Ich schreibe die folgenden Danksagungen nicht auf eine bestimmte Weise, wie es sich vielleicht etabliert haben sollte Danksagungen zu schreiben und verwende auch keine künstlichen Floskeln, wie „mein aufrichtigster Dank“ oder „aus tiefstem Herzen“, sondern drücke im Folgenden die Dinge nur so aus, wie ich sie auch wirklich meine.

Estmal, ich hatte selten im Studium das Gefühl irgendwo Kunde zu sein, aber bei dieser Bachelorarbeit und dem vorangegangenen Bachelorprojekt hatte ich genau diese Gefühl, obwohl die Verhältnisse eigentlich genau umgekehrt sein sollten. Die Umgang mit mir wahr echt unglaublich nett und unbürokratisch, was ich als keine Selbstverständlichkeit ansehe und sehr wertgeschätzt habe.

An erster Stelle will ich zu meinem Betreuer M.Sc. Tobias Seufert kommen, der netterweise auch bereits die Betreuung meines Bachelorprojektes übernommen hatte. Wie auch während des Bachelorprojektes, haben wir uns auch bei den Meetings während der Bachelorarbeit hervorragend verstanden. Dabei ging die Freundlichkeit und das Engagement seitens Tobias weit über das hinaus, was man bereits als eine gute Betreuung bezeichnen würde.

Es gibt verschiedene Typen von Menschen, es gibt Leute, die nur genauso viel tun, wie es die Anforderungen verlangen und nichts darüberhinaus tun, wenn es nicht einen eigenen Vorteil für sie hat und es gibt Personen, die sich für nichts zu Schade sind und dies aus einer Philanthropie oder Leidenschaft heraus tun, auch wenn es für sie keine Vorteile hat. Tobias¹ konnte ich während der langen Zeit, die er mein Bachelorprojekt und dann meine Bachelorarbeit betreut hat eindeutig als letzteren Typ Mensch einordnen.

Er war sich nie zu Schade für meine vielen Fragen während der Meetings, auch wenn ich meine Zeit ziemlich oft überzogen habe², er hat sich bei der Korrektur dieser Schriftlichen Ausarbeitung sogar die Mühe gemacht bei den einzelnen Problemstellen längere, wirklich hilfreiche Textkommentare zu verfassen und obendrauf auch noch Tippfehler usw. angemerkt und war sich nicht zu Schade die Rolle des Nachrichtenübermittlers zwischen mir und Prof. Dr. Scholl zu übernehmen. All dies war absolut keine Selbstverständlichkeit, vor allem wenn ich die Betreuung anderer Studenten, die ich kenne mit der vergleiche, die mir zu Teil wurde.

An den Kommentar zu meinem Betreuer Tobias will ich einen Kommentar zu meinem Gutachter Prof. Dr. Scholl anschließen. Wofür ich meinem Gutachter Prof. Dr. Scholl sehr dankbar bin, ist, dass er meine damals sehr ambitionierten Ideen für mögliche Funktionalitäten, die ich in den PicoC-Compiler für die Bachelorarbeit implementierten wollte runtergeschraubt hat. Man erlebt es äußerst selten im Studium, dass Studenten freiwillig weniger Arbeit gegeben wird. Bei den für die Bachelorarbeit zu implementierenden Funktionalitäten gab es bei der Implementierung viele unerwartete kleine Details, die ich vorher garnicht bedacht hatte, die in ihrer Masse unerwartet viel Zeit zum Implementieren gebraucht haben. Mit den von Prof. Dr. Scholl festgelegten Funktionalitäten für die Bachelorarbeit ist der Zeitplan jedoch ziemlich perfekt aufgegangen. Mit meinen ambitionierten Plänen wäre es bei der Bachelorarbeit dagegeben wohl mit der Zeit äußerst kritisch geworden. Das Prof. Dr. Scholl mir zu seinem eigenen Nachteil^{3 4} weniger Arbeit aufgebrummt hat empfand ich als ich eine äußerst nette Geste, die ich sehr geschätzt habe.

¹Wie auch Prof. Dr. Scholl. Hier geht es aber erstmal um Tobias.

²Wofür ich mich auch nochmal Entschuldigen will.

³Der PicoC-Compiler hätte schließlich mehr Funktionalitäten haben können.

⁴Vielleicht finde ich ja noch im nächsten Semester während des Betriebssysteme Tutorats noch etwas Zeit einige weitere Features einzubauen oder möglicherweise im Rahmen eines Masterprojektes 😊.

Wie mein Betreuer M.Sc. Tobias Seufert und wahrscheinlich auch mein Gutachter Prof. Dr. Scholl im Verlauf dieser Bachelorarbeit und des vorangegangenen Bachelorprojektes gemerkt haben, kann ich schon manchmal ziemlich eigensinnigen sein, bei der Weise, wie ich bestimmte Dinge umsetzen will. Ich habe es sehr geschätzt, dass mir das durchgehen gelassen wurde. Es ist, wie ich die Universitätswelt als Student erlebe bei Arbeitsvorgaben keine Selbstverständlichkeit, dass dem Studenten überhaupt die Freiheit und das Vertrauen gegeben wird diese auf seine eigenen Weise umzusetzen.

Vor allem, da mein eigenes Vorgehen größtenteils Vorteile für mich hatte, da ich auf diese Weise am meisten über Compilerbau gelernt hab und eher Nachteile für Prof. Dr. Scholl, da mein eigenes Vorgehen entsprechend mehr Zeit brauchte und ich daher als Bachelorarbeit keinen dazu passenden RETI-Emulator mit Graphischer Anzeige implementieren konnte, da die restlichen Funktionalitäten des PicoC-Compilers noch implementiert werden mussten.

Glücklicherweise gibt es aber doch noch einen passenden RETI-Emulator, der den PicoC-Compiler über seine Kommandozeilenargumente aufruft, um ein PicoC-Programm visuell auf einer RETI-CPU auszuführen. Für dessen Implementierung hat sich Michel Giehl netterweise zur Verfügung gestellt. Daher Danke auch an Michel Giehl, dass er sich mit meinem PicoC-Compiler auseinandergesetzt hat und diesen in seinen RETI-Emulator integriert hat, sodass am Ende durch unsere beiden Arbeiten ein anschauliches Lerntool für die kommenden Studentengenerationen entstehen konnte. Vor allem da er auch mir darin vertrauen musste, dass ich mit meinem PicoC-Compiler nicht irgendeinen Mist baue. Der RETI-Emulator von Michel Giehl ist unter [Link](#)⁵ zu finden.

Mir hat die Implementierung des PicoC-Compilers tatsächlich ziemlich viel Spaß gemacht, da Compilerbau auch in mein persönliches Interessengebiet fällt⁶. Das Aufschreiben dieser Schriftlichen Ausarbeitung hat mir dagegen eher weniger Spaß gemacht⁷. Wobei ich allerdings sagen muss, dass ich eine große Erleichterung verspüre das ganze Wissen über Compilerbau mal aufgeschrieben zu haben, damit ich mir keine Sorgen machen muss dieses ziemlich nützliche Wissen irgendwann wieder zu vergessen. Es hilft einem auch als Programmierer ungemein weiter zu wissen, wie ein Compiler unter der Haube funktioniert, da man sich so viel besser merken, wie eine bestimmte Funktionalität einer Programmiersprache zu verwenden ist. Manch eine Funktionalität einer Programmiersprache kann in der Verwendung ziemlich willkürlich erscheinen, wenn man die technische Umsetzung dahinter im Compiler nicht kennt.

Ich wollte mich daher auch noch dafür Bedanken, dass mir ein so ergiebiges und interessantes Thema als Bachelorarbeit vorgeschlagen wurde und vor allem, dass auch das Vertrauen in mich gesteckt wurde, dass ich am Ende auch einen funktionsfähigen, sauber programmierten und gut durchdachten Compiler implementiere.

Zum Schluss nochmal ein abschließendes Danke an meinen Betreuer M.Sc Seufert und meinen Gutachter Prof. Dr. Scholl für die Betreuung und Bereitstellung dieser interessanten Bachelorarbeit und des vorangegangenen Bachelorprojektes und Michel Giehl für das Integrieren des PicoC-Compilers in seinen RETI-Emulator.

⁵<https://github.com/michel-giehl/Reti-Emulator>.

⁶Womit nicht alle Studenten so viel Glück haben.

⁷Dieses ständige überlegen, wo man möglicherweise eine Erklärücke hat, ob man nicht was wichtiges ausgelassen hat usw.

Inhaltsverzeichnis

Abbildungsverzeichnis	I
Codeverzeichnis	II
Tabellenverzeichnis	III
Definitionsverzeichnis	V
Grammatikverzeichnis	VI
1 Motivation	1
1.1 RETI-Architektur	2
1.2 Die Sprache PicoC	4
1.3 Eigenheiten der Sprachen C und PicoC	5
1.4 Gesetzte Schwerpunkte	11
1.5 Über diese Arbeit	12
1.5.1 Still der Schriftlichen Ausarbeitung	13
1.5.2 Aufbau der Schriftlichen Arbeit	14
2 Einführung	16
2.1 Compiler und Interpreter	16
2.1.1 T-Diagramme	18
2.2 Formale Sprachen	21
2.2.1 Ableitungen	24
2.2.2 Präzedenz und Assoziativität	27
2.3 Lexikalische Analyse	28
2.4 Syntaktische Analyse	31
2.5 Code Generierung	40
2.5.1 Monadische Normalform	40
2.5.2 A-Normalform	42
2.5.3 Ausgabe des Maschinencodes	44
2.6 Fehlermeldungen	44
Literatur	A

Abbildungsverzeichnis

1.1	Schritte zum Ausführen eines Programmes mit dem GCC.	1
1.2	Stark vereinfachte Schritte zum Ausführen eines Programmes.	1
1.3	Speicherorganisation.	3
1.4	README.md im Github Repository der Bachelorarbeit.	12
2.1	Horizontale Übersetzungszwischenschritte zusammenfassen.	20
2.2	Vertikale Interpretierungszwischenschritte zusammenfassen.	21
2.3	Veranschaulichung von Linksassoziativität und Rechtsassoziativität.	28
2.4	Veranschaulichung von Präzedenz.	28
2.5	Veranschaulichung der Lexikalischen Analyse.	31
2.6	Veranschaulichung des Unterschieds zwischen Ableitungsbaum und Abstraktem Syntaxbaum.	38
2.7	Veranschaulichung der Syntaktischen Analyse.	39
2.8	Codebeispiel für das Trennen von Ausdrücken mit und ohne Nebeneffekten.	41
2.9	Codebeispiel für das Entfernen Komplexer Ausdrücke aus Operationen.	44

Codeverzeichnis

1.1	Beispiel für Spiralregel.	6
1.2	Ausgabe von Beispiel für Spiralregel.	6
1.3	Beispiel für unterschiedliche Ausführung.	7
1.4	Ausgabe des Beispiels für unterschiedliche Ausführung.	7
1.5	Beispiel mit Dereferenzierungsoperator.	7
1.6	Ausgabe des Beispiels mit Dereferenzierungsoperator.	7
1.7	Beispiel dafür, dass Struct kopiert wird.	8
1.8	Ausgabe von Beispiel, dass Struct kopiert wird.	8
1.9	Beispiel dafür, dass Zeiger auf Feld übergeben wird.	9
1.10	Ausgabe von Beispiel dafür, dass Zeiger auf Feld übergeben wird.	9
1.11	Beispiel für Deklaration und Definition.	10
1.12	Ausgabe von Beispiel für Deklaration und Definition.	10
1.13	Beispiel für Sichtbarkeitsbereichs.	11
1.14	Ausgabe von Beispiel für Sichtbarkeitsbereichs.	11

Tabellenverzeichnis

1.1 Präzedenzregeln von PicoC.	3
--	---

Definitionsverzeichnis

1.1	Imperative Programmierung	5
1.2	Strukturierte Programmierung	5
1.3	Prozedurale Programmierung	6
1.4	Call by Value	8
1.5	Call by Reference	9
1.6	Funktionsprototyp	9
1.7	Deklaration	10
1.8	Definition	10
1.9	Sichtbarkeitsbereich (bzw. engl. Scope)	11
2.1	Pipe-Filter Architekturpattern	16
2.2	Interpreter	17
2.3	Compiler	17
2.4	Maschinensprache	17
2.5	Immediate	18
2.6	Cross-Compiler	18
2.7	T-Diagram Programm	19
2.8	T-Diagram Übersetzer (bzw. eng. Translator)	19
2.9	T-Diagram Interpreter	19
2.10	T-Diagram Maschine	20
2.11	Symbol	21
2.12	Alphabet	21
2.13	Wort	21
2.14	Formale Sprache	22
2.15	Syntax	22
2.16	Semantik	22
2.17	Formale Grammatik	22
2.18	Chromsky Hierarchie	23
2.19	Reguläre Grammatik	23
2.20	Kontextfreie Grammatik	24
2.21	Wortproblem	24
2.22	1-Schritt-Ableitungsrelation	24
2.23	Ableitungsrelation	25
2.24	Links- und Rechtsableitungableitung	25
2.25	Linksrekursive Grammatiken	25
2.26	Formaler Ableitungsbaum	25
2.27	Mehrdeutige Grammatik	27
2.28	Assoziativität	28
2.29	Präzedenz	28
2.30	Pattern	29
2.31	Lexeme	29
2.32	Lexer (bzw. Scanner oder auch Tokenizer)	29
2.33	Literal	31
2.34	Konkrete Syntax	32
2.35	Konkrete Grammatik	32
2.36	Ableitungsbaum (bzw. Konkreter Syntaxbaum, engl. Derivation Tree)	33
2.37	Parser	33
2.38	Erkenner (bzw. engl. Recognizer)	34

2.39 Transformer	35
2.40 Visitor	36
2.41 Abstrakte Syntax	36
2.42 Abstrakte Grammatik	36
2.43 Abstrakter Syntaxbaum (bzw. engl. Abstract Syntax Tree, kurz AST)	36
2.44 Pass	40
2.45 Reiner Ausdruck (bzw. engl. pure expression)	41
2.46 Unreiner Ausdruck	41
2.47 Monadische Normalform (bzw. engl. monadic normal form)	41
2.48 Location	42
2.49 Atomarer Ausdruck	42
2.50 Komplexer Ausdruck	42
2.51 A-Normalform (ANF)	43
2.52 Fehlermeldung	45

Grammatikverzeichnis

2.1	Produktionen für einen Ableitungsbaum in EBNF	26
2.2	Produktionen für Ableitungsbaum in EBNF	38
2.3	Produktionen für Abstrakten Syntaxbaum in ASF	38

1 Motivation

Als Programmierer kommt man nicht drumherum einen **Compiler** zu nutzen, er ist geradezu **essentiell** für den Beruf oder das Hobby des Programmierens. Selbst in der Programmiersprache **Python**, welche als **interpretierte** Sprache bekannt ist, wird das in der Programmiersprache **Python** geschriebene Programm vorher zu **Bytecode** kompiliert, bevor dieser von der **Python Virtual Machine (PVM)** interpretiert wird. Die Programmiersprache **Python** und jegliche **andere Sprache** wird fortan als L_{Python} bzw. als $L_{Name\ der\ Sprache}$ bezeichnet wird.

Compiler, wie der **GCC**¹ oder **Clang**² werden üblicherweise über eine **Commandline-Schnittstelle** verwendet, welche es für den Benutzer **unkompliziert** macht ein Programm zu **Maschinencode** zu kompilieren. Das Programm muss hierzu in der Programmiersprache geschrieben sein, die der Compiler kompiliert³

Meist funktioniert das über schlichtes und einfaches **Angeben der Datei**, die das Programm enthält, welches kompiliert werden soll, z.B. im Fall des **GCC** über `> gcc program.c -o machine_code`⁴. Als Ergebnis erhält man im Fall des **GCC** die mit der Option `-o` selbst benannte Datei `machine_code`, welche dann z.B. unter **Unix-Systemen** über `> ./machine_code` **ausgeführt** werden kann, wenn das **Ausführungsrecht** gesetzt ist. Das gesamte gerade erläuterte Vorgehen ist in Abbildung 1.1 veranschaulicht.

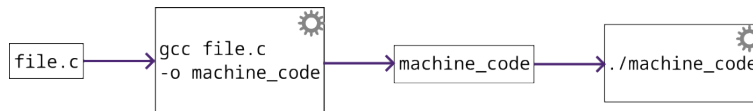


Abbildung 1.1: Schritte zum Ausführen eines Programmes mit dem GCC.

Der ganze Kompiliervorgang kann, wie er in Abbildung 1.2 dargestellt ist zu einer Box **Compiler** abstrahiert werden. Der Benutzer gibt ein **Programm** in der Sprache des Compilers rein und erhält **Maschinencode**, den er dann im besten Fall in eine andere Box hineingeben kann, welche die passende **Maschine** oder den passenden **Interpreter** in Form einer **Virtuellen Maschine** repräsentiert. Die Maschine bzw. der Interpreter kann den **Maschinencode** dann **ausführen**.

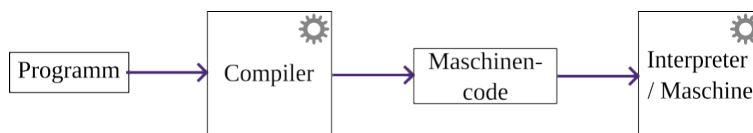


Abbildung 1.2: Stark vereinfachte Schritte zum Ausführen eines Programmes.

¹GCC, the GNU Compiler Collection - GNU Project.

²clang: C++ Compiler.

³Im Fall des **GCC** und **Clang** ist es die Programmiersprache L_C .

⁴Bei **mehreren Dateien** ist das ganze allerdings etwas komplizierter, weil der **GCC** beim Angeben aller `.c`-Dateien nacheinander `gcc program_1.c ... program_n.c` nicht darauf achtet doppelten Code zu entfernen. Beim **GCC** muss am besten mittels einer **Makefile** dafür gesorgt werden, dass jede Datei einzeln zu **Objectcode** (Definition ??) kompiliert wird. Das Kompilieren zu **Objectcode** geht mittels des Befehls `gcc -c program_1.c ... program_n.c` und alle **Objectdateien** können am Ende mittels des **Linkers** mit dem Befehl `gcc -o machine_code program_1.o ... program_n.o` zusammen gelinkt werden.

Der Programmierer muss für das Vorgehen in Abbildung 1.2 **nichts** über die **Theoretischen Grundlagen des Compilerbau** wissen, noch wie der Compiler **intern** umgesetzt ist. In dieser Bachelorarbeit soll diese **Compilerbox** allerdings geöffnet werden und anhand eines eigenen im Vergleich zum **GCC** im Funktionsumfang **reduzierten Compilers** gezeigt werden, wie so ein Compiler **unter der Haube** stark vereinfacht funktioniert.

Die konkrete **Aufgabe** besteht darin einen sogenannten **PicoC-Compiler** zu implementieren, der die **Programmiersprache** L_{PicoC} , welche eine Untermenge der Sprache L_C ist⁵ in eine zu **Lernzwecken** prädestinierte, **unkompliziert** gehaltene **Maschinensprache** L_{RETI} kompilieren kann. Im Unterkapitel 1.1 wird näher auf die **RETI-Architektur** eingegangen, die der Sprache L_{RETI} zu Grunde liegt und im Unterkapitel 1.2 wird näher auf die Sprache L_{PicoC} eingegangen, welche der **PicoC-Compiler** zur eben erwähnten Sprache L_{RETI} kompilieren soll.

1.1 RETI-Architektur

Die **RETI-Architektur** ist eine zu Lernzwecken für die Vorlesungen C. Scholl, „Betriebssysteme“ und P. D. C. Scholl, „Technische Informatik“ entwickelte **32-Bit** Architektur, die sich vor allem durch ihre einfache Zugänglichkeit kennzeichnet und deren **Maschinensprache** L_{RETI} als Zielsprache des **PicoC-Compilers** hergenommen wurde. In der Vorlesung P. D. C. Scholl, „Technische Informatik“ wird die **grundlegende RETI-Architektur** erklärt und in der Vorlesung C. Scholl, „Betriebssysteme“ wird diese Architektur erweitert, sodass diese mehr darauf angepasst ist, dass auch komplexere Konstrukte, wie ein **Betriebssystem**, **Interrupts**, **Prozesse**, **Funktionen** usw. auf nicht zu komplizierte Weise implementiert werden können.

Um den den **PicoC-Compiler** zu **testen** war es notwendig einen **RETI-Interpreter** zu implementieren, der genau die Variante der **RETI-Architektur** aus der Vorlesung C. Scholl, „Betriebssysteme“ **simuliert**.

Anmerkung

In dieser **Bachelorarbeit** wird **im Folgenden** bei der **Maschinensprache** L_{RETI} immer von der Variante, welche durch die **RETI-Architektur** aus der Vorlesung C. Scholl, „Betriebssysteme“ umgesetzt ist ausgegangen.

Die **Register** der **RETI-Architektur** werden in Tabelle 1.1 aufgezählt und erläutert. Die **Maschinenbefehle** und **Datenpfade** der **RETI-Architektur** sind im Kapitel ?? dokumentiert, da diese **nicht explizit** zum **Verständnis** der späteren Kapitel notwendig sind, aber zum **vollständigen Verständnis** notwendig sind, um die später auftauchenden **RETI-Befehle** usw. **zu verstehen**. Der **Aufbau** der **Maschinensprache** L_{RETI} ist durch Grammatik ?? und Grammatik ?? zusammengefasst beschrieben. Für genauere **Implementierungsdetails** ist allerdings auf die Vorlesungen P. D. C. Scholl, „Technische Informatik“ und C. Scholl, „Betriebssysteme“ zu verweisen.

⁵Die der **GCC** kompilieren kann.

Register Kürzel	Register Ausgeschrieben	Aufgabe
PC	Program Counter	Zeigt auf den Maschinenbefehl , der als nächstes ausgeführt werden soll.
ACC	Accumulator	Für Operanden von Operationen oder für temporäre Werte.
IN1	Indexregister 1	Hat dieselbe Aufgabe wie das ACC-Register.
IN2	Indexregister 2	Hat dieselbe Aufgabe wie das ACC-Register.
SP	Stackpointer	Zeigt immer auf die erste freie Speicherzelle am Ende des Stacks , wo als nächstes Speicher allokiert werden kann.
BAF	Beginn Aktive Funktion	Zeigt auf den Beginn des Stackframes der aktuell aktiven Funktion .
CS	Codesegment	Zeigt auf den Beginn des Codesegments . Die letzten 10 Bits werden verwendet, um 22 Bit Immediates aufzufüllen . Kann dadurch dazu verwendet werden, festzulegen welcher der 3 Peripheriegeräte ^a in der Memory Map ^b angesprochen werden soll.
DS	Datensegment	Zeigt auf den Beginn des Datensegments .

^a **EPROM**, **UART** und **SRAM**.

^b Da die **Memory Map** zum Verständnis der Bachelorarbeit **nicht wichtig** ist, wird diese **nicht** mehr als nötig im weiteren Verlauf **erläutert**.

Tabelle 1.1: Präzedenzregeln von PicoC.

Die **RETI-Architektur** ermöglicht bei der Ausführung von RETI-Programmen **Prozesse** zu nutzen. In Abbildung 1.3 ist der **Aufbau** eines **Prozesses** im **Hauptspeicher** der **RETI-Architektur** zu sehen. Das RETI-Programm nutzt dabei den **Stack** für **temporäre Zwischenergebnisse** von Berechnungen und zum **Anlegen der Stackframes** von **Funktionen**, welche die **Lokalen Variablen** und **Parameter** einer Funktion speichern. Das SP- und BAF-Register erfüllen dabei ihre zugeteilten Aufgaben für den Stack.

Der Abschnitt für die **Globalen Statischen Daten** ist allgemein dazu da Daten zu beherbergen, die für den Rest der Programmausführung **global** zugänglich sein sollen, aber auch für die **Lokalen Variablen** der **main-Funktion**. Das DS-Register markiert den **Anfang** des **Datensegments** und damit auch den **Anfang** der **Globalen Statischen Daten** und kann als **relativer Orientierungspunkt** beim **Zugriff** und **Abspeichern** Globaler Statischer Daten dienen. Das CS-Register wird als **relativer Orientierungspunkt** genutzt, an dem die **Ausführung** von RETI-Programmen **startet** und zur Bestimmung der **relativen Startadresse**, an welcher der **RETI-Code** einer bestimmten **Funktion** anfängt.

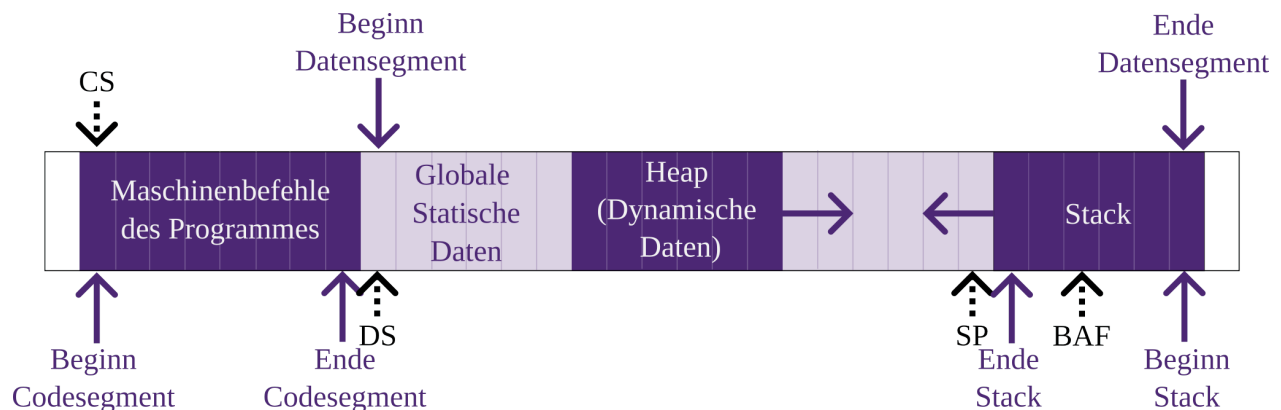


Abbildung 1.3: Speicherorganisation.

Die RETI-Architektur nutzt 3 verschiedene **Peripheriegeräte**, **EPROM**, **UART** und **SRAM**, die über eine **Memory Map**⁶ den über die **Datenpfade** der RETI-Architektur ?? ansprechbaren **Adressraum** von 2^{32} Adressen⁷ unter sich aufteilen.

Die **Ausführung** eines Programmes **startet** auf die einfachste Weise, indem es von einem **Startprogramm** im **EPROM**⁸ aufgerufen wird. Der **EPROM** wird beim Start einer **RETI-CPU** als **erstes** aufgerufen, da nach der **Memory Map** der erste **Adressraum** von 0 bis $2^{30} - 1$ dem **EPROM** zugeordnet ist und das PC-Register **initial** den Wert 0 hat, also als **erstes** das Programm ausgeführt wird, welches an **Adresse** 0 im EPROM anfängt.

Die **UART**⁹ ist eine **elektronische Schaltung** mit je nach Umsetzung mehr oder weniger **Pins**. Es gibt allerdings immer einen **RX**- und einen **TX**-Pin, für jeweils Empfangen¹⁰ und Versenden¹¹ von Daten. Jeder der **Pins** wird dabei mit einer anderen **Adresse** von 2^3 verschiedenen Adressen angesprochen. Jeweils 8-Bit können nach den **Datenpfaden der RETI-CPU** ?? auf einmal über einen Pin in ein **Register** der **UART** geschrieben werden, um **versandt** zu werden oder von einem Pin **empfangen** werden. Die **UART** kann z.B. genutzt werden, um Daten an einen sehr einfach gehaltenen **Monitor** zu senden, der diese dann anzeigt.

An letzter Stelle muss der **SRAM**¹² erwähnt werden, bei dem es sich um den **Hauptspeicher** der **RETI-CPU** handelt. Der **Zugriff auf den Hauptspeicher** ist deutlich schneller als z.B. auf ein **externes Speichermedium**, aber **langsamer** als der **Zugriff auf Register**.

1.2 Die Sprache PicoC

Die Sprache L_{PicoC} ist eine **Untermenge** der Sprache L_C , welche

- **Einzeilige Kommentare** `//` und **Mehrzeilige Kommentare** `/* comment */`.
- die **Basisdatentypen**¹³ `int`, `char` und `void`.
- die **Zusammengesetzten** Datentypen¹⁴ **Felder** (z.B. `int ar[3]`), **Verbunde** (z.B. `struct st {int attr1; attr2;}`) und **Zeiger** (z.B. `int *pntr`) und ihre zugehörigen **Operationen** `[i]`, `.attr` und `*` usw.
- `if(cond){ }-` und `else{ }-`**Anweisungen**¹⁵.
- `while(cond){ }-` und `do while(cond){ };`**Anweisungen**.
- **Arihmatische und Bitweise Ausdrücke**, welche mithilfe der **binären Operatoren** `+`, `-`, `*`, `/`, `%`, `&`, `|`, `^`, `<<`, `>>` und **unären Operatoren** `-`, `~` umgesetzt sind.¹⁶
- **Logische Ausdrücke**, welche mithilfe der **Relationen** `==`, `!=`, `<`, `>`, `<=`, `>=` und **Logischer Verknüpfungen** `!`, `&&`, `||` umgesetzt sind.

⁶Da die **Memory Map** zum Verständnis der Bachelorarbeit **nicht wichtig** ist, sondern nur bei der Umsetzung des **RETI-Interpreters**, wird diese **nicht näher erläutert** als notwendig.

⁷Von 0 bis $2^{32} - 1$.

⁸Kurz für **Erasable Programmable Read-Only Memory**.

⁹Kurz für **Universal Asynchronous Receiver Transmitter**.

¹⁰Engl. **R**eceiving, daher das **R**.

¹¹Engl. **T**ransmission, daher das **T**.

¹²Kurz für **Static random-access memory**.

¹³Bzw. **int** und **char** werden auch als **Primitive Datentypen** bezeichnet.

¹⁴Bzw. engl. **compound datatypes**.

¹⁵Was die Kombination von `if` und `else`, nämlich `else if(cond){ }` miteinschließt.

¹⁶Theoretisch sind die Operatoren `<<`, `>>` und `~` unnötig, da sie durch **Multiplikation** `*`, **Division** `/` und Anwendung des **Xor**-`^`-Operators auf eine Zahl, deren **binäre Repräsentation** ein Folge von `len` **gleicher Länge** ist ersetzt werden können.

- **Zuweisungen**, die mit dem **Zuweisungsoperator** = umgesetzt sind.
- **Funktionsdeklaration** (z.B. `int fun(int arg1[3], struct st arg2);`), **Funktionsdefinition** (z.B. `int fun(int arg1[3], struct st arg2){}`) und **Funktionsaufrufe** (z.B. `fun(ar, st_var)`).

beinhaltet. Die ausgegrauten • wurden bereits für das **Bachelorprojekt** umgesetzt und mussten für die **Bachelorarbeit** nur an die **neue Architektur** angepasst werden.

Der **Aufbau** der **Programmiersprache** L_C ist durch Grammatik ?? und Grammatik ?? zusammengekommen beschrieben.

1.3 Eigenheiten der Sprachen C und PicoC

Einige **Eigenheiten** der Programmiersprache L_C , die genauso ein Teil der Programmiersprache L_{PicoC} sind, da L_{PicoC} eine Untermenge von L_C ist und welche in der Implementierung des **PicoC-Compilers** in Kapitel ?? noch eine **wichtige Rolle** spielen werden im Folgenden genauer erläutert. Im Folgenden wird immer von der Programmiersprache L_{PicoC} gesprochen, da es in dieser Bachelorarbeit um diese geht und die folgenden Beispiele für die Ausgaben alle mithilfe des **PicoC-Compilers** und **RETI-Interpreters** **kompiliert** bzw. **ausgeführt** wurden, aber selbiges gilt genauso für L_C aus bereits erläuterten Grund.

Bei der Programmiersprache L_{PicoC} handelt es sich um eine **imperative** (Definition 1.1), **strukturierte** (Definition 1.2) und **prozedurale Programmiersprache** (Definition 1.3). Aufgrund dessen, dass es sich bei beiden um **Imperative Programmiersprachen** handelt ist es wichtig bei der Implementierung die **Reihenfolge** zu beachten und aufgrund dessen, dass es sich bei beiden um **Strukturierte** und **Prozedurale Programmiersprachen** handelt, ist es eine gute Methode bei der Implementierung auf **Blöcke**¹⁷ zu setzen zwischen denen **hin und her** gesprungen werden kann und welche in den einzelnen Implementierungsschritten die **notwendige Datenstruktur** darstellen um **Auswahl** zwischen Codestücken, **Wiederholung** von Codestücken und **Sprünge** zu Blöcken mit entsprechend zu bestimmten **Bezeichnern** (Definition ??) passenden **Labeln** (Definition ??) umzusetzen.

Definition 1.1: Imperative Programmierung



Wenn ein Programm aus einer **Folge von Befehlen** besteht, deren **Reihenfolge** auch bestimmt in welcher **Reihenfolge** diese **Befehle** auf einer **Maschine** ausgeführt werden.^a

^aThiemann, „Einführung in die Programmierung“.

Definition 1.2: Strukturierte Programmierung



Wenn ein Programm anstelle von z.B. `goto label`-Anweisungen **Kontrollstrukturen**, wie z.B. `if(cond) { } else { }`, `while(cond) { }` usw. verwendet, welche dem Programmcode **mehr Struktur** geben, weil die **Auswahl** zwischen Anweisungen und die **Wiederholung** von Anweisungen eine **klare und eindeutige Struktur** hat, welche bei Umsetzung mit einer `goto label`-Anweisung **nicht so eindeutig erkennbar** wäre und auch **nicht** unbedingt immer **gleich aufgebaut** wäre.^a

^aThiemann, „Einführung in die Programmierung“.

¹⁷Werden später im Kapitel ?? genauer erklärt.

Definition 1.3: Prozedurale Programmierung

Programme werden z.B. mittels **Funktionen** in **überschaubare Unterprogramme** bzw. **Prozeduren** aufgeteilt, die **aufrufbar** sind. Dies **vermeidet** einerseits **redundanten Code**, indem Code **wiederverwendbar** gemacht wird und andererseits erlaubt es z.B. Codestücke nach ihren Aufgaben zu **abstrahieren**, den Codestücken wird eine **Aufgabe zugeteilt**, sie werden zu **Unterprogrammen** gemacht und fortan über einen **Bezeichner aufgerufen**, was den Code deutlich **überschaubarer** macht. da man die Aufgabe eines Codestücks nun nur noch mit seinem **Bezeichner assoziieren** muss.^a

^aThiemann, „Einführung in die Programmierung“.

In L_C ist die Bestimmung des **Datentyps** einer Variable etwas **komplizierter** als in manch anderen Programmiersprachen. Der Grund liegt darin, dass die eckigen $[<i>]</i>$ -Klammern zur Festlegung der **Mächtigkeit** eines Feldes **hinter** der **Variable** stehen: $\langle \text{remaining-datatype} \rangle \langle i \rangle$, während andere Programmiersprachen die eckigen $[<i>]</i>$ -Klammern **vor** die Variable schreiben $\langle \text{remaining-datatype} \rangle [i] \langle \text{var} \rangle$.

Werden die eckigen $[<i>]</i>$ -Klammern **hinter** die Variable geschrieben, ist es **schwieriger** den **Datentyp abzulesen**, als auch ein **Programm zu implementieren** was diesen erkennt. Damit ein Programmierer den **Datentyp ablesen** kann, kann dieser die **Spiralregel** verwenden, die unter der Webseite *Clockwise/Spiral Rule* nachgelesen werden kann. Werden die eckigen $[<i>]</i>$ -Klammern **hinter** die Variable geschrieben, wirken diese zum verwechseln ähnlich zum $\langle \text{var} \rangle [i]$ -Operator für den **Zugriff auf den Index eines Feldes**. Wenn ein Ausdruck geschrieben wird, wie $\text{int ar}[1] = \{42\}$ wird, ist dieser vom Ausdruck $\text{var}[0] = 42$ nur durch den **Kontext** um $\text{var}[1]$ bzw. $\text{var}[0]$ rum zu unterscheiden.

In Code 1.1 ist ein Beispiel zu sehen, indem die Variable `complex_var` den **Datentyp „Feld der Mächtigkeit 1 von Feldern der Mächtigkeit 2 von Zeigern auf Felder der Mächtigkeit 2 von Verbunden vom Typ st“** hat. Ein Vorteil die eckigen $[<i>]</i>$ -Klammern **hinter** die Variable zu schreiben ist in der **markierten Zeile** in Code 1.1 zu sehen. Will man auf ein **Element dieses Datentyps** zugreifen ($\text{*complex_var}[0][1][1].\text{attr}$, so ist der Ausdruck fast genau **gleich aufgebaut**, wie der Ausdruck für den **Datentyp** `struct st (*complex_var[1][2])[2]`. Die **Ausgabe** des Beispiels in Code 1.1 ist in Code 1.2 zu sehen.

```
1 struct st {int attr;};
2
3 void main() {
4     struct st st_var[2] = {{.attr=314}, {.attr=42}};
5     struct st (*complex_var[1][2])[2] = {{&st_var, &st_var}};
6     print((*complex_var[0][1])[1].attr);
7 }
```

Code 1.1: Beispiel für Spiralregel.

```
1 42
```

Code 1.2: Ausgabe von Beispiel für Spiralregel.

In L_C ist die **Ausführbarkeit einer Operation** oder **wie** diese Operation ausgeführt wird davon abhängig, was für einen **Datentyp** die Variable in diesem **Kontext** der Operation hat. In dem Beispiel in Code 1.3 wird in Zeile 2 ein „**Feld der Mächtigkeit 1 von Feldern der Mächtigkeit 2**“ und Zeile 3 ein „**Zeiger auf Felder**“

der **Mächtigkeit 2**“ erstellt. In den **markierten Zeilen** wird zweimal in Folge die **gleiche Operation** `<var>[0][1]` ausgeführt, allerdings hat die Operation aufgrund der **unterschiedlichen Datentypen** der Variablen einen **unterschiedlichen Effekt**.

In Zeile 4 wird ein **normaler Zugriff** auf den **zweiten Eintrag** im **ersten Eintrag** des Felds `int ar[1][2] = {{314, 42}}` durchgeführt und in Zeile 5 wird allerdings erst dem **Zeiger** `int (*pntr)[2] = &ar[0]` **gefolgt** und dann ein Zugriff auf den **zweiten Eintrag** im **ersten Eintrag** des Felds `int ar[1][2] = {{314, 42}}` durchgeführt. **Beide Operationen** haben, wie in Code 1.4 zu sehen ist die **gleiche Ausgabe**.

```
1 void main() {
2   int ar[1][2] = {{314, 42}};
3   int (*pntr)[2] = &ar[0];
4   print(ar[0][1]);
5   print(pntr[0][1]);
6 }
```

Code 1.3: Beispiel für unterschiedliche Ausführung.

```
1 42 42
```

Code 1.4: Ausgabe des Beispiels für unterschiedliche Ausführung.

Eine weitere interessante Eigenheit, die tatsächlich nur in der **Untermenge** von L_C , die L_{PicoC} darstellt gültig ist¹⁸, ist dass die Operationen `<var>[0][1]` und `*(<var>+0)+1` aus Code 1.3 und Code 1.5 komplett **austauschbar** sind. Die Ausgabe in Code 1.4 ist folglich **identisch** zur Ausgabe in Code 1.6.

```
1 void main() {
2   int ar[1][2] = {{314, 42}};
3   int (*pntr)[2] = &ar[0];
4   print(*(&ar+0)+1);
5   print(*(&pntr+0)+1);
6 }
```

Code 1.5: Beispiel mit Dereferenzierungsoperator.

```
1 42 42
```

Code 1.6: Ausgabe des Beispiels mit Dereferenzierungsoperator.

In der Programmiersprache L_{PicoC} werden alle **Argumente** bei einem Funktionsaufruf nach der **Call By Value**-Strategie (Definition 1.4) übergeben. Ein Beispiel hierfür ist in Code 1.7 zu sehen. Hierbei wird

¹⁸In der Sprache L_C gibt es einen **Unterschied** bei der **Initialisierung** bei z.B. `int *var = "string"` und `int var[1] = "string"`, der allerdings nichts mit den Operatoren `*var` und `var[1]` zu tun hat, sondern mit der **Initialisierung**, bei der die Sprache L_C verwirrenderweise die eckigen Klammern `[]` genauso, wie beim **Operator für den Zugriff auf einen Feldindex**, hinter den Bezeichner schreibt (z.B. `var[1]`), obwohl es ein **Zusammengesetzter Datentyp** ist.

ein **Verbund** `struct st copyable_ar = {.ar={314, 314}};`¹⁹ an die Funktion `fun` übergeben. Hierzu wird der **Verbund** in den **Stackframe** der **aufgerufenen** Funktion `fun` **kopiert** und an den Parameter `fun` gebunden.

Wie an der Ausgabe in Code 1.7 zu sehen ist hat die **Zuweisung** `copyable_ar.ar[1] = 42` an den Parameter `struct st copyable_ar` in der **aufgerufenen** Funktion `fun` keinen Einfluss auf die übergebene **lokale Variable** `copyable_ar` der **aufzufendenden** Funktion. Der Eintrag an Index 1 im Feld bleibt bei 314.

Definition 1.4: Call by Value



*Es wird eine **Kopie** des **Ergebnisses** eines **Ausdrucks**, welcher ein **Argument** eines **Funktionsaufrufes** darstellt an den entsprechenden **Parameter** der **aufgerufenen** Funktion gebunden.*

*Das hat zur Folge, dass bei **Übergabe** einer Variable als **Argument** an eine Funktion, diese Variable bei **Änderungen** am entsprechenden **Parameter** der **aufgerufenen** Funktion in der **aufzufendenden** Funktion **unverändert** bleibt.^a*

^aBast, „Programmieren in C“.

```
1 struct st {int ar[2];};
2
3 int fun(struct st copyable_ar) {
4     copyable_ar.ar[1] = 42;
5 }
6
7 void main() {
8     struct st copyable_ar = {.ar={314, 314}};
9     print(copyable_ar.ar[1]);
10    fun(copyable_ar);
11    print(copyable_ar.ar[1]);
12 }
```

Code 1.7: Beispiel dafür, dass Struct kopiert wird.

```
1 314 314
```

Code 1.8: Ausgabe von Beispiel, dass Struct kopiert wird.

In der Programmiersprache L_{PicoC} gibt es **kein Call by Reference** (Definition 1.5), allerdings kann der **Effekt** von Call by Reference mittels **Zeigern simuliert** werden, wie es in Code 1.11 bei der Funktion `fun_declared_before` und dem Parameter `int *param` zu sehen ist. Genau dieser **Trick** wird bei **Feldern** verwendet, um **nicht** das gesamte Feld bei einem Funktionsaufruf in den **Stackframe** der **aufgerufenen** Funktion `fun` **kopieren** zu müssen.

Wie im Beispiel in Code 1.9 zu sehen ist, wird in der markierten Zeile ein Feld `int ar[2] = {314, 314}` an die Funktion `fun` übergeben. Wie in der **Ausgabe** in Code 1.10 zu sehen ist, hat sich der Eintrag an Index 1 im Feld nach dem **Funktionsaufruf** zu 42 geändert. Wird ein Feld direkt als Ausdruck `ar` ohne z.B. die eckigen `[]`-Klammern für einen **Indexzugriff** hingeschrieben wird die **Adresse** des Felds verwendet und **nicht** z.B. der **erste Eintrag** des Felds.

¹⁹Später wird darauf eingegangen, warum der Verbund den **Bezeichner** `copyable_ar` erhalten hat.

Eine Möglichkeit ein **Feld** als **Kopie** und **nicht** als **Referenz** zu übergeben ist es, wie in Code 1.7 das **Feld** als **Attribut** eines **Verbundes** zu übergeben, wie bei der Variable `copyable_ar`.

Definition 1.5: Call by Reference

*Es wird eine **implizite Referenz** einer Variable, welche ein **Argument eines Funktionsaufrufes** darstellt an den entsprechenden **Parameter der aufgerufenen Funktion** gebunden.*

***Implizit** meint hier, dass der Benutzer einer Programmiersprache mit Call by Reference **nicht mitbekommt**, dass er das **Argument selbst verändert** und **keine lokale Kopie des Arguments**.^a*

^aBast, „Programmieren in C“.

```
1 int fun(int ar[2]) {
2     ar[1] = 42;
3 }
4
5 void main() {
6     int ar[2] = {314, 314};
7     print(ar[1]);
8     fun(ar);
9     print(ar[1]);
10 }
```

Code 1.9: Beispiel dafür, dass Zeiger auf Feld übergeben wird.

```
1 314 42
```

Code 1.10: Ausgabe von Beispiel dafür, dass Zeiger auf Feld übergeben wird.

Ein Programm in der Programmiersprache L_{PicoC} wird von **oben-nach-unten** ausgewertet. Ein Problem tritt auf, wenn z.B. eine Funktion verwendet werden soll, die aber erst **unter** dem entsprechenden Funktionsaufruf **definiert** (Definition 1.8) wird. Es ist wichtig, dass der **Prototyp** (Definition 1.6) einer Funktion vorher durch die **Funktionsdefinition** bekannt ist, damit überprüft werden kann, ob die beim Funktionsaufruf **übergebenen Argumente** den gleichen **Datentyp** haben, wie die **Parameter des Prototyps** und ob die **Anzahl Argumente** mit der **Anzahl Parameter des Prototyps** übereinstimmt.

Allerdings lassen sich die Funktionen **nicht** immer so anordnen, dass jede in einem Funktionsaufruf referenzierte Funktion vorher definiert sein kann. Aus diesem Grund ist es möglich den **Prototyp** einer Funktion vorher zu **deklarieren** (Definition 1.7), wie es in den **markierten Zeile** im Beispiel in Code 1.11 zu sehen ist. Die **Ausgabe** des Beispiels ist in Code 1.12 zu sehen.

Definition 1.6: Funktionsprototyp

***Deklaration** einer Funktion, welche den **Funktionsbezeichner**, die **Datentypen** der einzelnen **Funktionsparameter**, die **Parameterreihenfolge** und den **Rückgabewert** einer Funktion spezifiziert. Es ist **nicht** möglich zwei **Funktionsprototypen** mit dem **gleichen Funktionsbezeichner** zu haben.^{a,b}*

^aDer **Funktionsprototyp** ist von der **FunktionsSignatur** zu unterscheiden, die in Programmiersprache wie **C++** und **Java** für die **Auflösung** von **Überladung** bei z.B. **Methoden** verwendet wird und sich in manchen Sprachen für den **Rückgabewert** interessiert und in manchen nicht, je nach Umsetzung. In solchen Sprachen ist es möglich mehrere **Metho-**

den oder **Funktionen** mit dem gleichen Bezeichner zu haben, solange sie sich durch die **Datentypen** von **Parametern**, die **Parameterreihenfolge**, manchmal auch **Sichtbarkeitsbereiche** und **Klassentypen** usw. unterscheiden.

^b What is the difference between function prototype and function signature?

Definition 1.7: Deklaration

Der **Datentyp** bzw. **Prototyp** einer **Variablen** bzw. **Funktion**, sowie der **Bezeichner** dieser **Variable** bzw. **Funktion** wird dem **Compiler** mitgeteilt.^{a b c}

^aÜber das **Schlüsselwort** **extern** lassen sich in der Programiersprache L_C Variablen **deklarieren**, ohne sie zu **definieren**.

^bVariablen in C und C++, Deklaration und Definition — Coder-Welten.de.

^cP. Scholl, „Einführung in Embedded Systems“.

Definition 1.8: Definition

Dem **Compiler** wird mitgeteilt, dass zu einem **bestimmten Zeitpunkt** in der Programmausführung **Speicher** für eine **Variable** **angelegt** werden soll und **wo**^a dieser angelegt werden soll. Eine **Funktion** ist definiert ihr eine **relative Anfangsadresse** im **Hauptspeicher** zugewiesen werden kann, aber welcher die **Maschinenbefehle** für diese Funktion abgespeichert werden können.^{b c}

^aIm Fall des PicoC-Compilers in den **Globalen Statischen Daten** oder auf dem **Stack**.

^bVariablen in C und C++, Deklaration und Definition — Coder-Welten.de.

^cP. Scholl, „Einführung in Embedded Systems“.

```

1 void fun_declared_before(int *param);
2
3 int fun_defined(int param) {
4     return param + 10;
5 }
6
7 void main() {
8     int res = fun_defined(22);
9     fun_declared_before(&res);
10    print(res);
11 }
12
13 void fun_declared_before(int *param) {
14     *param = *param + 10;
15 }
```

Code 1.11: Beispiel für Deklaration und Definition.

```

1 42
```

Code 1.12: Ausgabe von Beispiel für Deklaration und Definition.

In L_{PicoC} lässt sich eine **definierte Variable** nur innerhalb ihres **Sichtbarkeitsbereichs** (Definition 1.9) verwenden. **Lokale Variablen** und **Parameter** lassen sich nur innerhalb der **Funktion** in welcher sie deklariert bzw. definiert wurden verwenden. Der **Sichtbarkeitsbereich** von **Lokalen Variablen** und

Parametern erstreckt sich hierbei von der **öffnenden** `{`-Klammer bis zur **schließenden** `}`-Klammer der **Funktionsdefinition**, in welcher sie definiert wurden.

Verschiedene **Sichtbarkeitsbereiche** können dabei **identische** Bezeichner besitzen. Im Beispiel in Code 1.13 kommt der markierte **Bezeichner** `local_var` in 2 verschiedenen **Sichtbarkeitsbereichen** vor, doch bezeichnet er 2 **unterschiedliche Variablen**. Der **Parameter** `param` und die **Lokale Variable** `local_var` dürfen **nicht** den **gleichen Bezeichner** haben, da sie sich im gleichen **Sichtbarkeitsbereich** der Funktion `fun_scope` befinden. Die **Ausgabe** des Beispiels in Code 1.13 ist in Code 1.14 zu sehen.

Definition 1.9: Sichtbarkeitsbereich (bzw. engl. Scope)



*Bereich in einem Programm, in dem eine Variable **sichtbar** ist und **verwendet** werden kann.^a*

^aThiemann, „Einführung in die Programmierung“.

```
1 int fun_scope(int param) {
2     int local_var = 2;
3     print(param);
4     print(local_var);
5 }
6
7 void main() {
8     int local_var = 4;
9     fun_scope(local_var);
10 }
```

Code 1.13: Beispiel für Sichtbarkeitsbereichs.

```
1 4 2
```

Code 1.14: Ausgabe von Beispiel für Sichtbarkeitsbereichs.

1.4 Gesetzte Schwerpunkte

Ein **Schwerpunkt** dieser Bachelorarbeit ist es in **erster Linie** bei der Kompilierung der Programmiersprache L_{PicoC} in die Maschinensprache L_{RETI} die **Syntax** und **Semantik** der Sprache L_C identisch nachzuahmen. Der **PicoC-Compiler** soll die Sprache L_{PicoC} im Vergleich zu z.B. dem **GCC**²⁰ ohne merklichen Unterschied²¹ kompilieren können.

In **zweiter Linie** soll dabei möglichst immer so Vorgegangen werden, wie es die **RETI-Codeschnipsel** aus der Vorlesung C. Scholl, „Betriebssysteme“ vorgeben. Allerdings sollten diese bei **Inkonsistenzen** bezüglich der durch sie selbst vorgegebenen **Paradigmen** und anderen **Umstimmigkeiten** angepasst werden, da der **erstere Schwerpunkt** überwiegt.

²⁰Da die Sprache L_{PicoC} eine **Untermenge** von L_C ist, kann der **GCC** L_{PicoC} ebenfalls kompilieren, allerdings **nicht** in die gewünschte Maschinensprache L_{RETI} .

²¹Natürlich mit **Ausnahme** der sich unterscheidenden **Maschinensprachen** zu welchen kompiliert wird und der unterschiedlichen **Commandline-Optionen** und **Fehlermeldungen**.

Des Weiteren ist die **Laufzeit** bei Compilern zwar vor allem in der Industrie **nicht unwichtig**, aber bei **Compilern**, verglichen mit **Interpretern** weniger zu gewichten, da ein Compiler bei einem fertig implementierten Programm nur **einmal** Maschinencode generieren muss und dieser Maschinencode danach fortan ausgeführt wird. Beim einem **Compiler** ist daher eher zu priorisieren, dass der kompilierte **Maschinencode** möglichst **effizient** ist.

Beim **PicoC-Compiler** wurde daher eher darauf Wert gelegt **sauberen** und **strukturierten Code** zu schreiben, den interessierte Studenten bei Interesse selber nachvollziehen können und eine **unkomplizierte Bibliothek** mit **guter Dokumentation**²², nämlich das **Lark Parsing Toolkit**²³ für das **Parzen** (Definition 2.37) zu verwenden. Und wie man auch beim **Ausführen der Tests** (wie in Unterkapitel ?? beschrieben) sieht, macht die **Laufzeit** des Compilers für **übliche** und auch **längere Programme**, wie ein Student sie zu Lernzwecken mit dem Compiler kompilieren würde absolut **keine Probleme**.

1.5 Über diese Arbeit

Der Quellcode des **PicoC-Compilers** ist **öffentlich** unter [Link](#)²⁴ zu finden. In der Datei **README.md** (siehe Abbildung 1.4) ist unter „**Getting Started**“ ein kleines **Einführungstutorial** verlinkt. Unter „**Usage**“ ist eine **Dokumentation** über die verschiedenen **Command-line Optionen** und verschiedene **Funktionalitäten der Shell** verlinkt. Daneben finden sich noch weitere Links zu möglicherweise interessanten Dokumenten. Der **letzte Commit** vor der Abgabe der **Bachelorarbeit** ist unter [Link](#)²⁵ zu finden.

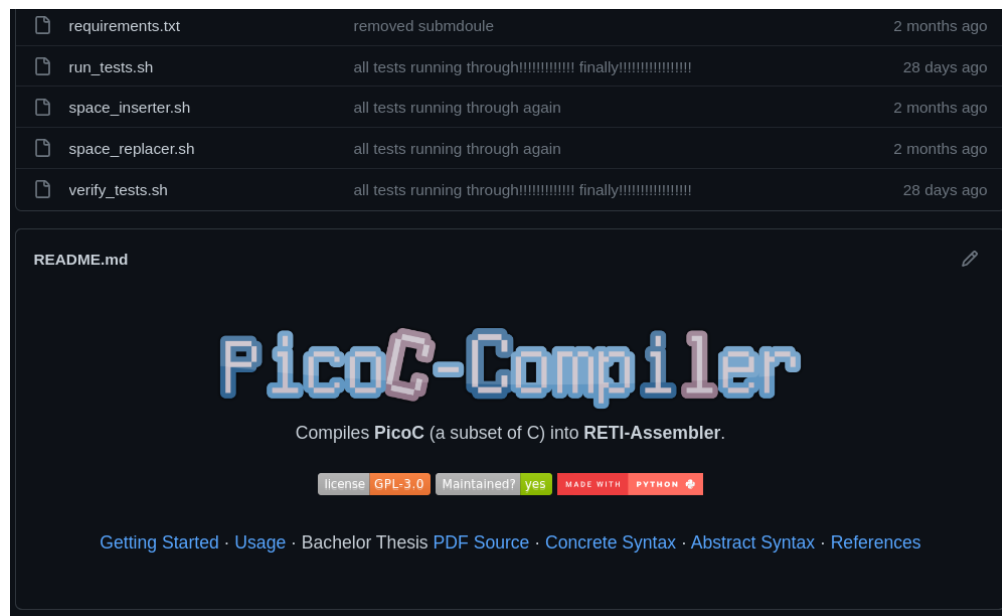


Abbildung 1.4: *README.md* im Github Repository der Bachelorarbeit.

Die **Schriftliche Ausarbeitung** der Bachelorarbeit wurde ebenfalls **veröffentlicht**, falls Studenten, die den **PicoC-Compiler** in Zukunft nutzen sich in der Tiefe dafür interessieren, wie dieser unter der Haube funktioniert. Die **Schriftliche Ausarbeitung** dieser Bachelorarbeit ist als **PDF** unter [Link](#)²⁶ zu finden. Die **PDF** der Schriftliche Ausarbeitung der Bachelorarbeit wird aus dem **Latexquellcode**, welcher unter

²² Welcome to Lark's documentation! — Lark documentation.

²³ Lark - a parsing toolkit for Python.

²⁴ <https://github.com/matthejue/PicoC-Compiler>.

²⁵ <https://github.com/matthejue/PicoC-Compiler/tree/bcafedffa9ff3075372554b14f1a1d369af68971>.

²⁶ https://github.com/matthejue/Bachelorarbeit_out/blob/main/Main.pdf.

Link²⁷ veröffentlicht ist automatisch mithilfe der **Github Action** Nemec, *copy_file_to_another_repo_action* und der **Makefile** Ueda, *Makefile for LaTeX* generiert.

Alle verwendeten **Latex Bibliotheken** sind unter Link²⁸ zu finden²⁹. Die Grafiken, die nicht mittels der Tikz Bibliothek in Latex erstellt wurden, wurden mithilfe des Vectorgraphikeditors **Inkscape**³⁰ erstellt. Falls Interesse besteht **Grafiken** aus der Schriftlichen Ausarbeitung der Bachelorarbeit zu verwenden, so sind diese zusammen mit den .svg-Dateien von **Inkscape** im Ordner `/figures` zu finden.

Alle weitere **verwendete Software**, wie verwendete **Python Bibliotheken**, **Vim/Neovim Plugins**, **Tmux Plugins** usw. sind in der `README.md` unter „References“ bzw. direkt unter Link³¹ zu finden.

Um die verschiedenen **Aspekte** dieser Schriftlichen Ausarbeitung der Bachelorarbeit besser erklären zu können, werden **Codebeispiele** verwendet. In diesem Kapitel **Motivation** werden Codebeispiele zur **Anschauung** verwendet und mithilfe des in den **PicoC-Compiler** integrierten **RETI-Interpreters Ausgaben** erzeugt, die in dieses Dokument **eingelassen** wurden. In Kapitel ?? werden kleine repräsentative **PicoC-Programme** in wichtigen **Zwischenstadien der Kompilierung** in Form von Codebeispielen gezeigt³².

Die **Codebeispiele** wurden alle mit dem **PicoC-Compiler** kompiliert und danach **nicht** mehr **verändert**, also genauso, wie der **PicoC-Compiler** sie kompiliert aus den Dateien in dieses Dokument eingelassen. Alle hier zur Repräsentation verwendeten **PicoC-Programme** lassen sich unter dem Link³³ finden. Mithilfe der im Ordner `/code_examples` beiliegenden `/Makefile` und dem Befehl `> make compile-all` lassen sich die Codebeispiele genauso **kompilieren**, wie sie hier dargestellt sind³⁴.

1.5.1 Still der Schriftlichen Ausarbeitung

In dieser **Schriftliche Ausarbeitung der Bachelorarbeit** sind die manche **Wörter** für einen besseren Lesefluss **hervorgehoben**. Es ist so gedacht, dass die **Hervorgehobenen Wörter** beim Lesen sichtbare **Ankerpunkte** darstellen an denen sich **orientiert** werden kann, aber auch damit der **Inhalt** eines vorher gelesener **Paragraphs** nochmal durch Überfliegen der Hervorgehobenen Wörter in **Erinnerung gerufen** werden kann.

Bei den **Erklärungen** wurden darauf geachtet bei jeder der verwendeten **Methodiken** und jeder **Designentscheidung** die Frage zu klären, „**warum** etwas genau so gemacht wurde und nicht anders“, denn wie es im Buch LeFever, *The Art of Explanation* auf eine deutlich ausführlichere Weise dargelegt wird, ist einer der **zentralen Fragen**, die ein Leser in erster Linie zum **wirklichen Verständnis** eines Themas beantwortet braucht³⁵ die Frage des „**warum**“.

Zum **Verweis auf Quellen** an denen sich z.B. bei der Formulierung von **Definitionen** orientiert wurde, wurden um den **Lesefluss** nicht zu stören **Fußnoten**³⁶ verwendet. Die meisten Definitionen wurden in **eigenen Worten** formuliert, damit die Definitionen selbst zueinander **konsistent** sind, wie auch das in Ihnen verwendete **Vokabular**. Wurde eine Definition **wörtlich** aus einer Quelle übernommen, so wurde die Definition oder der entsprechende Teil in „**Anführungszeichen**“ gesetzt. Beim Verweis auf Quellen **außerhalb** einer

²⁷<https://github.com/matthejue/Bachelorarbeit>.

²⁸https://github.com/matthejue/Bachelorarbeit/blob/master/content/Packete_und_Deklarationen.tex.

²⁹Jede einzelne verwendete Latex **Bibliothek** einzeln anzugeben wäre allerdings etwas zu aufwendig.

³⁰Developers, *Draw Freely — Inkscape*.

³¹https://github.com/matthejue/PicoC-Compiler/blob/new_architecture/doc/references.md.

³²Also die verschiedenen in den **Passes** generierten **Abstrakten Syntaxbäume**, sofern der **Pass** für den gezeigten Aspekt relevant ist.

³³https://github.com/matthejue/Bachelorarbeit/tree/master/code_examples.

³⁴Es wurde zu diesem Zweck die **Command-line Option** `-t`, `--thesis` erstellt, die bestimmte Kommentare **herausfiltert**, damit die generierten **Abstrakten Syntaxbäume** in den verschiedenen Zwischenstufen der Kompilierung **nicht** zu **überfüllt** mit Kommentaren sind.

³⁵Vor allem **Anfang**, wo der Leser **wenig** über das Thema **weiß**.

³⁶Das ist ein **Beispiel** für eine **Fußnote**.

Definitionsbox, wurde allerdings meistens, sofern die **Quelle** wirklich **relevant** war auf das **Zitieren über Fußnoten** verzichtet.

In den **sonstigen Fußnoten** befinden sich **Informationen**, die vielleicht beim **Verständnis** helfen oder **kleinere Details** enthalten, die bei **tiefgreifenderem Interesse** interessant sein könnten. Im Allgemeinen werden die **Informationen in den Fußnoten** allerdings **nicht** zum **Verständnis** der Bachelorarbeit **benötigt**.

Des Weiteren gibt es **Anmerkung**-Kästen, welche kleine **Anmerkungen** enthalten, die über **Konventionen** aufklären sollen, vor **Fallstricken warnen**, die leicht zur Verwirrung führen können oder Informationen bei **tiefergehendem Interesse** oder um **Überblick zu schaffen** enthalten. Der Inhalt dieser **Anmerkung**-Kästen ist allerdings zum Verständnis dieser Arbeit **nicht essentiell** wichtig.

Es wurde immer versucht möglichst **deutsche Fachbegriffe** zu verwenden, sofern sie einigermaßen **geläufig** sind und bei der Verwendung **nicht** eher **verwirren**. Bei dem z.B. auch im Deutschen geläufigen Fachbegriff „**Statement**“ war es eine schwierige Entscheidung, ob man nicht das deutsche Wort „**Anweisung**“ verwenden soll. Da es **nicht verwirrend** klingt wurde sich dazu entschieden überall das deutsche Wort „**Anweisung**“ zu verwenden.

1.5.2 Aufbau der Schriftlichen Arbeit

Der Hauptteil der **Schriftlichen Ausarbeitung** der Bachelorarbeit ist in 4 Kapitel unterteilt: **Motivation**, **Einführung**, ?? und ??. **Zusätzlich** gibt es noch den ??.

Im momentanen Kapitel **Motivation** wurde ein kurzer **Einstieg** in das Thema **Compilerbau** gegeben, die **zentrale Aufgabenstellung** der Bachelorarbeit erläutert und **Schwerpunkte** gesetzt, sowie auf **Eigenheiten der Sprache L_C** eingegangen, die für die **Implementierung** relevant sein werden.

Im Kapitel **Einführung** werden die notwendigen **Theoretischen Grundlagen** eingeführt, die zum Verständnis des Kapitels **Implementierung** notwendig sind. Das Kapitel soll darüberhinaus aber auch einen **Überblick** über das Thema Compilerbau geben, sodass nicht nur ein Grundverständnis für das eine **spezifische Vorgehen**, welches zur Implementierung des **PicoC-Compiler** verwendet wurde vermittelt wird, sondern auch ein **Vergleich** zu **anderen Vorgehensweisen** möglich ist. Die **Theoretischen Grundlagen** umfassen die wichtigsten Definitionen in Bezug zu Compilern und den verschiedenen **Phasen der Kompilierung**, welche durch die Unterkapitel **Lexikalische Analyse**, **Syntaktische Analyse** und **Code Generierung** repräsentiert sind.

Des Weiteren wurden für **T-Diagramme** und **Formale Sprachen** eigene Unterkapitel erstellt. Für **T-Diagramme** wurde ein eigenes Unterkapitel erstellt, da sie häufig in der Schriftlichen Ausarbeitung verwendet werden und die **T-Diagramm Notation** nicht allgemein bekannt ist. Für **Formale Sprachen** wurde ein eigenes Unterkapitel erstellt, da für den Gutachter Prof. Dr. Scholl das Thema **Formale Sprachen** eher **fachfremd** ist, aber dieses Thema einige zentrale und wichtige Fachbegriffe besitzt, bei denen es wichtig ist die genaue **Definition** zu haben. Generell wurde im Kapitel **Einführung** versucht an Erklärungen nicht zu sparren, damit aufgrund dessen, dass das Thema eher fachfremd für Prof. Dr. Scholl ist für das Kapitel **Implementierung** keine wichtigen Aspekte unverständlich bleiben.

Im Kapitel ?? werden die einzelnen Aspekte der Implementierung des **PicoC-Compilers**, unterteilt in die verschiedenen **Phasen der Kompilierung** nach denen das Kapitel **Einführung** ebenfalls unterteilt ist erklärt. Dadurch, dass Kapitel **Implementierung** und Kapitel **Einführung** eine **ähnliche Kapiteileinteilung** haben, ist es besonders einfach zwischen beiden hin und her zu wechseln. Wenn z.B. eine Definition im Kapitel **Einführung** gesucht wird, die zum Verständnis eines Aspekts in Kapitel **Implementierung** notwendig ist, so kann aufgrund der ähnlichen **Kapiteileinteilung** die entsprechende Definition analog im Kapitel **Einleitung** gefunden werden.

Im Kapitel ?? wird ein **Überblick** über die **wichtigsten Funktionalitäten** des PicoC-Compilers gegeben, indem anhand **kleiner Anleitungen** gezeigt wird wie man diese verwendet. Des Weiteren wird darauf eingegangen, wie die **Qualitätsicherung** für den **PicoC-Compiler** umgesetzt wurde, also wie gewährleistet wird, dass der **PicoC-Compiler** funktioniert. Zum Schluss wird noch auf **weitere Erweiterungsideen** eingegangen, die auch interessant zu implementieren wären.

Im ?? werden einige **Details der RETI-Architektur**, **Sonstigen Definitionen** und das Thema **Bootstrapping** angesprochen. Der **Appendix** dient als eine Lagerstätte für **Definitionen**, **Tabellen**, **Abbildungen** und ganze **Kapitel**, die bei **Interesse** zur **weiteren Vertiefung** da sind und zum Verständnis der anderen Kapitel **nicht notwendig** sind. Damit der **Rote Faden** in dieser Schriftlichen Ausarbeitung der Bachelorarbeit erkennbar bleibt und der **Lesefluss** nicht gestört wird, wurden alle diese Informationen in den **Appendix** ausgelagert.

Die **Sonstigen Defintionen** und das Thema **Bootstrapping** sind dazu da den Bogen von der **spezifischen** Implementierung des **PicoC-Compilers** wieder zum **allgemeinen Vorgehen** bei der Implementierung eines Compilers zu schlagen. Diese **Themen** und **Definitionen** passen nicht ins Kapitel ??, da diese selbst **nichts** mit der Implementierung des **PicoC-Compilers** zu tun haben und auch nichts ins **Kapitel Einführung**, da dieses nur **Theoretische Grundlagen** erklärt, die für das Kapitel **Implementierung** wichtig sind.

Generell wurde in der Schriftlichen Ausarbeitung immer versucht **Parallelen** zu Implementierung **echter** Compiler zu ziehen. Der Zweck des **PicoC-Compilers** ist es primär ein **Lerntool** zu sein, weshalb Methoden, wie **Liveness Analyse** (Definition ??) usw., die in **echten** Compilern zur Anwendung kommen **nicht umgesetzt** wurden, da sich an die **vorgegebenen Paradigmen** aus der Vorlesung C. Scholl, „Betriebssysteme“ gehalten werden sollte.

2 Einführung

In diesem Kapitel wird auf die Theoretischen Grundlagen eingegangen, die zum Verständnis der Implementierung in Kapitel ?? notwendig sind. Zuerst wird in Unterkapitel 2.1 genauer darauf eingegangen was ein Compiler und Interpreter eigentlich sind und damit in Verbindung stehende Begriffe erklärt. Danach wird in Unterkapitel 2.2 eine kleine Einführung zu einem der Grundpfeiler des Compilerbau den Formale Sprachen gegeben. Danach werden die einzelnen **Filter** der üblicherweise bei der Implementierung von Compilern genutzten Pipe-Architektur nacheinander erklärt. Die Filter beinhalten die **Lexikalische Analyse** 2.3, **Syntaktische Analyse** 2.4 und **Code Generierung** 2.5. Zum Schluss wird darauf eingegangen in welchen Situationen Fehlermeldungen auszugeben sind.

Definition 2.1: Pipe-Filter Architekturpattern

Ist ein **Architekturpattern**, welches aus **Pipes** und **Filtern** besteht, wobei der **Ausgang** eines **Filters** der **Eingang** des durch eine **Pipe** verbundenen adjazenten nächsten **Filters** ist, falls es einen gibt.

Ein **Filter** stellt einen Schritt dar, indem eine Eingabe **weiterverarbeitet** wird und **weitergereicht** wird. Bei der **Weiterverarbeitung** können Teile der Eingabe **entfernt**, **hinzugefügt** oder **vollständig ersetzt** werden.

Eine **Pipe** stellt ein **Bindeglied** zwischen zwei **Filtern** dar.^{a,b}



^aDas ein **Bindeglied** eine eigene Bezeichnung erhält, bedeutet allerdings nicht, dass es eine eigene wichtige **Aufgabe** erfüllt. Wie bei vielen **Pattern**, soll mit dem Namen des **Pattern**, in diesem Fall durch das **Pipe** die Anlehnung an z.B. die **Pipes aus Unix**, z.B. `cat /proc/bus/input/devices | less` zum Ausdruck gebracht werden. Und so banal es klingt, sollen manche Bezeichnungen von Pattern auch einfach nur gut klingen.

^bWestphal, „Softwaretechnik“.

2.1 Compiler und Interpreter

Die wohl wichtigsten zu klärenden Begriffe, sind die eines **Compilers** (Definition 2.3) und eines **Interpreters** (Definition 2.2), da das Schreiben eines Compilers von der **PicoC-Sprache** L_{PicoC} in die **RETI-Sprache** L_{RETI} das Thema dieser Bachelorarbeit ist und die Definition eines **Interpreters** genutzt wird, um zu definieren was ein **Compiler** ist. Des Weiteren wurde zur **Qualitätssicherung** ein **RETI-Interpreter** implementiert, um mithilfe des **GCC**¹ und von **Tests** die **Beziehungen** in 2.3.1 zu belegen (siehe Subkapitel ??).

¹Sammlung von Compilern für Linux bzw. GNU-Linux, steht für **GNU Compiler Collection**

Definition 2.2: Interpreter

Interpretiert die Befehle^a oder Anweisungen eines Programmes P direkt.

Auf die Implementierung bezogen arbeitet ein Interpreter auf den compilerinternen **Sub-Bäumen** des **Abstrakten Syntaxbaumes** (wird später eingeführt unter Definition 2.43) und führt je nach Komposition der **Knoten** des Abstrakten Syntaxbaumes, auf die er während des Darüber-Iterierens stösst unterschiedliche Anweisungen aus.^b

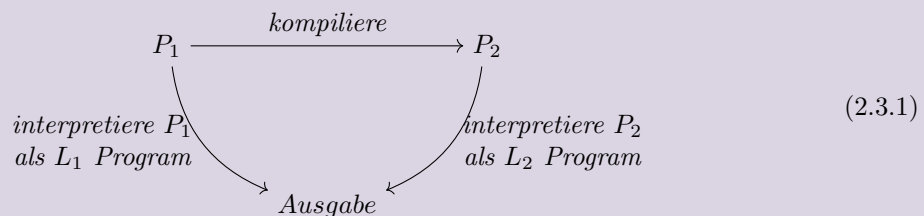
^aMaschinensprache kann genauso interpretiert werden, wie auch eine Programmiersprache.

^bG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Definition 2.3: Compiler

Kompiliert ein beliebiges Program P_1 , welches in einer Sprache L_1 geschrieben ist, in ein Program P_2 , welches in einer Sprache L_2 geschrieben ist.

Wobei **Kompilieren** meint, dass ein beliebiges Program P_1 in der Sprache L_1 so in die Sprache L_2 zu einem Program P_2 übersetzt wird, dass bei beiden Programmen, wenn sie von **Interpretern** ihrer jeweiligen Sprachen L_1 und L_2 **interpretiert** werden, sie die gleiche **Ausgabe** haben, wie es in Diagramm 2.3.1 dargestellt ist. Also beide Programme P_1 und P_2 die gleiche **Semantik** (Definition 2.16) haben und sich nur **syntaktisch** (Definition 2.15) durch die Sprachen L_1 und L_2 , in denen sie geschrieben stehen unterscheiden.^a



^aG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Üblicherweise kompiliert ein **Compiler** ein **Program**, das in einer **Programmiersprache** geschrieben ist zu **Maschinencode**, der in **Maschinensprache** (Definition 2.4) geschrieben ist, aber es gibt z.B. auch **Transpiler** (Definition ??) oder **Cross-Compiler** (Definition 2.6). Des Weiteren sind **Maschinensprache** und **Assemblersprache** (Definition ??) voneinander zu unterscheiden.

Definition 2.4: Maschinensprache

Programmiersprache, deren mögliche Programme die **hardwarenaheste Repräsentation** eines möglicherweise zuvor hierzu kompilierten bzw. assemblierten Programmes darstellen. Jeder Maschinenbefehl entspricht einer bestimmten **Aufgabe**, die die CPU im **vereinfachten Fall** in einem **Zyklus** der **Fetch-** und **Execute-Phase**, genauer gesagt in der **Execute-Phase** übernehmen kann oder allgemein in einer **geringen konstanten** Anzahl von Fetch- und Execute Phasen im **Komplexeren Fall**. Die Maschinenbefehle sind meist so entworfen, dass sie sich innerhalb bestimmter **Wortbreiten**, die **Zweierpotenzen** sind kodieren lassen. Im einfachsten Fall innerhalb einer **Speicherzelle** des **Hauptspeichers**.^{a,b}

^aViele Prozessorarchitekturen erlauben es allerdings auch z.B. **zwei** Maschinenbefehle in **eine** Speicherzelle des Hauptspeichers zu komprimieren, wenn diese zwei Maschinenbefehle keine Operanden mit zu großen **Immediates** (Definition 2.5) haben.

^bC. Scholl, „Betriebssysteme“.

Der **Maschinencode**, den ein üblicher Compiler einer Programmiersprache generiert, enthält seine Folge von Maschinenbefehlen üblicherweise in **binärer Repräsentation**, da diese in erster Linie für die Maschine, die binär arbeitet verständlich sein sollen und nicht für den Programmierer.

Der **PicoC-Compiler**, der den Zweck erfüllt für Studenten ein **Anschauungs- und Lernwerkzeug** zu sein, generiert allerdings Maschinencode, der die Maschinenbefehle bzw. RETI-Befehle in **menschenlesbarer Form** mit ausgeschriebenem RETI-Operationen, RETI-Registern und Immediates (Definition 2.5) enthält. Für den **RETI-Interpreter** ist es ebenfalls nicht notwendig, dass der Maschinencode, den der PicoC-Compiler generiert, in binärer Darstellung ist, denn es ist für den RETI-Interpreter ebenfalls leichter diese einfach direkt in menschenlesbarer Form zu interpretieren, da der RETI-Interpreter nur die sichtbare Funktionsweise einer RETI-CPU **simulieren** soll und nicht deren mögliche interne Umsetzung².

Definition 2.5: Immediate

Konstanter Wert, der als Teil eines Maschinenbefehls gespeichert ist und dessen Wertebereich dementsprechend auch durch die Anzahl an Bits, die ihm innerhalb dieses Maschinenbefehls zur Verfügung gestellt sind beschränkt ist. Der Wertebereich ist beschränkter als bei sonstigen Werten innerhalb des Hauptspeichers, denen eine ganze Speicherzelle des Hauptspeichers zur Verfügung steht.^a

^aLjohhuh, *What is an immediate value?*

Definition 2.6: Cross-Compiler

Kompiliert auf einer Maschine M_1 ein Program, dass in einer Sprache L_w geschrieben ist für eine andere Maschine M_2 , wobei beide Maschinen M_1 und M_2 unterschiedliche Maschinensprachen B_1 und B_2 haben.^{a,b}

^aBeim **PicoC-Compiler** handelt es sich um einen **Cross-Compiler** C_{PicoC}^{Python} , der in der Sprache L_{Python} geschrieben ist und die Sprache L_{PicoC} kompiliert.

^bEarley und Sturgis, „A formalism for translator interactions“.

Ein **Cross-Compiler** ist entweder notwendig, wenn eine Zielmaschine M_2 nicht ausreichend **Rechenleistung** hat, um ein Programm in der Wunschsprache L_w selbst **zeitnah** zu kompilieren oder wenn noch kein Compiler C_w für die **Wunschsprache** L_w und andere Programmiersprachen L_o , in denen man Programmieren wollen würde existiert, der unter der **Maschinensprache** B_2 einer Zielmaschine M_2 läuft.³

2.1.1 T-Diagramme

Um die Architektur von **Compilern** und **Interpretern** übersichtlich darzustellen eignen sich **T-Diagramme**, deren Spezifikation aus der Wissenschaftlichen Publikation Earley und Sturgis, „A formalism for translator interactions“ entnommen ist besonders gut, da diese optimal darauf **zugeschnitten** sind die Eigenheiten von Compilern in ihrer Art der Darstellung unterzubringen.

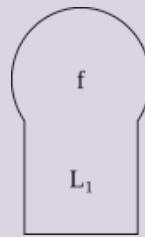
Die **Notation** setzt sich dabei aus den **Blöcken** für ein Program (Definition 2.7), einen Übersetzer (Definition 2.8), einen Interpreter (Definition 2.9) und eine Maschine (Definition 2.10) zusammen.

²Eine **RETI-CPU** zu bauen, die menschenlesbaren Maschinencode in z.B. **UTF-8 Kodierung** ausführen kann, wäre dagegen unnötig kompliziert und aufwändig, da Hardware **binär** arbeitet und man dieser daher lieber direkt die binär kodierten Maschinenbefehle übergibt, anstatt z.B. eine unnötig **platzverbrauchenden** UTF-8 Codierung zu verwenden, die nur in sehr vielen Schritten einen Befehl verarbeiten kann, da die Register und Speicherzellen des Hauptspeichers üblicherweise nur **32- bzw. 64-Bit Breite** haben.

³Die an vielen Universitäten und Schulen eingesetzten programmierbaren Roboter von **Lego Mindstorms** nutzen z.B. einen **Cross-Compiler**, um für den programmierbaren Microcontroller eine **C-ähnliche Sprache** in die Maschinensprache des Microcontrollers zu kompilieren, da der Microcontroller selbst nicht genug Rechenleistung besitzt, um ein Programm selbst zeitnah zu kompilieren.

Definition 2.7: T-Diagramm Programm

Repräsentiert ein **Programm**, dass in der **Sprache** L_1 geschrieben ist und die **Funktion** f berechnet.^a



^aEarley und Sturgis, „A formalism for translator interactions“.

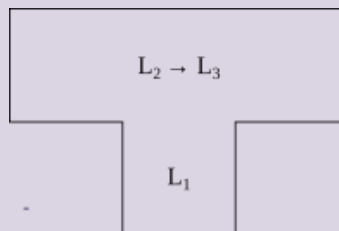
Anmerkung

Es ist bei **T-Diagrammen** nicht notwendig beim entsprechenden **Platzhalter**, in den man die genutzte **Sprache** schreibt, den **Namen der Sprache** an ein L dranzuhängen, weil hier immer eine **Sprache** steht. Es würde in Definition 2.7 also reichen einfach eine 1 hinzuschreiben.

Definition 2.8: T-Diagramm Übersetzer (bzw. eng. Translator)

Repräsentiert einen **Übersetzer**, der in der **Sprache** L_1 geschrieben ist und **Programme** von der **Sprache** L_2 in die **Sprache** L_3 kompiliert.

Für den **Übersetzer** gelten genauso, wie für einen **Compiler**^a die **Beziehungen** in 2.3.1.^b

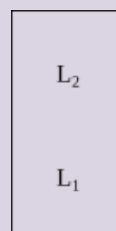


^aZwischen den Begriffen **Übersetzung** und **Kompilierung** gibt es einen kleinen Unterschied, **Übersetzung** ist **kleinschrittiger** als **Kompilierung** und ist auch zwischen **Passes** möglich, **Kompilierung** beinhaltet dagegen bereits alle **Passes** in einem Schritt. **Kompilieren** ist also auch **Übersetzen**, aber **Übersetzen** ist nicht immer auch **Kompilieren**.

^bEarley und Sturgis, „A formalism for translator interactions“.

Definition 2.9: T-Diagramm Interpreter

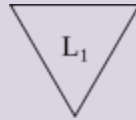
Repräsentiert einen **Interpreter**, der in der **Sprache** L_1 geschrieben ist und **Programme** in der **Sprache** L_2 interpretiert.^a



^aEarley und Sturgis, „A formalism for translator interactions“.

Definition 2.10: T-Diagram Maschine

Repräsentiert eine **Maschine**, welche ein **Programm** in **Maschinensprache** L_1 ausführt.^{a,b}



^aWenn die Maschine **Programme** in einer höheren Sprache als **Maschinensprache** ausführt, ist es auch erlaubt diese Notation zu verwenden, dann handelt es sich um eine **Abstrakte Maschine**, wie z.B. die **Python Virtual Machine** (PVM) oder **Java Virtual Machine** (JVM).

^bEarley und Sturgis, „A formalism for translator interactions“.

Aus den verschiedenen **Blöcken** lassen sich **Kompositionen** bilden, indem man sie **adjazent** zueinander platziert. Allgemein lässt sich grob sagen, dass **vertikale Adjazenz** für **Interpretation** und **horizontale Adjazenz** für **Übersetzung** steht.

Sowohl **horizontale** als auch **vertikale Adjazenz** lassen sich, wie man in den Abbildungen 2.1 und 2.2 erkennen kann zusammenfassen.

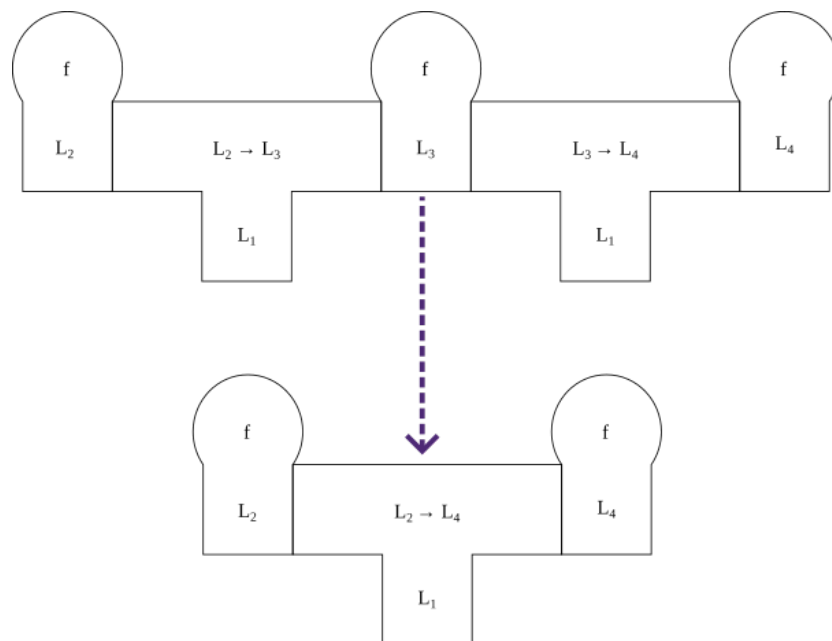


Abbildung 2.1: Horizontale Übersetzungsschritte zusammenfassen.

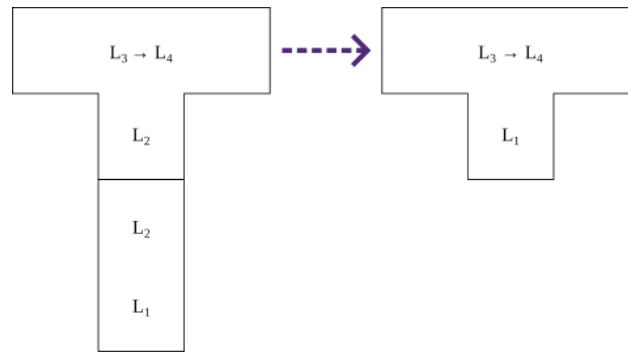


Abbildung 2.2: Vertikale Interpretierungszwischenschritte zusammenfassen.

2.2 Formale Sprachen

Das **Kompilieren** eines Programmes hat viel mit dem Thema **Formaler Sprachen** (Definition 2.14) zu tun, da bereits das Kompilieren an sich das Übersetzen eines Programmes aus der Sprache L_1 in eine Sprache L_2 ist. Aus diesem Grund ist es wichtig die **Grundlagen Formaler Sprachen**, was die Begriffe **Symbol** (Definition 2.11), **Alphabet** (Definition 2.12), **Wort** (Definition 2.13) beinhaltet vorher eingeführt zu haben.

Definition 2.11: Symbol

„Ein Symbol ist ein **Element** eines **Alphabets** Σ .“^a

^aNebel, „Theoretische Informatik“.

Definition 2.12: Alphabet

„Ein Alphabet ist eine **endliche, nicht-leere** Menge aus **Symbolen** (Definition 2.11).“^a

^aNebel, „Theoretische Informatik“.

Definition 2.13: Wort

„Ein Wort $w = a_1 \dots a_n \in \Sigma^*$ ist eine **endliche Folge** von **Symbolen** aus einem **Alphabet** Σ .

Es gibt es die **Konkatenation** $w_1 w_2 = a_1 \dots a_n b_1 \dots b_n$ von **Wörtern** $w_1 = a_1 \dots a_n$ und $w_2 = b_1 \dots b_n$ und die **Länge** eines **Wortes** $|w|$.

Ein wichtiges Wort ist das **leere Wort** ε für das gilt: $|\varepsilon| = 0$ und $\forall w \in \Sigma^* : \varepsilon w = w \varepsilon = w$. Es handelt sich bei ε also um das **Neutrale Element** bei der **Konkatenation** von **Wörtern**.

Bei Σ^* handelt es sich um **Kleenesche Hülle** eines **Alphabets** Σ , es ist die **Sprache aller Wörter**, welche durch beliebige **Konkatenation** von **Symbolen** aus dem **Alphabet** Σ gebildet werden können, wobei $\varepsilon \in \Sigma^*$. Dies ist die **größte Sprache** über Σ und **jede Sprache** über Σ ist eine **Teilmenge** davon.“^a

^aNebel, „Theoretische Informatik“.

Definition 2.14: Formale Sprache

„Eine **Formale Sprache** ist eine Menge von **Wörtern** (Definition 2.13) über dem **Alphabet** Σ (Definition 2.12).“^a

Das Adjektiv „**formal**“ kann dabei weggelassen werden, wenn der **Kontext** indem die **Sprache** verwendet wird **eindeutig** ist, da man das Adjektiv „**formal**“ nur verwendet um den Unterschied zum im normalen Sprachgebrauch verwendeten Begriff einer **Sprache** herauszustellen.

^aNebel, „Theoretische Informatik“.

Bei der Übersetzung eines Programmes von einer Sprache L_1 zur Sprache L_2 muss die **Semantik** (Definition 2.16) **gleich** bleiben. Beide Sprachen L_1 und L_2 haben eine **Grammatik** (Definition 2.17), welche diese beschreibt und können verschiedene **Syntaxen** (Definition 2.15) haben.

Definition 2.15: Syntax

Bezeichnet alles was mit dem **Aufbau** von Wörtern einer **Formalen Sprache** zu tun hat. Eine **Formale Grammatik**, aber auch in **Natürlicher Sprache** ausgedrückte Regeln können die **Syntax** einer Sprache beschreiben. Es kann auch mehrere **verschiedene Syntaxen** für die **gleiche Sprache** geben^{a, b}.

^aZ.B. die **Konkrete** und **Abstrakte Syntax**, die später eingeführt werden.

^bThiemann, „Einführung in die Programmierung“.

Definition 2.16: Semantik

Die **Semantik** bezeichnet alles was mit der **Bedeutung** von Wörtern einer **Formalen Sprache** zu tun hat.^a

^aThiemann, „Einführung in die Programmierung“.

Definition 2.17: Formale Grammatik

„Eine **Formale Grammatik** beschreibt wie **Wörter** einer **Sprache** abgeleitet werden können.

Das Adjektiv „**formal**“ kann dabei weggelassen werden, wenn der **Kontext** indem die **Grammatik** verwendet wird **eindeutig** ist, da man das Adjektiv „**formal**“ nur verwendet um den Unterschied zum im normalen Sprachgebrauch verwendeten Begriff einer **Grammatik** herauszustellen.

Eine **Grammatik** wird durch das Tupel $G = \langle N, \Sigma, P, S \rangle$ dargestellt, wobei“:

- $N \hat{=}$ **Nicht-Terminalsymbole**.
- $\Sigma \hat{=}$ **Terminalsymbole**, wobei $N \cap \Sigma = \emptyset$.^{a, b}
- $P \hat{=}$ Menge von **Produktionsregeln** $w \rightarrow v$, wobei $w, v \in (N \cup \Sigma)^*$ und $w \notin \Sigma^*$.^{c, d}
- $S \hat{=}$ **Startsymbol**, wobei $S \in N$.

„Zusätzlich ist es praktisch **Nicht-Terminalsymbole** N , **Terminalsymbole** Σ und das **leere Wort** ε allgemein als Menge der **Grammatiksymbole** $C = N \cup \Sigma \cup \varepsilon$ zu definieren.

Es ist möglich **zwei Grammatiken** G_1 und G_2 in einer **Vereinigungsgrammatik** $G_1 \uplus G_2 = \langle N_1 \cup N_2 \cup \{S\}, \Sigma, P_1 \cup P_2 \cup \{S ::= S_1 \mid S_2\}, S \rangle$ zu vereinigen.“^{e, f}

^aWeil mit ihnen **terminiert** wird.

^bKann auch als **Alphabet** bezeichnet werden.

^c w muss **mindestens** ein **Nicht-Terminalsymbol** enthalten.

^dBzw. $w, v \in C^*$ und $w \notin \Sigma^*$.

^eDie Produktion $S ::= S_1 \mid S_2$ kann hierbei durch **beliebige** andere Produktionen ersetzt werden, welche die beiden Grammatiken **miteinander verbinden**.

^fNebel, „Theoretische Informatik“.

Die gerade definierten **Formale Sprachen** lassen sich des Weiteren in Klassen der **Chromsky Hierarchie** (Definition 2.18) einteilen.

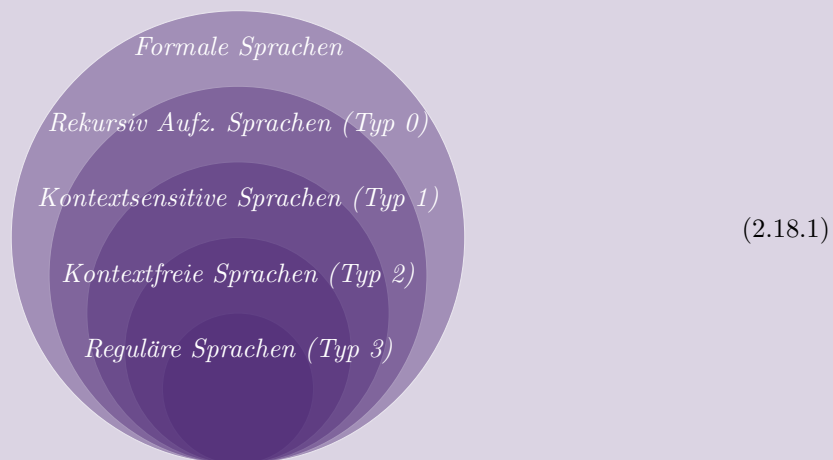
Definition 2.18: Chromsky Hierarchie



Die **Chromsky Hierarchie** ist eine Hierarchie in der **Formale Sprachen** nach der **Komplexität** ihrer **Formalen Grammatiken** in verschiedene **Klassen** unterteilt werden. Jede dieser Klassen hat verschiedene **Eigenschaften**, wie **Entscheidungsprobleme**, die in dieser Klasse **entscheidbar** bzw. **unentscheidbar** sind usw.

Eine Sprache L_i ist in der **Chromsky Hierarchie** vom Typ $i \in \{0, \dots, 3\}$, falls sie von einer Grammatik dieses Typs i erzeugt wird.

Zwischen den Sprachmengen **benachbarter Klassen** in Abbildung 2.18.1 besteht eine **echte Teilmen-genbeziehung**: $L_3 \subset L_2 \subset L_1 \subset L_0$. Jede **Reguläre Sprache** ist auch eine **Kontextfreie Sprache**, aber nicht jede **Kontextfreie Sprache** ist auch eine **Reguläre Sprache**.^a



^aNebel, „Theoretische Informatik“.

Für diese Bachelorarbeit sind allerdings nur die **Spracheklassen** der **Chromsky-Hierarchie** relevant, die von **Regulären** (Definition 2.19) und **Kontextfreien Grammatiken** (Definition 2.20) beschrieben werden.

Definition 2.19: Reguläre Grammatik



„Ist eine Grammatik für die gilt, dass **alle Produktionen** eine der Formen:

$$A \rightarrow cB, \quad A \rightarrow c, \quad A \rightarrow \varepsilon \quad (2.19.1)$$

haben, wobei A, B **Nicht-Terminalsymbole** sind und c ein **Terminalsymbol** ist^{a,b}.“^c

^aDiese Definition einer **Regulären Grammatik** ist **rechtsregulär**, es ist auch möglich diese Definition **linksregulär** zu

formulieren, aber diese Details sind für die Bachelorarbeit nicht relevant.

^bDadurch, dass die **linke** Seite immer nur ein **Nicht-Terminalsymbol** sein darf ist jede **Reguläre Grammatik** auch eine **Kontextfrei Grammatik**.

^cNebel, „Theoretische Informatik“.

Definition 2.20: Kontextfreie Grammatik

„Ist eine Grammatik für die gilt, dass **alle Produktionen** die Form:

$$A \rightarrow v \quad (2.20.1)$$

haben, wobei A ein **Nicht-Terminalsymbol** ist und v ein beliebige Folge von **Grammatiksymbolen**^a ist.“^b

^aAlso eine beliebige Folge von **Nicht-Terminalsymbolen** und **Terminalsymbolen**.

^bNebel, „Theoretische Informatik“.

Ob sich ein Programm überhaupt kompilieren lässt entscheidet sich anhand des **Wortproblems** (Definition 2.21). In einem **Compiler** oder **Interpreter** ist das Wortproblem üblicherweise immer **entscheidbar**. Wenn das Programm ein **Wort** der **Sprache** ist, die der Compiler kompiliert, so klappt das Kompilieren, ist es **kein Wort** der **Sprache**, die der Compiler kompiliert, wird eine **Fehlermeldung** ausgegeben.

Definition 2.21: Wortproblem

Ein Entscheidungsproblem, bei dem man zu einem **Wort** $w \in \Sigma^*$ und einer **Sprache** L als **Eingabe** 1 oder 0^a **ausgibt**, je nachdem, ob dieses Wort w Teil der Sprache L ist $w \in L$ oder nicht $w \notin L$.^b

Das Wortproblem kann durch die folgende **Indikatorfunktion**^c zusammengefasst werden:

$$\mathbb{1}_L : \Sigma^* \rightarrow \{0, 1\} : w \mapsto \begin{cases} 1 & \text{falls } w \in L \\ 0 & \text{sonst} \end{cases} \quad (2.21.1)$$

^aBzw. „ja“ oder „nein“ usw., es muss nicht umgedingt 1 oder 0 sein.

^bNebel, „Theoretische Informatik“.

^cAuch **Charakteristische Funktion** genannt.

2.2.1 Ableitungen

Um sicher zu wissen, ob ein Compiler ein **Programm**⁴ kompilieren kann, ist es möglich das Programm mithilfe der **Grammatik** der **Sprache** des Compilers **abzuleiten**. Hierbei wird zwischen der **1-Schritt-Ableitungsrelation** (Definition 2.22) und der normalen **Ableitungsrelation** (Definition 2.23) unterscheiden.

Definition 2.22: 1-Schritt-Ableitungsrelation

„Eine **binäre Relation** \Rightarrow zwischen Wörtern aus $(N \cup \Sigma)^*$, die alle möglichen Wörter $(N \cup \Sigma)^*$ in Relation zueinander setzt, die sich nur durch das **einmalige** Anwenden einer Produktionsregel voneinander unterscheiden.“

Es gilt $u \Rightarrow v$ **genau dann wenn** $u = w_1 x w_2$, $v = w_1 y w_2$ **und** es eine Regel $x \rightarrow y \in P$ gibt, wobei $w_1, w_2, x, y \in (N \cup \Sigma)^*$ „^a“

^aNebel, „Theoretische Informatik“.

⁴Bzw. **Wort**.

Definition 2.23: Ableitungsrelation

„Eine **binäre Relation** \Rightarrow^* , welche der **reflexive, transitive Abschluss** der **1-Schritt-Ableitungsrelation** \Rightarrow ist. Auf der **rechten Seite** der Ableitungsrelation \Rightarrow^* steht also ein Wort aus $(N \cup \Sigma)^*$, welches durch **beliebig häufiges** Anwenden von Produktionsregeln entsteht.

Es gilt $u \Rightarrow^* v$ **genau dann wenn** $u = w_1 \Rightarrow \dots \Rightarrow w_n = v$, wobei $n \geq 1$ und $w_1, \dots, w_n \in (N \cup \Sigma)^*$.“^a

^aNebel, „Theoretische Informatik“.

Beim Ableiten kann auf verschiedene Weisen vorgegangen werden, dasselbe **Programm**⁵ kann z.B. über eine **Linksableitung** als auch eine **Rechtsableitung** (Definition 2.24) abgeleitet werden. Das ist später bei den verschiedenen **Ansätzen** für das **Parsen** eines Programmes in Unterkapitel 2.4 relevant.

Definition 2.24: Links- und Rechtsableitung

„In jedem **Ableitungsschritt** wird bei **Typ-3- und Typ-2-Grammatiken** auf das am **weitesten links** (**Linksableitung**) bzw. **rechts** (**Rechtsableitung**) stehende **Nicht-Terminalsymbol** eine Produktionsregel angewandt, bei **Typ-1- und Typ-0-Grammatiken** ist es statt einem **Nicht-Terminalsymbol** die **linke Seite** einer Produktion.

Mit diesem Vorgehen kann man jedes ableitbare Wort generieren, denn dieses Vorgehen entspricht **Tiefensuche** von **links-nach-rechts**.“^a

^aNebel, „Theoretische Informatik“.

Manche der **Ansätze** für das **Parsen** eines Programmes haben ein Problem, wenn die Grammatik, die zur Entscheidung des **Wortproblems** für das Programm verwendet wird eine **Linksrekursive Grammatik** (Definition 2.25) ist⁶.

Definition 2.25: Linksrekursive Grammatiken

Eine **Grammatik** ist **linksrekursiv**, wenn sie ein **Nicht-Terminalsymbol** enthält, dass **linksrekursiv** ist.

Ein **Nicht-Terminalsymbol** ist **linksrekursiv**, wenn das **linkeste Symbol** in einer seiner **Produktionen** es selbst ist oder zu sich selbst gemacht werden kann durch eine Folge von Ableitungen:

$$A \Rightarrow^* Aa,$$

wobei a eine beliebige Folge von **Terminalsymbolen** und **Nicht-Terminalsymbolen** ist.^a

^aParsing Expressions · Crafting Interpreters.

Um herauszufinden, ob eine Grammatik **mehrdeutig** (Definition 2.27) ist, werden **Ableitungen** als **Formale Ableitungsbäume** (Definition 2.26) dargestellt. **Formale Ableitungsbäume** werden im Unterkapitel 2.4 nochmal relevant, da in der **Syntaktischen Analyse** Ableitungsbäume (Definition 2.36) als eine **compiler-interne Datenstruktur** umgesetzt werden.

Definition 2.26: Formaler Ableitungsbaum

Ist ein Baum, in dem die Syntax eines **Wortes**^a nach den **Produktionen** der zugehörigen Grammatik, die angewendet werden mussten um das Wort abzuleiten **hierarchisch** zergliedert dargestellt wird.

⁵Bzw. **Wort**.

⁶Für den im **PicoC-Compiler** verwendeten **Earley Parsers** stellt dies allerdings **kein** Problem dar.

Das Adjektiv „**formal**“ kann dabei weggelassen werden, wenn der **Kontext** indem der **Ableitungsbaum** verwendet wird **eindeutig** ist, da man das Adjektiv „**formal**“ nur verwendet um den Unterschied zum **compilerinternen Ableitungsbaum** herauszustellen, der den Formalen Ableitungsbaum als **Datentstruktur** zur einfachen Weiterverarbeitung umsetzt.

Den Knoten dieses Baumes sind **Grammatiksymbole** $C = N \cup \Sigma \cup \varepsilon$ (Definition 2.17) zugeordnet. Die **Inneren Knoten** des Baumes sind **Nicht-Terminalsymbole** N und die **Blätter** sind entweder **Terminalsymbole** Σ oder das **leere Wort** ε .^b

^aZ.B. **Programmcode**.

^bNebel, „Theoretische Informatik“.

In Abbildung 2.26.2 ist ein Beispiel für einen **Formalen Ableitungsbaum** zu sehen, der sich aus der **Ableitung** 2.26.1 nach den im **Dialekt der Erweiterter Backus-Naur-Form des Lark Parsing Toolkit** (Definition ??) angegebenen **Produktionen** 2.1 einer **Grammatik** $G = \langle N, \Sigma, P, add \rangle$ ergibt.

DIG_NO_0	$::=$	"1" "2" "3" "4" "5" "6"	L_Lex
		"7" "8" "9"	
DIG_WITH_0	$::=$	"0" DIG_NO_0	
NUM	$::=$	"0" $DIG_NO_0 DIG_WITH_0^*$	
ADD_OP	$::=$	"+"	
MUL_OP	$::=$	"*"	
mul	$::=$	$mul MUL_OP NUM$ NUM	L_Parse
add	$::=$	$add ADD_OP mul$ mul	

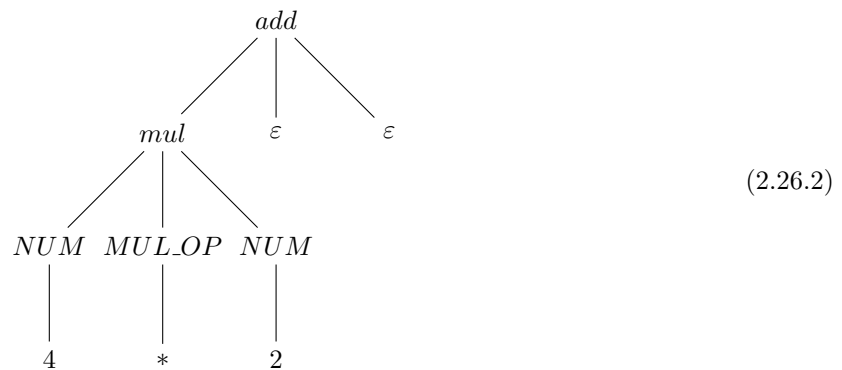
Grammatik 2.1: Produktionen für einen Ableitungsbaum in EBNF

Anmerkung

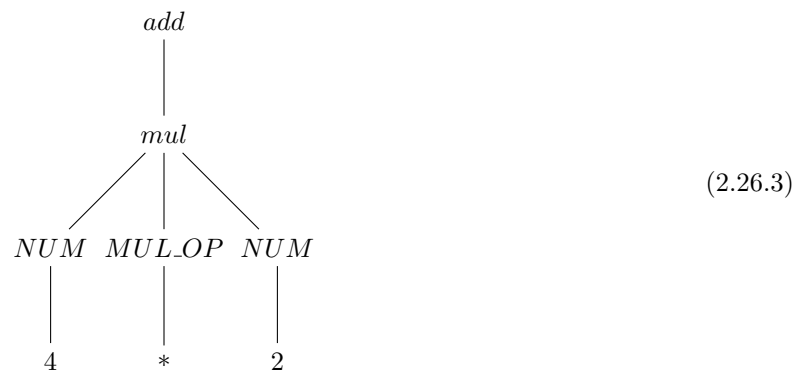
Werden die **Produktionen** einer Grammatik in z.B. **EBNF** angegeben, wie in Grammatik ??, wird die Angabe dieser Produktionen auch oft als **Grammatik** bezeichnet, obwohl Grammatiken eigentlich durch ein Tupel $G = \langle N, \Sigma, P, S \rangle$ dargestellt sind.

$$add \Rightarrow mul \Rightarrow mul \ MUL_OP \ NUM \Rightarrow NUM \ MUL_OP \ NUM \Rightarrow^* "4" \ "*" \ "2" \quad (2.26.1)$$

Bei Ableitungsbäumen gibt es **keine** einheitliche **Regelung**, wie damit umgegangen wird, wenn die **Alternativen** einer Produktion unterschiedliche viele **Nicht-Terminalsymbole** enthalten. Es gibt einmal die Möglichkeit, wie im Ableitungsbaum 2.26.2 von der **Maximalzahl** auszugehen und beim Nicht-Erreichen der Maximalzahl entsprechend der **Differenz zur Maximalzahl** viele **Blätter** mit dem **leeren Wort** ε hinzuzufügen.



Eine andere Möglichkeit ist, wie im Ableitungsbaum 2.26.3 nur die vorhandenen **Nicht-Terminalsymbole** als Kinder hinzuzufügen⁷.



Für einen Compiler ist es notwendig, dass die **Konkrete Grammatik** keine **Mehrdeutige Grammatik** (Definition 2.27) ist, denn sonst können unter anderem die **Präcedenzregeln** der verschiedenen **Operatoren** nicht gewährleistet werden, wie später in Unterkapitel ?? an einem Beispiel demonstriert wird.

Definition 2.27: Mehrdeutige Grammatik



„Eine Grammatik ist **mehrdeutig**, wenn es ein Wort $w \in L(G)$ gibt, das mehrere **Ableitungsbäume** zulässt“.^{a,b}

^aAlternativ, wenn es für w **mehrere** unterschiedliche **Linksableitungen** gibt.

^bNebel, „Theoretische Informatik“.

2.2.2 Präcedenz und Assoziativität

Will man die **Operatoren** aus einer **Programmiersprache** in einer **Konkreten Grammatik** ausdrücken, die **nicht mehrdeutig** ist, so lässt sich das nach einem klaren Schema machen, wenn die **Assoziativität** (Definition 2.28) und **Präcedenz** (Definition 2.29) dieser **Operatoren** festgelegt ist. Dieses Schema wird in Unterkapitel ?? genauer erklärt.

⁷Diese Option wurde beim **PicoC-Compiler** gewählt.

Definition 2.28: Assoziativität

„Bestimmt, welcher Operator aus einer Reihe **gleicher** Operatoren **zuerst** ausgewertet wird.“

Es wird grundsätzlich zwischen **linksassoziativen** Operatoren, bei denen der **linke Operator** vor dem **rechten Operator** ausgewertet wird und **rechtsassoziativen** Operatoren, bei denen es genau anders rum ist unterschieden.^a

^aParsing Expressions · Crafting Interpreters.

Bei **Assoziativität** ist z.B. der **Multiplikationsoperator** `*` ein Beispiel für einen **linksassoziativen** Operator und ein **Zuweisungsoperator** `=` ein Beispiel für einen **rechtsassoziativen** Operator. Dies ist in Abbildung 2.3 mithilfe von Klammern `()` veranschaulicht.

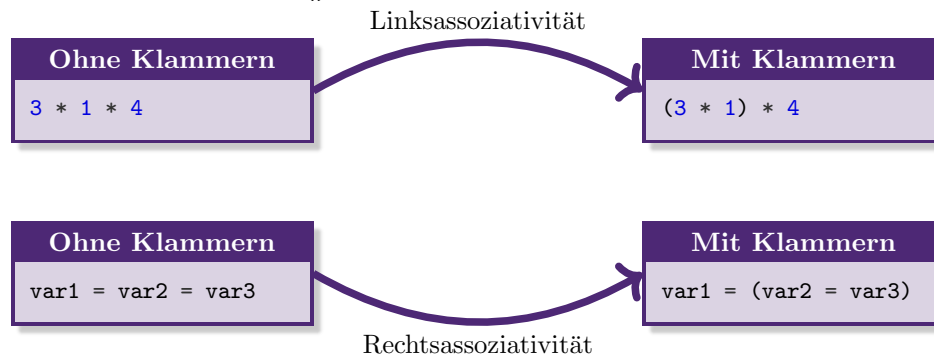


Abbildung 2.3: Veranschaulichung von Linksassoziativität und Rechtsassoziativität.

Definition 2.29: Präzedenz

„Bestimmt, welcher Operator **zuerst** in einem Ausdruck, der eine Mischung **verschiedener** Operatoren enthält, ausgewertet wird. Operatoren mit einer **höheren Präzedenz**, werden **vor** Operatoren mit **niedrigerer Präzedenz** ausgewertet.“^a

^aParsing Expressions · Crafting Interpreters.

Bei **Präzedenz** ist die Mischung der Operatoren für **Subtraktion** `-` und für **Multiplikation** `*` ein Beispiel für den Einfluss von Präzedenz. Dies ist in Abbildung 2.4 mithilfe der Klammern `()` veranschaulicht. Im Beispiel in Abbildung 2.4 ist bei den beiden **Subtraktionsoperatoren** `-` nacheinander und dem darauffolgenden **Multiplikationsoperator** `*` sowohl **Assoziativität** als auch **Präzedenz** im Spiel.



Abbildung 2.4: Veranschaulichung von Präzedenz.

2.3 Lexikalische Analyse

Die **Lexikalische Analyse** bildet üblicherweise den ersten Filter innerhalb des **Pipe-Filter Architektur-patterns** (Definition 2.1) bei der Implementierung von Compilern. Die Aufgabe der lexikalischen Analyse

ist vereinfacht gesagt in einem Eingabewort⁸ endliche Folgen Symbolen⁹ zu finden, die durch bestimmte **Pattern** (Definition 2.30) erkannt werden, die durch eine **reguläre Grammatik** spezifiziert sind. Diese Folgen endlicher Symbole werden auch **Lexeme** (Definition 2.31) genannt.

Definition 2.30: Pattern

Beschreibung aller möglichen **Lexeme**, die eine Menge \mathbb{P}_T bilden und einem bestimmten **Token** T zugeordnet werden. Die Menge \mathbb{P}_T ist eine möglicherweise unendliche Menge von **Wörtern**, die sich mit den Produktionen einer **regulären Grammatik** G_{Lex} einer **regulären Sprache** L_{Lex} beschreiben lassen^a, die für die Beschreibung eines **Tokens** T zuständig sind.^b

^aAls Beschreibungswerkzeug können aber auch z.B. reguläre Ausdrücke hergenommen werden.

^bThiemann, „Compilerbau“.

Definition 2.31: Lexeme

Ein **Lexeme** ist ein **Teilwort** aus dem **Eingabewort**, welches von einem **Pattern** für eines der **Token** T einer **Sprache** L_{Lex} erkannt wird.^a

^aThiemann, „Compilerbau“.

Diese **Lexeme** werden vom **Lexer** (Definition 2.32) im **Eingabewort** identifiziert und **Tokens** T zugeordnet. Das jeweils nächste **Lexeme** fängt dabei genau nach dem letzten Symbol des **Lexemes** an, das zuletzt vom **Lexer** erkannt wurde. Die **Tokens** (Definition 2.32) sind es, die letztendlich an die **Syntaktische Analyse** weitergegeben werden.

Ein **Lexeme** ist dabei nicht immer das gleiche wie der **Tokenwert**, denn z.B. im Fall von L_{PicoC} kann der Wert 99 durch zwei verschiedene **Literale** (Definition 2.33) dargestellt werden, einmal als ASCII-Zeichen 'c', das dann als Tokenwert den entsprechenden Wertes aus der ASCII-Tabelle hat und des Weiteren auch in Dezimalschreibweise als 99¹⁰. Zu einem **Lexeme** wie z.B. 'c' wäre das entsprechende **Token** dazu $\text{Token}('CHAR', '99')$. Bei einem **Lexeme** wie z.B. '99' wäre das entsprechende **Token** $\text{Token}('NUM', '99')$, bei dem das **Lexeme** mit dem **Tokenwert** übereinstimmt. Der **Tokenwert** ist der letztendlich verwendete **Wert** an sich, unabhängig von der Darstellungsform.

Definition 2.32: Lexer (bzw. Scanner oder auch Tokenizer)

Ein **Lexer** ist eine **partielle Funktion** $\text{lex} : \Sigma^* \rightarrow (N \times W)^*$, welche ein **Wort** bzw. **Lexeme** aus Σ^* auf ein **Token** T mit einem **Tokennamen** N und einem **Tokenwert** W abbildet, falls dieses **Wort** sich unter der **regulären Grammatik** G_{Lex} , der **regulären Sprache** L_{Lex} ableiten lässt bzw. einem der **Pattern** der Sprache L_{Lex} entspricht.^a

^aThiemann, „Compilerbau“.

Ein **Lexer** ist im Allgemeinen eine **partielle Funktion**, da es Zeichenfolgen geben kann, die von keinem **Pattern** eines **Tokens** der Sprache L_{Lex} erkannt werden. In Bezug auf eine Implementierung, wird, wenn der Lexer Teil der Implementierung eines Compilers ist, in diesem Fall eine **Fehlermeldung** ausgegeben.

Anmerkung

Um Verwirrung vorzubeugen ist es wichtig folgende Unterscheidung hervorzuheben:

Wenn von **Symbolen** die Rede ist, so werden in der **Lexikalischen Analyse**, der **Syntaktischen**

⁸Z.B. dem Inhalt einer Datei, welche in **UTF-8** kodiert ist.

⁹Also **Teilwörter** des **Eingabeworts**.

¹⁰Die Programmiersprache L_{Python} erlaubt es z.B. dieser Wert auch mit den Literalen **0b1100011** und **0x63** darzustellen.

Analyse und der **Code Generierung**, auf diesen verschiedenen Ebenen unterschiedliche Konzepte als Symbole bezeichnet.

In der Lexikalischen Analyse sind einzelne **Zeichen eines Zeichensatzes** die Symbole.

In der Syntaktischen Analyse sind die **Tokennamen** die Symbole.

In der Code Generierung sind die **Bezeichner** (Definition ??) von **Variablen, Konstanten und Funktionen** die Symbole^a.

^aDas ist der Grund, warum die **Tabelle**, in der Informationen zu **Bezeichnern** gespeichert werden, in Kapitel ?? **Symboltabelle** genannt wird.

Eine weitere Aufgabe der **Lexikalischen Analyse** ist es jegliche für die Weiterverarbeitung unwichtigen Symbole, wie Leerzeichen `␣`, Newline `\n`¹¹ und Tabs `\t` aus dem Eingabewort herauszufiltern. Das geschieht mittels des **Lexers**, der allen für die **Syntaktische Analyse** unwichtigen Zeichen das leere Wort ϵ zuordnet. Das ist auch im Sinne der Definition, denn $\epsilon \in (N \times W)^*$ ist immer der Fall bei der **Kleeneschen Hülle** Σ^* , wobei $\Sigma = N \times W$. Nur das, was für die **Syntaktische Analyse** wichtig ist, soll weiterverarbeitet werden, alles andere wird herausgefiltert.

Der Grund warum nicht einfach nur die **Lexeme** an die **Syntaktische Analyse** weitergegeben werden und der Grund für die Aufteilung des **Tokens** in **Tokenname** und **Tokenwert**, ist, weil z.B. die Bezeichner von Variablen, Konstanten und Funktionen beliebige Zeichenfolgen sein können, wie `my_fun`, `my_var` oder `my_const` und es auch viele verschiedenen Zahlen gibt, wie 42, 314 oder 12. Die **Tokennamen** sind **Überbegriffe** für beliebige Bezeichner von Variablen, Konstanten und Funktionen und beliebige Zahlen.

Für die zuvor als Beispiel genannten **Bezeichner** und **Zahlen** wären z.B. `NAME`¹² und `NUM`¹³ passende **Tokennamen**¹⁴, bzw. wenn man sich nicht Kurzformen sucht `IDENTIFIER` und `NUMBER`. Für **Lexeme**, wie `if` oder `}` sind die **Tokennamen** genau die Bezeichnungen, die man diesen Zeichenfolgen geben würde, nämlich `IF` und `RBRACE`.

Die **Konkrete Grammatik** G_{Lex} , die zur Beschreibung der Token T der Sprache L_{Lex} verwendet wird ist üblicherweise **regulär**, da ein typischer **Lexer** immer nur **ein Symbol** vorausschaut¹⁵, sich **nichts merkt**, also unabhängig davon, **was** für Symbole und **wie oft** bestimmte Symbole **davor** aufgetaucht sind funktioniert. Auch für den PicoC-Compiler lässt sich aus der im **Dialekt der Backus-Naur-Form des Lark Parsing Toolkit** (Definition ??) spezifizierten Grammatik ?? schlussfolgern, dass die **Sprache** des **PicoC-Compilers** für die **Lexikalische Analyse** L_{PicoC_Lex} **regulär** ist, da alle ihre **Produktionen** die Definition 2.19 erfüllen.

Produktionen mit **Alternative**, wie z.B. $DIG_WITH_0 ::= "0" \mid DIG_NO_0$ sind **unproblematisch**, denn sie können immer auch als $\{DIG_WITH_0 ::= "0", DIG_WITH_0 ::= DIG_NO_0\}$ ausgedrückt werden und z.B. $DIG_WITH_0^*$, $(LETTER \mid DIG_WITH_0 \mid "_")^+$ und $"_".^*\sim$ in Grammatik ?? können alle zu **Alternativen** umgeschrieben werden, womit diese **Alternativen** wie gerade gezeigt umgeformt werden können, um ebenfalls **regulär** zu sein. Somit existiert mit der Grammatik ?? eine **reguläre Grammatik**, welche die **Sprache** L_{PicoC_Lex} beschreibt und damit ist die **Sprache** L_{PicoC_Lex} nach der **Chomsky Hierarchie** (Definition 2.18) **regulär**.

¹¹In Unix Systemen wird für Newline das ASCII Symbol **line feed**, in Windows hingegen die ASCII Symbole **carriage return** und **line feed** nacheinander verwendet. Das wird aber meist durch die verwendete Programmiersprache, die man zur Implementierung des Lexers nutzt wegabstrahlt.

¹²Für z.B. `my_fun`, `my_var` und `my_const`.

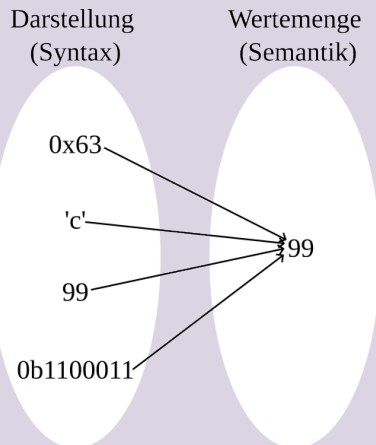
¹³Für z.B. 42, 314 und 12.

¹⁴Diese **Tokennamen** wurden im **PicoC-Compiler** verwendet, da man beim Programmieren möglichst **kurze** und **leicht verständliche** Bezeichner für seine Knoten haben will, damit unter anderem **mehr Code** in eine Zeile passt.

¹⁵Man nennt das auch einem **Lookahead** von 1

Definition 2.33: Literal

Eine von möglicherweise vielen weiteren *Darstellungsformen* (als *Zeichenkette*) für ein und denselben *Wert* eines *Datentyps*.^a



^aThiemann, „Einführung in die Programmierung“.

Um eine Gesamtübersicht über die **Lexikalische Analyse** zu geben, ist in Abbildung 2.5 die Lexikalische Analyse an einem Beispiel veranschaulicht.

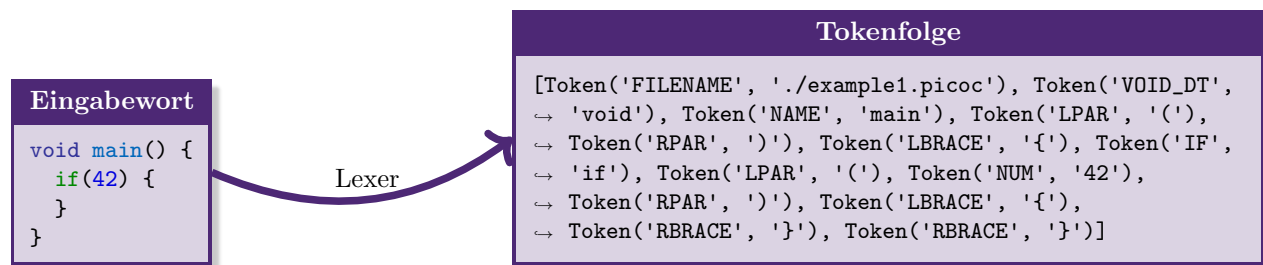


Abbildung 2.5: Veranschaulichung der Lexikalischen Analyse.

Anmerkung

Das Symbol \hookrightarrow zeigt im Code der Tokens in Abbildung 2.5 und in den folgenden Codes einen **Zeilenumbruch** an, wenn eine **Zeile zu lang** ist.

2.4 Syntaktische Analyse

In der **Syntaktischen Analyse** ist für einige Sprachen eine **Kontextfreie Grammatik** G_{Parse} notwendig, um diese Sprachen zu beschreiben, da viele Programmiersprachen z.B. für **Funktionsaufrufe** `fun(arg)` und **Codeblöcke** `if(1){}` syntaktische Mittel verwenden, die es notwendig machen sich zu merken, wieviele öffnende runde Klammern '(' bzw. öffnende geschweifte Klammern '{' es momentan gibt, die noch nicht durch eine entsprechende schließende runde Klammer ')' bzw. schließende geschweifte Klammer '}' geschlossen wurden. Dies lässt sich nicht mehr mit einer **Regulären Grammatik** (Definition 2.19) beschreiben, sondern es braucht eine **Kontextfreie Grammatik** (Definition 2.20) hierfür, die es erlaubt zwischen **zwei Terminalsymbolen** ein **Nicht-Terminalsymbol** abzuleiten.

Für den PicoC-Compiler lässt sich aus der Grammatik ?? schlussfolgern, dass die **Sprache** des **PicoC-Compilers** für die **Syntaktische Analyse** L_{PicoC_Parse} **kontextfrei**, aber nicht mehr **regulär** ist, da **alle** ihre **Produktionen** die Definition für **Kontextfreie Grammatiken 2.20** erfüllen, aber **nicht** die Definition für **Reguläre Grammatiken 2.19**.

Dass die Grammatik **kontextfrei** ist lässt sich auch sehr leicht erkennen, weil **alle Produktionen** auf der **linken Seite** des $::=$ -Symbols immer nur ein **Nicht-Terminalsymbol** haben und auf der **rechten Seite** eine **beliebige Folge** von **Grammatiksymbolen**¹⁶. Dass diese Grammatik aber nicht **regulär** sein kann, lässt sich sehr einfach an z.B. der Produktion $if_stmt ::= "if" ("logic_or")^* exec_part$ erkennen, bei der das **Nicht-Terminalsymbol** $logic_or$ von den **Terminalsymbolen** für **öffnende Klammer** { und **schließende Klammer** } eingeschlossen sein muss, was mit einer **Regulären Grammatik** nicht ausgedrückt werden kann.

Somit existiert mit der Grammatik ?? eine **Kontextfreie Grammatik** und **nicht Reguläre Grammatik**, welche die **Sprache** L_{PicoC_Parse} beschreibt und damit ist die **Sprache** L_{PicoC_Parse} nach der **Chomsky Hierarchie** (Definition 2.18) **kontextfrei**, aber **nicht regulär**.

Die **Syntax**, in welcher ein **Programm** aufgeschrieben ist, wird auch als **Konkrete Syntax** (Definition 2.34) bezeichnet. In einem Zwischenschritt, dem **Parsen** wird aus diesem Programm mithilfe eines **Parsers** (Definition 2.37) ein **Ableitungsbaum** (Definition 2.36) generiert, der als Zwischenstufe hin zum einem **Abstrakten Syntaxbaum** (Definition 2.43) dient. Beim Compilerbau ist es förderlich kleinschrittig vorzugehen, deshalb erst die Generierung des **Ableitungsbaumes** und dann erst des **Abstrakten Syntaxbaumes**.

Definition 2.34: Konkrete Syntax



*Steht für alles, was mit dem **Aufbau** von nach einer **Konkreten Grammatik** (Definition 2.35) **abgeleiteten Wörtern**^a zu tun hat.*

*Die **Konkrete Syntax** ist die Teilmenge der **gesamten Syntax** einer Sprache, welche für die **Lexikalische** und **Syntaktische Analyse** relevant ist. In der **gesamten Syntax** einer Sprache^b kann es z.B. Wörter geben, welche die gesamte Syntax **nicht einhalten**, die allerdings **korrekt** nach der **Konkreten Grammatik** abgeleitet sind^c.*

*Ein **Programm** in seiner **Textrepräsentation**, wie es in einer **Textdatei** nach der Konkreten Grammatik $G_{Lex} \uplus G_{Parse}$ ^d abgeleitet steht, bevor man es kompiliert, ist in **Konkreter Syntax** aufgeschrieben.^e*

^aBzw. **Programmen**.

^bVor allem bei **Programmiersprachen**.

^cWenn ein Programm z.B. **nicht deklarierte Variablen** hat und aufgrund dessen **nicht kompiliert** werden kann, hält dieses die gesamte Syntax **nicht** ein, kann allerdings so nach der **Konkreten Grammatik** abgeleitet werden.

^d**Vereinigungsgrammatik** wie in Definition 2.17 erklärt.

^eG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Um einen kurzen Begriff für die **Grammatik** zu haben, welche die **Konkrete Syntax** einer Sprache beschreibt, wird diese im Folgenden als **Konkrete Grammatik** (Definition 2.35) bezeichnet.

Definition 2.35: Konkrete Grammatik



*Grammatik, die eine **Konkrete Syntax** einer Sprache beschreibt und die Grammatiken G_{Lex} und G_{Parse} miteinander vereinigt: $G_{Lex} \uplus G_{Parse}$ ^a.*

*In der **Konkreten Grammatik** entsprechen die **Terminalsymbole** den **Tokennamen**, der in der **Lexikalischen Analyse** generierten **Tokens**^b und **Nicht-Terminalsymbole** entsprechen bei einem*

¹⁶Also eine **beliebige Folge** von **Nicht-Terminalsymbolen** und **Terminalsymbolen**.

Ableitungsbaum den Stellen, wo ein *Teilbaum* eingehängt ist.

^aVereinigungsgrammatik wie in Definition 2.17 erklärt.

^bWobei das **Lark Parsing Toolkit**, welches später bei der **Implementierung** verwendet wird eine spezielle **Metasyntax** zur Spezifikation von Grammatiken nutzt, bei der für bestimmten häufig genutzte **Terminalsymbolen** ein **Tokenwert** in die Grammatik geschrieben wird.

Definition 2.36: Ableitungsbaum (bzw. Konkreter Syntaxbaum, engl. Derivation Tree)

Compilerinterne Datenstruktur für den Formalen Ableitungsbaum (Definition 2.26) eines in Konkreter Syntax geschriebenen Programmes.

Die **Blätter**, die beim Formalen Ableitungsbaum **Terminalsymbole** einer **Konkreten Grammatik** $G_{Lex} \uplus G_{Parse}$ ^a sind, sind in dieser Datenstruktur **Tokens**. In dieser Datenstruktur werden allerdings nur die **Ableitungen** eines Formalen Ableitungsbauemes dargestellt, die sich aus den **Produktionen** einer Grammatik G_{Parse} ergeben. Die **Tokens** sind in der Syntaktischen Analyse ein **atomarer Grundbaustein**^b, daher sind die **Ableitungen** der Grammatik G_{Lex} uninteressant.^c

^aVereinigungsgrammatik wie in Definition 2.17 erklärt.

^bNicht mehr weiter teilbar.

^cJSON parser - Tutorial — Lark documentation.

Die **Konkrete Grammatik** nach der **Ableitungsbaum** konstruiert ist, wird optimalerweise immer so definiert, dass sich möglichst **einfach** aus dem **Ableitungsbaum** ein **Abstrakter Syntaxbaum** konstruieren lässt.

Definition 2.37: Parser

Ein **Parser** ist ein Programm, dass aus einem Eingabewort^a, welches in **Konkreter Syntax** geschrieben ist eine compilerinterne Datenstruktur, den **Ableitungsbaum** generiert, was auch als **Parsen** bezeichnet wird^b.^c

^aZ.B. wiederum ein **Programm**.

^bEs gibt allerdings auch alternative Definitionen, denen nach ein Parser in Bezug auf Compilerbau ein Programm ist, dass ein Eingabewort von **Konkreter Syntax** in **Abstrakte Syntax** übersetzt. Im Folgenden wird allerdings die Definition 2.37 verwendet.

^cJSON parser - Tutorial — Lark documentation.

Anmerkung

An dieser Stelle könnte möglicherweise eine Verwirrung entstehen, welche Rolle dann überhaupt ein **Lexer** hier spielt.

In Bezug auf Compilerbau ist ein **Lexer** ein Teil eines **Parsers**. Der **Lexer** ist ausschließlich für die **Lexikalische Analyse** verantwortlich und entspricht z.B., wenn man bei einem Wanderausflug verschiedene Insekten entdeckt, dem Nachschlagen in einem Insekten**lexikon** und dem Aufschreiben, welchen Insekten man in welcher **Reihenfolge** begegnet ist. Zudem kann man bestimmte **Sehenswürdigkeiten** an denen man während des Ausflugs vorbeikommt ebenfalls festhalten, da es eine Rolle spielen kann in welchem örtlichen **Kontext** man den Insekten begegnet ist^a.

Der **Parser** vereinigt sowohl die **Lexikalische Analyse**, als auch einen Teil der **Syntaktischen Analyse** in sich und entspricht, um auf das Beispiel zurückzukommen, dem Darstellen von **Beziehungen** zwischen den Insektenbeugnungen in einer für die **Weiterverarbeitung tauglichen Form**^b.

In der Weiterverarbeitung kann der **Interpreter** das interpretieren und daraus bestimmte Schlüsse

ziehen und ein **Compiler** könnte es vielleicht in eine für Menschen leichter entschlüsselbare Sprache kompilieren.

^aDas würde z.B. der Rolle eines **Semikolon** ; in der Sprache L_{PicoC} entsprechen.

^bZ.B. gibt es bestimmte **Wechselbeziehungen** zwischen Insekten, Insekten beeinflussen sich gegenseitig und ihre Umwelt.

Die vom **Lexer** im Eingabewort identifizierten **Token** werden in der **Syntaktischen Analyse** vom **Parser** als **Wegweiser** verwendet, da je nachdem, in welcher Reihenfolge die **Token** auftauchen, dies einer anderen Ableitung in der **Grammatik** G_{Parse} entspricht. Dabei wird in der Grammatik L_{Parse} nach dem **Tokennamen** unterschieden und nicht nach dem Tokenwert, da es nur von Interesse ist, ob an einer bestimmten Stelle z.B. eine **Zahl** steht und nicht, welchen konkreten Wert diese **Zahl** hat. Der **Tokenwert** ist erst später in der **Code Generierung** in 2.5 wieder relevant.

Ein **Parser** ist genauer gesagt ein erweiterter **Erkenner** (Definition 2.38), denn ein Parser löst das **Wortproblem** (Definition 2.21) für die **Sprache**, in der das Programm, welches kompiliert werden soll geschrieben ist und konstruiert parallel dazu oder im Nachgang aus den Informationen, die während der Ausführung des Erkennungsalgorithmus¹⁷ gesichert wurden den **Ableitungsbaum**.

Definition 2.38: Erkenner (bzw. engl. Recognizer)



*Entspricht einem **Kellerautomaten**^a, in dem **Wörter** bestimmter **Kontextfreier Sprachen** erkannt werden. Der **Erkenner** ist ein **Algorithmus**, der erkennt, ob ein **Eingabewort** sich mit den Produktionen der **Konkreten Grammatik** einer Sprache ableiten lässt, also ob er bzw. es Teil der Sprache ist, die von der **Konkreten Grammatik** beschrieben wird oder nicht. Das vom **Erkenner** gelöste Problem ist auch als **Wortproblem** (Definition 2.21) bekannt.^b*

^a**Automat** mit dem **Kontextfreie Grammatiken** erkannt werden.

^bThiemann, „Compilerbau“.

Anmerkung

Für das **Parsen** gibt es grundsätzlich **drei** verschiedene Ansätze:

- **Top-Down Parsing:** Der **Ableitungsbaum** wird von **oben-nach-unten** generiert, also von der **Wurzel** zu den **Blättern**. Dementsprechend fängt die Generierung des **Ableitungsbaumes** mit dem **Startsymbol** der **Konkreten Grammatik** an und wendet in jedem Schritt eine **Linksableitung** auf die **Nicht-Terminalsymbole** an, bis man **Terminalsymbole** hat, die sich zum gewünschten **Eingabewort** abgeleitet haben oder sich herausstellt, dass dieses nicht abgeleitet werden kann.^a

Der Grund, warum die **Linksableitung** verwendet wird und nicht z.B. die **Rechtsableitung**, ist, weil das **Eingabewort** von **links nach rechts** eingelesen wird, was gut damit zusammenpasst, dass die **Linksableitung** die **Blätter** von **links-nach-rechts** generiert.

Welche der **Produktionen** für ein **Nicht-Terminalsymbol** angewandt wird, wenn es mehrere **Alternativen** gibt, wird entweder durch **Backtracking** oder durch **Vorausschauen** gelöst.

Eine sehr einfach zu implementierende Technik für **Top-Down Parser** ist hierbei der **Rekursive Abstieg** (Definition ??).

Mit dieser Methode ist das Parsen **Linksrekursiver Grammatiken** (Definition 2.25) allerdings nicht möglich, ohne die Konkrete Grammatik vorher umgeformt zu haben und jegliche **Linksrekursion** aus der Konkreten Grammatik entfernt zu haben, da diese zu **Unendlicher Rekursion**

¹⁷Bzw. engl. recognition algorithm.

führt.

Rekursiver Abstieg kann mit **Backtracking** verbunden werden, um auch Konkrete Grammatiken parsen zu können, die nicht **LL(k)** (Definition ??) sind. Dabei werden meist nach dem Prinzip der **Tiefensuche** alle **Produktionen** für ein **Nicht-Terminalsymbol** solange durchgegangen bis der gewünschte Inputstring abgeleitet ist oder alle **Alternativen** für einen Schritt abgesucht sind, bis man wieder beim ersten Schritt angekommen ist und da auch alle **Alternativen** abgesucht sind, was dann bedeutet, dass das **Eingabewort** sich **nicht** mit der verwendeten Konkreten Grammatik ableiten lässt.^b

Wenn man eine **LL(k)**-Grammatik hat, kann man auf **Backtracking verzichten** und es reicht einfach nur immer k **Token** im Eingabewort **vorauszuschauen**. **Mehrdeutige Grammatiken** sind dadurch ausgeschlossen, weil **LL(k)** keine **Mehrdeutigkeit** zulässt.^c

- **Bottom-Up Parsing:** Es wird mit dem **Eingabewort** gestartet und versucht **Rechtsableitungen** entsprechend der **Produktionen** einer **Konkreten Grammatik** rückwärts anzuwenden, bis man beim **Startsymbol** landet.^d
- **Chart Parsing:** Es wird **Dynamische Programmierung** verwendet und **partielle Zwischenergebnisse** werden in einer **Tabelle** (bzw. einem **Chart**) gespeichert und können **wiederverwendet** werden. Das macht das Parsen **Kontextfreier Grammatiken** effizienter, sodass es nur noch **polynomielle** Zeit braucht, da **Backtracking** nicht mehr notwendig ist^e. **Chart Parser** können dabei **top-down** oder **bottom-up** Ansätze umsetzen. Da die **Implementierung** von **Chart Parsern** fundamental anders ist als bei **Top-Down** und **Bottom-Up Parsern**, wird diese **Kategorie** von Parsern nochmal **speziell unterschieden** und nicht gesagt, es sei ein **Top-Down Parser** oder **Bottom-Up Parser**, der **Dynamische Programmierung** verwendet.

^aWhat is Top-Down Parsing?

^bDiese Form von Parsing wurde im **PicoC-Compiler** implementiert, als dieser noch auf dem Stand des **Bachelorprojektes** war, bevor er durch den nicht selbst implementierten **Earley Parser** von **Lark** (siehe Webseite *Lark - a parsing toolkit for Python*) ersetzt wurde.

^cDiese Art von Parser ist im **RETI-Interpreter** implementiert, da die **RETI-Sprache** eine besonders simple **LL(1) Grammatik** besitzt. Diese Art von **Parser** wird auch als **Predictive Parser** oder **LL(k) Recursive Descent Parser** bezeichnet, wobei **Recursive Descent** das englische Wort für **Rekursiven Abstieg** ist.

^dWhat is Bottom-up Parsing?

^eDer **Earley Parser**, den **Lark** und damit der **PicoC-Compiler** verwendet fällt unter diese Kategorie.

Der **Abstrakte Syntaxbaum** wird mithilfe von **Transformern** (Definition 2.39) und **Visitors** (Definition 2.40) generiert und ist das Endprodukt der **Syntaktischen Analyse**, welches an die **Code Generierung** weitergegeben wird. Wenn man die gesamte **Syntaktische Analyse** betrachtet, so übersetzt diese ein Programm von der **Konkreten Syntax** in die **Abstrakte Syntax** (Definition 2.41).

Definition 2.39: Transformer



Ein Programm, das von **unten-nach-oben**^a nach dem Prinzip der **Breitensuche** alle Knoten des **Ableitungsbaums** besucht und beim Antreffen eines bestimmten Knoten des **Ableitungsbaumes** je nach Kontext einen entsprechenden Knoten des **Abstrakten Syntaxbaumes** erzeugt und diesen anstelle des Knotens des **Ableitungsbaumes** setzt und so Stück für Stück den **Abstrakten Syntaxbaum** konstruiert.^b

^aIn der **Informatik** wachsen Bäume von **oben-nach-unten**, von der **Wurzel** zur den **Blättern**.

^bTransformers & Visitors — Lark documentation.

Definition 2.40: Visitor

Ein Programm, das von **unten-nach-oben**^a, nach dem Prinzip der Breitensuche alle Knoten des **Ableitungsbaumes** besucht und beim Antreffen eines bestimmten **Knoten** des Ableitungsbaumes, diesen **in-place** mit anderen Knoten **tauscht** oder **manipuliert**, um den Ableitungsbaum für die weitere Verarbeitung durch z.B. einen Transformer zu vereinfachen.^{b,c}

^aIn der **Informatik** wachsen Bäume von **oben-nach-unten**, von der **Wurzel** zur den **Blättern**.

^bKann theoretisch auch zur Konstruktion eines **Abstrakten Syntaxbaumes** verwendet werden, wenn z.B. eine externe Klasse verwendet wird, welches für die Konstruktion des **Abstrakten Syntaxbaumes** verantwortlich ist. Aber dafür ist ein **Transformer** besser geeignet.

^c*Transformers & Visitors — Lark documentation.*

Definition 2.41: Abstrakte Syntax

Steht für alles, was mit dem **Aufbau** von **Abstrakten Syntaxbäumen** zu tun hat.

Ein **Abstrakter Syntaxbaum**, der zur Kompilierung eines Wortes^a generiert wurde befindet sich in **Abstrakter Syntax**.^b

^aZ.B. **Programmcode**.

^bG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Um einen kurzen Begriff für die **Grammatik**, welche die **Abstrakte Syntax** einer **Sprache** beschreibt zu haben, wird diese im Folgenden als **Abstrakte Grammatik** (Definition 2.42) bezeichnet.

Definition 2.42: Abstrakte Grammatik

Grammatik, die eine **Abstrakte Syntax** beschreibt, also beschreibt was für Arten von **Kompositionen** mit den **Knoten** eines **Abstrakten Syntaxbaumes** möglich sind.

Jene Produktionen, die in der **Konkreten Grammatik** für die Umsetzung von **Präzedenz** notwendig waren, sind in der **Abstrakten Grammatik** abgeflacht. Dadurch sind die **Kompositionen**, welche die Knoten im **Abstrakten Syntaxbaum** bilden können **syntaktisch** meist näher an der Syntax von **Maschinenbefehlen**.

Definition 2.43: Abstrakter Syntaxbaum (bzw. engl. Abstract Syntax Tree, kurz AST)

Ist ein **compilerinterne Datenstruktur**, welche eine **Abstraktion** eines dazugehörigen **Ableitungsbaumes** darstellt, in dessen Aufbau auch das Erfordernis eines **leichten Zugriffs** und einer **leichten Weiterverarbeitbarkeit** eingeflossen ist. Bei der Betrachtung eines **Knoten**, der für einen Teil des Programms steht, soll man möglichst schnell die Fragen beantworten können, welche **Funktionalität** der Sprache dieser umsetzt, welche **Bestandteile** er hat und welche **Funktionalität** der Sprache diese Bestandteile umsetzen usw.

Die Knoten des **Abstrakten Syntaxbaumes** enthalten dabei verschiedene **Attribute**, welche wichtigen Informationen für den Kompilervorang und Fehlermeldungen enthalten.^a

^aG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Anmerkung

In dieser Bachelorarbeit wird häufig von der „**Abstrakten Syntax**“, der „**Abstrakten Grammatik**“ bzw. dem „**Abstrakten Syntaxbaum**“ einer „**Sprache**“ *L* gesprochen. Gemeint ist hier mit der Sprache *L* **nicht** die Sprache, welche durch die **Abstrakte Grammatik**, nach welcher der **Abstrakte**

Syntaxbaum abgeleitet ist beschrieben wird. Es ist damit **immer** die Sprache gemeint, die **kompiliert** werden soll^a und zu deren Zweck der **Abstrakte Syntaxbaum** überhaupt konstruiert wird. Für die tatsächliche Sprache, die durch die **Abstrakte Grammatik** beschrieben wird, interessiert man sich nie wirklich explizit. Diese **Konvention** wurde aus dem Buch G. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)* übernommen.

^aBzw. es ist die **Sprache**, welche durch die **Konkrete Grammatik** beschrieben wird.

Im **Abstrakten Syntaxbaum** können theoretisch auch die **Token** aus der **Lexikalischen Analyse** weiterverwendet werden, allerdings ist dies **nicht empfehlenswert**. Es ist zum empfehlen die **Token** durch eigene entsprechende Knoten umzusetzen, damit der **Zugriff** auf Knoten des Abstrakten Syntaxbaumes immer **einheitlich** erfolgen kann und auch, da manche **Token** des Abstrakten Syntaxbaum noch **nicht optimal benannt** sind. Manche „Symbole“ werden in der **Lexikalischen Analyse** mehrfach verwendet, wie z.B. das Symbol `-` in L_{PicoC} , welches für die **binäre Subtraktionsoperation** als auch die **unäre Minusoperation** verwendet wurde. Der verwendete **Tokenname** dieses Symbols lautet im **PicoC-Compiler** `SUB_MINUS`. Da in der **Syntaktischen Analyse** beide Operationen nur in **bestimmten Kontexten** vorkommen, **lassen** sie sich **unterscheiden** und dementsprechend können für beide Operationen jeweils zwei **seperate Knoten** erstellt werden. Im Fall des **PicoC-Compilers** sind es die Knoten `Sub()` und `Minus()`.

Im Gegensatz zum **Formalen Ableitungsbaum**, ergibt es beim **Abstrakten Syntaxbaum** keinen Sinn zusätzlich einen **Formalen Abstrakten Syntaxbaum** zu unterscheiden, da das Konzept eines **Abstrakten Syntaxbaumes** ohne eine Datenstruktur zu sein für sich allein gesehen keine Anwendung hat. Wenn von Abstrakten Syntaxbäumen die Rede ist, ist immer eine **Datenstruktur** gemeint.

Die **Abstrakte Grammatik** nach der ein **Abstrakter Syntaxbaum** konstruiert ist wird optimalerweise immer so definiert, dass der **Abstrakte Syntaxbaum** in den darauffolgenden Verarbeitungsschritten¹⁸ möglichst **einfach weiterverarbeitet** werden kann.

Auf der **linken** Seite in Abbildung 2.6 wird das Beispiel 2.26.2 aus Unterkapitel 2.2.1 fortgeführt. Dieses Beispiel stellt den **Arithmetischen Ausdruck** $4 * 2$ in Bezug auf die Konkrete Grammatik 2.2¹⁹, welche die **höhere Präzedenz** der **Multiplikation** `*` berücksichtigt in einem **Ableitungsbaum** dar. Allerdings handelt es sich bei diesem Ableitungsbaum **nicht** um einen **Formalen Ableitungsbaum**, sondern um eine **compilerinterne Datenstruktur** für einen solchen. Dementsprechend sind die **Blätter** nun **Tokens**, die mithilfe der Grammatik L_{Lex} generiert wurden, womit die Darstellung von **Ableitungen** sich auf die Grammatik L_{Parse} beschränkt.

Auf der **rechten** Seite in Abbildung 2.6 wird der Ableitungsbaum zu einem Abstrakten Syntaxbaum **abstrahiert**, der nach der Abstrakten Grammatik 2.3 konstruiert ist. Die **Abstrakte Grammatik** ist hierbei in **Abstrakter Syntaxform** (Definition ??) angegeben. In der Abstrakten Grammatik 2.3 sind jegliche Produktionen **wegabstrahiert**, die in der **Konkreten Grammatik 2.2** so umgesetzt sind, damit diese **Präzidenz** beachtet und nicht **mehrdeutig** ist. Aus diesem Grund gibt es nur noch einen **allgemeinen Knoten** für **binäre Operationen** $BinOp(\langle exp \rangle, \langle bin_op \rangle, \langle exp \rangle)$.

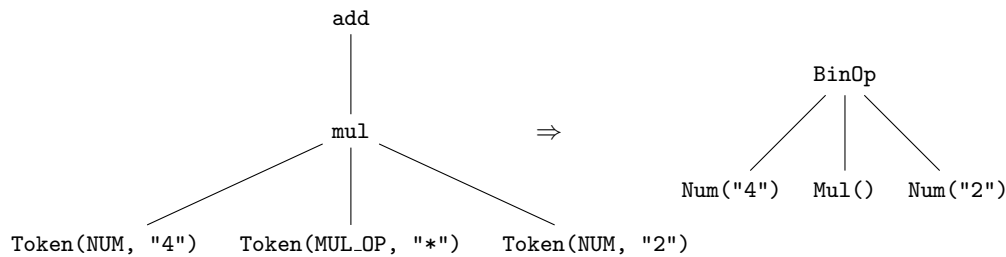
¹⁸Den verschiedenen **Passes**.

¹⁹Die **Konkrete Grammatik** ist hierbei im **Dialekt der Erweiterter Backus-Naur-Form** des Lark Parsing Toolkit (Definition ??) angegeben.

DIG_NO_0	$::=$	"1" "2" "3" "4" "5" "6"	L_Lex
		"7" "8" "9"	
DIG_WITH_0	$::=$	"0" DIG_NO_0	
NUM	$::=$	"0" $DIG_NO_0 DIG_WITH_0^*$	
ADD_OP	$::=$	"+"	
MUL_OP	$::=$	"*"	
mul	$::=$	$mul MUL_OP NUM$ NUM	L_Parse
add	$::=$	$add ADD_OP mul$ mul	

Grammatik 2.2: Produktionen für Ableitungsbaum in EBNF

bin_op	$::=$	$Add()$ $Mul()$
exp	$::=$	$BinOp(\langle exp \rangle, \langle bin_op \rangle, \langle exp \rangle)$ $Num(\langle str \rangle)$

Grammatik 2.3: Produktionen für Abstrakten Syntaxbaum in ASF**Abbildung 2.6:** Veranschaulichung des Unterschieds zwischen Ableitungsbaum und Abstraktem Syntaxbaum.

Die **Baumdatenstruktur** des **Ableitungsbaumes** und **Abstrakten Syntaxbaumes** ermöglicht es die Operationen, die ein Compiler bzw. Interpreter bei der Weiterverarbeitung des Programmes ausführen muss möglichst **effizient** auszuführen und auf **unkomplizierte** Weise direkt zu erkennen, welche er ausführen muss.

Um eine Gesamtübersicht über die **Syntaktische Analyse** zu geben, sind in Abbildung 2.7 die einzelnen **Zwischenschritte** von den Tokens der **Lexikalischen Analyse** zum **Abstrakten Syntaxbaum** anhand des fortgeführten Beispiels aus Subkapitel 2.3 veranschaulicht. In Abbildung 2.7 werden die Darstellungen des **Ableitungsbaumes** und des **Abstrakten Syntaxbaumes** verwendet, wie sie vom **PicoC-Compiler** ausgegeben werden. In der Darstellung des **PicoC-Compilers** stellen die verschiedenen **Eindrückungen** die verschiedenen **Ebenen** dieser Bäume dar. Die Bäume **wachsen** von der **Wurzel** von **links-nach-rechts** zu den **Blättern**.

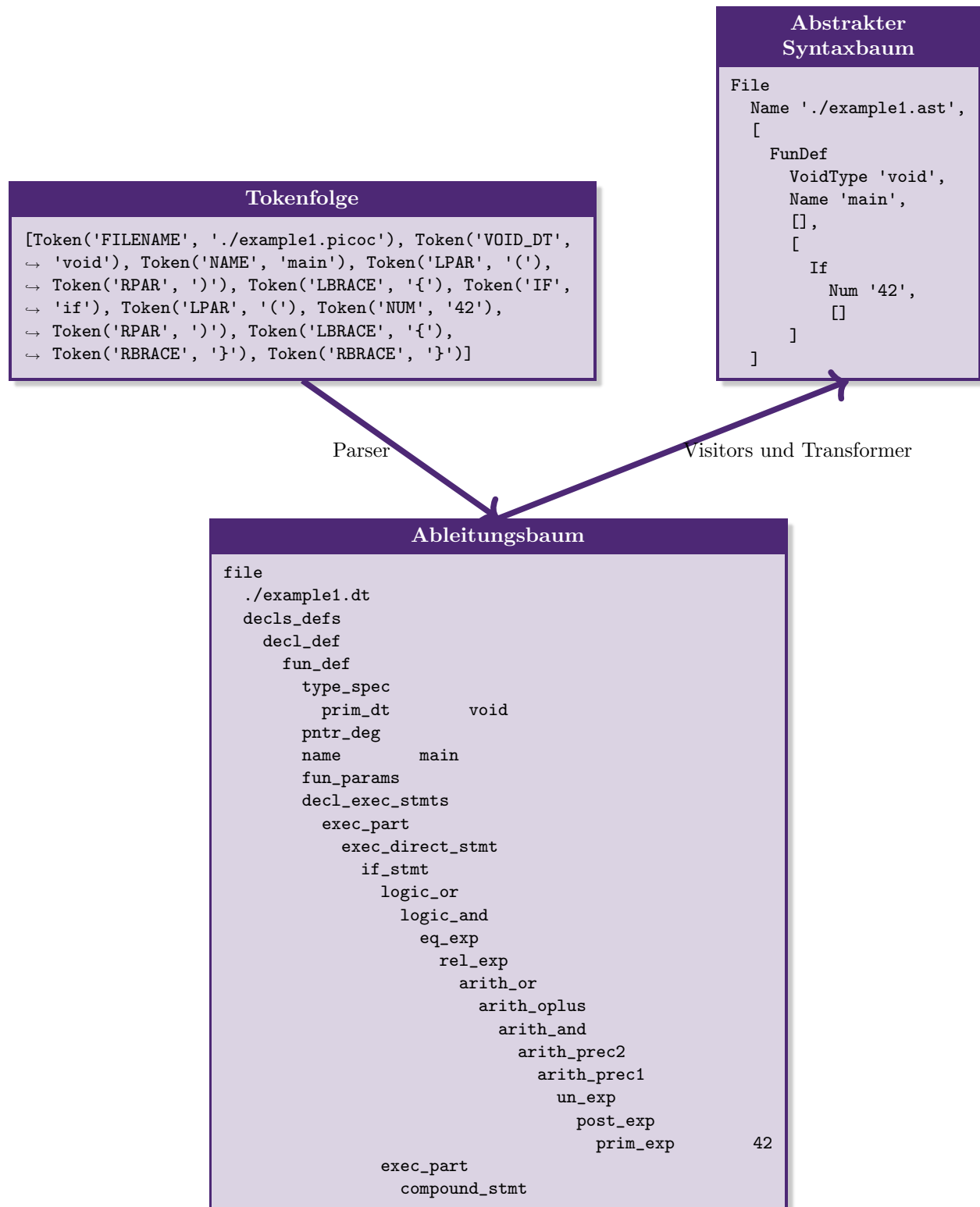


Abbildung 2.7: Veranschaulichung der Syntaktischen Analyse.

2.5 Code Generierung

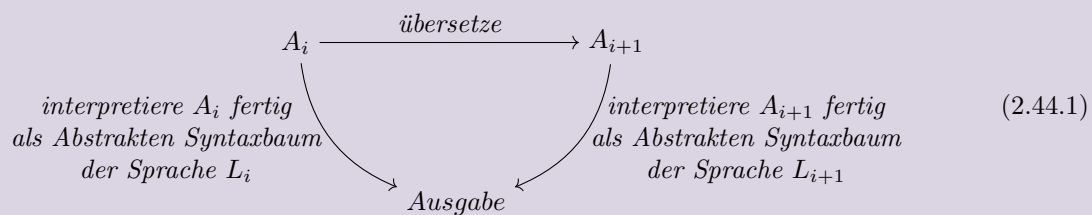
In der **Code Generierung** steht man nun dem Problem gegenüber einen **Abstrakten Syntaxbaum** einer Sprache L_1 in den **Abstrakten Syntaxbaum** einer Sprache L_2 umformen zu müssen. Dieses Problem lässt sich vereinfachen, indem man das Problem in mehrere Schritte unterteilt, die man **Passes** (Definition 2.44) nennt. So wie es auch schon mit dem **Ableitungsbaum** in der Syntaktischen Analyse gemacht wurde, den man als Zwischenstufe zum **Abstrakten Syntaxbaum** konstruiert hatte. Aus dem Ableitungsbaum konnte dann unkompliziert und einfach mit **Transformern** und **Visitors** ein **Abstrakter Syntaxbaum** generiert werden.

Definition 2.44: Pass



*Einzelner Übersetzungsschritt in einem Kompilervorgang von einem beliebigen **Abstrakten Syntaxbaum** A_i einer Sprache L_i zu einem **Abstrakten Syntaxbaum** A_{i+1} einer Sprache L_{i+1} , der meist eine bestimmte **Teilaufgabe** übernimmt, die sich mit keiner **Teilaufgabe** eines anderen **Passes** überschneidet und möglichst wenig **Ähnlichkeit** mit den **Teilaufgaben** anderer **Passes** haben sollte.^{ab}*

*Für jeden **Pass** und für einen beliebigen **Abstrakten Syntaxbaum** A_i gilt ähnlich, wie bei einem **vollständigen Compiler** in 2.44.1, dass:*



*wobei man hier so tut, als gäbe es zwei **Interpreter** für die zwei Sprachen L_i und L_{i+1} , welche den jeweiligen **Abstrakten Syntaxbaum** A_i bzw. A_{i+1} fertig interpretieren.^{cd}*

^aEin **Pass** kann mit einem **Transpiler** ?? (Definition ??) verglichen werden, da sich die zwei Sprachen L_i und L_{i+1} aufgrund der **Kleinschrittigkeit** meist auf einem ähnlichen **Abstraktionslevel** befinden. Der Unterschied ist allerdings, dass ein **Transpiler** zwei Programme, die in L_i bzw. L_{i+1} geschrieben sind kompiliert. Ein **Pass** ist dagegen immer **kleinschrittig** und operiert ausschließlich auf **Abstrakten Syntaxbäumen**, ohne Parsing usw.

^bDer Begriff kommt aus dem **Englischen** von „passing over“, da der gesamte **Abstrakte Syntaxbaum** in einem **Pass** durchlaufen wird.

^c**Interpretieren** geht immer von einem Programm in **Konkreter Syntax** aus, wobei der **Abstrakte Syntaxbaum** ein **Zwischenschritt** bei der **Interpretierung** ist.

^dG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Die von den **Passes** umgeformten **Abstrakten Syntaxbäume** sollten dabei mit jedem **Pass** der **Syntax** von **Maschinenbefehlen** immer ähnlicher werden, bis es schließlich nur noch Maschinenbefehle sind.

2.5.1 Monadische Normalform

Hat man es mit einer Sprache zu tun, welche **Unreine Ausdrücke** (Definition 2.46) besitzt, so ist es sinnvoll einen **Pass** einzuführen, der **Reine** (Definition 2.45) und **Unreine Ausdrücke** voneinander **trennt**. Das wird erreicht, indem man aus den Unreinen Ausdrücken **vorangestellte Anweisungen** macht, die man **vor** den jeweiligen reinen Ausdruck, mit dem sie **gemischt** waren stellt. Der Unreine Ausdruck muss als **erstes** ausgeführt werden, für den Fall, dass der **Effekt**, denn ein **Unreiner Ausdruck** hatte den **Reinen Ausdruck**, mit dem er gemischt war in irgendeinerweise beeinflussen könnte.

Definition 2.45: Reiner Ausdruck (bzw. engl. pure expression)

Ein **Reiner Ausdruck** ist ein Ausdruck, der **rein** ist. Das bedeutet, dass dieser Ausdruck **keine Nebeneffekte** erzeugt. Ein **Nebeneffekt** ist eine **Bedeutung**, die ein Ausdruck hat, die sich **nicht** mit **RETI-Code** darstellen lässt.^{a,b}

^aSondern z.B. **intern** etwas am **Kompilierprozess** ändert.

^bG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Definition 2.46: Unreiner Ausdruck

Ein **Unreiner Ausdruck** ist ein Ausdruck, der kein **Reiner Ausdruck** ist.

Auf diese Weise sind alle **Anweisungen** und **Ausdrücke** in **Monadischer Normalform** (Definition 2.47).

Definition 2.47: Monadische Normalform (bzw. engl. monadic normal form)

Eine **Anweisung** oder **Ausdruck** ist in **Monadischer Normalform**, wenn es oder er nach einer **Konkreten Grammatik** in **Monadischer Normalform** abgeleitet wurde.

Eine **Konkrete Grammatik** ist in **Monadischer Normalform**, wenn sie **reine Ausdrücke** und **unreine Ausdrücke nicht miteinander mischt**, sondern voneinander **trennt**.^a

Eine **Abstrakte Grammatik** ist in **Monadischer Normalform**, wenn die **Konkrete Grammatik** für welche sie definiert wurde in **Monadischer Normalform** ist.

^aG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Ein **Beispiel** für dieses Vorgehen ist in Abbildung 2.8 zu sehen, wo der Einfachheit halber auf die Darstellung in **Abstrakter Syntax** verzichtet wurde und die Codebeispiele in der entsprechenden **Konkreten Syntax**²⁰ aufgeschrieben wurden.

In der Abbildung 2.8 ist der Ausdruck mit dem **Nebeneffekt** eine Variable zu **allokieren**: `int var`, mit dem Ausdruck für eine **Zuweisung** `exp = 5 % 4` gemischt, daher muss der **Unreine** Ausdruck als eigenständige Anweisung **vorangestellt** werden.

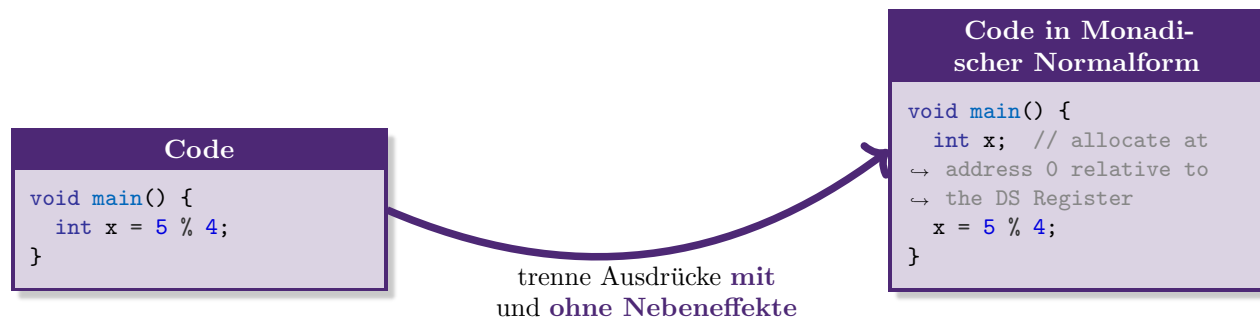


Abbildung 2.8: Codebeispiel für das Trennen von Ausdrücken mit und ohne Nebeneffekten.

Die Aufgabe eines solchen **Passes** ist es, den **Abstrakten Syntaxbaum** der **Syntax** von **Maschinenbefehlen** anzunähern, indem Subbäume vorangestellt werden, die keine Entsprechung in **RETI-Knoten** haben. Somit wird eine **Separation** von Subbäumen, die keine Entsprechung in **RETI-Knoten** haben und denen, die

²⁰Für deren Kompilierung die **Abstrakte Syntax** überhaupt definiert wurde.

eine haben bewerkstelligt wird. Ein **Reiner Ausdruck** ist **Maschinenbefehlen** ähnlicher als ein Ausdruck, indem ein **Reiner** und **Unreiner Ausdruck** gemischt sind. Somit sparrt man sich in der Implementierung **Fallunterscheidungen**, indem die **Reinen Ausdrücke** direkt in **RETI-Code** übersetzt werden können und **nicht** unterschieden werden muss, ob darin **Unreine Ausdrücke** vorkommen.

2.5.2 A-Normalform

Im Falle dessen, dass es sich bei der **Sprache** L_1 um eine **höhere Programmiersprache** und bei L_2 um **Maschinensprache** handelt, ist es fast unerlässlich einen **Pass** einzuführen, der **Komplexe Ausdrücke** (Definition 2.50) aus **Anweisungen** und **Ausdrücken** entfernt. Das wird erreicht, indem man aus den Komplexen Ausdrücken **vorangestellte** Anweisungen macht, in denen die **Komplexen Ausdrücke temporären Locations** zugewiesen werden (Definition 2.48) und dann anstelle des **Komplexen Ausdrucks** auf die jeweilige **temporäre Location** zugegriffen wird.

Sollte in der **Anweisung**, in welcher der **Komplexe Ausdruck** einer **temporären Location** zugewiesen wird, der Komplexe Ausdruck **Teilausdrücke** enthalten, die **komplex** sind, muss die gleiche Prozedur erneut für die **Teilausdrücke** angewandt werden, bis **Komplexe Ausdrücke** nur noch in Anweisungen zur Zuweisung an Locations auftauchen, aber die Komplexen Ausdrücke nur **Atomare Ausdrücke** (Definition 2.49) enthalten.

Sollte es sich bei dem **Komplexen Ausdruck** um einen **Unreinen Ausdruck** handeln, welcher nur einen **Nebeneffekt** ausführt und sich nicht in **RETI-Befehle** übersetzt, so wird aus diesem eine **vorangestellte Anweisung** gemacht, welches einfach nur den **Nebeneffekt** dieses **Unreinen Ausdrucks** ausführt.

Definition 2.48: Location

*Kollektiver Begriff für **Variablen**, **Attribute** bzw. **Elemente** von Variablen bestimmter Datentypen, **Speicherbereiche auf dem Stack**, die **temporäre Zwischenergebnisse** speichern und **Register**.*

*Im Grunde genommen alles, was mit einem **Programm zu tun** hat und irgendwo **gespeichert** ist oder als **Speicherort** dient.^a*

^aG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Auf diese Weise sind alle **Anweisungen** und **Ausdrücke** in **A-Normalform** (Definition 2.51). Wenn eine **Konkrete Grammatik** in **A-Normalform** ist, ist diese auch automatisch in **Monadischer Normalform** (Definition 2.51), genauso, wie ein **Atomarer Ausdruck** auch ein **Reiner Ausdruck** ist (nach Definition 2.49).

Definition 2.49: Atomarer Ausdruck

*Ein **Atomarer Ausdruck** ist ein Ausdruck, der ein **Reiner Ausdruck** ist und der in eine **Folge von RETI-Befehlen** übersetzt werden kann, die **atomar** ist, also **nicht** mehr weiter in kleinere Folgen von RETI-Befehlen **zerkleinert** werden kann, welche die **Übersetzung** eines anderen Ausdrucks sind.*

*Also z.B. im Fall der Sprache L_{PicoC} entweder eine **Variable** `var`, eine **Zahl** `12`, ein **ASCII-Zeichen** `'c'` oder ein **Zugriff auf eine Location**, wie z.B. `stack(1)`.^a*

^aG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Definition 2.50: Komplexer Ausdruck

*Ein **Komplexer Ausdruck** ist ein **Ausdruck**, der **nicht atomar** ist, wie z.B. `5 % 4`, `-1`, `fun(12)` oder `int var`.^{ab}*

^a`int var` ist eine **Allokation**.

^bG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Definition 2.51: A-Normalform (ANF)

Eine **Anweisung** oder ein **Ausdruck** ist in **A-Normalform**, wenn es oder er nach einer **Konkreten Grammatik** in **A-Normalform** abgeleitet wurde.

Eine **Konkrete Grammatik** ist in **A-Normalform**, wenn sie in **Monadischer Normalform** ist und wenn alle **Komplexen Ausdrücke** nur **Atomare Ausdrücke** enthalten und einer **Location** zugewiesen sind.

Eine **Abstrakte Grammatik** ist in **A-Normalform**, wenn die **Konkrete Grammatik** für welche sie definiert wurde in **A-Normalform** ist.^{a b c}

^a*A-Normalization: Why and How (with code)*.

^bBolingbroke und Peyton Jones, „Types are calling conventions“.

^cG. Siek, *Course Webpage for Compilers (P423, P523, E313, and E513)*.

Ein **Beispiel** für dieses Vorgehen ist in Abbildung 2.9 zu sehen, wo der Einfachheit halber auf die Darstellung in **Abstrakter Syntax** verzichtet wurde und die Codebeispiele in der entsprechenden **Konkreten Syntax**²¹ aufgeschrieben wurden.

Der **PicoC-Compiler** nutzt, anders als es geläufig ist keine **Register** und **Graph Coloring** (Definition ??) inklusive **Liveness Analysis** (Definition ??) usw., um Werte von Variablen, temporäre Zwischenergebnisse usw. abzuspeichern, sondern immer nur den **Hauptspeicher**, wobei **temporäre Zwischenergebnisse** auf den **Stack** gespeichert werden.²²

Aus diesem Grund verwendet das Beispiel in Abbildung 2.9 eine andere Definition für **Komplexe** und **Atomare Ausdrücke**, da dieses Beispiel, um später keine Verwirrung zu erzeugen der Art nachempfunden ist, wie im **PicoC-ANF Pass** der **Abstrakte Syntaxbaum** umgeformt wird. Weil beim PicoC-Compiler **temporäre Zwischenergebnisse** auf den **Stack** gespeichert werden, wird nur noch ein **Zugriffen auf den Stack**, wie z.B. `stack('1')` als **Atomarer Ausdruck** angesehen. Dementsprechend werden **Ausdrücke** für **Zahl 4**, **Variable** `var` und **ASCII-Zeichen** `'c'` nun ebenfalls zu den **Komplexen Ausdrücken** gezählt.

Im Fall, dass **Register** für z.B. **temporäre Zwischenergebnisse** genutzt werden und der **Maschinenbefehlssatz** es erlaubt **zwei Register** miteinander zu verrechnen²³, ist es möglich **Ausdrücke** für **Zahl 4**, **Variable** `var` und **ASCII-Zeichen** `'c'` als **atomar** zu definieren, da sie mit einem **Maschinenbefehl** verarbeitet werden können²⁴. Werden allerdings keine **Register** für **Zwischenergebnisse** genutzt werden, braucht man **mehrere Maschinenbefehle**, um die Zwischenergebnisse vom **Stack** zu holen, zu **verrechnen** und das Ergebnis wiederum auf den **Stack** zu **speichern** und das **SP-Register** **anzupassen**. Daher werden die **Ausdrücke** für **Zahl 4**, **Variable** `var` und **ASCII-Zeichen** `'c'` als **Komplexe Ausdrücke** gewertet, da sie niemals in einem **Maschinenbefehl** miteinander verrechnet werden können.

Die Anweisungen `4, x`, usw. für sich sind in diesem Fall **Anweisungen**, bei denen ein **Komplexer Ausdruck** einer **Location**, in diesem Fall einer **Speicherzelle des Stack** zugewiesen wird, da `4, x` usw. in diesem Fall auch als **Komplexe Ausdrücke** zählen. Auf das Ergebnis dieser **Komplexen Ausdrücke** wird mittels `stack(2)` und `stack(1)` zugegriffen, um diese im **Komplexen Ausdruck** `stack(2) % stack(1)` miteinander

²¹Für deren Kompilierung die **Abstrakte Syntax** überhaupt definiert wurde.

²²Die in diesem **Paragraph** erwähnten **Begriffe** werden nur grob erläutert, da sie für den **PicoC-Compiler** keine Rolle spielen. Aber sie wurden erwähnt, damit in dieser **Bachelorarbeit** auch das übliche Vorgehen Erwähnung findet und vom Vorgehen beim **PicoC-Compiler** abgegrenzt werden kann.

²³Z.B. **Addieren** oder **Subtraktion** von zwei **Registerinhalten**.

²⁴Mit dem **RETI-Befehlssatz** wäre das durchaus möglich, durch z.B. `MULT ACC IN2`.

zu verrechnen und wiederum einer Speicherzelle des Stack zuzuweisen.

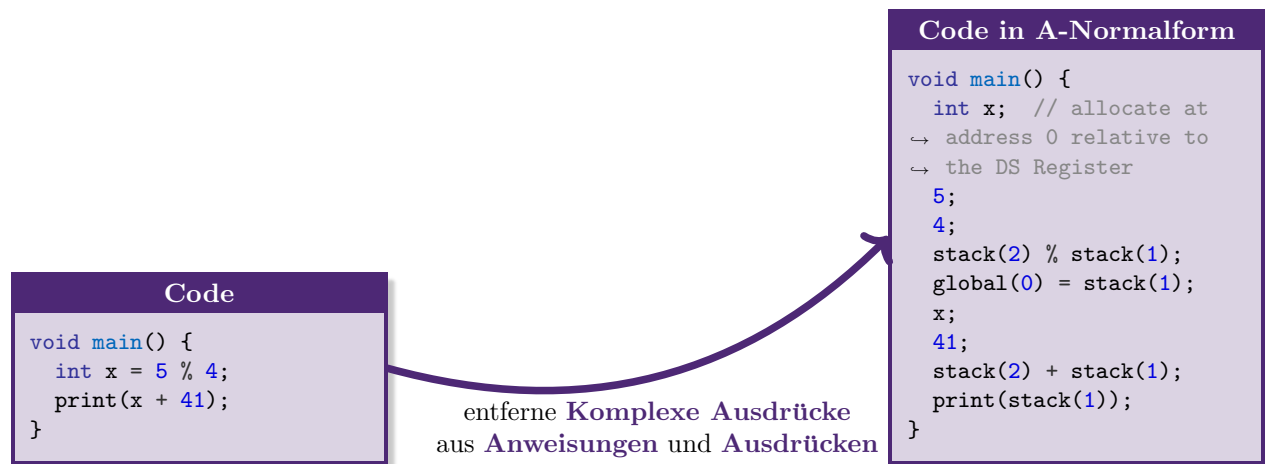


Abbildung 2.9: Codebeispiel für das Entfernen Komplexer Ausdrücke aus Operationen.

Ein solcher **Pass** hat vor allem in erster Linie die Aufgabe den **Abstrakten Syntaxbaum** der **Syntax** von **Maschinenbefehlen** besonders dadurch anzunähern, dass er die Anweisungen **weniger komplex** macht und diese dadurch den ziemlich **einfachen Maschinenbefehlen** syntaktisch ähnlicher sind. Des Weiteren **vereinfacht** dieser Pass die **Implementierung** der nachfolgenden Passes enorm, da Anweisungen z.B. nur noch die Form `global(rel_addr) = stack(1)` haben, die viel **einfacher verarbeitet** werden kann.

Alle weiteren denkbaren **Passes** sind zu **spezifisch** auf bestimmte **Anweisungen** und **Ausdrücke** ausgelegt, als das sich zu diesen allgemein etwas mit einer **Theorie** dahinter sagen lässt. Alle **Passes**, die zur Implementierung des **PicoC-Compilers** geplant und ausgedacht wurden sind im Unterkapitel ?? definiert.

2.5.3 Ausgabe des Maschinencodes

Nachdem alle **Passes** durchgearbeitet wurden ist es notwendig aus dem finalen **Abstrakten Syntaxbaum** den eigentlichen **Maschinencode** in **Konkreter Syntax** zu generieren. In üblichen Compilern wird hier für den **Maschinencode** eine **binäre Repräsentation** gewählt. Da der **PicoC-Compiler** vor allem zu **Lernzwecken** konzipiert ist, wird bei diesem der **Maschinencode** allerdings in einer **menschenlesbaren Repräsentation** ausgegeben. Der Weg von der **Abstrakten Syntax** zur **Konkreten Syntax** ist allerdings wesentlich einfacher, als der Weg von der **Konkreten Syntax** zur **Abstrakten Syntax**, für die eine gesamte **Syntaktische Analyse**, die eine **Lexikalische Analyse** beinhaltet durchlaufen werden musste.

Jeder **Knoten** des **Abstrakten Syntaxbaumes** erhält dazu eine Methode, welche hier `to_string` genannt wird, die eine **Textrepräsentation** seiner selbst und all seiner Knoten mit an den richtigen Stellen passend gesetzten **Semikolons** ; usw. ausgibt. Dabei wird nach dem Prinzip der **Tiefensuche** der gesamte **Abstrakte Syntaxbaum** durchlaufen und die Methode `to_string` zur Ausgabe der **Textrepräsentation** der verschiedenen Knoten aufgerufen, die immer wiederum die Methode `to_string` ihrer Kinder aufrufen und die zurückgegebene Textrepräsentation passend **zusammenfügen** und selbst **zurückgeben**.

2.6 Fehlermeldungen

Wenn bei einem Compiler ein **unerwünschtes Verhalten** der folgenden **Kategorien**²⁵ eintritt:

²⁵Errors in C/C++ - GeeksforGeeks.

1. in der **Lexikalischen** oder **Syntaktischen Analyse** tritt eine Fall ein, der **nicht** in der **Syntax** der Sprache des Compilers abgedeckt ist, z.B.:
 - der **Lexer** kann für eine Zeichenfolge **kein** passendes **Pattern** eines Tokens finden, welches diese erkennt. Der **Lexer** ist genaugenommen ein **Teil des Parsers** und ist damit bereits durch den nachfolgenden Punkt „Parser“ abgedeckt. Um die unterschiedlichen Ebenen, **Lexikalische** und **Syntaktische Analyse** gesondert zu betrachten wurde der Lexer an dieser Stelle ebenfalls kurz eingebracht.
 - der **Parser**²⁶ entscheidet das **Wortproblem** für ein **Eingabeprogramm**²⁷ mit 0, also das **Eingabeprogramm** lässt sich **nicht** durch die Konkrete Grammatik des Compilers **ableiten**.
2. in den **Passes** tritt eine Fall ein, der **nicht** in der **Syntax** der Sprache des Compilers abgedeckt ist, z.B.:
 - eine **Variable** wird **verwendet**, obwohl sie noch **nicht deklariert** ist.
 - bei einem **Funktionsaufruf** werden **mehr** Argumente oder Argumente des **falschen Datentyps** übergeben, als in der **Funktionsdeklaration** oder **Funktionsdefinition** angegeben ist.
3. Während der **Laufzeit** des Compilers tritt ein Ereignis ein, das **nicht** durch die **Semantik** der Sprache des Compilers abgedeckt ist oder das **Betriebssystem** nicht erlaubt, z.B.:
 - eine **nicht erlaubte Operation**, wie **Division durch 0** (z.B. 42 / 0) soll ausgeführt werden.
 - **Segmentation Fault**: Wenn auf **Speicher zugegriffen** wird, der vom **Betriebssystem geschützt** ist.

oder während des des **Linkens** (Definition ??) etwas nicht zusammenpasst, wie z.B.:

- es gibt **keine** oder **mehr als eine** **main-Funktion**.
- eine Funktion, die in einer **Objektdatei** (Definition ??) benötigt wird, wird von **keiner** anderen oder **mehr als einer** Objektdatei bereitgestellt.

wird eine **Fehlermeldung** (Definition 2.52) ausgegeben.

Definition 2.52: Fehlermeldung



*Benachrichtigung beliebiger Form, die einen **Grund** angibt weshalb ein Programm **nicht weiter ausgeführt** werden kann^a. Das **Ausgeben** einer Fehlermeldung kann dabei auf **verschiedene Weisen** erfolgen, wie z.B.*

- über **stdout** oder **stderr** im einem **Terminal Emulator** oder **richtigen Terminal**^b.
- über eine **Dialogbox** in einer **Graphischen Benutzerfläche**^c oder **Zeichenorientierten Benutzerschnittstelle**^d.
- in ein **Register** oder an eine **spezielle Adresse** des **Hauptspeichers** wird ein **Wert geschrieben**.
- **Logdatei**^e auf einem **Speichermedium**.

^aDieses Programm kann z.B. ein **Compiler** sein oder ein **Programm**, dass dieser **Compiler selbst kompiliert** hat.

²⁶Bzw. der **Erkenner** innerhalb des Parsers.

²⁷Bzw. **Wort**.

^bNur unter **Linux**, **Windows** hat sowas nicht.

^cIn engl. **G**raphical **U**ser **I**nterface, kurz **GUI**.

^dIn engl. **T**ext-based **U**ser **I**nterface, kurz **TUI**.

^eIn engl. **log file**.

Literatur

Online

- *A-Normalization: Why and How (with code)*. URL: <https://matt.might.net/articles/a-normalization/> (besucht am 23.07.2022).
- *clang: C++ Compiler*. URL: <http://clang.org/> (besucht am 29.07.2022).
- *Clockwise/Spiral Rule*. URL: <https://c-faq.com/decl/spiral.anderson.html> (besucht am 29.07.2022).
- Developers, Inkscape Website. *Draw Freely — Inkscape*. URL: <https://inkscape.org/> (besucht am 03.08.2022).
- *Errors in C/C++ - GeeksforGeeks*. URL: <https://www.geeksforgeeks.org/errors-in-cc/> (besucht am 10.05.2022).
- *GCC, the GNU Compiler Collection - GNU Project*. URL: <https://gcc.gnu.org/> (besucht am 13.07.2022).
- *JSON parser - Tutorial — Lark documentation*. URL: https://lark-parser.readthedocs.io/en/latest/json_tutorial.html (besucht am 09.07.2022).
- Ljohhuh. *What is an immediate value?* 4. Apr. 2018. URL: <https://reverseengineering.stackexchange.com/q/17671> (besucht am 13.04.2022).
- *Parsing Expressions · Crafting Interpreters*. URL: <https://www.craftinginterpreters.com/parsing-expressions.html> (besucht am 09.07.2022).
- *Transformers & Visitors — Lark documentation*. URL: <https://lark-parser.readthedocs.io/en/latest/visitors.html> (besucht am 09.07.2022).
- *Variablen in C und C++, Deklaration und Definition — Coder-Welten.de*. URL: <https://www.coderwelten.de/einstieg/variablen-in-c-3.html> (besucht am 11.08.2022).
- *Welcome to Lark's documentation! — Lark documentation*. URL: <https://lark-parser.readthedocs.io/en/latest/> (besucht am 31.07.2022).
- *What is Bottom-up Parsing?* URL: <https://www.tutorialspoint.com/what-is-bottom-up-parsing> (besucht am 22.06.2022).
- *What is the difference between function prototype and function signature?* SoloLearn. URL: <https://www.sololearn.com/Discuss/171026/what-is-the-difference-between-function-prototype-and-function-signature/> (besucht am 18.07.2022).
- *What is Top-Down Parsing?* URL: <https://www.tutorialspoint.com/what-is-top-down-parsing> (besucht am 22.06.2022).

Bücher

- G. Siek, Jeremy. *Course Webpage for Compilers (P423, P523, E313, and E513)*. 28. Jan. 2022. URL: <https://iucompilercourse.github.io/IU-Fall-2021/> (besucht am 28.01.2022).
- LeFever, Lee. *The Art of Explanation: Making your Ideas, Products, and Services Easier to Understand*. 1. Aufl. Wiley, 20. Nov. 2012.

Artikel

- Earley, J. und Howard E. Sturgis. „A formalism for translator interactions“. In: *CACM* (1970). DOI: [10.1145/355598.362740](https://doi.org/10.1145/355598.362740).

Vorlesungen

- Bast, Hannah. „Programmieren in C“. Vorlesung. Vorlesung. Universität Freiburg, 2020. URL: <https://ad-wiki.informatik.uni-freiburg.de/teaching/ProgrammierenCplusplusSS2020> (besucht am 09.07.2022).
- Nebel, Bernhard. „Theoretische Informatik“. Vorlesung. Vorlesung. Universität Freiburg, 2020. URL: http://gki.informatik.uni-freiburg.de/teaching/ss20/info3/index_de.html (besucht am 09.07.2022).
- Scholl, Christoph. „Betriebssysteme“. Vorlesung. Vorlesung. Universität Freiburg, 2020. URL: https://abs.informatik.uni-freiburg.de/src/teach_main.php?id=157 (besucht am 09.07.2022).
- Scholl, Philipp. „Einführung in Embedded Systems“. Vorlesung. Vorlesung. Universität Freiburg, 2021. URL: <https://earth.informatik.uni-freiburg.de/uploads/es-2122/> (besucht am 09.07.2022).
- Scholl, Prof. Dr. Christoph. „Technische Informatik“. Vorlesung. Vorlesung. Universität Freiburg, 3. Aug. 2022. (Besucht am 03.08.2022).
- Thiemann, Peter. „Compilerbau“. Vorlesung. Vorlesung. Universität Freiburg, 2021. URL: <http://proglang.informatik.uni-freiburg.de/teaching/compilerbau/2021ws/> (besucht am 09.07.2022).
- — „Einführung in die Programmierung“. Vorlesung. Vorlesung. Universität Freiburg, 2018. URL: <http://proglang.informatik.uni-freiburg.de/teaching/info1/2018/> (besucht am 09.07.2022).
- Westphal, Dr. Bernd. „Softwaretechnik“. Vorlesung. Vorlesung. Universität Freiburg, 2021. URL: <https://swt.informatik.uni-freiburg.de/teaching/SS2021/swtv1> (besucht am 19.07.2022).

Sonstige Quellen

- Bolingbroke, Maximilian C. und Simon L. Peyton Jones. „Types are calling conventions“. In: *Proceedings of the 2nd ACM SIGPLAN symposium on Haskell - Haskell '09*. the 2nd ACM SIGPLAN symposium. Edinburgh, Scotland: ACM Press, 2009, S. 1. ISBN: 978-1-60558-508-6. DOI: [10.1145/1596638.1596640](https://doi.org/10.1145/1596638.1596640). URL: <http://portal.acm.org/citation.cfm?doid=1596638.1596640> (besucht am 23.07.2022).

- *Lark - a parsing toolkit for Python*. 26. Apr. 2022. URL: <https://github.com/lark-parser/lark> (besucht am 28.04.2022).
- Nemec, Devin. *copy_file_to_another_repo_action*. original-date: 2020-08-24T19:25:58Z. 27. Juli 2022. URL: https://github.com/dmnemec/copy_file_to_another_repo_action (besucht am 03.08.2022).
- Ueda, Takahiro. *Makefile for LaTeX*. original-date: 2018-07-06T15:01:24Z. 10. Mai 2022. URL: <https://github.com/tueda/makefile4latex> (besucht am 03.08.2022).