
ALBERT LUDWIGS UNIVERSITÄT FREIBURG

TECHNISCHE FAKULTÄT

PicoC-Compiler

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

BACHELORARBEIT

Abgabedatum: 28th April 2022

Author:
Jürgen Mattheis

Gutachter:
Prof. Dr. Scholl

Betreuung:
M.Sc. Seufert

Eine Bachelorarbeit am Lehrstuhl für
Betriebssysteme

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.

Inhaltsverzeichnis

1	Motivation	7
1.1	PicoC und RETI	7
1.2	Aufgabenstellung	7
1.3	Eigenheiten der Sprache C	7
1.4	Richtlinien	7
2	Einführung	8
2.1	Compiler und Interpreter	8
2.1.1	T-Diagramme	12
2.2	Grammatiken	12
2.3	Grundlagen	12
2.3.1	Mehrdeutige Grammatiken	13
2.3.2	Präzidenz und Assoziativität	13
2.4	Lexikalische Analyse	13
2.5	Syntaktische Analyse	15
2.6	Code Generierung	21
2.7	Fehlermeldungen	21
3	Implementierung	22
3.1	Architektur	22
3.2	Lexikalische Analyse	22
3.2.1	Verwendung von Lark	22
3.2.2	Basic Parser	22
3.3	Syntaktische Analyse	22
3.3.1	Verwendung von Lark	22
3.3.2	Umsetzung von Präzidenz	22
3.3.3	Derivation Tree Generierung	23
3.3.4	Early Parser	23
3.3.5	Derivation Tree Vereinfachung	23
3.3.6	Abstrakt Syntax Tree Generierung	23
3.4	Code Generierung	23
3.4.1	Passes	23
3.4.2	Umsetzung von Pointern und Arrays	24
3.4.3	Umsetzung von Structs	24
3.4.4	Umsetzung von Funktionen	24
3.4.5	Umsetzung kleinerer Details	24
3.5	Fehlermeldungen	24
3.5.1	Error Handler	24
4	Ergebnisse und Ausblick	25
4.1	Funktionsumfang	25
4.2	Qualitätssicherung	25
4.3	Kommentierter Kompilervorgang	25
4.4	Erweiterungsideen	25
A	Appendix	26
A.1	Konkrete und Abstrakte Syntax	26

A.2	Bedienungsanleitungen	26
A.2.1	PicoC-Compiler	26
A.2.2	Showmode	26
A.2.3	Entwicklertools	26

Abbildungsverzeichnis

2.1 Veranschaulichung der Lexikalischen Analyse	15
2.2 Veranschaulichung der Syntaktischen Analyse	20

Tabellenverzeichnis

3.1 Präzidenzregeln von PicoC	23
---	----

Definitionen

2.1	Interpreter	8
2.2	Compiler	8
2.3	Maschinensprache	9
2.4	Assemblersprache (bzw. engl. Assembly Language)	9
2.5	Assembler	9
2.6	Objectcode	9
2.7	Linker	10
2.8	Immediate	10
2.9	Transpiler (bzw. Source-to-source Compiler)	10
2.10	Cross-Compiler	10
2.11	Bootstrapping	11
2.12	T-Diagram	12
2.13	Sprache	12
2.14	Chromsky Hierarchie	12
2.15	Grammatik	12
2.16	Reguläre Sprachen	12
2.17	Ableitung	12
2.18	Links- und Rechtsableitung	12
2.19	Linksrekursive Grammatiken	12
2.20	Ableitungsbaum	13
2.21	Mehrdeutige Grammatik	13
2.22	Assoziativität	13
2.23	Präzidenz	13
2.24	Wortproblem	13
2.25	LL(k)-Grammatik	13
2.26	Kontextfreie Sprachen	13
2.27	Pattern	14
2.28	Lexeme	14
2.29	Lexer (bzw. Scanner)	14
2.30	Literal	15
2.31	Konkrete Syntax	16
2.32	Derivation Tree (bzw. Parse Tree)	16
2.33	Parser	16
2.34	Recognizer (bzw. Erkennen)	17
2.35	Transformer	18
2.36	Visitor	18
2.37	Abstrakte Syntax	19
2.38	Abstrakt Syntax Tree	19
2.39	Pass	21
2.40	Fehlermeldung	21
3.1	Symboltabelle	23

1 Motivation

1.1 PicoC und RETI

1.2 Aufgabenstellung

1.3 Eigenheiten der Sprache C

1.4 Richtlinien

2 Einführung

2.1 Compiler und Interpreter

Der wohl wichtigsten zu klärenden Begriffe, sind die eines **Compilers** (Definition 2.2) und eines **Interpreters** (Definition 2.1), da das Schreiben eines Compilers von der **PicoC-Sprache** in die **RETI-Sprache** das Thema dieser Bachelorarbeit ist und die Definition eines **Interpreters** genutzt wird, um zu definieren was ein **Compiler** ist. Des Weiteren wurde zur **Qualitätsicherung** ein **RETI-Interpreter** implementiert, um mithilfe des **GCC**¹ und von **Tests** die **Beziehungen** in 2.1 zu belegen (siehe Subkapitel 4.2).

Definition 2.1: Interpreter

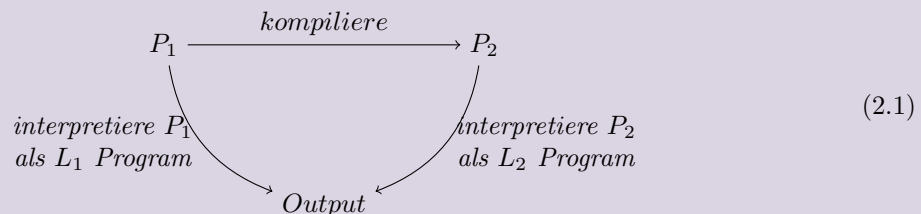
*Interpretiert die **Instructions** bzw. **Statements** eines Programmes P direkt.*

*Auf die Implementierung bezogen arbeitet ein Interpreter auf den compilerinternen **Sub-Bäumen** des **Abstract Syntax Tree** (Definition 2.38) und führt je nach Komposition der **Nodes** des Abstract Syntax Tree, auf die er während des Darüber-Iterierens stösst unterschiedliche Anweisungen aus.*

Definition 2.2: Compiler

***Kompiliert** ein 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 das Program P_1 in das Program P_2 übersetzt und bei beiden Programmen, wenn sie von **Interpreter** ihrer jeweiligen Sprachen L_1 und L_2 **interpretiert** werden, der gleiche **Output** rauskommt. Also beide Programme P_1 und P_2 die gleiche **Semantik** haben und sich nur **syntaktisch** durch die Sprachen L_1 und L_2 , in denen sie geschrieben stehen unterscheiden.*



Üblicherweise kompiliert ein **Compiler** ein **Program**, dass in einer **Programmiersprache** geschrieben ist zu **Maschinenncode**, der in **Maschinensprache** (Definition 2.3) geschrieben ist, aber es gibt z.B. auch **Transpiler** (Definition 2.9) oder **Cross-Compiler** (Definition 2.10). Des Weiteren sind **Maschinensprache** und **Assemblersprache** (Definition 2.4) voneinander zu unterscheiden.

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

Definition 2.3: Maschinensprache

Programmiersprache, deren mögliche Programme die **hardwarenahe 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 und allgemein in einer **geringen konstanten** Anzahl von Fetch- und Execute Phasen im **komplexeren Fall**. Die Maschinenbefehle sind meist so designed, dass sie sich innerhalb bestimmter **Wortbreiten**, die 2er Potenzen sind codieren lassen. Im einfachsten Fall innerhalb einer **Speicherzelle** des **Hauptspeichers**.^a.

^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 2.8) haben.

Definition 2.4: Assemblersprache (bzw. engl. Assembly Language)

Eine sehr **hardwarenahe** Programmiersprache, deren **Instructions** eine starke Entsprechung zu bestimmten Maschinenbefehlen bzw. Folgen von Maschinenbefehlen^a haben. Viele **Instructions** haben eine ähnliche Struktur mit **Operation** <Operanden>, mit einer **Operation**, die einem **Opcode** eines Maschinenbefehls bezeichnet und keinen oder mehreren **Operanden**, wie die späteren Maschinenbefehle, denen sie entsprechen. Allerdings gibt es oftmals noch viel „syntaktischen Zucker“ innerhalb^b der Instructions und drumherum^c.

^aInstructions der Assemblersprache, die mehreren Maschinenbefehlen entsprechen werden auch als **Pseudo-Instructions** bezeichnet und entsprechen dem, was man im allgemeinen als Macro bezeichnet.

^bZ.B. erlaubt die Assemblersprache des **GCC** für die **X86_64-Architektur** für manche Operanden die Syntax **n(%r)**, die einen **Speicherzugriff** mit **Offset** *n* zur Adresse, die im **Register** *%r* steht durchführt, wobei z.B. die Klammern () usw. nur „syntaktischer Zucker“ sind und natürlich nicht mitcodiert werden.

^cZ.B. sind in **X86_64** die Instructions in **Blöcken** untergebracht, die ein **Label** haben und zu denen mittels **jmp <label>** gesprungen werden kann. Ein solches Konstrukt, was vor allem auch noch relativ beliebig wählbare Bezeichner verwendet kann keine direkte Entsprechung in einem Prozessor und Hauptspeicher haben.

Ein **Assembler** (Definition 2.5) ist in üblichen Compilern in einer bestimmten Form meist schon integriert sein, da Compiler üblicherweise direkt **Maschinenencode** bzw. **Objectcode** (Definition 2.6) erzeugen. Ein **Compiler** soll möglichst viel von seiner internen Funktionsweise und der damit verbundenen Theorie für den Benutzer abstrahieren und dem Benutzer daher standardmäßig einfach nur das als Output liefern, was er eigentlich haben will, nämlich den **Maschinenencode** bzw. **Objectcode**, der direkt ausführbar ist bzw. wenn er später mit dem **Linker** (Definition 2.7) zu Maschiencode zusammengesetzt wird ausführbar ist.

Definition 2.5: Assembler

Übersetzt im allgemeinen **Assemblercode**, der in **Assemblersprache** geschrieben ist zu **Maschinenencode** der in **Maschinensprache** geschrieben ist.

Definition 2.6: Objectcode

Bei komplexeren Compilern, die es erlauben den Programmcode in **mehrere Dateien** aufzuteilen wird **Objectcode** erzeugt, der neben der Folge von Maschinenbefehlen in binärer Repräsentation auch noch Informationen für den **Linker** enthält, die im späteren **Maschiencode** nicht mehr enthalten sind, sobald der **Linker** die Objekte Dateien zum Maschiencode zusammengesetzt hat.

Definition 2.7: Linker

Programm, das **Objektcode** aus mehreren Objektdateien zu ausführbarem **Maschinencode** in eine ausführbare Datei oder Bibliotheksdatei **linkt**, sodass unter anderem kein vermeidbarer **doppelter** Code darin vorkommt.

Der **Maschinencode**, denn 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 2.8) enthält. Für den **RETI-Interpreter** ist es ebenfalls nicht notwendig, dass der Maschinencode, denn 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 eines RETI-Prozessors **simulieren** soll und nicht dessen mögliche interne Umsetzung².

Definition 2.8: Immediate

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

^aLjohhuh, What is an immediate value?

Definition 2.9: Transpiler (bzw. Source-to-source Compiler)

Kompiliert zwischen Sprachen, die ungefähr auf dem **gleichen** Level an **Abstraktion** arbeiten^a.

^aDie im aktuellen Zeitgeist in Mode gekommene Sprache **TypeScript** will als Obermenge von JavaScript und wird daher zu **JavaScript** transpiliert.

Definition 2.10: 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.

Ein **Cross-Compiler** ist entweder notwendig, wenn noch kein oder niemals ein Compiler C_w für die **Wunschsprache** L_w existiert bzw. existieren wird, der unter der **Maschinensprache** B_2 einer Zielmaschine M_2 läuft oder die Zielmaschine M_2 nicht ausreichend **Rechenleistung** hat, um ein Programm in der Wunschsprache L_w selbst zeitnah zu kompilieren.³

²Einen **RETI-Prozessor** zu bauen, der menschenlesbaren Maschinencode in z.B. **UTF-8 Codierung** ausführen kann, wäre dagegen unnötig kompliziert und aufwändig, da Hardware **binär** arbeitet und man dieser daher lieber direkt die binär codierten 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.

Hat die Zielmaschine M_2 allerdings ausreichend Rechenleistung, um Programme, die in L_w geschrieben sind, selbst in ihre Maschinensprache B_2 zu kompilieren und es fehlt lediglich ein Compiler C_w , der das übernimmt, so ist eine mögliche Lösung, um diesen **Compiler** C_w zu erhalten, in einem ersten Schritt einen **minimalen Compiler** $C_{w_{min}}$ zu schreiben, der eine reduzierte **Wunschsprache** $L_{w_{min}}$ in die **Maschinensprache** B_2 der Zielmaschine M_2 kompiliert.

Diesen **minimalen Compiler** $C_{w_{min}}$ könnte man nun auf dem aufwendigen Weg komplett in der **Assemblersprache** A_2 (Definition 2.4) der **Maschinensprache** B_2 schreiben. Oder man verwendet einen anderen Compiler C_o , der eine Sprache L_o kompiliert und schreibt diesen **minimalen Compiler** $C_{w_{min}}$ in dieser Sprache L_o und kompiliert diesen dann mit dem Compiler C_o . Dieser andere Compiler C_o kann entweder ein **Cross-Compiler** sein, der auf einer anderen Maschine M_1 läuft oder ein Compiler, der bereits auf der Zielmaschine M_2 mittels **Bootstrapping** (Definition 2.11) zum laufen gebracht wurde.

Hat man diesen **minimalen Compiler** $C_{w_{min}}$, kann man Programme in der Wunschsprache schreiben und mithilfe des **minimalen Compilers** kompilieren, sodass diese Programme auf der Zielmaschine laufen.

Nun kann man den **minimalen Compiler** $C_{w_{min}}$, den man gerade eben in einer **Programmiersprache** L_o implementiert hat, in der **Wunschsprache** L_w selbst implementieren und dann mit dem **minimale Compiler** für ebendiese Wunschsprache selbst kompilieren. Was man als Output bekommt ist ein **minimaler Compiler**, der aber auf der **Zielmaschine** läuft und die **Wunschsprache** in die **Maschinensprache** der Zielmaschine kompiliert.

Aufbauend auf diesem **minimalen Compiler**, der auf der **Zielmaschine** läuft, kann man nun auf der Zielmaschine selbst **iterativ** den minimalen Compiler schrittweise zu einem umfangreicheren Compiler, der mehr Funktionalitäten unterstützt weiterentwickeln und braucht die ursprüngliche Maschine, auf dem man die allererste Version des minimalen Compilers implementiert hat nicht mehr. Dieses Vorgehen wird auch als **Bootstrapping** (Definition 2.11) bezeichnet.⁴

Definition 2.11: Bootstrapping

*Es gibt zwei verschiedene Formen von **Bootstrapping**:*

2.11.1: Wenn man ein **Programm**, das auf einer **Maschine** M_1 läuft, auch auf einer **Zielmaschine** M_2 zum laufen bringt.

2.11.2: Wenn man einen Compiler mithilfe von **früheren Versionen** seiner selbst schreibt. Man schreibt den erweiterten Compiler in der Sprache, welche von der früheren Version des Compilers kompiliert wird und schafft es so iterativ immer umfangreichere Compiler zu bauen.^a

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

Die Form von Bootstrapping in 2.11.2 bildet im Falle dessen, dass es sich beim dem Programm um einen Compiler für eine Programmiersprache handelt die Lösung für ein Problem, dass auf das **Henne-Ei-Problem**⁵ reduziert werden kann.

Wenn es auf der Zielmaschine M_2 noch gar keinen Compiler für die Programmiersprache gibt, in der der Compiler für ebendiese Programmiersprache geschrieben steht, so liegt eine **zirkulare Abhängigkeit vor**, die man nur auflösen kann, indem eine **externe Entität** zur Hilfe nimmt.

⁴Der Begriff hat seinen Ursprung in der englischen Redewendung „pulling yourself up by your own bootstraps“, was im deutschen ungefähr der aus den Lügengeschichten des Freiherrn von Münchhausen bekannten Redewendung „sich am eigenen Schopf aus dem Sumpf ziehen“entspricht.

⁵Beschreibt die Situation, wenn ein System sich selbst als Abhängigkeit hat, damit es überhaupt einen Anfang für dieses System geben kann. Dafür steht das Problem mit der Henne und dem Ei sinnbildlich, da hier die Frage ist, wie das ganze seinen Anfang genommen hat, da beides zirkular voneinander abhängt.

Eine Möglichkeit ist, indem man einen **Cross-Compiler** auf der Maschine M_1 nutzt, der den **Compiler** der Programmiersprache für die andere Maschine M_2 kompiliert bzw. **bootstraped**.

2.1.1 T-Diagramme

Definition 2.12: T-Diagramm

2.2 Grammatiken

2.3 Grundlagen

Definition 2.13: Sprache

Definition 2.14: Chomsky Hierarchie

Definition 2.15: Grammatik

Definition 2.16: Reguläre Sprachen

Definition 2.17: Ableitung

Definition 2.18: Links- und Rechtsableitung

Definition 2.19: 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.

2.3.1 Mehrdeutige Grammatiken

Definition 2.20: Ableitungsbaum

Definition 2.21: Mehrdeutige Grammatik

2.3.2 Präzidenz und Assoziativität

Definition 2.22: Assoziativität

Definition 2.23: Präzidenz

Definition 2.24: Wortproblem

Definition 2.25: LL(k)-Grammatik

Eine Grammatik ist **LL(k)** für $k \in \mathbb{N}$, falls jeder Ableitungsschritt eindeutig durch die nächsten k **Symbole** des **Eingabeworts** bzw. in Bezug zu Compilerbau **Token** des **Inputstrings** zu bestimmen ist^a. Dabei steht **LL** für *left-to-right* und *leftmost-derivation*, da das **Eingabewort** von *links nach rechts* geparsed und immer **Linksableitungen** genommen werden müssen^b, damit die obige Bedingung mit den *nächsten* k Symbolen gilt.

^aDas wird auch als **Lookahead** von k bezeichnet.

^bWobei sich das mit den **Linksableitungen** automatisch ergibt, wenn man das Eingabewort von **links-nach-rechts** parsed und jeder der nächsten k **Ableitungsschritte** eindeutig sein soll.

Definition 2.26: Kontextfreie Sprachen

2.4 Lexikalische Analyse

Die **Lexikalische Analyse** bildet üblicherweise die erste Ebene innerhalb der **Pipe Architektur** bei der Implementierung von Compilern. Die Aufgabe der lexikalischen Analyse ist vereinfacht gesagt, in einem Inputstring, z.B. dem Inhalt einer Datei, welche in **UTF-8** codiert ist, Folgen endlicher Symbole (auch **Wörter** genannt) zu finden, die bestimmte **Pattern** (Definition 2.27) matchen, die durch eine **reguläre Grammatik** spezifiziert sind.

Definition 2.27: Pattern

Beschreibung aller möglichen **Lexeme** einer Menge \mathbb{P}_T , die 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.

^bWhat is the difference between a token and a lexeme?

Diese Folgen endlicher Symbole werden auch **Lexeme** (Definition 2.28) genannt.

Definition 2.28: Lexeme

Ein **Lexeme** ist ein **Wort** aus dem Inputstring, welches das **Pattern** für eines der **Token** T einer **Sprache** L_{Lex} matched.^a

^aWhat is the difference between a token and a lexeme?

Diese **Lexeme** werden vom **Lexer** im **Inputstring** identifiziert und **Tokens** T zugeordnet (Definition 2.29). Die **Tokens** sind es, die letztendlich an die **Syntaktische Analyse** weitergegeben werden.

Definition 2.29: Lexer (bzw. Scanner)

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

^alecture-notes-2021.

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

Eine weitere Aufgabe der **Lexikalischen Analyse** ist es jegliche für die Weiterverarbeitung unwichtigen Symbole, wie Leerzeichen `␣`, Newline `\n`⁶ und Tabs `\t` aus dem Inputstring 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 $\epsilon \in \Sigma^*$. 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 Überbegriffe bzw. Tokennamen für beliebige Bezeichner von Variablen, Konstanten und Funktionen und beliebige Zahlen sind aber trotz allem z.B. **Zahl** und **Bezeichner**.

Ein **Lexeme** ist damit aber nicht das gleiche, wie der **Tokenwert**, denn z.B. im Falle von PicoC kann z.B. der Wert 99 durch zwei verschiedene Literale dargestellt werden, einmal als ASCII-Zeichen `'c'` und des Weiteren

⁶In 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 wegabstrahiert.

auch in Dezimalschreibweise als 99⁷. Der **Tokenwert** ist jedoch der letztendliche Wert an sich, unabhängig von der Darstellungsform.

Die **Grammatik** G_{Lex} , die zur Beschreibung der Token T einer regulären Sprache L_{Lex} verwendet wird, ist üblicherweise **regulär**, da ein typischer **Lexer** immer nur **ein Symbol** vorausschaut⁸, unabhängig davon, was für Symbole davor aufgetaucht sind. Die übliche Implementierung eines **Lexers** merkt sich nicht, was für Symbole davor aufgetaucht sind.

Um Verwirrung verzubäuen ist es wichtig folgende Unterscheidung hervorzuheben: Wenn von **Symbolen** die Rede ist, so werden in der **Lexikalischen Analyse**, der **Syntaktischen 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 von Variablen, Konstanten und Funktionen** die Symbole^a.

^aDas ist der Grund, warum die Tabelle, in der Informationen zu Identifiern gespeichert werden aus Kapitel 3 Symboltabelle genannt wird.

Definition 2.30: Literal

*Eine von möglicherweise vielen weiteren **Darstellungsformen** für ein und denselben **Wert**.*

Um eine Gesamtübersicht über die **Lexikalische Analyse** zu geben, ist in Abbildung 2.1 die Lexikalische Analyse an einem Beispiel veranschaulicht.

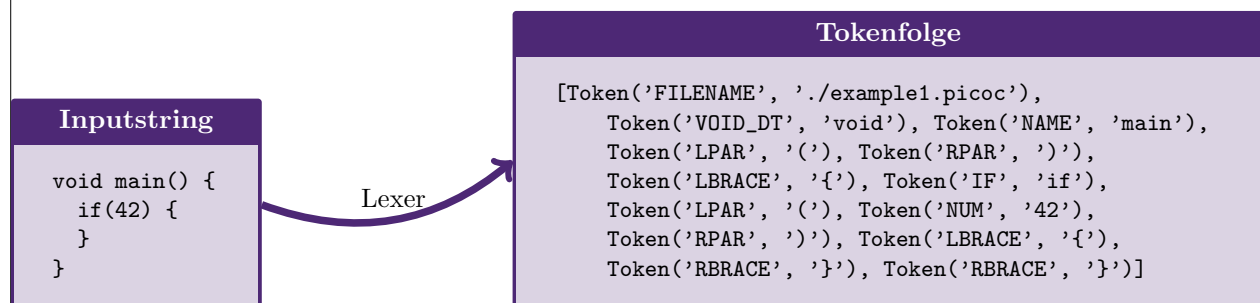


Abbildung 2.1: Veranschaulichung der Lexikalischen Analyse

2.5 Syntaktische Analyse

In der **Syntaktischen Analyse** ist für einige Sprachen eine **Kontextfreie Grammatik** G_{Parse} notwendig, um diese Sprache 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 Klammern '(' bzw. öffnende geschweifte Klammern '{' es momentan gibt, die noch nicht durch eine

⁷Die Programmiersprache Python erlaubt es z.B. diesern Wert auch mit den Literalen `0b1100011` und `0x63` darzustellen.

⁸Man nennt das auch einem **Lookahead** von 1.

entsprechende schließende Klammer '}' bzw. schließende geschweifte Klammer '}' geschlossen wurden.

Die **Syntax**, in welcher der **Inputstring** aufgeschrieben ist, wird auch als **Konkrete Syntax** (Definition 2.31) bezeichnet. In einem Zwischenschritt, dem **Parsen** wird aus diesem Inputstring mithilfe eines **Parsers** (Definition 2.33), ein **Derivation Tree** (Definition 2.32) generiert, der als Zwischenstufe hin zum einem **Abstrakt Syntax Tree** (Definition 2.38) dient. Für einen ordentlichen Code ist es vor allem im Compilerbau förderlich kleinschrittig vorzugehen, deshalb erst die Generierung des **Derivation Tree** und dann der **Abstrakt Syntax Tree**.

Definition 2.31: Konkrete Syntax

Syntax einer Sprache, die durch die Grammatiken G_{Lex} und G_{Parse} zusammengekommen beschrieben wird.

Ein Programm in seiner Textrepräsentation, wie es in einer Textdatei nach den Produktionen der Grammatiken G_{Lex} und G_{Parse} abgeleitet steht, bevor man es kompiliert, ist in Konkreter Syntax aufgeschrieben.^a

^aCourse Webpage for Compilers (P423, P523, E313, and E513).

Definition 2.32: Derivation Tree (bzw. Parse Tree)

Compilerinterne Darstellung eines in Konkreter Syntax geschriebenen Inputstrings als Baumdatenstruktur, in der Nichtterminalsymbole die Inneren Knoten des Baumes und Terminalsymbole die Blätter des Baumes bilden. Jede Produktions der Grammatik G_{Parse} , die ein Teil der Konkrete Syntax ist, wird zu einem eigenen Knoten.

Der Derivation Tree wird optimalerweise immer so konstruiert bzw. die Konkrete Syntax immer so definiert, dass sich möglichst einfach ein Abstrakt Syntax Tree daraus konstruieren lässt.

Definition 2.33: Parser

Ein Programm, dass eine Eingabe in eine für die Weiterverarbeitung taugliche Form bringt.

2.33.1: *In Bezug auf Compilerbau ist ein Parser ein Programm, dass einen Inputstring von Konkreter Syntax in die compilerinterne Darstellung eines Derivation Tree übersetzt, was auch als Parsen bezeichnet wird.^{a, b}*

^aEs gibt allerdings auch alternative Definitionen, denen nach ein Parser in Bezug auf Compilerbau ein Programm ist, dass einen Inputstring von Konkreter Syntax in Abstrakte Syntax übersetzt. Im Folgenden wird allerdings die obige Definition 2.33.1 verwendet.

^bCompiler Design - Phases of Compiler.

An dieser Stelle könnte möglicherweise eine Begriffsverwirrung entstehen, ob ein **Lexer** nach der obigen Definition nicht auch ein **Parser** ist.

In Bezug auf Compilerbau ist ein **Lexer** ein Teil eines Parsers. Der Parser vereinigt sowohl die **Lexikalische Analyse**, als auch einen Teil der **Syntaktischen Analyse** in sich. Aber für sich isoliert, ohne Bezug zu Compilerbau betrachtet, ist ein Lexer nach Definition 2.33 ebenfalls ein Parser. Aber im Compilerbau hat **Parser** eine spezifischere Definition und hier überwiegt beim **Lexer** seine Funktionalität, dass er den Inputstring lexikalisch weiterverarbeitet, um ihn als Lexer zu bezeichnen, der Teil eines Parsers ist.

Die vom **Lexer** im Inputstring identifizierten **Token** werden in der **Syntaktischen Analyse** vom **Parser** (Definition 2.33) 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 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 2.6 relevant.

Ein **Parser** ist genauer gesagt ein erweiterter **Recognizer** (Definition 2.34), denn ein Parser löst das **Wortproblem** (Definition 2.24) für die **Sprache**, die durch die **Konkrete Syntax** beschrieben wird und konstruiert parallel dazu oder im Nachgang aus den Informationen, die während der Ausführung des Recognition Algorithmus gesichert wurden den **Derivation Tree**.

Definition 2.34: Recognizer (bzw. Erkenner)

*Entspricht dem Maschinenmodell eines **Automaten**. Im Bezug auf Compilerbau entspricht der **Recognizer** einem **Kellerautomaten**, in dem **Wörter** bestimmter **Kontextfreier Sprachen** erkannt werden. Der **Recognizer** erkennt, ob ein Inputstring bzw. **Wort** sich mit den Produktionen der **Konkrete Syntax** ableiten lässt, also ob er bzw. es Teil der Sprache ist, die von der **Konkreten Syntax** beschrieben wird oder nicht^a.*

^aDas vom **Recognizer** gelöste Problem ist auch als **Wortproblem** bekannt.

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

- **Top-Down Parsing:** Der **Derivation Tree** wird von **oben-nach-unten** generiert, also von der **Wurzel** zu den **Blättern**. Dementsprechend fängt die Generierung des **Derivation Tree** mit dem **Startsymbol** der **Grammatik** an und wendet in jedem Schritt eine **Linksableitung** auf die **Nicht-Terminalsymbole** an, bis man **Terminalsymbole** hat und der gewünschte **Inputstring** abgeleitet wurde oder es sich herausstellt, dass dieser nicht abgeleitet werden kann.^a

Der Grund, warum die **Linksableitung** verwendet wird und nicht z.B. die **Rechtsableitung** ist, weil der das **Eingabewert** bzw. der **Inputstring** 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**. Dabei wird jedem **Nicht-Terminalsymbol** eine **Prozedur** zugeordnet, welche die Produktionsregeln dieses **Nicht-Terminalsymbols** umsetzt. Prozeduren rufen sich dabei wechselseitig gegenseitig entsprechend der **Produktionsregeln** auf, falls eine entsprechende Produktionsregel eine **Rekursion** enthält.

Rekursiver Abstieg kann mit **Backtracking** verbunden werden, um auch Grammatiken parsen zu können, die nicht **LL(k)** (Definition 2.25) sind. Dabei werden meist nach dem **Depth-First-Search Prinzip** 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. Mit dieser Methode ist das Parsen **Linksrekursiver Grammatiken** (Definition 2.19) allerdings nicht möglich, ohne die Grammatik vorher umgeformt zu haben und jegliche **Linksrekursion** aus der **Grammatik** entfernt zu haben, da diese zu **Unendlicher Rekursion** führt^b

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

- **Bottom-Up Parsing:** Es wird mit dem **Eingabewort** bzw. **Inputstring** gestartet und versucht **Rechtsableitungen**, entsprechend der **Produktionen** der **Konkreten Syntax** rückwärts anzuwenden bis man beim **Startsymbol** landet.^d
- **Chart Parser:** 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

^aWhat is Top-Down Parsing?

^bDiese Art von Parser 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 *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. Dieser **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 **Abstrakt Syntax Tree** wird mithilfe von **Transformern** (Definition 2.35) und **Visitors** (Definition 2.36) generiert und ist das Endprodukt der **Syntaktischen Analyse**. Wenn man die gesamte **Syntaktische Analyse** betrachtet, so übersetzt diese einen Inputstring von der **Konkreten Syntax** in die **Abstrakte Syntax** (Definition 2.37).

Die **Baumdatenstruktur** des **Derivation Tree** und **Abstrakt Syntax Tree** ermöglicht es die Operationen, die der Compiler bei der Weiterverarbeitung des Inputstrings ausführen muss möglichst **effizient** auszuführen.

Definition 2.35: Transformer

Ein Programm, dass von **unten-nach-oben**, nach dem **Breadth First Search** Prinzip alle Knoten des **Derivation Tree** besucht und in Bezug zu Compilerbau, beim Antreffen eines bestimmten Knoten des **Derivation Tree** einen entsprechenden Knoten des **Abstrakt Syntax Tree** erzeugt und diesen anstelle des Knotens des **Derivation Tree** setzt und so Stück für Stück den **Abstrakt Syntax Tree** konstruiert.

Definition 2.36: Visitor

Ein Programm, dass von **unten-nach-oben**, nach dem **Breadth First Search** Prinzip alle Knoten des **Derivation Tree** besucht und in Bezug zu Compilerbau, beim Antreffen eines bestimmten **Knoten** des **Derivation Tree**, diesen **in-place** mit anderen Knoten **tauscht** oder **manipuliert**, um den **Derivation Tree** für die weitere Verarbeitung durch z.B. einen **Transformer** zu vereinfachen.

Kann theoretisch auch zur Konstruktion eines **Abstrakt Syntax Tree** verwendet werden, wenn z.B. eine externe Klasse verwendet wird, welches für die Konstruktion des **Abstrakt Syntax Tree** verantwortlich ist, aber dafür ist ein **Transformer** besser geeignet.

Definition 2.37: Abstrakte Syntax

Syntax, die beschreibt, was für Arten von **Komposition** bei den **Knoten** eines **Abstrakt Syntax Trees** möglich sind.^a

Jene Produktionen, die in der **Konkreten Syntax** für die Umsetzung von **Präzidenz** notwendig waren, sind in der **Abstrakten Syntax** abgeflacht.

^aCourse Webpage for Compilers (P423, P523, E313, and E513).

Definition 2.38: Abstrakt Syntax Tree

Compilerinterne Darstellung eines Programs, in welcher sich anhand der Knoten auf dem Pfad von der Wurzel zu einem **Blatt** nicht mehr direkt nachvollziehen lässt, durch welche **Produktionen** dieses Blatt abgeleitet wurde.

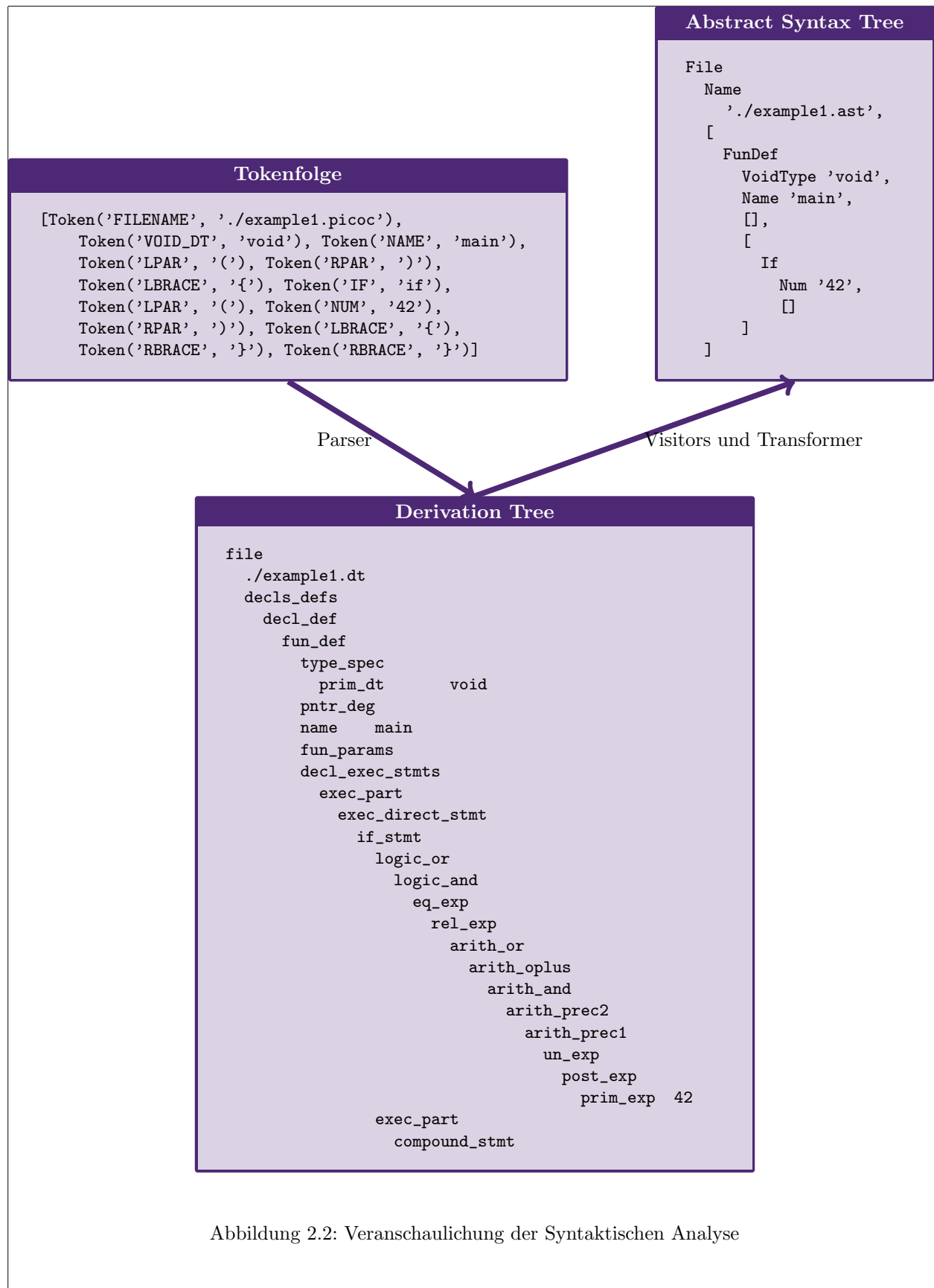
Der **Abstrakt Syntax Tree** hat einmal den Zweck, dass die Kompositionen, die die Knoten bilden können **semantisch** näher an den **Instructions eines Assemblers** dran sind und, dass man mit ihm bei der Betrachtung eines **Knoten**, der für einen Teil des Programms steht, möglichst schnell die Frage beantworten kann, welche **Funktionalität** der Sprache dieser umsetzt, welche **Bestandteile** er hat und welche Funktionalität der Sprache diese Bestandteile umsetzen usw.^a

^aCourse Webpage for Compilers (P423, P523, E313, and E513).

Je weiter **unten**⁹ und **links** ein Knoten im **Abstrakt Syntax Tree** liegt, desto eher wird dieser Knoten komplett abgearbeitet sein, da in der **Code Generierung** die Knoten nach dem **Depth First Search** Prinzip abgearbeitet werden.

Um eine Gesamtübersicht über die **Syntaktische Analyse** zu geben, ist in Abbildung 2.2 die Syntaktische mit dem Beispiel aus Subkapitel 2.4 fortgeführt.

⁹In der Informatik wachsen **Bäume** von **oben-nach-unten**. Die **Wurzel** ist also **oben**.



2.6 Code Generierung

Definition 2.39: Pass

2.7 Fehlermeldungen

Definition 2.40: Fehlermeldung

3 Implementierung

3.1 Architektur

3.2 Lexikalische Analyse

3.2.1 Verwendung von Lark

3.2.2 Basic Parser

3.3 Syntaktische Analyse

3.3.1 Verwendung von Lark

3.3.2 Umsetzung von Präzidenz

Die **PicoC Sprache** hat dieselben **Präzidenzregeln** implementiert, wie die **Sprache C**¹. Die **Präzidenzregeln** von **PicoC** sind in Tabelle 3.3.2 aufgelistet.

¹[C Operator Precedence - cppreference.com](http://c.operatorprecedence-cppreference.com).

Präzidenz	Operator	Beschreibung	Assoziativität
1	a() a[] a.b	Funktionsaufruf Indexzugriff Attributzugriff	Links, dann rechts →
2	-a !a ~a *a &a	Unäres Minus Logisches NOT und Bitweise NOT Dereferenz und Referenz, auch Adresse-von	Rechts, dann links ←
3	a*b a/b a%b	Multiplikation, Division und Modulo	Links, dann rechts →
4	a+b a-b	Addition und Subtraktion	
5	a<b a<=b a>b a>=b	Kleiner, Kleiner Gleich, Größer, Größer gleich	
6	a==b a!=b	Gleichheit und Ungleichheit	
7	a&b	Bitweise UND	
8	a^b	Bitweise XOR (exclusive or)	
9	a b	Bitweise ODER (inclusive or)	
10	a&& b	Logisches UND	
11	a b	Logisches ODER	
12	a=b	Zuweisung	Rechts, dann links ←
13	a,b	Komma	Links, dann rechts →

Tabelle 3.1: Präzidenzregeln von PicoC

3.3.3 Derivation Tree Generierung

3.3.4 Early Parser

3.3.5 Derivation Tree Vereinfachung

3.3.6 Abstrakt Syntax Tree Generierung

ASTNode

PicoC Nodes

RETI Nodes

3.4 Code Generierung

3.4.1 Passes

PicoC-Shrink Pass

PicoC-Blocks Pass

PicoC-Mon Pass

Definition 3.1: Symboltabelle

RETI-Blocks Pass

RETI-Patch Pass

RETI Pass

3.4.2 Umsetzung von Pointern und Arrays

3.4.3 Umsetzung von Structs

3.4.4 Umsetzung von Funktionen

3.4.5 Umsetzung kleinerer Details

3.5 Fehlermeldungen

3.5.1 Error Handler

4 Ergebnisse und Ausblick

4.1 Funktionsumfang

4.2 Qualitätssicherung

4.3 Kommentierter Kompiliervorgang

4.4 Erweiterungsideen

A Appendix

A.1 Konkrete und Abstrakte Syntax

A.2 Bedienungsanleitungen

A.2.1 PicoC-Compiler

A.2.2 Showmode

A.2.3 Entwicklertools

Literatur

Online

- *C Operator Precedence* - *cppreference.com*. URL: https://en.cppreference.com/w/c/language/operator_precedence (besucht am 27.04.2022).
- *Compiler Design - Phases of Compiler*. URL: https://www.tutorialspoint.com/compiler_design/compiler_design_phases_of_compiler.htm (besucht am 19.06.2022).
- *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).
- *lecture-notes-2021*. 20. Jan. 2022. URL: <https://github.com/Compiler-Construction-Uni-Freiburg/lecture-notes-2021/blob/56300e6649e32f0594bbbd046a2e19351c57dd0c/material/lexical-analysis.pdf> (besucht am 28.04.2022).
- Ljohhuh. *What is an immediate value?* 4. Apr. 2018. URL: <https://reverseengineering.stackexchange.com/q/17671> (besucht am 13.04.2022).
- *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 a token and a lexeme?* NewbeDEV. URL: <http://newbedev.com/what-is-the-difference-between-a-token-and-a-lexeme> (besucht am 17.06.2022).
- *What is Top-Down Parsing?* URL: <https://www.tutorialspoint.com/what-is-top-down-parsing> (besucht am 22.06.2022).