Programming Languages

http://www.cse.iitd.ac.in/~sak/courses/pl/2017-18.index.html

S. Arun-Kumar

Department of Computer Science and Engineering
I. I. T. Delhi, Hauz Khas, New Delhi 110 016.

April 21, 2018
## Contents

1. **The Programming Languages Geneology**
   - Page 3

2. **Introduction**
   - Page 13

3. **Scanning or Lexical Analysis**
   - 3.1 Regular Expressions
     - Page 41
   - 3.2 Nondeterministic Finite Automata (NFA)
     - Page 59
   - 3.3 Deterministic Finite Automata (DFA)
     - Page 92

4. **Parsing or Syntax Analysis**
   - 4.1 Grammars
     - Page 105
   - 4.2 Context-Free Grammars
     - Page 109
   - 4.3 Ambiguity
     - Page 121
   - 4.4 Shift-Reduce Parsing
     - Page 140
   - 4.5 Bottom-Up Parsing
     - Page 176
   - 4.6 Recursive Descent Parsing
     - Page 194
   - 4.7 Specification of Syntax: Extended Backus-Naur Form
     - 4.7.1 The Extended Backus-Naur Form (EBNF)
       - Page 201
14.2 FL with type rules ................................................................. 379

15 Logic Programming and Prolog 392
1. The Programming Languages Geneology
The Landscape of General PLs
The Usage of General PLs

![Diagram showing the usage of general programming languages over time, with labels and categories like Scientific & Numerical Computations, Business Data Processing, Structured Programming, and more.](image-url)
The Major Features of General PLs
FORTRAN

- The very first high-level programming language
- Still used in scientific computation
- Static memory allocation
- Very highly compute oriented
- Runs very fast because of static memory allocation
- Parameter passing by reference
COBOL

- A business oriented language
- Extremely verbose
- Very highly input-oriented
- Meant to manage large amounts of data on disks and tapes and generate reports
- Not computationally friendly
LisP

- First functional programming language
- Introduced lists and list-operations as the only data-structure
- Introduced symbolic computation
- Much favoured for AI and NLP programming for more than 40 years
- The first programming language whose interpreter could be written in itself.
ALGOL-60

• Introduced the Backus-Naur Form (BNF) for specifying syntax of a programming language
• Formal syntax defined by BNF (an extension of context-free grammars)
• First language to implement recursion
• Introduction of block-structure and nested scoping
• Dynamic memory allocation
• Introduced the call-by-name parameter mechanism
Pascal

• ALGOL-like language meant for teaching structured programming
• Introduction of new data structures – records, enumerated types, subrange types, recursive data-types
• Its simplicity led to its “dialects” being adopted for expressing algorithms in pseudo-code
• First language to be ported across a variety of hardware and OS platforms – introduced the concepts of virtual machine and intermediate code (bytecode)
ML

- First strongly and statically typed functional programming language
- Created the notion of an inductively defined type to construct complex types
- Powerful pattern matching facilities on complex data-types.
- Introduced type-inference, thus making declarations unnecessary except in special cases
- Its module facility is inspired by the algebraic theory of abstract data types
- The first language to introduce functorial programming between algebraic structures and modules
2. Introduction
Introduction to Compiling

• Translation of programming languages into executable code
• But more generally any large piece of software requires the use of compiling techniques.
• The processes and techniques of designing compilers is useful in designing most large pieces of software.
• Compiler design uses techniques from theory, data structures and algorithms.
Software Examples

Some examples of other software that use compiling techniques

- Almost all user-interfaces require scanners and parsers to be used.
- All XML-based software require interpretation that uses these techniques.
- All mathematical text formatting requires the use of scanning, parsing and code-generation techniques (e.g. \texttt{\LaTeX}).
- Model-checking and verification software are based on compiling techniques.
- Synthesis of hardware circuits requires a description language and the final code that is generated is an implementation either at the register-transfer level or gate-level design.
Books and References


Source and Target

In general a **compiler** for a a **source** language $S$ written in some language $C$ translates code to a **target** language $T$.

**Source** $S$ could be

- a programming language, or
- a description language (e.g. Verilog, VHDL), or
- a markup language (e.g. XML, HTML, SGML, \LaTeX)

**Target** $T$ could be

- another programming language, assembly language or machine language, or
- a language for describing various objects (circuits etc.), or
- a low level language for execution, display, rendering etc.

We will be primarily concerned with compiling from a **high-level** programming language (source) to **low-level** code.
The Compiling Process

In general the process of compiling involves at least three languages

1. The language $S$ of source programs in which the users of the compiler write code.

2. The language $C$ in which the compiler itself is written. The assumption is that unless the compiler itself is written in machine language there is already a compiler or an interpreter for $C$.

3. The language $T$ into which the compiler translates the user programs.

Besides these three languages there could be several other intermediate languages $I_1, I_2, \ldots$ (also called intermediate representations) into which the source program could be translated in the process of compiling or interpreting the source programs written in $S$. In modern compilers, for portability, modularity and reasons of code improvement, there is usually at least one intermediate representation.
Compiling as Translation

Except in the case of a source to source translation (for example, a Pascal to C translator which translates Pascal programs into C programs), we may think of the process of compiling high-level languages as one of transforming programs written in $S$ into programs of lower-level languages such as the intermediate representation or the target language. By a low-level language we mean that the language is in many ways closer to the architecture of the target language.
Phases of a Compiler

A compiler or translator is a fairly complex piece of software that needs to be developed in terms of various independent modules. In the case of most programming languages, compilers are designed in phases. The various phases may be different from the various passes in compilation.
The Big Picture: 1

stream of characters

stream of tokens

SCANNER
The Big Picture: 2

SCANNER

stream of characters

stream of tokens

PARSER

parse tree
The Big Picture: 3

- **Scanner**: stream of characters
- **Parser**: stream of tokens, parse tree
- **Semantic Analyzer**: abstract syntax tree
The Big Picture: 4

- **SCANNER**
  - stream of characters

- **PARSER**
  - stream of tokens
  - parse tree

- **SEMANTIC ANALYZER**
  - abstract syntax tree

- **I.R. CODE GENERATOR**
  - intermediate representation
The Big Picture: 5

- SCANNER
  - stream of characters
- PARSER
  - stream of tokens
  - parse tree
- SEMANTIC ANALYZER
  - abstract syntax tree
- I.R. CODE GENERATOR
  - intermediate representation
- OPTIMIZER
  - optimized intermediate representation
The Big Picture: 6

- **Scanner**: stream of characters
- **Parser**: stream of tokens
- **Semantic Analyzer**: parse tree
- **I.R. Code Generator**: abstract syntax tree
- **Optimizer**: intermediate representation
- **Code Generator**: target code
The Big Picture: 7

Diagram:
- **Scanner**: Stream of characters → Stream of tokens
- **Parser**: Parse tree
- **Semantic Analyzer**: Abstract syntax tree
- **I.R. Code Generator**: Intermediate representation
- **Optimizer**: Optimized intermediate representation
- **Code Generator**: Target code
- **Error-Handler**
- **Symbol Table Manager**
The Big Picture: 8

Scanner → Parser → Semantic Analysis → Symbol Table

<table>
<thead>
<tr>
<th>Scanner</th>
<th>Parser</th>
<th>Semantic Analysis</th>
<th>Symbol Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR</td>
<td>Optimization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Scanning or Lexical Analysis

Lexical Analysis
Language

• Every language is built from a finite alphabet of symbols. The alphabet of a programming language consists of the symbols of the ASCII set.

• Each language has a vocabulary consisting of words. Each word is a string of symbols drawn from the alphabet.

• Each language has a finite set of punctuation symbols, which separate phrases, clauses and sentences.

• The phrases, clauses and sentences of a programming language are expressions, commands, functions, procedures and programs.
Lexical Analysis

**lex-i-cal:** relating to words of a language

- A *source program* (usually a file) consists of a stream of characters.
- Given a stream of characters that make up a *source program* the compiler must first break up this stream into groups of meaningful words, and other symbols.
- Each such group of characters is then classified as belonging to a certain *token type*.
- Certain sequences of characters are not tokens and are completely ignored (or skipped) by the compiler.
Tokens and Non-tokens

Tokens Typical tokens are

- **Identifiers**: Names of variables, constants, procedures, functions etc.
- **Keywords/Reserved words**: `void`, `public`, `main`
- **Operators**: `+`, `*`, `/`
- **Punctuation**: ``, `:`, `.`
- **Brackets**: `()`, `[]`, `begin`, `end`, `case`, ` esac`

Non-tokens Typical non-tokens

- **whitespace**: sequences of tabs, spaces, new-line characters,
- **comments**: compiler ignores comments
- **preprocessor directives**: `#include ...`, `#define ...
- **macros** in the beginning of C programs
Scanning: 1

- Takes a stream of characters and identifies tokens from the lexemes.

- Eliminates comments and redundant whitespace.

- Keeps track of line numbers and column numbers and passes them as parameters to the other phases to enable error-reporting to the user.
Scanning: 2

- **Whitespace**: A sequence of space, tab, newline, carriage-return, form-feed characters etc.

- **Lexeme**: A sequence of non-whitespace characters delimited by whitespace or special characters (e.g. operators like +, -, *).

- **Examples of lexemes**.
  - reserved words, keywords, identifiers etc.
  - Each comment is usually a single lexeme
  - preprocessor directives
Scanning: 3

- Token: A sequence of characters to be treated as a single unit.
- Examples of tokens.
  - Reserved words (e.g. begin, end, struct, if etc.)
  - Keywords (integer, true etc.)
  - Operators (+, &&, ++ etc)
  - Identifiers (variable names, procedure names, parameter names)
  - Literal constants (numeric, string, character constants etc.)
  - Punctuation marks (:, , etc.)
Scanning: 4

- Identification of tokens is usually done by a Deterministic Finite-state automaton (DFA).
- The set of tokens of a language is represented by a large regular expression.
- This regular expression is fed to a lexical-analyser generator such as Lex, Flex or JLex.
- A giant DFA is created by the Lexical analyser generator.
Lexical Rules

• Every programming language has **lexical rules** that define how a token is to be defined.

  **Example.** In most programming languages identifiers satisfy the following rules.

  1. An identifier consists of a sequence of letters (A ... Z, a ... z), digits (0 ... 9) and the underscore (_) character.
  2. The first character of an identifier must be a letter.

• Any two tokens are separated by some **delimiters** (usually whitespace) or **non-tokens** in the source program.
3.1. Regular Expressions

Lecture 03
Regular Expressions
Specifying Lexical Rules

We require compact and simple ways of specifying the lexical rules of a language. In particular,

- there are an infinite number of legally correct programs in any programming language and
- we require finite descriptions/specifications of the lexical rules so that they can cover the infinite number of legal programs.

One way of specifying the lexical rules of a programming language is to use regular expressions.
Regular Expressions

- Each regular expression is a finite sequence of symbols.
- A regular expression may be used to describe an infinite collection of strings.

The regular expression used to define the set of possible identifiers as defined by the rules is

\[ [A-Za-z][A-Za-z0-9_]^* \]
Concatenations

Consider a (finite) alphabet (of symbols) $A$.

- Any set of strings built up from the symbols of $A$ is called a language.
- Given any two strings $x$ and $y$ in a language, $x.y$ or simply $xy$ is the concatenation of the two strings.

**Example** Given the strings $x = \text{Mengesha}$ and $y = \text{Mamo}$, $x.y = \text{MengeshaMamo}$ and $y.x = \text{MamoMengesha}$.

- Given two languages $X$ and $Y$, then $X.Y$ or simply $XY$ is the concatenation of the languages.

**Example** Let $X = \{\text{Mengesha, Gemechis}\}$ and $Y = \{\text{Mamo, Bekele, Selassie}\}$

$XY = \{\text{MengeshaMamo, MengeshaBekele, MengeshaSelassie, GemechisMamo, GemechisBekele, GemechisSelassie}\}$
Note on the Concept of “language”.

Unfortunately we have too many related but slightly different concepts, each of which is simply called a “language”. Here is a clarification of the various concepts that we use.

- Every language has a non-empty finite set of symbols called letters. This non-empty finite set is called the alphabet.
- Each word is a finite sequence of symbols called letters.
- The words of a language usually constitute its vocabulary. Certain sequences of symbols may not form a word in the vocabulary. A vocabulary for a natural language is defined by a dictionary, whereas for a programming language it is usually defined by formation rules.
- A phrase, clause or sentence is a finite sequence of words drawn from the vocabulary.
- Every natural language or programming language is a finite or infinite set of sentences.
- In the case of formal languages, the formal language is the set of words that can be formed using the formation rules. The language is also said to be generated by the formation rules.

There are a variety of languages that we need to get familiar with.

Natural languages. These are the usual languages such as English, Hindi, French, Tamil which we employ for daily communication and in teaching, reading and writing.

Programming languages. These are the languages such as C, Java, SML, Perl, Python etc. that are used to write computer programs in.

Formal languages. These are languages which are generated by certain formation rules.

Meta-languages. These are usually natural languages used to explain concepts related to programming languages or formal languages. We are using English as the meta-language to describe and explain concepts in programming languages and formal languages.
In addition, we do have the concept of a **dialect** of a natural language or a programming language. For example the natural languages like Hindi, English and French do have several dialects. A dialect (in the case of natural languages) is a particular form of a language which is peculiar to a specific region or social group. *Creole* (spoken in Mauritius) is a dialect of French, Similarly *Brij, Awadhi* are dialects of Hindi. A dialect (in the case of programming languages) is a version of the programming language. There are many dialects of C and C++. Similarly *SML-NJ* and *poly-ML* are dialects of Standard ML. The notion of a dialect does not really exist for formal languages.

Closer home to what we are discussing, the language of regular expressions is a **formal language** which describes the rules for forming the words of a programming language. Each regular expression represents a finite or infinite set of words in the vocabulary of a programming language. We may think of the language of regular expressions also as a **functional programming language** for describing the vocabulary of a programming language. It allows us to generate words belonging to the vocabulary of a programming language.

Any formally defined language also defines an algebraic system of operators applied on a **carrier set**. Every operator in any algebraic system has a pre-defined **arity** which refers to the number of operands it requires. In the case of regular expressions, the operators are concatenation and alternation are 2-ary operators (binary operators), whereas the Kleene closure and plus closure are 1-ary operators (unary). In addition the letters of the alphabet, which are con-
Simple Language of Regular Expressions

We consider a simple language of regular expressions over a finite alphabet \( A \) of symbols. Each regular expression \( r \) denotes a set of strings \( \mathcal{L}(r) \). \( \mathcal{L}(r) \) is also called the language specified by the regular expression \( r \).

**Symbol** For each symbol \( a \) in \( A \), the regular expression \( a \) denotes the set \( \{a\} \).

**Concatenation** For any two regular expressions \( r \) and \( s \), their concatenation \( rs \) or \( r.s \) denotes the concatenation of the languages specified by \( r \) and \( s \). That is,

\[
\mathcal{L}(rs) = \mathcal{L}(r)\mathcal{L}(s)
\]
Epsilon and Alternation

**Epsilon** $\epsilon$ denotes the language with a single element the **empty** string ("" ) i.e.

$$L(\epsilon) = \{ "" \}$$

**Alternation** Given any two regular expressions $r$ and $s$, $r | s$ is the set union of the languages specified by the individual expressions $r$ and $s$ respectively.

$$L(r | s) = L(r) \cup L(s)$$

**Example** $L(\text{Menelik}|\text{Selassie}|\epsilon) = \{ \text{Menelik}, \text{Selassie}, "" \}$. 
String Repetitions

For any string $x$, we may use concatenation to create a string $y$ with as many repetitions of $x$ as we want, by defining repetitions by induction.

\[
x^0 = ""
\]
\[
x^1 = x
\]
\[
x^2 = x.x
\]
\[
\vdots
\]
\[
x^{n+1} = x.x^n = x^n.x
\]
\[
\vdots
\]

Then

\[
x^* = \{x^n \mid n \geq 0\}
\]
String Repetitions Example

Example. Let $x = \text{Selassie}$. Then

\[
x^0 = ""
\]
\[
x^1 = \text{Selassie}
\]
\[
x^2 = \text{SelassieSelassie}
\]
\[
\vdots
\]
\[
x^5 = \text{SelassieSelassieSelassieSelassieSelassieSelassieSelassie}
\]
\[
\vdots
\]

Then $x^*$ is the language consisting of all strings that are finite repetitions of the string $\text{Selassie}$
Language Iteration

The * operator can be extended to languages in the same way. For any language $X$, we may use concatenation to create a another language $Y$ with as many repetitions of the strings in $X$ as we want, by defining repetitions by induction.

$$
X^0 = \"\" \\
X^1 = X \\
X^2 = X.X \\
\vdots \\
X^{n+1} = X.X^n = X^n.X \\
\vdots
$$

Then

$$
X^* = \bigcup_{n \geq 0} X^n
$$
Language Iteration Example

Example Let \( X = \{\text{Mengesha, Gemechis}\} \). Then

\[
X^0 = \{''\}
\]
\[
X^1 = \{\text{Mengesha, Gemechis}\}
\]
\[
X^2 = \{\text{MengeshaMengesha, GemechisMengesha,}
    \text{MengeshaGemechis, GemechisGemechis}\}
\]
\[
X^3 = \{\text{MengeshaMengeshaMengesha,}
    \text{GemechisMengeshaMengesha,}
    \text{MengeshaGemechisMengesha,}
    \text{GemechisGemechisMengesha,}
    \text{MengeshaMengeshaGemechis,}
    \text{GemechisMengeshaGemechis,}
    \text{MengeshaGemechisGemechis,}
    \text{GemechisGemechisGemechis}\}
\]

\[\vdots\]

\[
X^{n+1} = X.X^n
\]

\[\vdots\]
Kleene Closure

Given a regular expression \( r \), \( r^n \) specifies the \( n \)-fold iteration of the language specified by \( r \).

Given any regular expression \( r \), the **Kleene closure** of \( r \), denoted \( r^* \) specifies the language \((\mathcal{L}(r))^*\).

In general

\[
    r^* = r^0 \mid r^1 \mid \ldots \mid r^{n+1} \mid \ldots
\]

denotes an infinite union of languages.

**Question 1.** For what regular expression \( r \) will \( r^* \) specify a finite set?

**Question 2.** How many strings will be in the language specified by \((a \mid b \mid c)^n\)?

**Question 3.** Give an informal description of the language specified by \((a \mid b \mid c)^*\)?

**Question 4.** Give a regular expression which specifies the language \( \{a^k \mid k > 100\} \).
Plus Closure

The **Kleene closure** allows for **zero or more iterations** of a language. The **+-closure** of a language $X$ denoted by $X^+$ and defined as

$$X^+ = \bigcup_{n>0} X^n$$

denotes **one or more iterations** of the language $X$.

Analogously we have that $r^+$ specifies the language $(\mathcal{L}(r))^+$. Notice that for any language $X$, $X^+ = X.X^*$ and hence for any regular expression $r$ we have

$$r^+ = r.r^*$$

We also have the identity

$$r^* = \epsilon \mid r^+$$

**Question 5.** Simplify the expression $r^*.r^*$, i.e. give a simpler regular expression which specifies the same language.

**Question 6.** Simplify the expression $r^+.r^+$. 
Range Specifications

We may specify ranges of various kinds as follows.

- \([a-c] = a \mid b \mid c\). Hence the expression of Question 3 may be specified as \([a-c]^*\).

- Multiple ranges: \([a-c0-3] = [a-c] \mid [0-3]\)

**Question 6.** Try to understand what the regular expression for identifiers really specifies.

**Question 7.** Modify the regular expression so that all identifiers start only with upper-case letters.

**Question 8.** Give regular expressions to specify

- real numbers in *fixed decimal point notation*
- real numbers in *floating point notation*
- real numbers in both *fixed decimal point notation* as well as *floating point notation*.
Notes on bracketing and precedence of operators

In general regular expressions could be *ambiguous* (in the sense that the same expression may be interpreted to refer to different languages. This is especially so in the presence of

- multiple binary operators
- some unary operators used in prefix form while some others are used in post-fix form. The Kleene-closure and plus closure are operators in postfix form. We have not introduced any prefix unary operator in the language of regular expressions.

All expressions may be made unambiguous by specifying them in a fully parenthesised fashion. However, that leads to too many parentheses and is often hard to read. Usually rules for precedence of operators is defined and we may use the parentheses “(“ and “)” to group expressions over-riding the precedence conventions of the language.

For the operators of regular expressions we will use the precedence convention that $|$ has a lower precedence than $.$ and that all unary operators have the highest precedence.

**Example 3.1** The language of arithmetic expressions over numbers uses the “BDMAS” convention that brackets have the highest precedence, followed by division and multiplication and the operations of addition and subtraction have the lowest precedence.

**Example 3.2** The regular expression $r.s|t.u$ is ambiguous because we do not know beforehand whether it represents $(r.s)|(t.u)$ or $r.(s|t).u$ or even various other possibilities. By specifying that the operator $|$ has lower precedence than $.$ we are disambiguating the expression to mean $(r.s)|(t.u)$.

**Example 3.3** The language of arithmetic expressions can also be extended to include the unary post-fix operation in which case an expression such as $-a!$ becomes ambiguous. It could be interpreted to mean either $(-a)!$ or $-(a!)$. In the absence of a well-known convention it is best adopt parenthesisation to disambiguate the expression.
Besides the ambiguity created by multiple binary operators, there are also ambiguities created by the same operator and in deciding in what order two or more occurrences of the same operator need to be evaluated. A classic example is the case of subtraction in arithmetic expressions.

Example 3.4 The arithmetic expression \( a - b - c \), in the absence of any well-defined convention could be interpreted to mean either \((a - b) - c\) or \(a - (b - c)\) and the two interpretations would yield different values in general. The problem does not exist for operators such addition and multiplication on numbers, because these operators are associative. Hence even though \(a + b + c\) may be interpreted in two different ways, both interpretations yield identical values.

Example 3.5 Another non-associative operator in arithmetic which often leaves students confused is the exponentiation operator. Consider the arithmetic expression \(a^b^c\). For \(a = 2, b = 3, c = 4\) is this expression to be interpreted as \(a^{(b^c)}\) or as \((a^b)^c\)?
3.2. Nondeterministic Finite Automata (NFA)

Nondeterministic Finite Automata (NFA)
Nondeterministic Finite Automata

A regular expression is useful in defining a *finite state automaton*. An automaton is a machine (simple program) which can be used to recognize all valid lexical tokens of a language. A **nondeterministic finite automaton (NFA)** $\mathcal{N}$ over a finite alphabet $\mathcal{A}$ consists of

- a finite set $Q$ of states,
- an initial state $q_0 \in Q$,
- a finite subset $F \subseteq Q$ of states called the final states or accepting states, and
- a **transition relation** $\rightarrow \subseteq Q \times (\mathcal{A} \cup \{\varepsilon\}) \times Q$. Equivalently

  $$\rightarrow: Q \times (\mathcal{A} \cup \{\varepsilon\}) \rightarrow 2^Q$$

  is a function that for each source state $q \in Q$ and symbol $a \in \mathcal{A}$ associates a set of target states.
Nondeterminism and Automata

• In general the automaton reads the input string from left to right.
• It reads each input symbol only once and executes a transition to new state.
• The $\varepsilon$ transitions represent going to a new target state without reading any input symbol.
• The NFA may be nondeterministic because of
  – one or more $\varepsilon$ transitions from the same source state different target states,
  – one or more transitions on the same input symbol from one source state to two or more different target states,
  – choice between executing a transition on an input symbol and a transition on $\varepsilon$ (and going to different states).
Acceptance of NFA

• For any alphabet $A$, $A^*$ denotes the set of all (finite-length) strings of symbols from $A$.

• Given a string $x = a_1 a_2 \ldots a_n \in A^*$, an accepting sequence is a sequence of transitions

$$q_0 \xrightarrow{\varepsilon} \cdots \xrightarrow{a_1} \xrightarrow{\varepsilon} \cdots q_1 \xrightarrow{\varepsilon} \cdots \xrightarrow{a_2} \cdots \xrightarrow{\varepsilon} \cdots \xrightarrow{a_n} \xrightarrow{\varepsilon} \cdots q_n$$

where $q_n \in F$ is an accepting state.

• Since the automaton is nondeterministic, it is also possible that there exists another sequence of transitions

$$q_0 \xrightarrow{\varepsilon} \cdots \xrightarrow{a_1} \xrightarrow{\varepsilon} \cdots q_1' \xrightarrow{\varepsilon} \cdots \xrightarrow{a_2} \cdots \xrightarrow{\varepsilon} \cdots \xrightarrow{a_n} \xrightarrow{\varepsilon} \cdots q_n'$$

where $q_n'$ is not a final state.

• The automaton accepts $x$, if there is an accepting sequence for $x$. 
Language of a NFA

- The language **accepted** or **recognized** by a NFA is the set of strings that can be accepted by the NFA.
- $\mathcal{L}(N)$ is the language accepted by the NFA $N$. 
Construction of NFAs

- We show how to construct an NFA to accept a certain language of strings from the regular expression specification of the language.
- The method of construction is by *induction on the structure* of the regular expression. That is, for each regular expression operator, we show how to construct the corresponding automaton assuming that the NFAs corresponding to individual components of expression have already been constructed.
- For any regular expression $r$ the corresponding NFA constructed is denoted $N_r$. Hence for the regular expression $r|s$, we construct the NFA $N_{r|s}$ using the NFAs $N_r$ and $N_s$ as the building blocks.
- Our method requires only one initial state and one final state for each automaton. Hence in the construction of $N_{r|s}$ from $N_r$ and $N_s$, the initial states and the final states of $N_r$ and $N_s$ are not initial or final unless explicitly used in that fashion.
Constructing NFA

- We show the construction only for the most basic operators on regular expressions.
- For any regular expression $r$, we construct a NFA $N_r$ whose initial state is named $r_0$ and final state $r_f$.
- The following symbols show the various components used in the depiction of NFAs.

![Diagram of NFA components](image)
Regular Expressions to NFAs:1

We may also express the automaton in tabular form as follows:

<table>
<thead>
<tr>
<th>$N_a$</th>
<th>Input Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>a</td>
</tr>
<tr>
<td>$a_0$</td>
<td>${a_f}$</td>
</tr>
<tr>
<td>$a_f$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

Notice that all the cells except one have empty targets.
Regular Expressions to NFAs:2

\[ N_\varepsilon \]

<table>
<thead>
<tr>
<th>State</th>
<th>Input Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\varepsilon_0)</td>
<td>(\emptyset) (\emptyset) (\cdots) (\emptyset) ({\varepsilon_f})</td>
</tr>
<tr>
<td>(\varepsilon_f)</td>
<td>(\emptyset) (\emptyset) (\cdots) (\emptyset) (\emptyset)</td>
</tr>
</tbody>
</table>
Regular Expressions to NFAs:3

\[
N_{r|s} \quad \text{Input Symbol}
\]

<table>
<thead>
<tr>
<th>State</th>
<th>a</th>
<th>\cdots</th>
<th>\varepsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r</td>
<td>s_0)</td>
<td>\emptyset</td>
<td>\cdots</td>
</tr>
<tr>
<td>(r_0)</td>
<td>\cdots</td>
<td>\cdots</td>
<td>\cdots</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>(r_f)</td>
<td>\cdots</td>
<td>\cdots</td>
<td>{r</td>
</tr>
<tr>
<td>(s_0)</td>
<td>\cdots</td>
<td>\cdots</td>
<td>\cdots</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>(s_f)</td>
<td>\cdots</td>
<td>\cdots</td>
<td>{r</td>
</tr>
<tr>
<td>(r</td>
<td>s_f)</td>
<td>\emptyset</td>
<td>\cdots</td>
</tr>
</tbody>
</table>
Regular Expressions to NFAs:4

Notice that the initial state of $N_{r,s}$ is $r_0$ and the final state is $s_f$ in this case.
Regular Expressions to NFAs: 5

```
<table>
<thead>
<tr>
<th>( N_r^* )</th>
<th>Input Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>a</td>
</tr>
<tr>
<td>( r^*_0 )</td>
<td>( \emptyset )</td>
</tr>
<tr>
<td>( r_0 )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( {r^<em>_0, r^</em>_f} )</td>
</tr>
<tr>
<td>( r^*_f )</td>
<td>( \emptyset )</td>
</tr>
<tr>
<td>( r^*_f )</td>
<td>( \emptyset )</td>
</tr>
</tbody>
</table>
```
Regular expressions vs. NFAs

• It is obvious that for each regular expression $r$, the corresponding NFA $N_r$ is correct by construction i.e.

$$\mathcal{L}(N_r) = \mathcal{L}(r)$$

• Each regular expression operator
  – adds at most 2 new states and
  – adds at most 4 new transitions

• Every state of each $N_r$ so constructed has
  – either 1 outgoing transition on a symbol from $A$
  – or at most 2 outgoing transitions on $\varepsilon$

• Hence $N_r$ has at most $2|r|$ states and $4|r|$ transitions.
Example

We construct a NFA for the regular expression \((a|b)^*abb\).

- Assume the alphabet \(A = \{a, b\}\).
- We follow the steps of the construction as given in Constructing NFA to Regular Expressions to NFAs:5
- For ease of understanding we use the regular expression itself (subscripted by \(0\) and \(f\) respectively) to name the two new states created by the regular expression operator.
Example:-6

Steps in NFA for \((a|b)^*abb\)
Example:-5

Steps in NFA for \((a|b)^*abb\)
Example:-4

Steps in NFA for \((a|b)^*abb\)
Example:-3

Steps in NFA for \((a|b)^*abb\)
Example:-2

Steps in NFA for $(a|b)^*abb$
Example:- 1

Steps in NFA for \((a|b)^*abb\)
Example-final

Steps in NFA for \((a|b)^*abb\)
Extensions

We have provided constructions for only the most basic operators on regular expressions. Here are some extensions you can attempt

1. Show how to construct a NFA for ranges and multiple ranges of symbols

2. Assuming \( N_r \) is a NFA for the regular expression \( r \), how will you construct the NFA \( N_{r+} \).

3. Certain languages like Perl allow an operator like \( r\{k,n\} \), where

\[
\mathcal{L}(r\{k,n\}) = \bigcup_{k \leq m \leq n} \mathcal{L}(r^m)
\]

Show to construct \( N_{r\{k,n\}} \) given \( N_r \).

4. Consider a new regular expression operator \(^\wedge\) defined by

\[
\mathcal{L}(^\wedge r) = A^* - \mathcal{L}(r)
\]

What is the automaton \( N_{^\wedge r} \) given \( N_r \)?
Lecture 05
Scanning Using NFAs
Scanning and Automata

- **Scanning** is the only phase of the compiler in which every character of the source program is read.
- The scanning phase therefore needs to be defined *accurately* and *efficiently*.
- **Accuracy** is achieved by regular expression specification of the tokens.
- **Efficiency** implies that the input should **not** be read more than once.
Nondeterminism and Token Recognition

- The three kinds of nondeterminism in the NFA construction are depicted in the figure below.

(i) It is difficult to know which $\varepsilon$ transition to pick without reading any further input.

(ii) For two transitions on the same input symbol $a$, it is difficult to know which of them would reach a final state on further input.

(iii) Given an input symbol $a$ and a $\varepsilon$ transition on the current state, it is impossible to decide which one to take without looking at further input.
Nondeterministic Features

- In general it is impossible to recognize tokens in the presence of nondeterminism without backtracking.
- Hence NFAs are not directly useful for scanning because of the presence of nondeterminism.
- The nondeterministic feature of the construction of $N_r$ for any regular expression $r$ is in the $\varepsilon$ transitions.
- The $\varepsilon$ transitions in any automaton refer to the fact that no input character is consumed in the transition.
- Backtracking usually means algorithms involving them are very complex and hence inefficient.
- To avoid backtracking, the automaton should be made deterministic.
From NFA to DFA

• Since the only source of nondeterminism in our construction are the $\varepsilon$, we need to eliminate them without changing the language recognized by the automaton.

• Two consecutive $\varepsilon$ transitions are the same as one. In fact any number of $\varepsilon$ transitions are the same as one. So as a first step we compute all finite sequences of $\varepsilon$ transitions and collapse them into a single $\varepsilon$ transition.

• Two states $q, q'$ are equivalent if there are only $\varepsilon$ transitions between them. This is called the $\varepsilon$-closure of states.
Given a set $T$ of states, then $T_\varepsilon = \varepsilon\text{-closure}(T)$ is the set of states which either belong to $T$ or can be reached from states belonging to $T$ only through a sequence of $\varepsilon$ transitions.

**Algorithm 1 $\varepsilon$-Closure**

Require: $T$ a set of states of the NFA

Ensure: $T_\varepsilon = \varepsilon\text{-closure}(T)$.

1. $U := T$
2. repeat
3. $U_{old} := U$
4. $U := U_{old} \cup \{q' \mid q' \notin U, \exists q \in U_{old} : q \xrightarrow{\varepsilon} q'\}$
5. until $U = U_{old}$
6. $T_\varepsilon = U$
7. return $T_\varepsilon$
Analysis of $\varepsilon$-Closure

- $U$ can only grow in size through each iteration
- The set $U$ cannot grow beyond the total set of states $Q$ which is finite. Hence the algorithm always terminates for any NFA $N$.
- Time complexity: $O(|Q|)$. 
Recognition using NFA

The following algorithm may be used to recognize a string using a NFA.

**Algorithm 2 Recognition using NFA**

**Require:** A string $x \in A^*$.

**Ensure:** Boolean $S := \varepsilon\text{-}\text{CLOSURE}(\{q_0\})$.

$a := \text{nextchar}(x)$

**while** $a \neq \text{end\_of\_string}$ **do**

$S := \varepsilon\text{-}\text{CLOSURE}(S \xrightarrow{a})$

$a := \text{nextchar}(x)$

**end while**

return $S \cap F \neq \emptyset$

In the above algorithm we extend our notation for targets of transitions to include sets of sources. Thus

$$S \xrightarrow{a} = \{q' \mid \exists q \in S : q \xrightarrow{a} q'\}$$
Analysis of Recognition using NFA

• Even if $\varepsilon$-closure is computed as a call from within the algorithm, the time taken to recognize a string is bounded by $O(|x| \cdot |Q_{N_r}|)$ where $|Q_{N_r}|$ is the number of states in $N_r$.

• The space required for the automaton is at most $O(|r|)$.

• Given that $\varepsilon$-closure of each state can be pre-computed knowing the NFA, the recognition algorithm can run in time linear in the length of the input string $x$ i.e. $O(|x|)$.

• Knowing that the above algorithm is deterministic once $\varepsilon$-closures are pre-computed one may then work towards a Deterministic automaton to reduce the space required.
Lecture 06
Conversion of NFAs to DFAs
Deterministic Finite Automata

- A deterministic finite automaton (DFA) is a NFA in which
  1. there are no transitions on $\varepsilon$ and
  2. $\rightarrow$ yields a *at most one* target state for each source state and symbol from $A$ i.e. $\rightarrow$ is no longer a relation but a *function* of the form

$$\rightarrow: Q \times A \rightarrow Q$$

- Clearly if every regular expression had a DFA which accepts the same language, all backtracking could be avoided.
Transition Tables of NFAs

We may think of a finite-state automaton as being defined by a 2-dimensional table of size $|Q| \times |A|$ in which for each state and each letter of the alphabet there is a set of possible target states defined. In the case of a non-deterministic automaton,

1. for each state there could be $\varepsilon$ transitions to
   (a) a set consisting of a single state or
   (b) a set consisting of more than one state.

2. for each state $q$ and letter $a$, there could be
   (a) an empty set of states or
   (b) a set consisting of a single state or
   (c) a set consisting of more than one state or
Transition Tables of DFAs

In the case of a deterministic automaton

1. there are no $\varepsilon$ transitions, and

2. for each state $q$ and letter $a$
   
   (a) either there is no transition
   
   (b) or there is a transition to a unique state $q'$.

The recognition problem for the same language of strings becomes simpler and would work faster (it would have no back-tracking) if the NFA could be converted into a DFA accepting the same language.
NFA to DFA

Let \( N = \langle Q_N, A \cup \{\varepsilon\}, s_N, F_N, \rightarrow_N \rangle \) be a NFA with

- \( Q_N \) the set of states of the NFA
- \( A \) the alphabet
- \( s_N \in Q_N \) the start state of the NFA
- \( F_N \) the final or accepting states of the NFA and
- \( \rightarrow_N \subseteq Q_N \times A \times Q_N \) the transition relation.

We would like to construct a DFA \( D = \langle Q_D, A, s_D, F_D, \rightarrow_D \rangle \) where

- \( Q_D \) the set of states of the DFA
- \( A \) the alphabet
- \( s_D \in Q_D \) the start state of the DFA
- \( F_D \) the final or accepting states of the DFA and
- \( \rightarrow_D \subseteq Q_D \times A \times Q_D \) the transition function of the DFA.

We would like \( \mathcal{L}(N) = \mathcal{L}(D) \)
The Subset Construction

- The $\varepsilon$-closure of each NFA state is a set of NFA states with “similar” behaviour, since they make their transitions on the same input symbols though with different numbers of $\varepsilon$s.

- Each state of the DFA refers to a subset of states of the NFA which exhibit “similar” behaviour. Similarity of behaviour refers to the fact that they accept the same input symbols. The behaviour of two different NFA states may not be “identical” because they may have different numbers of $\varepsilon$ transitions for the same input symbol.

- A major source of non-determinism is the presence of $\varepsilon$ transitions. The use of $\varepsilon$-Closure creates a cluster of similar states.

- Since the notion of acceptance of a string by an automaton, implies finding an accepting sequence even though there may be other non-accepting sequences, the non-accepting sequences may be ignored and those non-accepting states may be clustered with the accepting states of the NFA. So two different states reachable by the same sequence of symbols may be also though to be similar.
Algorithm 3 Construction of DFA from NFA

Require: NFA $N = \langle Q_N, A \cup \{\varepsilon\}, s_N, F_N, \rightarrow_N \rangle$

Ensure: DFA $D = \langle Q_D, A, s_D, F_D, \rightarrow_D \rangle$ with $\mathcal{L}(N) = \mathcal{L}(D)$

1. $s_D := \varepsilon$-CLOSURE$\{s_N\}$;
2. $Q_D := \{s_D\}$; $F_D := \emptyset$; $\rightarrow_D := \emptyset$
3. $U := \{s_D\}$ \{$U$ is the set of unvisited states of the DFA\}
4. while $U \neq \emptyset$ do
5. Choose any $q_D \in U$; $U := U - \{q_D\}$
6. for all $a \in A$ do
7. $q'_D := \varepsilon$-CLOSURE$\{q_D \xrightarrow{a} \rightarrow_N \}$ \{Note: $q_D \subseteq Q_N$\}
8. $\rightarrow_D := \rightarrow_D \cup \{q_D \xrightarrow{a} q'_D\}$
9. if $q'_D \cap F_N \neq \emptyset$ then
10. $F_D := F_D \cup \{q'_D\}$
11. end if
12. if $q'_D \notin Q_D$ then
13. $Q_D := Q_D \cup \{q'_D\}$
14. $U := U \cup \{q'_D\}$
15. end if
16. end for
17. end while
Example-NFA

Consider the NFA constructed for the regular expression \((a|b)^*abb\).

and apply the NFA to DFA construction algorithm
Determinising

\[ N(a|b)^{abb} \]

\[ D(a|b)^{abb} \]

\[ EC_0 = \varepsilon - \text{CLOSURE}(0) = \{0, 1, 2, 3, 7\} \]

2 \xrightarrow{a}_N 4 \text{ and } 7 \xrightarrow{a}_N 8. \text{ So } EC_0 \xrightarrow{a}_D \varepsilon - \text{CLOSURE}(4, 8) = EC_{4,8}. \text{ Similarly}

\[ EC_0 \xrightarrow{b}_D \varepsilon - \text{CLOSURE}(5) = EC_5 \]

\[ EC_{4,8} = \varepsilon - \text{CLOSURE}(4, 8) = \{4, 6, 7, 1, 2, 3\} \]

\[ EC_5 = \varepsilon - \text{CLOSURE}(5) = \{5, 6, 7, 1, 2, 3\} \]

\[ EC_5 \xrightarrow{a}_D \varepsilon - \text{CLOSURE}(4, 8) = EC_{4,8} \text{ and } EC_5 \xrightarrow{b}_D \varepsilon - \text{CLOSURE}(5) \]

\[ EC_{4,8} \xrightarrow{a}_D \varepsilon - \text{CLOSURE}(4, 8) = EC_{4,8} \text{ and } EC_{4,8} \xrightarrow{b}_D \varepsilon - \text{CLOSURE}(5, 9) = EC_{5,9} \]

\[ EC_{5,9} = \varepsilon - \text{CLOSURE}(5, 9) = \{5, 6, 7, 1, 2, 3, 9\} \]

\[ EC_{5,9} \xrightarrow{a}_D \varepsilon - \text{CLOSURE}(4, 8) = EC_{4,8} \text{ and } EC_{5,9} \xrightarrow{b}_D \varepsilon - \text{CLOSURE}(5, 10) = EC_{5,10} \]

\[ EC_{5,10} = \varepsilon - \text{CLOSURE}(5, 10) = \{5, 6, 7, 1, 2, 3, 10\} \]

\[ EC_{5,10} \xrightarrow{a}_D \varepsilon - \text{CLOSURE}(4, 8) \text{ and } EC_{5,10} \xrightarrow{b}_D \varepsilon - \text{CLOSURE}(5) \]
Final DFA

\[ D_{(a|b)^*abb} \]

| State   | Input Symbol | \[ D_{(a|b)^*abb} \] |
|---------|--------------|-----------------------|
| \(EC_0\) |              | \(EC_{4,8}\), \(EC_5\) |
| \(EC_{4,8}\) |              | \(EC_{4,8}\), \(EC_{5,9}\) |
| \(EC_5\) |              | \(EC_{4,8}\), \(EC_5\) |
| \(EC_{5,9}\) |              | \(EC_{4,8}\), \(EC_{5,10}\) |
| \(EC_{5,10}\) |              | \(EC_{4,8}\), \(EC_5,9\) |
Scanning: 5

The Big Picture
Lexical Analysis: Problems

1. Write a regular expression to specify all numbers in binary form that are multiples of 4.

2. Write regular expressions to specify all numbers in binary form that are not multiples of 4.

3. Each comment in the C language
   - begins with the characters “/ /” and ends with the newline character, or
   - begins with the characters “/ *” and ends with “* /” and may run across several lines.

   (a) Write a regular expression to recognize comments in the C language.
   (b) Transform the regular expression into a NFA.
   (c) Transform the NFA into a DFA.
   (d) Explain why most programming languages do not allow nested comments.

   (e) **modified C comments.** If the character sequences “/ /”, “/ *” and “* /” are allowed to appear in 'quoted' form as “/ /”, “/ *” and “* /” respectively within a C comment, then give
      - a modified regular expression for C comments
      - a NFA for these modified C comments
      - a corresponding DFA for modified C comments

4. Many systems such as Windows XP and Linux recognize commands, filenames and folder names by their shortest unique prefix. Hence given the 3 commands `chmod`, `chgrp` and `chown`, their shortest unique prefixes are respectively `chm`, `chg` and `cho`. A user can type the shortest unique prefix of the command and the system will automatically complete it for him/her.

   (a) Draw a DFA which recognizes all prefixes that are at least as long as the shortest unique prefix of each of the above commands.
   (b) Suppose the set of commands also includes two more commands `cmp` and `cmpdir`, state how you will include such commands also in your DFA where one command is a prefix of another.
4. Parsing or Syntax Analysis

4.1. Grammars

Formal languages: Definition, Recognition, Generation

There are three different processes used in dealing with a formal language.

**Definition** Regular expressions is another formal language used to represent a formal language of tokens.

**Recognition** Automata are the standard mechanism used to recognize words/phrases of a formal language. An automaton is used to determine whether a given word/phrase is a member of the formal language defined in some other way.

**Generation** Grammars are used to define the generation of the words/phrases of a formal language.
Parsing Or Syntax Analysis
Non-regular language

Consider the following two languages over an alphabet $A = \{a, b\}$.

$$R = \{a^n b^n | n < 100\}$$
$$P = \{a^n b^n | n > 0\}$$

- $R$ may be finitely represented by a regular expression (even though the actual expression is very long).
- However, $P$ cannot actually be represented by a regular expression.
- A regular expression is not powerful enough to represent languages which require parenthesis matching to arbitrary depths.
- All high level programming languages require an underlying language of expressions which require parentheses to be nested and matched to arbitrary depth.
4.2. Context-Free Grammars

Grammars

Definition 4.1 A grammar $G = \langle N, T, P, S \rangle$ consists of

- a set $N$ of nonterminal symbols, or variables,
- a start symbol $S \in N$,
- a set $T$ of terminal symbols or the alphabet,
- a set $P$ of productions or rewrite rules where each rule is of the form $\alpha \rightarrow \beta$ for $\alpha, \beta \in (N \cup T)^*$

Definition 4.2 Given a grammar $G = \langle N, T, P, S \rangle$, any $\alpha \in (N \cup T)^*$ is called a sentential form. Any $x \in T^*$ is called a sentence.

Note. Every sentence is also a sentential form. We use small greek letters ($\alpha, \beta$ etc.) to denote sentential forms and small roman letters at the end of the roman alphabet ($x, y, z$ etc.) to denote sentences.
Context-Free Grammars: Definition

Definition 4.3 A grammar $G = \langle N, T, P, S \rangle$ is called context-free if

- each production is of the form $X \rightarrow \alpha$, where
  - $X \in N$ is a nonterminal and
  - $\alpha \in (N \cup T)^*$ is a sentential form.
- The production is terminal if $\alpha$ is a sentence.
CFG: Example 1

\[ G = \langle \{S\}, \{a, b\}, P, S \rangle, \]  

where \( S \rightarrow ab \) and \( S \rightarrow aSb \) are the only productions in \( P \).

Derivations look like this:

- \[ S \Rightarrow ab \]

- \[ S \Rightarrow aSb \Rightarrow aabb \]

- \[ S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaabbb \]

- \[ S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \]

The first three derivations are complete while the last one is partial
Derivations

Definition 4.4 A (partial) derivation (of length \( n \in \mathbb{N} \)) in a context-free grammar is a finite sequence of the form

\[
\alpha_0 \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \cdots \alpha_n
\]  

where each \( \alpha_i \in (\mathbb{N} \cup T)^* \) (\( 0 \leq i \leq n \)) is a sentential form where \( \alpha_0 = S \) and \( \alpha_{i+1} \) is obtained by applying a production rule to a non-terminal symbol in \( \alpha_i \) for \( 0 \leq i < n \).

Notation. \( S \Rightarrow^* \alpha \) denotes that there exists a derivation of \( \alpha \) from \( S \).

Definition 4.5 A derivation is complete if \( \alpha_n \in T^* \). Then \( \alpha_n \) is said to have been generated by the grammar.
Language Generation

Definition 4.6

Definition 4.7 *The language generated by a context-free grammar is the set of sentences that can be generated.*

Example 4.8 $\mathcal{L}(G)$, the language generated by $G$ is $\{a^n b^n | n > 0\}$.

Actually can be proved by induction on the length and structure of derivations.
Regular Grammars

Definition 4.9 A production rule of a context-free grammar is

Right Linear: if it is of the form $X \rightarrow a$ or $X \rightarrow aY$

Left Linear: if it is of the form $X \rightarrow a$ or $X \rightarrow Ya$

where $a \in T$ and $X, Y \in N$.

Definition 4.10 A regular grammar is a context-free grammar whose productions are either only right linear or only left linear.
CFG: Empty word

\[ G = \langle \{ S \}, \{ a, b \}, P, S \rangle, \text{ where } S \rightarrow SS \mid aSb \mid \varepsilon \]
generates all sequences of matching nested parentheses, including the empty word \( \varepsilon \).

A leftmost derivation might look like this:

\[
S \Rightarrow SS \Rightarrow SSS \Rightarrow SS \Rightarrow aSbS \Rightarrow abS \Rightarrow abaSb \ldots
\]

A rightmost derivation might look like this:

\[
S \Rightarrow SS \Rightarrow SSS \Rightarrow SS \Rightarrow SaSb \Rightarrow Sab \Rightarrow aSbab \ldots
\]

Other derivations might look like \textit{God alone knows what!}

\[
S \Rightarrow SS \Rightarrow SSS \Rightarrow SS \Rightarrow \ldots
\]

Could be quite confusing!
CFG: Derivation trees 1

Derivation sequences

• put an artificial order in which productions are fired.
• instead look at trees of derivations in which we may think of productions as being fired in parallel.
• There is then no highlighting in red to determine which copy of a non-terminal was used to get the next member of the sequence.
• Of course, generation of the empty word $\varepsilon$ must be shown explicitly in the tree.
CFG: Derivation trees 2

Derivation tree of \textit{abaabb}
Another Derivation tree of \textit{abaabb}
Yet another Derivation tree of $abaabb$
4.3. Ambiguity

Ambiguity Disambiguation
Ambiguity: 1

\[ G_1 = \langle \{ E, I, C \}, \{ y, z, 4, *, + \}, P_1, \{ E \} \rangle \] where \( P_1 \) consists of the following productions.

\[
E \rightarrow I \mid C \mid E + E \mid E * E \\
I \rightarrow y \mid z \\
C \rightarrow 4
\]

Consider the sentence \( y + 4 * z \).
Ambiguity: 2

\[ G_1 = \langle \{E, I, C\}, \{y, z, 4, *, +\}, P_1, \{E\} \rangle \] where \( P_1 \) consists of the following productions.

\[
\begin{align*}
E & \to I \mid C \mid E + E \mid E \ast E \\
I & \to y \mid z \\
C & \to 4
\end{align*}
\]

Consider the sentence \( y + 4 \ast z \).
Ambiguity: 3

$$G_1 = \langle \{E, I, C\}, \{y, z, 4, *, +\}, P_1, \{E\} \rangle$$ where $$P_1$$ consists of the following productions.

$$E \rightarrow I \mid C \mid E+E \mid E*E$$

$$I \rightarrow y \mid z$$

$$C \rightarrow 4$$

Consider the sentence $$y + 4 * z$$.
Ambiguity: 4

\[ G_1 = \langle \{E, I, C\}, \{y, z, 4, \ast, +\}, P_1, \{E\} \rangle \]

where \( P_1 \) consists of the following productions.

\[
E \rightarrow I \mid C \mid E+E \mid E*E \\
I \rightarrow y \mid z \\
C \rightarrow 4
\]

Consider the sentence \( y + 4 * z \).
Ambiguity: 5

\[ G_1 = \langle \{E, I, C\}, \{y, z, 4, *, +\}, P_1, \{E\} \rangle \]
where \( P_1 \) consists of the following productions.

\[
E \rightarrow I \mid C \mid E + E \mid E * E \\
I \rightarrow y \mid z \\
C \rightarrow 4
\]

Consider the sentence \( y + 4 * z \).
Left-most Derivation 1

Left-most derivation of $y+4*z$ corresponding to the first derivation tree.

$$
E \quad \Rightarrow \\
E+E \quad \Rightarrow \\
I+E \quad \Rightarrow \\
y+E \quad \Rightarrow \\
y+E*E \quad \Rightarrow \\
y+C*E \quad \Rightarrow \\
y+4*E \quad \Rightarrow \\
y+4*I \quad \Rightarrow \\
y + 4 * z
$$
Left-most Derivation 2

Left-most derivation of $y+4*z$ corresponding to the second derivation tree.

\[
E \Rightarrow E*E \\
E*E \Rightarrow E+E*E \\
E+E*E \Rightarrow I+E*E \\
I+E*E \Rightarrow y+E*E \\
y+E*E \Rightarrow y+C*E \\
y+C*E \Rightarrow y+4*E \\
y+4*E \Rightarrow y+4*I \\
y+4*I \Rightarrow y+4*z
\]
Right-most Derivation 1

Right-most derivation of $y+4*z$ corresponding to the first derivation tree.

\[
\begin{align*}
E & \Rightarrow \\
E+E & \Rightarrow \\
E+E*E & \Rightarrow \\
E+E*I & \Rightarrow \\
E+E*z & \Rightarrow \\
E+C*z & \Rightarrow \\
E+4*z & \Rightarrow \\
I+4*z & \Rightarrow \\
y + 4*z &
\end{align*}
\]
Right-most Derivation 2

Right-most derivation of $y+4*z$ corresponding to the second derivation tree.

$$
E \quad \Rightarrow \\
E*E \quad \Rightarrow \\
E*I \quad \Rightarrow \\
E*z \quad \Rightarrow \\
E+E*z \quad \Rightarrow \\
E+C*z \quad \Rightarrow \\
E+4*z \quad \Rightarrow \\
I+4*z \quad \Rightarrow \\
y + 4*z
$$
Characterizing Ambiguity

The following statements are equivalent.

- A CFG is *ambiguous* if some sentence it generates has more than one *derivation tree*
- A CFG is *ambiguous* if there is a some sentence it generates with more than one *left-most derivation*
- A CFG is *ambiguous* if there is a some sentence it generates with more than one *right-most derivation*
Disambiguation

The only way to remove ambiguity (without changing the language generated) is to change the grammar by introducing some more non-terminal symbols and changing the production rules. Consider the grammar $G'_1 = \langle N', \{y, z, 4, *, +\}, P', \{E\} \rangle$ where $N' = N \cup \{T, F\}$ with the following production rules $P'$.

$$
E \to E + T \mid T \\
T \to T \ast F \mid F \\
F \to I \mid C \\
I \to y \mid z \\
C \to 4
$$

and compare it with the grammar $G_1$.
Left-most Derivation 1’

The left-most derivation of $y+4*z$ is then as follows.

\[
E \quad \Rightarrow \\
E+T \quad \Rightarrow \\
I+T \quad \Rightarrow \\
y+T \quad \Rightarrow \\
y+T*F \quad \Rightarrow \\
y+T*F \quad \Rightarrow \\
y+F*F \quad \Rightarrow \\
y+C*F \quad \Rightarrow \\
y+4*F \quad \Rightarrow \\
y+4*I \quad \Rightarrow \\
y + 4 * z
\]
Left-most Derivations

Compare it with the Left-most Derivation 1.

\[ G_1. \quad E \Rightarrow E + E \Rightarrow I + E \Rightarrow y + E \Rightarrow y + E * E \Rightarrow y + C * E \Rightarrow y + 4 * E \Rightarrow y + 4 * I \Rightarrow y + 4 * z \]

\[ G'_1. \quad E \Rightarrow E + T \Rightarrow I + T \Rightarrow y + T \Rightarrow y + T * F \Rightarrow y + T * F \Rightarrow y + F * F \Rightarrow y + C * F \Rightarrow y + 4 * F \Rightarrow y + 4 * I \Rightarrow y + 4 * z \]

There is no derivation in \( G'_1 \) corresponding to Left-most Derivation 2 (Why not?).
Right-most Derivation 1’

Right-most derivation of \( y+4*z \) corresponding to the *first* derivation tree.

\[
\begin{align*}
E & \Rightarrow \\
E+T & \Rightarrow \\
E+T*F & \Rightarrow \\
E+T*I & \Rightarrow \\
E+T*z & \Rightarrow \\
E+C*z & \Rightarrow \\
E+4*z & \Rightarrow \\
F+4*z & \Rightarrow \\
I+4*z & \Rightarrow \\
+z* & \Rightarrow \\
y+4*z & \\
\end{align*}
\]

Compare it with the **Right-most Derivation 1**.

There is no derivation corresponding to **Right-most Derivation 2**.
Disambiguation by Parenthesization

Another method of disambiguating a language is to change the language generated, by introducing suitable bracketing mechanisms.

**Example 4.11** Compare the following fully parenthesized grammar $G_2$ (which has the extra terminal symbols ( and )) with the grammar $G_1$ without parentheses

$$
E \rightarrow I \mid C \mid (E+E) \mid (E\ast E) \\
I \rightarrow y \mid z \\
C \rightarrow 4
$$

Though unambiguous, the language defined by this grammar is different from that of the original grammar without parentheses.
Associativity and Precedence

The grammar $G'_1$ implements

**Precedence.** $\ast$ has higher precedence than $+$.  

**Associativity.** $\ast$ and $+$ are both left associative operators.
Context-Free Grammars: Problems

1. Two context-free grammars are considered equivalent if they generate the same language. Prove that $G_1$ and $G'_1$ are equivalent.

2. Palindromes. A palindrome is a string that is equal to its reverse i.e. it is the same when read backwards (e.g. aabbaa and abaabaaba are both palindromes). Design a grammar for generating all palindromes over the terminal symbols a and b.

3. Matching brackets.
   
   (a) Design a context-free grammar to generate sequences of matching brackets when the set of terminals consists of three pairs of brackets {, [, ], {, }}.
   
   (b) If your grammar is ambiguous give two rightmost derivations of the same string and draw the two derivation trees. Explain how you will modify the grammar to make it unambiguous.
   
   (c) If your grammar is not ambiguous prove that it is not ambiguous.

4. Design an unambiguous grammar for the expression language on integers consisting of expressions made up of operators +, -, *, /, % and the bracketing symbols ( and ), assuming the usual rules of precedence among operators that you have learned in school.

5. Modify the above grammar to include the exponentiation operator $^\ast$ which has a higher precedence than the other operators and is right-associative.

6. How will you modify the grammar above to include the unary minus operator $-$ where the unary minus has a higher precedence than other operators?

7. The language specified by a regular expression can also be generated by a context-free grammar.
   
   (a) Design a context-free grammar to generate all floating-point numbers allowed by the C language.
   
   (b) Design a context-free grammar to generate all numbers in binary form that are not multiples of 4.
   
   (c) Write a regular expression to specify all numbers in binary form that are multiples of of 3.

8. Prove that the $G'_1$ is indeed unambiguous.

9. Prove that the grammar of fully parenthesized expressions is unambiguous.

10. Explain how the grammar $G'_1$ implements left associativity and precedence.
4.4. Shift-Reduce Parsing

Introduction to Parsing
Overview of Parsing

Since

• parsing requires the checking whether a given token stream conforms to the rules of the grammar and

• since a context-free grammar may generate an infinite number of different strings

any parsing method should be guided by the given input (token) string, so that a deterministic strategy may be evolved.
Parsing Methods

Two kinds of parsing methods

**Top-down parsing** Try to **generate** the given input sentence from the start symbol of the grammar by applying the production rules.

**Bottom-up parsing** Try to **reduce** the given input sentence to the start symbol by applying the rules in **reverse**

In general top-down parsing requires long *look-aheads* in order to do a deterministic guess from the given input token stream. On the other hand bottom-up parsing yields better results and can be automated by software tools.
Reverse of Right-most Derivations

The result of a Bottom-Up Parsing technique is usually to produce a reverse of the right-most derivation of a sentence.

Example For the ambiguous grammar $G_1$ and corresponding to the right-most derivation 2 we get

\[
\begin{align*}
y + 4 \ast z & \Leftarrow \\
I + 4 \ast z & \Leftarrow \\
E + 4 \ast z & \Leftarrow \\
E + C \ast z & \Leftarrow \\
E + E \ast z & \Leftarrow \\
E \ast z & \Leftarrow \\
E \ast I & \Leftarrow \\
E \ast E & \Leftarrow \\
E & \Leftarrow 
\end{align*}
\]
Bottom-Up Parsing Strategy

The main problem is to match parentheses of arbitrary nesting depths. This requires a stack data structure to do the parsing so that unbounded nested parentheses and varieties of brackets may be matched. Our basic parsing strategy is going to be based on a technique called *shift-reduce* parsing.

**shift.** Refers to moving the next token from the input token stream into a *parsing* stack.

**reduce.** Refers to applying a production rule in reverse i.e. given a production $X \rightarrow \alpha$ we reduce any occurrence of $\alpha$ in the parsing stack to $X$. 
r1. $E \rightarrow E \, T$

r2. $E \rightarrow T$

r3. $T \rightarrow T \, D$

r4. $T \rightarrow D$

r5. $D \rightarrow a \mid b \mid (E)$

$$a \rightarrow a \, / \, b$$
Principle:
Reduce whenever possible.
Shift only when reduce is impossible.
Parsing: 3

r1. $E \rightarrow E \cdot T$
r2. $E \rightarrow T$
r3. $T \rightarrow T \downarrow D$
r4. $T \rightarrow D$
r5. $D \rightarrow \textbf{a} \mid \textbf{b} \mid ( \textbf{E} )$

Reduce by r4
r1. $E \rightarrow E \cdot T$
r2. $E \rightarrow T$
r3. $T \rightarrow T \uparrow D$
r4. $T \rightarrow D$
r5. $D \rightarrow a \mid b \mid (E)$

Reduce by r2
Parsing: 5

r1. E → E − T
r2. E → T
r3. T → T (/) D
r4. T → D
r5. D → a | b | () E)

a / b

Shift

E
Parsing: 6

\[
\begin{align*}
&\text{r1. } E \rightarrow E \cdot T \\
&\text{r2. } E \rightarrow T \\
&\text{r3. } T \rightarrow T \bigcup D \\
&\text{r4. } T \rightarrow D \\
&\text{r5. } D \rightarrow a \bigcup b \bigcup (E) \\
\end{align*}
\]
Parsing: 7

r1. \[ E \rightarrow E - T \]
r2. \[ E \rightarrow T \]
r3. \[ T \rightarrow T \mid D \]
r4. \[ T \rightarrow D \]
r5. \[ D \rightarrow a \mid b \mid ( E ) \]

Reduce by r5
Parsing: 8

r1. \[ E \rightarrow E \cdot T \]
r2. \[ E \rightarrow T \]
r3. \[ T \rightarrow T \cdot D \]
r4. \[ T \rightarrow D \]
r5. \[ D \rightarrow a \mid b \mid (E) \]

Reduce by r4
Parsing: 8a

r1. $E \rightarrow E ~ T$

r2. $E \rightarrow T$

r3. $T \rightarrow T \lor ~ D$

r4. $T \rightarrow D$

r5. $D \rightarrow a ~ \mid ~ b ~ \mid ~ ( ~ E ~ )$

Reduce by $r4$
Parsing: 9a

r1. \[ E \rightarrow E - T \]
r2. \[ E \rightarrow T \]
r3. \[ T \rightarrow T \cup D \]
r4. \[ T \rightarrow D \]
r5. \[ D \rightarrow a \mid b \mid (E) \]

Reduce by r1
Parsing: 10a

r1. E → E T
r2. E → T
r3. T → T D
r4. T → D
r5. D → a | b | ( E )

Shift
### Parsing: 11a

<table>
<thead>
<tr>
<th>Rule</th>
<th>Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>E → E T</td>
</tr>
<tr>
<td>r2</td>
<td>E → T</td>
</tr>
<tr>
<td>r3</td>
<td>T → T D</td>
</tr>
<tr>
<td>r4</td>
<td>T → D</td>
</tr>
<tr>
<td>r5</td>
<td>D → a</td>
</tr>
</tbody>
</table>

#### Diagram:

```
  E → a | b | ( E )
     ↓  Shift
      E
  b
     /    |
     D
```

---

**QXIT**
Parsing: 12a

r1. \( E \rightarrow E \downarrow T \)

r2. \( E \rightarrow T \)

r3. \( T \rightarrow T \downarrow D \)

r4. \( T \rightarrow D \)

r5. \( D \rightarrow a \mid b \mid (E) \)

Reduce by r5
Parsing: 13a

r1. $E \rightarrow E - T$

r2 $E \rightarrow T$

r3 $T \rightarrow T U D$

r4 $T \rightarrow D$

r5 $D \rightarrow a \mid b \mid (E)$

Reduce by r4
Parsing: 14a

r1. $E \rightarrow E \cdot T$

r2. $E \rightarrow T$

r3. $T \rightarrow T \downarrow D$

r4. $T \rightarrow D$

r5. $D \rightarrow a \mid b \mid (E)$

Get back!

Reduce by r2

Stuck!
Parsing: 14b

r1. \( E \rightarrow E \cdot T \)

r2. \( E \rightarrow T \)

r3. \( T \rightarrow T \cdot D \)

r4. \( T \rightarrow D \)

r5. \( D \rightarrow a \mid b \mid (E) \)

Get back!

Reduce by r2
Parsing: 13b

r1. $E \rightarrow E \cdot T$

r2. $E \rightarrow T$

r3. $T \rightarrow T \cup D$

r4. $T \rightarrow D$

r5. $D \rightarrow a \mid b \mid ( \ E \ )$

Get back!

Reduce by r4
Parsing: 12b

r1. $E \rightarrow E + T$

r2. $E \rightarrow T$

r3. $T \rightarrow T ( D$

r4. $T \rightarrow D$

r5. $D \rightarrow a \mid b \mid ( E )$

Get back!

Reduce by r5
Parsing: 11b

r1.  $E \rightarrow E \cdot T$

r2.  $E \rightarrow T$

r3.  $T \rightarrow T \text{ (a)} D$

r4.  $T \rightarrow D$

r5.  $D \rightarrow \text{ (b)} \mid (\text{ (E) }$

Get back! Shift
Parsing: 10b

r1. \( E \rightarrow E \rightarrow T \)

r2 \( E \rightarrow T \)

r3 \( T \rightarrow T \rightarrow D \)

r4 \( T \rightarrow D \)

r5 \( D \rightarrow a \mid b \mid (E) \)

Get back! Shift
Parsing: 9b

r1. E → E T
r2. E → T
r3. T → T D
r4. T → D
r5. D → a | b | ( E )

Reduce by r1

Get back to where you once belonged!
Parsing: 8b

Principle:

Reduce whenever possible, but depending upon lookahead.

Shift instead of reduce here!
 Parsing: 8

r1. E → E - T
r2. E → T
r3. T → T / D
r4. T → D
r5. D → a | b | ( E )

Reduce by r4
Parsing: 9

r1. E → E T
r2. E → T
r3. T → T D
r4. T → D
r5. D → a | b | ( E )
Parsing: 10

r1. $E \rightarrow E - T$

r2 $E \rightarrow T$

r3 $T \rightarrow T \cup D$

r4 $T \rightarrow D$

r5 $D \rightarrow a | b | (E)$

Shift

\[
\begin{array}{c}
\text{b} \\
\text{/} \\
T \\
- \\
E
\end{array}
\]
 Parsing: 11

r1. $E \rightarrow E - T$

r2. $E \rightarrow T$

r3. $T \rightarrow T \downarrow D$

r4. $T \rightarrow D$

r5. $D \rightarrow a | b | ( E )$

Reduce by r5
Parsing: 12

r1. $E \rightarrow E - T$
r2. $E \rightarrow T$
r3. $T \rightarrow T \lor D$
r4. $T \rightarrow D$
r5. $D \rightarrow a \mid b \mid (E)$

Reduce by r3
Parsing: 13

r1. $E \rightarrow E \cdot T$

r2. $E \rightarrow T$

r3. $T \rightarrow T \cdot D$

r4. $T \rightarrow D$

r5. $D \rightarrow a \mid b \mid (E)$

Reduce by r1
4.5. Bottom-Up Parsing

Bottom-Up Parsing
Parse Trees: 0

r1. $E \rightarrow E - T$
r2. $E \rightarrow T$
r3. $T \rightarrow T / D$

r4. $T \rightarrow D$

r5. $D \rightarrow a \mid b \mid (E)$

shift-reduce parsing: 0
Parse Trees: 1

r1. $E \rightarrow E + T$

r2. $E \rightarrow T$

r3. $T \rightarrow T \cdot D$

r4. $T \rightarrow D$

r5. $D \rightarrow a \mid b \mid (E)$

shift-reduce parsing: 1
Parse Trees: 2

r1. $E \rightarrow E \ E$ $T$

r2. $E \rightarrow T$

r3. $T \rightarrow T \ D$

r4. $T \rightarrow D$

r5. $D \rightarrow a \mid b \mid (E)$

shift-reduce parsing: 2
Parse Trees: 3

r1. $E \rightarrow E - T$

r2. $E \rightarrow T$

r3. $T \rightarrow T / D$

r4. $T \rightarrow D$

r5. $D \rightarrow a | b | (E)$

shift-reduce parsing: 3
Parse Trees: 3a

\[
\begin{align*}
  &r1. \quad E \rightarrow E - T \\
  &r2. \quad E \rightarrow T \\
  &r3. \quad T \rightarrow T / D \\
  &r4. \quad T \rightarrow D \\
  &r5. \quad D \rightarrow a \mid b \mid (E)
\end{align*}
\]
Parse Trees: 3b

r1. E → E T
r2. E → T
r3. T → T | D
   r4. T → D
   r5. D → a | b | ( E )

shift-reduce parsing
Parse Trees: 4

r1. E → E − T
r2. E → T
r3. T → T ( D
r4. T → D
r5. D → a | b | ( E )
Parse Trees: 5

r1. $E \rightarrow E - T$

r2. $E \rightarrow T$

r3. $T \rightarrow T \lor D$

r4. $T \rightarrow D$

r5. $D \rightarrow a \mid b \mid (E)$

shift-reduce parsing
Parse Trees: 5a

r1. $E \rightarrow E - T$

r2. $E \rightarrow T$

r3. $T \rightarrow T / D$

r4. $T \rightarrow D$

r5. $D \rightarrow a | b | (E)$

shift-reduce parsing
Parse Trees: 5b

r1. $E \rightarrow E - T$

r2. $E \rightarrow T$

r3. $T \rightarrow T / D$

r4. $T \rightarrow D$

r5. $D \rightarrow a | b | ( E )$

shift-reduce parsing
Parse Trees: 6

r1. \( E \rightarrow E - T \)
r2. \( E \rightarrow T \)
r3. \( T \rightarrow T / D \)
r4. \( T \rightarrow D \)
r5. \( D \rightarrow \text{a} | \text{b} | (E) \)

shift-reduce parsing
Parse Trees: 7

r1. E $\rightarrow$ E − T
r2. E $\rightarrow$ T
r3. T $\rightarrow$ T / D
r4. T $\rightarrow$ D
r5. D $\rightarrow$ a | b | ( E )

shift-reduce parsing
Parse Trees: 8

r1. E → E − T
r2. E → T
r3. T → T / D
r4. T → D
r5. D → a | b | ( E )

shift-reduce parsing
Parsing: Summary: 1

- All high-level languages are designed so that they may be parsed in this fashion with only a single token look-ahead.
- Parsers for a language can be automatically constructed by parser-generators such as Yacc, Bison, ML-Yacc and CUP in the case of Java.
- Shift-reduce conflicts if any, are automatically detected and reported by the parser-generator.
- Shift-reduce conflicts may be avoided by suitably redesigning the context-free grammar.
Parsing: Summary: 2

- Very often shift-reduce conflicts may occur because of the prefix problem. In such cases many parser-generators resolve the conflict in favour of shifting.

- There is also a possibility of reduce-reduce conflicts. This usually happens when there is more than one nonterminal symbol to which the contents of the stack may reduce.

- A minor reworking of the grammar to avoid redundant non-terminal symbols will get rid of reduce-reduce conflicts.

The Big Picture
Recursive Descent Parsing

- Suitable for grammars that are LL(1)
- A set of (mutually) recursive procedures
- Has a single procedure/function for each non-terminal symbol
- Allows for syntax errors to be pinpointed more accurately than most other parsing methods
Caveats with RDP: Left Recursion

Any direct or indirect left-recursion in the grammar can lead to infinite recursive calls during which no input token is consumed and there is no return from the recursion. In particular,

- Production rules cannot be left-recursive i.e. they should not be of the form $A \rightarrow A\alpha$. This would result in an infinite recursion with no input token consumed.

- A production cannot even be *indirectly* left recursive. For instance the following is *indirect* left-recursion of cycle length 2.

**Example 4.12**

\[
A \rightarrow B\beta \\
B \rightarrow A\alpha
\]

*where* $\alpha, \beta \in (N \cup T)^*$. 

- In general it should be impossible to have derivation sequences of the form $A \Rightarrow A_1\alpha_1 \cdots \Rightarrow A_{n-1}\alpha_{n-1} \Rightarrow A\alpha_n$ for nonterminal symbols $A, A_1, \ldots, A_{n-1}$ for any $n > 0$. 
Caveats with RDP: Left Factoring

For RDP to succeed without backtracking, for each input token and each non-terminal symbol there should be only one rule applicable;

**Example 4.13** A set of productions of the form

\[ A \rightarrow aB\beta \mid aC\gamma \]

where \( B \) and \( C \) stand for different phrases would lead to non-determinism. The normal practice then would be to left-factor the two productions by introducing a new non-terminal symbol \( A' \) and rewrite the rule as

\[ A \rightarrow aA' \]

\[ A' \rightarrow B\beta \mid C\gamma \]

provided \( B \) and \( C \) generate terminal strings with different first symbols (otherwise more left-factoring needs to be performed).
Left Recursion

The grammar used in shift-reduce parsing is clearly left-recursive in both the nonterminals $E$ and $T$ and hence is not amenable to recursive-descent parsing. The grammar may then have to be modified as follows:

$$
E \rightarrow TE' \\
E' \rightarrow -TE' \mid \varepsilon \\
T \rightarrow DT' \\
T' \rightarrow \slash DT' \mid \varepsilon \\
D \rightarrow a \mid b \mid (E)
$$

Now this grammar is no longer left-recursive and may then be parsed by a recursive-descent parser.
4.7. Specification of Syntax: Extended Backus-Naur Form

Specification of Syntax: EBNF
4.7.1. The Extended Backus-Naur Form (EBNF)

The EBNF specification of a programming language is a collection of rules that defines the (context-free) grammar of the language. It specifies the formation rules for the correct grammatical construction of the phrases of the language.

Start symbol. The rules are written usually in a “top-down fashion” and the very first rule gives the productions of the start symbol of the grammar.

Non-terminals. Uses English words or phrases to denote non-terminal symbols. These words or phrases are suggestive of the nature or meaning of the constructs.

Metasymbols.

- Sequences of constructs enclosed in “{” and “}” denote zero or more occurrences of the construct (c.f. Kleene closure on regular expressions).
- Sequences of constructs enclosed in “[” and “]” denote that the enclosed constructs are optional i.e. there can be only zero or one occurrence of the sequence.
- Constructs are enclosed in “(” and “)” to group them together.
- “|” separates alternatives.
- “::=” defines the productions of each non-terminal symbol.
- “.” terminates the possibly many rewrite rules for a non-terminal.

Terminals. Terminal symbol strings are usually enclosed in double-quotes when written in monochrome (we shall additionally colour-code them).
Balanced Parentheses: CFG

Example 4.14 A context-free grammar for balanced parentheses (including the empty string) over the terminal alphabet \{\(,\), \[\], \{, \}\} could be given as \(BP_3 = \langle \{S\}, \{(\,), [\,], \{, \}\}, P, \{S\} \rangle\), where \(P\) consists of the productions

\[
  \begin{align*}
  S & \rightarrow \epsilon, \\
  S & \rightarrow (S)S, \\
  S & \rightarrow [S]S, \\
  S & \rightarrow \{S\}S
  \end{align*}
\]
Balanced Parentheses: EBNF

Example 4.15 $BP_3$ may be expressed in EBNF as follows:

$BracketSeq ::= \{Bracket\}.$

$Bracket ::= LeftParen BracketSeq RightParen \mid LeftSqbracket BracketSeq RightSqbracket \mid LeftBrace BracketSeq RightBrace.$

$LeftParen ::= \"(\".$

$RightParen ::= \\"\).\"

$LeftSqbracket ::= \"[\".$

$RightSqbracket ::= \\"\].\"

$LeftBrace ::= \"\{".$

$RightBrace ::= \\"\}\".$
EBNF in EBNF

EBNF has its own grammar which is again context-free. Hence EBNF (4.7.1) may be used to define EBNF in its own syntax as follows:

\[
\begin{align*}
\text{Syntax} & ::= \{ \text{Production} \}.
\text{Production} & ::= \text{NonTerminal} \ "::=\" \ \text{PossibleRewrites} \ "\".
\text{PossibleRewrites} & ::= \text{Rewrite} \ \{ \"\" \text{Rewrite} \}.
\text{Rewrite} & ::= \text{Symbol} \ \{ \text{Symbol} \}.
\text{Symbol} & ::= \text{NonTerminal} \mid \text{Terminal} \mid \text{GroupRewrites}.
\text{GroupRewrites} & ::= \"\" \ \text{PossibleRewrites} \ \"\" \mid
\"\" \ \text{PossibleRewrites} \ \"\" \mid
\"(" \ \text{PossibleRewrites} \ ")\".
\text{NonTerminal} & ::= \text{Letter} \ \{ \text{Letter} \mid \text{Digit} \}.
\text{Terminal} & ::= \text{Character} \ \{ \text{Character} \}.
\end{align*}
\]
EBNF: Character Set

The character set used in EBNF is described below.

Character ::= Letter | Digit | SpecialChar
Letter ::=UpperCase | LowerCase
LowerCase ::= “a” | “b” | “c” | “d” | “e” | “f” | “g” | “h” | “i” | “j” | “k” | “l” | “m” | “n” | “o” | “p” | “q” | “r” | “s” | “t” | “u” | “v” | “w” | “x” | “y” | “z”
Digit ::= “0” | “1” | “2” | “3” | “4” | “5” | “6” | “7” | “8” | “9”
SpecialChar ::= “!” | “” | “#” | “$” | “%” | “&” | “’” | “(” | “)” | “*” | “+” | “,” | “-” | “.” | “/” | “:” | “;” | “<” | “=” | “>” | “?” | “@” | “[” | “\” | “]” | “^” | “_” | “|” | “{” | “|” | “}” | “~”
Syntax Diagrams

• EBNF was first used to define the grammar of ALGOL-60 and the syntax was used to design the parser for the language.

• EBNF also has a diagrammatic rendering called syntax diagrams or railroad diagrams. The grammar of SML has been produced by a set of syntax diagrams.

• Pascal has been defined using both the text-version of EBNF and through syntax diagrams.

• While the text form of EBNF helps in parsing, the diagrammatic rendering is only for the purpose of readability.

• EBNF is a specification language that almost all modern programming languages use to define the grammar of the programming language.
Syntax Specifications

• BNF of C
• BNF of Java
• EBNF of Pascal
• Pascal Syntax diagrams
• BNF of Standard ML
• BNF of Datalog
• BNF of Prolog
Syntax of Standard ML

Tobias Nipkow and Larry Paulson

PROGRAMS AND MODULES

Program

TopLevelDeclaration ;

TopLevelDeclaration

Expression

ObjectDeclaration

SignatureDeclaration

FunctorDeclaration

ObjectDeclaration

Declaration

structure Ident : Signature = Structure

and

local ObjectDeclaration in ObjectDeclaration end

;
**SignatureDeclaration**

\[
\text{signature} \quad \text{Ident} \quad \text{=} \quad \text{Signature} \\
\text{FunctorDeclaration} \quad \text{functor} \quad \text{FunctorBinding} \quad \text{FunctorArguments} \quad \text{Signature} \quad \text{Structure} \\
\text{FunctorBinding} \quad \text{Ident} \quad \text{FunctorArguments} \quad \text{Signature} \quad \text{Structure} \\
\text{FunctorArguments} \quad \text{Ident} \quad \text{Signature} \\
\text{Structure} \quad \text{struct} \quad \text{ObjectDeclaration} \quad \text{end} \\
\text{let} \quad \text{ObjectDeclaration} \quad \text{in} \quad \text{Structure} \quad \text{end} \\
\text{Signature} \quad \text{sig} \quad \text{Specification} \quad \text{end} \]
DECLARATIONS

Declaration

val<br>rec<br>fun<br>type<br>datatype<br>abstype<br>with<br>exception<br>local<br>open<br>infix<br>infixr<br>nonfix
EXPRESSIONS

Expression

InfixExpression

Expression : Type
Expression andalso Expression
Expression handle Match
raise Expression
if Expression then Expression else Expression
while Expression do Expression
case Expression of Match
fn Match

InfixExpression

AtomicExpression

InfixExpression InfixOperator InfixExpression
**Syntax Diagrams of SML: 7**

**AtomicExpression**

**Match and Patterns**

**Match**

**Pattern**

**Pattern**

**AtomicPattern**
Syntax Diagrams of SML: 8

FieldPattern

...
TYPES

Type

LEXICAL MATTERS: IDENTIFIERS, CONSTANTS, COMMENTS

CompoundIdent

CompoundName

Name

InfixOperator

(any Ident that has been declared to be infix)
Syntax Diagrams of SML: 10

1. Constant
   Numeral
   .
   Digit
   E
   Numeral
   "
   any printable character except \ and "
   StringEscape \\

2. StringEscape
   n
   t
   one of \@ABCDEFGHIJKLMNOPQRSTUVWXYZ\`~
   Digit
   Digit
   Digit
   \
   \
   Space
   Tab
   Newline
   Formfeed

3. Numeral
   Digit
   .

4. TypeVar
   .
   AlphanumericIdent
Syntax Diagrams of SML: 11

**Ident**
- AlphanumericIdent
  - one of \%\&\$\+\-\./\<=\>?\^\~\`\|\*/\-
- Label
  - Ident
- Digit

**AlphanumericIdent**
- Letter
  - Letter
  - Digit
  - `-`
  - ``, `'

**Digit**
- one of \d\i\z\n
**Letter**
- one of \A\B\C\D\E\F\G\H\I\J\K\L\M\N\O\P\Q\R\S\T\U\V\W\X\Y\Z\a\b\c\d\e\f\g\h\i\j\k\l\m\n\o\p\q\r\s\t\u\v\w\x\y\z\

**Comment**
- any text that does not include (* or *) as a substring
Exercise 4.1

1. Translate all the context-free grammars that we have so far seen into EBNF specifications.
2. Specify the language of regular expressions over a non-empty finite alphabet \( A \) in EBNF.
3. Given a textual EBNF specification write an algorithm to render each non-terminal as a syntax diagram.
5. Attributes & Semantic Analysis
Attributes & Semantic Analysis
Semantic Analysis: 0

1. Every programming language can be used to program any computable function, assuming of course, it has
   - unbounded memory, and
   - unbounded time

2. Context-free grammars are not powerful enough to represent all computable functions.

   **Example 5.1** The language \( \{ a^n b^n c^n | n > 0 \} \) is not context-free.

3. Semantic analysis is an essential step to generating IR-code, since it requires the computation of certain *bits and pieces of information* called attributes (which include information to be entered into the symbol table or useful for error-handling)

4. These attributes are usually *context-sensitive* in nature. They need to be computed and if necessary propagated during parsing from wherever they are available.
Semantic Analysis: 1

The parser of a programming language provides the framework within which the IR-code or even the target code is to be generated.

The parser also provides a structuring mechanism that divides the task of code generation into bits and pieces determined by the individual nonterminals and production rules.

The parser provides the framework from within which the semantic analysis (which includes the bits and pieces of information that are required for code generation) is performed.
Semantic Analysis: 2

- There are context-sensitive aspects of a program that cannot be represented/enforced by a context-free grammar definition. Examples include
  - **type consistency** between declaration and use.
  - **correspondence** between formal and actual parameters (example 5.1 is an abstraction where $a^n$ represents a function or procedure declaration with $n$ formal parameters and $b^n$ and $c^n$ represent two calls to the same procedure in which the number of actual parameters should equal $n$).
  - **scope** and **visibility** issues with respect to identifiers in a program.
5.1. Syntax-Directed Translation

Attributes

An attribute can represent anything we choose e.g.

- a string
- a number (e.g. size of an array or the number of formal parameters of a function)
- a type
- a memory location
- a procedure to be executed
- an error message to be displayed

The value of an attribute at a parse-tree node is defined by the semantic rule associated with the production used at that node.
Syntax-Directed Definitions (SDD)

Syntax-Directed definitions are high-level specifications which specify the evaluation of

1. various attributes

2. various procedures such as
   - transformations
   - generating code
   - saving information
   - issuing error messages

They hide various implementation details and free the compiler writer from explicitly defining the order in which translation, transformations, and code generation take place.
Kinds of Attributes

There are two kinds of attributes that one can envisage.

**Synthesized attributes** A synthesized attribute is one whose value depends upon the values of its immediate children in the concrete parse tree.

A syntax-directed definition that uses only synthesized attributes is called an *S-attributed* definition. See example

**Inherited attributes** An inherited attribute is one whose value depends upon the values of the attributes of its parents or siblings in the parse tree.

Inherited attributes are convenient for expressing the dependence of a language construct on the *context* in which it appears.
What is Syntax-directed?

- A syntax-directed definition is a generalisation of a context-free grammar in which each grammar symbol has an associated set of attributes, partitioned into two subsets called synthesized and inherited attributes.

- The various attributes are computed by so-called semantic rules associated with each production of the grammar which allows the computation of the various attributes.

- These semantic rules are in general executed during bottom-up (SR) parsing at the stage when a reduction needs to be performed by the given rule and top-down (RDP) parsing in the procedure before the next call or return from the procedure.

- A parse tree showing the various attributes at each node is called an annotated parse tree.
Forms of SDDs

In a syntax-directed definition, each grammar production rule $X \rightarrow \alpha$ has associated with it a set of semantic rules of the form $b = f(a_1, \ldots, a_k)$ where $a_1, \ldots, a_k$ are attributes belonging to $X$ and/or the grammar symbols of $\alpha$.

**Definition 5.2** Given a production $X \rightarrow \alpha$, an attribute $a$ is **synthesized**: a synthesized attribute of $X$ (denoted $X.a$) or **inherited**: an inherited attribute of one of the grammar symbols of $\alpha$ (denoted $B.a$ if $a$ is an attribute of $B$).

In each case the attribute $a$ is said to depend upon the attributes $a_1, \ldots, a_k$. 
Attribute Grammars

• An attribute grammar is a syntax-directed definition in which the functions in semantic rules can have no side-effects.

• The attribute grammar also specifies how the attributes are propagated through the grammar, by using *graph dependency* between the productions.

• In general *different occurrences* of the *same* non-terminal symbol in each production will be distinguished by appropriate subscripts when defining the semantic rules associated with the rule.

The following example illustrates the concept of a syntax-directed definition using synthesized attributes.
Attribute Grammars: Example

Determining the values of arithmetic expressions. Consider a simple attribute \textit{val} associated with an expression

\[
E_0 \rightarrow E_1 - T \quad \triangleright \quad E_0.val := E_1.val - T.val
\]

\[
E \rightarrow T 
\quad \triangleright \quad E.val := T.val
\]

\[
T_0 \rightarrow T_1 / F 
\quad \triangleright \quad T_0.val := T_1.val / F.val
\]

\[
T \rightarrow F 
\quad \triangleright \quad T.val := F.val
\]

\[
F \rightarrow (E) 
\quad \triangleright \quad F.val := E.val
\]

\[
F \rightarrow n 
\quad \triangleright \quad F.val := n.val
\]

Note: The attribute \textit{n.val} is the value of the numeral \textit{n} computed during scanning (lexical analysis).
Attributes: Basic Assumptions

- Terminal symbols are assumed to have only synthesized attributes. Their attributes are all supplied by the lexical analyser during scanning.
- The start symbol of the grammar can have *only* synthesized attributes.
- In the case of LR parsing with its special start symbol, the start symbol *cannot have any* inherited attributes because
  1. it does not have any parent nodes in the parse tree and
  2. it does not occur on the right-hand side of any production.
5.1.1. Synthesized Attributes

Synthesized Attributes: 0

Evaluating the expression \((4 - 1)/2\) generated by the grammar for subtraction and division
Synthesized Attributes: 1
Synthesized Attributes: 2
Synthesized Attributes: 3

```
ETF
/
/
(n)

ETF
/
/
(n)

ET
-
T
F
n

ET
-
T
F
n

ET
-
T
F
n

ET
-
T
F
n

ET
-
T
F
n
```

Synthesized Attributes

4 3 2 1
Synthesized Attributes: 4

Synthesized Attributes
Synthesized Attributes: 5

Synthesized Attributes
4 3 2 1
Synthesized Attributes: 6
Synthesized Attributes: 7
Synthesized Attributes: 8
Synthesized Attributes: 9
Synthesized Attributes: 10
Synthesized Attributes: 11
Synthesized Attributes: 12
Synthesized Attributes: 13
Synthesized Attributes: 14
An Attribute Grammar

\[ E_0 \rightarrow E_1 - T \quad \triangleright \quad E_0.val := sub(E_1.val, T.val) \]

\[ E \rightarrow T \quad \triangleright \quad E.val := T.val \]

\[ T_0 \rightarrow T_1 / F \quad \triangleright \quad T_0.val := div(T_1.val, F.val) \]

\[ T \rightarrow F \quad \triangleright \quad T.val := F.val \]

\[ F \rightarrow (E) \quad \triangleright \quad F.val := E.val \]

\[ F \rightarrow n \quad \triangleright \quad F.val := n.val \]
Evaluation of Synthesized Attributes

During parsing synthesized attributes are evaluated as follows:

**Bottom-up Parsers**

1. Keep an attribute value stack along with the parsing stack.
2. Just before applying a reduction of the form \( Z \rightarrow Y_1 \ldots Y_k \), compute the attribute values of \( Z \) from the attribute values of \( Y_1, \ldots, Y_k \) and place them in the same position on the attribute value stack corresponding to the one where the symbol \( Z \) will appear on the parsing stack as a result of the reduction.

**Top-down Parsers** In any production of the form \( Z \rightarrow Y_1 \ldots Y_k \), the parser makes recursive calls to procedures corresponding to the symbols \( Y_1 \ldots Y_k \). In each case the attributes of the non-terminal symbols \( Y_1 \ldots Y_k \) are computed and returned to the procedure for \( Z \). Compute the synthesized attributes of \( Z \) from the attribute values returned from the recursive calls.
5.1.2. Inherited Attributes

Inherited Attributes: 0

C-style declarations generating \( \text{int} \ x, y, z. \)

\[
D \rightarrow T \ L \\
T \rightarrow \text{int} \ | \ \text{float} \\
L \rightarrow L, I \ | \ I \\
I \rightarrow x \ | \ y \ | \ z
\]
Inherited Attributes: 1

C-style declarations generating \(\text{int } x, y, z\).

\[
D \rightarrow T \ L \\
L \rightarrow L, I | I \\
T \rightarrow \text{int } | \text{float} \\
I \rightarrow x | y | z
\]
Inherited Attributes: 2

C-style declarations generating \texttt{int x, y, z}.

\[
D \rightarrow T \ L \\
L \rightarrow L, I \mid I \\
T \rightarrow \text{int} \mid \text{float} \\
I \rightarrow x \mid y \mid z
\]
Inherited Attributes: 3

C-style declarations generating int x, y, z.

\[
D \rightarrow TL \\
L \rightarrow LI | I \\
I \rightarrow x | y | z
\]
Inherited Attributes: 4

C-style declarations generating \texttt{int} \( x, y, z \).

\[
D \rightarrow T \ L \\
T \rightarrow \texttt{int} \mid \texttt{float} \\
L \rightarrow L, I \mid I \\
I \rightarrow x \mid y \mid z
\]
Inherited Attributes: 5

C-style declarations generating \[ \text{int } x, y, z. \]

\[
\begin{align*}
D & \rightarrow T \ L \\
T & \rightarrow \text{int} \mid \text{float} \\
L & \rightarrow L, I \mid I \\
I & \rightarrow x \mid y \mid z
\end{align*}
\]
Inherited Attributes: 6

C-style declarations generating \texttt{int x, y, z}.

\[
D \rightarrow TL \\
L \rightarrow LI | I \\
I \rightarrow x | y | z
\]
Inherited Attributes: 7

C-style declarations generating \( \text{int } x, y, z \).

\[
D \rightarrow T \ L \quad T \rightarrow \text{int } | \text{float} \\
L \rightarrow L, I \mid I \quad I \rightarrow x \mid y \mid z
\]
Attribute Grammar: Inherited

\[ D \to TL \quad \triangleright \quad L.in := T.type \]

\[ T \to \text{int} \quad \triangleright \quad T.type := \text{int}.\text{int} \]

\[ T \to \text{float} \quad \triangleright \quad T.type := \text{float}.\text{float} \]

\[ L_0 \to L_1,I \quad \triangleright \quad L_1 := L_0.in \]

\[ L \to I \quad \triangleright \quad I.in := L.in \]

\[ I \to \text{id} \quad \triangleright \quad \text{id}.type := I.in \]
L-attributed Definitions

Definition 5.3 A grammar is L-attributed if for each production of the form $Y \rightarrow X_1 \ldots X_k$, each inherited attribute of the symbol $X_j$, $1 \leq j \leq k$ depends only on

1. the inherited attributes of the symbol $Y$ and
2. the synthesized or inherited attributes of $X_1, \ldots, X_{j-1}$.
Why L-attributedness?

Intuitively, if $X_{j}.inh$ is an inherited attribute then

- it cannot depend on any synthesized attribute $Y.syn$ of $Y$ because it is possible that the computation of $Y.syn$ requires the value of $X_{j}.inh$ leading to circularity in the definition.
- if the value of $X_{j}.inh$ depends upon the attributes of one or more of the symbols $X_{j+1}, \ldots, X_{k}$ then the computation of $X_{j}.inh$ cannot be performed just before the reduction by the rule $Y \rightarrow X_{1} \ldots X_{k}$ during parsing. Instead it may have to be postponed till the end of parsing.
- it could depend on the synthesized or inherited attributes of any of the symbols $X_{1} \ldots X_{j-1}$ since they would already be available on the attribute value stack.
- it could depend upon the inherited attributes of $Y$ because these inherited attributes can be computed from the attributes of the symbols lying below $X_{1}$ on the stack, provided these inherited attributes of $Y$ are also L-attributed.
A Non L-attributed Definition

Our attribute grammar for C-style declarations is definitely L-attributed. However consider the following grammar for declarations in Pascal and ML.

\[
\begin{align*}
D & \rightarrow L:T \triangleright L.in := T.type \\
T & \rightarrow \text{int} \triangleright T.type := \text{int}.\text{int} \\
T & \rightarrow \text{real} \triangleright T.type := \text{real}.\text{real} \\
L_0 & \rightarrow L_1,I \triangleright L_1 := L_0.in \\
L & \rightarrow I \triangleright I.in := L.in \\
I & \rightarrow \text{id} \triangleright \text{id}.\text{type} := I.in
\end{align*}
\]

In the first semantic rule the symbol \( L.in \) is inherited from a symbol to its right viz. \( T.type \) and hence is not L-attributed.
Evaluating Non-L-attributed Definitions

In many languages like ML which allow higher order functions as values, a definition not being L-attributed may not be of serious concern. But in most other languages it is serious enough to warrant changing the grammar of the language so as to replace inherited attributes by corresponding synthesized ones. The language of the grammar of Pascal and ML declarations can be generated as follows:

\[
\begin{align*}
D & \rightarrow \text{id}L \quad \triangleright \quad \text{addtype}(\text{id}, L.\text{type}) \\
L & \rightarrow :T \quad \triangleright \quad L.\text{in} := T.\text{type} \\
L & \rightarrow ,\text{id} L \quad \triangleright \quad L_0.\text{type} := L_1.\text{type}; \\
\end{align*}
\]

\[
\begin{align*}
\text{addtype}(\text{id}L_1.\text{type}) \\
T & \rightarrow \text{int} \quad \triangleright \quad T.\text{type} := \text{int}.\text{int} \\
T & \rightarrow \text{real} \quad \triangleright \quad T.\text{type} := \text{real}.\text{real}
\end{align*}
\]
Dependency Graphs

In general, the attributes required to be computed during parsing could be synthesized or inherited and further it is possible that some synthesized attributes of some symbols may depend on the inherited attributes of some other symbols. In such a scenario it is necessary to construct a dependency graph of the attributes of each node of the parse tree.
Dependency Graph Construction

**Algorithm 4** Attribute Dependency Graph Construction

**Require:** A parse tree of a CFG and the list of attributes

**Ensure:** A dependency graph

for all nodes \( n \) of the parse tree do
  for all attributes \( a \) of node \( n \) do
    Create an attribute node \( n.a \)
  end for
end for

for all nodes \( n \) of the parse tree do
  for all semantic rules \( a := f(b_1, \ldots, b_k) \) do
    for all \( i : 1 \leq i \leq k \) do
      Create a directed edge \( b_i \rightarrow a \)
    end for
  end for
end for
6. Abstract Syntax
Abstract Syntax Trees

The construction of ASTs from concrete parse trees is another example of a transformation that can be performed using a syntax-directed definition that has no side-effects. Hence we define it using an attribute grammar.
Abstract Syntax: 0

\[ E \rightarrow E - T \mid T \]
\[ T \rightarrow T / F \mid F \]
\[ F \rightarrow n \mid (E) \]

Suppose we want to evaluate an expression \((4 - 1)/2\). What we actually want is a tree that looks like this:
Evaluation: 0

```
/  \\  
\   \  
4    1
```

---
Evaluation: 1
Evaluation: 2
Evaluation: 3
But what we actually get during parsing is a tree that looks like ...
Abstract Syntax: 1

...THIS!
Abstract Syntax

Shift-reduce parsing produces a concrete syntax tree from the rightmost derivation. The syntax tree is concrete in the sense that

- It contains a lot of redundant symbols that are important or useful only during the parsing stage.
  - punctuation marks
  - brackets of various kinds

- It makes no distinction between operators, operands, and punctuation symbols

On the other hand the abstract syntax tree (AST) contains no punctuation and makes a clear distinction between an operand and an operator.
Abstract Syntax: Imperative Approach

We use attribute grammar rules to construct the abstract syntax tree (AST) from the parse tree. But in order to do that we first require two procedures for tree construction.

**makeLeaf(literal)**: Creates a node with label `literal` and returns a pointer or a reference to it.

**makeBinaryNode(opr, opd1, opd2)**: Creates a node with label `opr` (with fields which point to `opd1` and `opd2`) and returns a pointer or a reference to the newly created node.

Now we may associate a *synthesized* attribute called `ptr` with each terminal and nonterminal symbol which points to the root of the subtree created for it.
Abstract Syntax Trees: Imperative

\[ E_0 \rightarrow E_1 - T \quad \Rightarrow \quad E_0.ptr := makeBinaryNode(-, E_1.ptr, T.ptr) \]

\[ E \rightarrow T \quad \Rightarrow \quad E.ptr := T.ptr \]

\[ T_0 \rightarrow T_1 / F \quad \Rightarrow \quad T_0.ptr := makeBinaryNode(/, T_1.ptr, F.ptr) \]

\[ T \rightarrow F \quad \Rightarrow \quad T.ptr := F.ptr \]

\[ F \rightarrow (E) \quad \Rightarrow \quad F.ptr := E.ptr \]

\[ F \rightarrow n \quad \Rightarrow \quad F.ptr := makeLeaf(n.val) \]

The Big Picture
Abstract Syntax: Functional Approach

We use attribute grammar rules to construct the abstract syntax tree (AST) functionally from the parse tree. But in order to do that we first require two functions/constructors for tree construction.

makeLeaf(literal) : Creates a node with label literal and returns the AST.
makeBinaryNode(opr, opd1, opd2) : Creates a tree with root label opr (with sub-trees opd1 and opd2).

Now we may associate a synthesized attribute called ast with each terminal and nonterminal symbol which points to the root of the subtree created for it.
Abstract Syntax: Functional

\[
\begin{align*}
E_0 & \rightarrow E_1 - T & \triangleright E_0.ast := \text{makeBinaryNode}(\_, E_1.ast, T.ast) \\
E & \rightarrow T & \triangleright E.ast := T.ast \\
T_0 & \rightarrow T_1 / F & \triangleright T_0.ast := \text{makeBinaryNode}(/, T_1.ast, F.ast) \\
T & \rightarrow F & \triangleright T.ast := F.ast \\
F & \rightarrow (E) & \triangleright F.ast := E.ast \\
F & \rightarrow n & \triangleright F.ast := \text{makeLeaf}(n.val)
\end{align*}
\]

The Big Picture
7. Symbol Table
Symbol Table
Symbol Table: 1

- The store house of context-sensitive and run-time information about every identifier in the source program.
- All accesses relating to an identifier require to first find the attributes of the identifier from the symbol table.
- Usually organized as a hash table – provides fast access.
- Compiler-generated temporaries may also be stored in the symbol table.
Symbol Table: 2

Attributes stored in a symbol table for each identifier:

- type
- size
- scope/visibility information
- base address
- addresses to location of auxiliary symbol tables (in case of records, procedures, classes)
- address of the location containing the string which actually names the identifier and its length in the string pool
Symbol Table:3

- A symbol table exists throughout the compilation and run-time.
- Major operations required of a symbol table:
  - insertion
  - search
  - deletions are purely logical (depending on scope and visibility) and not physical
- Keywords are often stored in the symbol table before the compilation process begins.
Symbol Table: 4

Accesses to the symbol table at every stage of the compilation process,

**Scanning:** Insertion of new identifiers.

**Parsing:** Access to the symbol table to ensure that an operand exists (declaration before use).

**Semantic analysis:**

- Determination of types of identifiers from declarations
- Type checking to ensure that operands are used in type-valid contexts.
- Checking scope, visibility violations.
Symbol Table: 5

**IR generation:** Memory allocation and relative\(^a\) address calculation.

**Optimization:** All memory accesses through symbol table

**Target code:** Translation of relative addresses to absolute addresses in terms of word length, word boundary etc.

The Big picture

\(^a\text{i.e. relative}\) to a base address that is known only at run-time
8. Intermediate Representation
Intermediate Representation
Intermediate Representation

Intermediate representations are important for reasons of portability i.e. platform (hardware and OS) independence.

• *(more or less) independent* of specific features of the high-level language.
  
  Example. Java byte-code for any high-level language.

• *(more or less) independent* of specific features of any particular target architecture (e.g. number of registers, memory size)
  
  – number of registers
  
  – memory size
  
  – word length
IR Properties: 1

1. It is fairly **low-level** containing instructions common to all target architectures and assembly languages.
   How low can you stoop? . . .

2. It contains some fairly **high-level** instructions that are common to most high-level programming languages.
   How high can you rise?

3. To ensure **portability**
   - an **unbounded** number of variables and memory locations
   - no commitment to **Representational Issues**

4. To ensure **type-safety**
   - memory locations are also typed according to the data they may contain,
   - no commitment is made regarding word boundaries, and the structure of individual data items.

Next
IR: Representation?

- No commitment to word boundaries or byte boundaries
- No commitment to representation of
  - `int` vs. `float`,
  - `float` vs. `double`,
  - `packed` vs. `unpacked`,
  - strings – where and how?

Back to IR Properties:1
IR: How low can you stoop?

- most arithmetic and logical operations, load and store instructions etc.
- so as to be interpreted easily,
- the interpreter is fairly small,
- execution speeds are high,
- to have fixed length instructions (where each operand position has a specific meaning).

Back to IR Properties:1
IR: How high can you rise?

- typed variables,
- temporary variables instead of registers,
- array-indexing,
- random access to record fields,
- parameter-passing,
- pointers and pointer management
- no limits on memory addresses

Back to IR Properties:1
A typical instruction set: 1

Three address code: A suite of instructions. Each instruction has at most 3 operands.

- an opcode representing an operation with at most 2 operands
- two operands on which the binary operation is performed
- a target operand, which accumulates the result of the (binary) operation.

If an operation requires less than 3 operands then one or more of the operands is made null.
A typical instruction set: 2

- Assignments (LOAD-STORE)
- Jumps (conditional and unconditional)
- Procedures and parameters
- Arrays and array-indexing
- Pointer Referencing and Dereferencing
A typical instruction set: 2.1

- Assignments (LOAD-STORE)
  - \( x := y \ bop \ z \), where \( bop \) is a binary operation
  - \( x := uop \ y \), where \( uop \) is a unary operation
  - \( x := y \), load, store, copy or register transfer

- Jumps (conditional and unconditional)

- Procedures and parameters

- Arrays and array-indexing

- Pointer Referencing and Dereferencing
A typical instruction set: 2.2

- Assignments (LOAD-STORE)
- Jumps (conditional and unconditional)
  - goto L – Unconditional jump,
  - x relop y goto L – Conditional jump, where relop is a relational operator
- Procedures and parameters
- Arrays and array-indexing
- Pointer Referencing and Dereferencing
A typical instruction set: 2.3

• Assignments (LOAD-STORE)
• Jumps (conditional and unconditional)
• Procedures and parameters
  – \texttt{call p n}, where \texttt{n} is the number of parameters
  – \texttt{return y}, return value from a procedures call
  – \texttt{param x}, parameter declaration
• Arrays and array-indexing
• Pointer Referencing and Dereferencing
A typical instruction set: 2.4

- Assignments (LOAD-STORE)
- Jumps (conditional and unconditional)
- Procedures and parameters
- Arrays and array-indexing
  - \( x := a[i] \) – array indexing for \( r\)-value
  - \( a[j] := y \) – array indexing for \( l\)-value

Note: The two opcodes are different depending on whether \( l\)-value or \( r\)-value is desired. \( x \) and \( y \) are always simple variables

- Pointer Referencing and Dereferencing
A typical instruction set: 2.5

- Assignments (LOAD-STORE)
- Jumps (conditional and unconditional)
- Procedures and parameters
- Arrays and array-indexing
- Pointer Referencing and Dereferencing
  - \( x := \hat{y} \) – referencing: set \( x \) to point to \( y \)
  - \( x := *y \) – dereferencing: copy contents of location pointed to by \( y \) into \( x \)
  - \( *x := y \) – dereferencing: copy r-value of \( y \) into the location pointed to by \( x \)

Picture
Pointers

\[
x := \ast y \\
@z := \ast y \\
(*) := y \\
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]
IR: Generation Basics

- Can be generated by recursive traversal of the abstract syntax tree.
- Can be generated by syntax-directed translation as follows:
  For every non-terminal symbol \( N \) in the grammar of the source language there exist two attributes
  \( N\.place \), which denotes the address of a temporary variable where the result of the execution of the generated code is stored
  \( N\.code \), which is the actual code segment generated.
- In addition a global counter for the instructions generated is maintained as part of the generation process.
- It is independent of the source language but can express target machine operations without committing to too much detail.
IR: Infrastructure 1

Given an abstract syntax tree $T$, with $T$ also denoting its root node.

**$T$.place** address of *temporary* variable where result of execution of the $T$ is stored.

$newtemp$ returns a *fresh* variable name and also installs it in the symbol table along with relevant information

**$T$.code** the actual sequence of instructions generated for the tree $T$.

$newlabel$ returns a *label* to mark an instruction in the generated code which may be the *target* of a jump.

$emit$ emits an instructions (regarded as a *string*).
IR: Infrastructure 2

Colour and font coding of IR code generation process.

- **Green**: Nodes of the Abstract Syntax Tree
- **Brown**: Characters and strings of the Intermediate Representation
- **Red**: Variables and data structures of the language in which the IR code generator is written
- **Blue**: Names of relevant procedures used in IR code generation.
- **Black**: All other stuff.
IR: Example 1

$E \rightarrow id$

$E\text{.place} := id\text{.place};$
$E\text{.code} := \text{emit}()$

$E_0 \rightarrow E_1 - E_2$

$E_0\text{.place} := \text{newtemp};$
$E_0\text{.code} := E_1\text{.code}$
  $\| E_2\text{.code}$
  $\| \text{emit}(E_0\text{.place} := E_1\text{.place} - E_2\text{.place})$
IR: Example 2

\[ S \rightarrow id := E \]

\[ S.code := E.code \]
\[ || \text{emit(id.place:=E.place)} \]

\[ S_0 \rightarrow \text{while } E \text{ do } S_1 \]

\[ S_0.begin := \text{newlabel}; \]
\[ S_0.after := \text{newlabel}; \]
\[ S_0.code := \text{emit(S_0.begin:) } \]
\[ || \text{E.code} \]
\[ || \text{emit(if E.place= 0 goto S_0.after)} \]
\[ || S_1.code \]
\[ || \text{emit(goto S_0.begin)} \]
\[ || \text{emit(S_0.after:)} \]
IR: Example 3

\[ S \rightarrow id := E \]

\[ S.code := E.code \]
\[ \text{|| emit} (id.place := E.place) \]

\[ S_0 \rightarrow \text{while } E \text{ do } S_1 \]

\[ S_0.begin := \text{newlabel;} \]
\[ S_0.after := \text{newlabel;} \]
\[ S_0.code := \text{emit}(S_0.begin:) \]
\[ \text{|| } E.code \]
\[ \text{|| } \text{emit}(\text{if } E.place = 0 \text{ goto } S_0.after) \]
\[ \text{|| } S_1.code \]
\[ \text{|| } \text{emit}(\text{goto } S_0.begin) \]
\[ \text{|| } \text{emit}(S_0.after:) \]
IR: Generation End

While generating the intermediate representation, it is sometimes necessary to generate jumps into code that has not been generated as yet (hence the address of the label is unknown). This usually happens while processing

- **forward** jumps
- **short-circuit** evaluation of boolean expressions

It is usual in such circumstances to either fill up the empty label entries in a second pass over the the code or through a process of *backpatching* (which is the maintenance of lists of jumps to the same instruction number), wherein the blank entries are filled in once the sequence number of the target instruction becomes known.
9. The Pure Untyped Lambda Calculus: Basics
Pure Untyped $\lambda$-Calculus: Syntax

The language $\Lambda$ of pure untyped $\lambda$-terms is the smallest set of terms built up from an infinite set $V$ of variables and closed under the following productions

$$L, M, N ::= x \quad \text{Variable}$$
$$| \quad \lambda x[L] \quad \text{Abstraction}$$
$$| \quad (L M) \quad \text{Application}$$

where $x \in V$.

- A **Variable** denotes a possible binding in the external environment.
- An **Abstraction** denotes a function which takes a formal parameter.
- An **Application** denotes the application of a function to an actual parameter.
Free and Bound Variables

**Definition 9.1** For any term $N$ the set of free variables and the set of all variables are defined by induction on the structure of terms.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$FV(N)$</th>
<th>$Var(N)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>${x}$</td>
<td>${x}$</td>
</tr>
<tr>
<td>$\lambda x[L]$</td>
<td>$FV(L) - {x}$</td>
<td>$Var(L) \cup {x}$</td>
</tr>
<tr>
<td>$(L M)$</td>
<td>$FV(L) \cup FV(M)$</td>
<td>$Var(L) \cup Var(M)$</td>
</tr>
</tbody>
</table>

- The set of **bound variables** $BV(N) = Var(N) - FV(N)$.
- The same variable name may be used with different bindings in a single term (e.g. $(\lambda x[x] \lambda x[(x y)])$)
- The brackets “[” and “]” delimit the **scope** of the bound variable $x$ in the term $\lambda x[L]$.
- $\Lambda_0 \subseteq \Lambda$ is the set of **closed** $\lambda$-terms (i.e. terms with no free variables).
Notational Conventions

To minimize use of brackets unambiguously

1. \( \lambda x_1 x_2 \ldots x_m[L] \) denotes \( \lambda x_1[\lambda x_2[\ldots \lambda x_m[L] \ldots]] \) i.e. \( L \) is the scope of each of the variables \( x_1, x_2, \ldots x_m \).

2. \( (L_1 L_2 \ldots L_m) \) denotes \( (\ldots (L_1 L_2) \ldots L_m) \) i.e. application is left-associative.
Substitution

Definition 9.2 For any terms \( L, M \) and \( N \) and any variable \( x \), the substitution of the term \( N \) for a variable \( x \) is defined as follows:

- \( \{N/x\}x \equiv N \)
- \( \{N/x\}y \equiv y \) if \( y \neq x \)
- \( \{N/x\}\lambda x[L] \equiv \lambda x[L] \)
- \( \{N/x\}\lambda y[L] \equiv \lambda y[\{N/x\}L] \) if \( y \neq x \) and \( y \notin FV(N) \)
- \( \{N/x\}\lambda y[L] \equiv \lambda z[\{N/x\}\{z/y\}L] \) if \( y \neq x \) and \( y \in FV(N) \) and \( z \) is 'fresh'
- \( \{N/x\}(L M) \equiv (\{N/x\}L \{N/x\}M) \)

- In the above definition it is necessary to ensure that the free variables of \( N \) continue to remain free after substitution i.e. none of the free variables of \( N \) should be "captured" as a result of the substitution.
- The phrase "\( z \) is 'fresh'" may be taken to mean \( z \notin FV(N) \cup Var(L) \).
- \( z \) could be fresh even if \( z \in BV(N) \)
Compatibility

Definition 9.3 A binary relation $\rho \subseteq \Lambda \times \Lambda$ is said to be compatible if $L \rho M$ implies

1. for all variables $x$, $\lambda x[L] \rho \lambda x[M]$ and
2. for all terms $N$, $(L N) \rho (M N)$ and $(N L) \rho (N M)$.

Example 9.4

1. $\equiv_\alpha$ is a compatible relation
2. $\rightarrow^1_\beta$ is by definition a compatible relation.
\(\alpha\)-equivalence

**Definition 9.5 (\(\alpha\)-equivalence)** \(\equiv_\alpha \subseteq \Lambda \times \Lambda\) is the compatible closure of the relation \(\{ (\lambda x[L] \equiv_\alpha \lambda y[y/x](\{y/x\}L)) \mid y \not\in FV(L) \}\).

- Here again if \(y \in FV(L)\) it must not be captured by a change of bound variables.
Untyped $\lambda$-Calculus: Basic $\beta$-Reduction

Definition 9.6

- Any (sub-)term of the form $(\lambda x[L] \ M)$ is called a $\beta$-redex
- Basic $\beta$-reduction is the relation on $\Lambda$

$$
\rightarrow_\beta \overset{df}{=} \{((\lambda x[L] \ M), \{M/x\}L') \mid L' \equiv_\alpha L, L', L, M \in \Lambda\} 
$$

- It is usually represented by the axiom

$$
(\lambda x[L] \ M) \rightarrow_\beta \{M/x\}L'
$$

where $L' \equiv_\alpha L$. 

(2)
Untyped $\lambda$-Calculus: 1-step $\beta$-Reduction

Definition 9.7 A 1-step $\beta$-reduction $\rightarrow^1_\beta$ is the smallest relation (under the $\subseteq$ ordering) on $\Lambda$ such that

\[
\begin{align*}
\beta_1 & \quad \frac{L \rightarrow_\beta M}{L \rightarrow^1_\beta M} \\
\beta_1 \text{Abs} & \quad \frac{L \rightarrow^1_\beta M}{\lambda x[L] \rightarrow^1_\beta \lambda x[M]} \\
\beta_1 \text{AppL} & \quad \frac{L \rightarrow^1_\beta M}{(L \ N) \rightarrow^1_\beta (M \ N)} \\
\beta_1 \text{AppR} & \quad \frac{L \rightarrow^1_\beta M}{(N \ L) \rightarrow^1_\beta (N \ M)}
\end{align*}
\]

- $\rightarrow^1_\beta$ is the compatible closure of basic $\beta$-reduction to all contexts.
- We will often omit the superscript $^1$ as understood.
Untyped $\lambda$-Calculus: $\beta$-Reduction

Definition 9.8

- For all integers $n \geq 0$, $n$-step $\beta$-reduction $\to^n_\beta$ is defined by induction on 1-step $\beta$-reduction

\[
\begin{array}{c}
\text{1-step $\beta$-reduction} \\
L \to_\beta M
\end{array}
\]

- $\beta$-reduction $\to^*_\beta$ is the reflexive-transitive closure of 1-step $\beta$-reduction. That is,

\[
\begin{array}{c}
\text{1-step $\beta$-reduction} \\
L \to_\beta M
\end{array}
\]
Untyped $\lambda$-Calculus: Normalization

Definition 9.9

- A term is called a $\beta$-normal form ($\beta$-nf) if it has no $\beta$-redexes.
- A term is weakly normalising ($\beta$-WN) if it reduces to a $\beta$-normal form.
- A term $L$ is strong normalising ($\beta$-SN) if it has no infinite reduction sequence $L \rightarrow^1_\beta L_1 \rightarrow^1_\beta L_2 \rightarrow^1_\beta \cdots$
Untyped $\lambda$-Calculus: Examples

Example 9.10

1. $K \overset{df}{=} \lambda x\ y[x], \ I \overset{df}{=} \lambda x[x], \ S \overset{df}{=} \lambda x\ y\ z[\((x\ z)\ (y\ z)\)], \ \omega \overset{df}{=} \lambda x[\((x\ x)\)]$ are all $\beta$-nfs.

2. $\Omega \overset{df}{=} (\omega\ \omega)$ has no $\beta$-nf. Hence it is neither weakly nor strongly normalising.

3. $(K\ (\omega\ \omega))$ cannot reduce to any normal form because it has no finite reduction sequences. All its reductions are of the form

\[
(K\ (\omega\ \omega)) \rightarrow^1_\beta (K\ (\omega\ \omega)) \rightarrow^1_\beta (K\ (\omega\ \omega)) \rightarrow^1_\beta \cdots
\]

or at some point it could transform to

\[
(K\ (\omega\ \omega)) \rightarrow^1_\beta \lambda y[(\omega\ \omega)] \rightarrow^1_\beta \lambda y[(\omega\ \omega)] \rightarrow^1_\beta \cdots
\]

4. $(K\ (\omega\ \Omega))$ is weakly normalising because it can reduce to the normal form $\omega$ but it is not strongly normalising because it also has an infinite reduction sequence

\[
((K\ \omega)\ \Omega) \rightarrow^1_\beta ((K\ \omega)\ \Omega) \rightarrow^1_\beta \cdots
\]
Examples of Strong Normalization

Example 9.11

1. \(((K \omega) \omega)\) is strongly normalising because it reduces to the normal form \(\omega\) in two \(\beta\)-reduction steps.

2. Consider the term \(((S K) K)\). Its reduction sequences go as follows:

\[
(S K) K \rightarrow^1_\beta \lambda z[(((K z) (K z)))] \rightarrow^1_\beta \lambda z[z] \equiv I
\]
10. Notions of Reduction
Reduction

For any function such as \( p = \lambda x[3.x.x + 4.x + 1] \),

\[(p\ 2) = 3.2.2 + 4.2 + 1 = 21\]

However there is something *asymmetric* about the identity, in the sense that while \((p\ 2)\) deterministically produces \(3.2.2 + 4.2 + 1\) which in turn simplifies deterministically to 21, it is not possible to deterministically infer that 21 came from \((p\ 2)\). It would be more accurate to refer to this sequence as a *reduction sequence* and capture the asymmetry as follows:

\[(p\ 2) \rightsquigarrow 3.2.2 + 4.2 + 1 \rightsquigarrow 21\]

And yet they are *behaviourally* equivalent and mutually substitutable in all contexts (*referentially transparent*).

1. Reduction (specifically \(\beta\)-reduction) captures this asymmetry.
2. Since reduction produces behaviourally *equal* terms we have the following notion of equality.
Untyped λ-Calculus: β-Equality

Definition 10.1 β-equality or β-conversion (denoted $\equiv_{\beta}$) is the smallest equivalence relation containing β-reduction ($\rightarrow_{\beta}^*$).

The following are equivalent definitions.

1. $\equiv_{\beta}$ is the reflexive-symmetric-transitive closure of 1-step β-reduction.

2. $\equiv_{\beta}$ is the smallest relation defined by the following rules.

\[
\begin{align*}
\text{β Basis} &\quad L \rightarrow_{\beta}^* M \\
\text{β Reflexivity} &\quad L \equiv_{\beta} M \\
\text{β Symmetry} &\quad L \equiv_{\beta} M \\
\text{β Transitivity} &\quad L \equiv_{\beta} M, \quad M \equiv_{\beta} N \\
\end{align*}
\]
Compatibility

Definition 10.2 A binary relation $\rho \subseteq \Lambda \times \Lambda$ is said to be compatible if $L \rho M$ implies

1. for all variables $x$, $\lambda x[L] \rho \lambda x[M]$ and
2. for all terms $N$, $(L N) \rho (M N)$ and $(N L) \rho (N M)$.

Example 10.3

1. $\equiv_\alpha$ is a compatible relation
2. $\rightarrow^{1}_\beta$ is by definition a compatible relation.
Compatibility of Beta-reduction and Beta-Equality

Theorem 10.4 \( \beta \text{-reduction} \rightarrow^* \beta \text{ and } \beta \text{-equality } =^\beta \) are both compatible relations.
Proof of theorem 10.4

Proof: \( (\rightarrow^*_\beta) \) Assume \( L \rightarrow^*_\beta M \). By definition of \( \beta\)-reduction \( L \rightarrow^n_\beta M \) for some \( n \geq 0 \). The proof proceeds by induction on \( n \)

**Basis.** \( n = 0 \). Then \( L \equiv M \) and there is nothing to prove.

**Induction Hypothesis (IH).**

\[
\text{The proof holds for all } k, \ 0 \leq k \leq m \text{ for some } m \geq 0.
\]

**Induction Step.** For \( n = m + 1 \), let \( L \equiv L_0 \rightarrow^m_\beta L_m \rightarrow^1_\beta M \). Then by the induction hypothesis and the compatibility of \( \rightarrow^1_\beta \) we have

\[
\begin{align*}
&\text{for all } x \in V, \ \lambda x[L] \rightarrow^m_\beta \lambda x[L_m], \quad \lambda x[L_m] \rightarrow^1_\beta \lambda x[M], \\
&\text{for all } N \in \Lambda, \ (L N) \rightarrow^m_\beta (L_m N), \quad (L_m N) \rightarrow^1_\beta (M N), \\
&\text{for all } N \in \Lambda, \ (N L) \rightarrow^m_\beta (N L_m), \quad (N L_m) \rightarrow^1_\beta (N M) \\
\end{align*}
\]

By definition of \( \rightarrow^m_\beta \)

\[
\begin{align*}
&\lambda x[L] \rightarrow^m_\beta \lambda x[M], \\
&(L N) \rightarrow^m_\beta (M N) \\
&(N L) \rightarrow^m_\beta (N M)
\end{align*}
\]

End \( (\rightarrow^*_\beta) \)

\( (=\beta) \) Assume \( L =_\beta M \). We proceed by induction on the length of the proof of \( L =_\beta M \) using the definition of \( \beta\)-equality.

**Basis.** \( n = 1 \). Then either \( L \equiv M \) or \( L \rightarrow^*_\beta M \). The case of reflexivity is trivial and the case of \( L \rightarrow^*_\beta M \) follows from the previous proof.

**Induction Hypothesis (IH).**

\[
\text{For all terms } L \text{ and } M, \text{ such that the proof of } L =_\beta M \text{ requires less than } n \text{ steps for } n > 1, \text{ the compatibility result holds.}
\]

**Induction Step.** Suppose the proof requires \( n \) steps and the last step is obtained by use of either \( =_\beta \text{ Symmetry} \) or \( =_\beta \text{ Transitivity} \) on some previous steps.

\( (=\beta \text{ Symmetry}) \). Then the \( (n - 1)\)-st step proved \( M =_\beta L \). By the induction hypothesis and then by applying \( =_\beta \text{ Symmetry} \) to each case we get

\[
\begin{align*}
&\text{for all variables } x, \quad \lambda x[M] =_\beta \lambda x[L] \quad \lambda x[L] =_\beta \lambda x[M], \\
&\text{for all terms } N, \quad (M N) =_\beta (L N) \quad (L N) =_\beta (M N) \\
&\text{for all terms } N, \quad (N M) =_\beta (N L) \quad (N L) =_\beta (N M)
\end{align*}
\]
Case ($\beta$ Transitivity). Suppose $L =_\beta M$ was inferred in the $n$-th step from two previous steps which proved $L =_\beta P$ and $P =_\beta M$ for some term $P$. Then again by induction hypothesis and then applying $=_\beta$ Transitivity we get

<table>
<thead>
<tr>
<th>By $=_\beta$ Transitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda x[L] =_\beta \lambda x[P]$</td>
</tr>
<tr>
<td>$\lambda x[P] =_\beta \lambda x[M]$</td>
</tr>
<tr>
<td>$\lambda x[L] =_\beta \lambda x[M]$</td>
</tr>
<tr>
<td>$(L N) =_\beta (P N)$</td>
</tr>
<tr>
<td>$(P N) =_\beta (M N)$</td>
</tr>
<tr>
<td>$(L N) =_\beta (M N)$</td>
</tr>
<tr>
<td>$(N L) =_\beta (N P)$</td>
</tr>
<tr>
<td>$(N P) =_\beta (N M)$</td>
</tr>
<tr>
<td>$(N L) =_\beta (N P)$</td>
</tr>
</tbody>
</table>

$End (=_\beta)$

QED
**Eta reduction**

Given any term $M$ and a variable $x \notin FV(M)$, the syntax allows us to construct the term $\lambda x[(M\ x)]$ such that for every term $N$ we have

$$(\lambda x[(M\ x)]\ N) \rightarrow^1_\beta (M\ N)$$

In other words,

$$(\lambda x[(M\ x)]\ N) =^\beta (M\ N) \text{ for all terms } N$$

We say that the two terms $\lambda x[(M\ x)]$ and $M$ are **extensionally** equivalent i.e. they are *syntactically distinct* but there is no way to distinguish between their *behaviours*.

So we define **basic $\eta$-reduction** as the relation

$$\lambda x[(L\ x)] \rightarrow^\eta L \text{ provided } x \notin FV(L)$$

(3)
Eta-Reduction and Eta-Equality

The following notions are then defined similar to the corresponding notions for \( \beta \)-reduction.

- **1-step \( \eta \)-reduction** \( \to^1_\eta \) is the closure of basic \( \eta \)-reduction to all contexts,
- \( \to^n_\eta \) is defined by induction on 1-step \( \eta \)-reduction
- **\( \eta \)-reduction** \( \to^*_\eta \) is the reflexive-transitive closure of 1-step \( \eta \)-reduction.
- the notions of strong and weak \( \eta \) normal forms \( \eta \)-nf.
- the notion of \( \eta \)-equality or \( \eta \)-conversion denoted by \( =_\eta \).
Exercise 10.1

1. Prove that $\eta$-reduction and $\eta$-equality are both compatible relations.

2. Prove that $\eta$-reduction is strongly normalising.

3. Define basic $\beta\eta$-reduction as the application of either (2) or (3). Now prove that $\rightarrow^1_{\beta\eta}$, $\rightarrow^*_{\beta\eta}$ and $=_{\beta\eta}$ are all compatible relations.

11. Confluence Definitions
Confluence
Reduction Relations

**Definition 11.1** For any binary relation $\rho$ on $\Lambda$

1. $\rho^1$ is the compatible closure of $\rho$
2. $\rho^+$ is the transitive closure of $\rho^1$
3. $\rho^*$ is the reflexive-transitive-closure of $\rho^1$ and is a preorder
4. $((\rho^1) \cup (\rho^1)^{-1})^*$ (denoted $=\rho$) is the reflexive-symmetric-transitive closure of $\rho^1$ and is an equivalence relation.
5. $=\rho$ is also called the equivalence generated by $\rho$.

We will often use $\rightarrow$ (suitably decorated) as a reduction relation instead of $\rho$. Then $\rightarrow^1$, $\rightarrow^+$, $\rightarrow^*$ and $\leftrightarrow^*$ denote respectively the compatible closure, the transitive closure, the reflexive transitive closure and the equivalence generated by $\rightarrow$. 
The Diamond Property

Definition 11.2 Let \( \rho \) be any relation on terms. \( \rho \) has the diamond property if for all \( L, M, N \),

\[
\begin{array}{ccc}
M & M \\
\rho & \rho \\
L & \Rightarrow & \exists P \\
\rho & \rho \\
N & N \\
\end{array}
\]
Reduction Relations: Termination

Let $\rightarrow$ be a reduction relation, $\rightarrow^*$ the least preorder containing $\rightarrow$ and $\leftarrow^*$ the least equivalence relation containing $\rightarrow^*$. Then

**Definition 11.3** $\rightarrow$ is terminating iff there is no infinite sequence of the form

$$L_0 \rightarrow L_1 \rightarrow \cdots$$
Reduction: Local Confluence

**Definition 11.4** \(\text{def:localconfluence} \rightarrow\) is **locally confluent** if for all \(L, M, N\),

\[
N \leftarrow L \rightarrow M \Rightarrow \exists P : N \rightarrow^* P \leftarrow^* M
\]

which we denote by

\[
\begin{array}{c}
\text{M} \\
\downarrow^* \quad \downarrow^* \\
\text{L} & \Rightarrow \exists \quad \text{P} \\
\downarrow \quad \downarrow \\
\text{N}
\end{array}
\]
Reduction: Semi-confluence

Definition 11.5 $\rightarrow$ is semi-confluent if for all $L, M, N$,

$$N \leftarrow L \rightarrow^* M \Rightarrow \exists P : N \rightarrow^* P \leftarrow^* M$$

which we denote by

\[ M \]

\[ \uparrow \]

\[ \Rightarrow \exists \]

\[ L \]

\[ \leftrightarrow^* \]

\[ N \]

\[ \downarrow \]

\[ \leftrightarrow^* \]
Reduction: Confluence

Definition 11.6 \( \rightarrow \) is confluent if for all \( L, M, N \),
\[
N \leftarrow L \rightarrow^* M \Rightarrow \exists P : N \rightarrow^* P \leftarrow^* M
\]
which we denote as

Fact 11.7 Any confluent relation is also semi-confluent.
Reduction: Church-Rosser

Definition 11.8 $\rightarrow$ is Church-Rosser if for all $L, M$,

$L \leftrightarrow^{*} M \Rightarrow \exists P : L \rightarrow^{*} P \leftarrow^{*} M$

which we denote by

$L \leftrightarrow^{*} M \Rightarrow \exists P : L \rightarrow^{*} P \leftarrow^{*} M$
Equivalence Characterization

Lemma 11.9

1. $\leftrightarrow^*$ is the least equivalence containing $\to$.

2. $\leftrightarrow^*$ is the least equivalence containing $\to^*$. 

3. $L \leftrightarrow^* M$ if and only if there exists a finite sequence $L \equiv M_0, M_1, \ldots M_m \equiv M$, $m \geq 0$ such that for each $i$, $0 \leq i < m$, $M_i \to M_{i+1}$ or $M_{i+1} \to M_i$. We represent this fact more succinctly as

$$L \equiv \alpha M_0 \to / \leftarrow M_1 \to / \leftarrow \cdots \to / \leftarrow M_m \equiv \alpha M$$ (4)
Proof of lemma 11.9

*Proof:*

1. Just prove that $\leftrightarrow^*$ is a subset of every equivalence that contains $\rightarrow$.
2. Use induction on the length of proofs to prove this part
3. For the last part it is easy to see that the existence of the “chain equation” (4) implies $L \leftrightarrow^* M$ by transitivity. For the other part use induction on the length of the proof.

QED

12. The Church-Rosser Property
Parallel Beta Reduction

Definition 12.1 The parallel-$\beta$ or $\parallel\beta$ reduction is the smallest relation for which the following rules hold.

\[\parallel\beta_1\quad L \rightarrow^1_\beta L\]

\[\parallel\beta_1\text{App}\quad \frac{L \rightarrow^1_\beta L', M \rightarrow^1_\beta M'}{\langle L \, M \rangle \rightarrow^1_\beta \langle L' \, M' \rangle}\]

\[\parallel\beta_1\text{Abs1}\quad \frac{L \rightarrow^1_\beta L'}{\lambda x[L] \rightarrow^1_\beta \lambda x[L']}\]

\[\parallel\beta_1\text{Abs2}\quad \frac{L \rightarrow^1_\beta L', M \rightarrow^1_\beta M}{\langle \lambda x[L] \, M \rangle \rightarrow^1_\beta \{M'/x\}}\]
Parallel Beta: The Diamond Property

Lemma 12.2
1. $L \xrightarrow{\beta} L' \Rightarrow L \xrightarrow{\\beta} L'$.

2. $L \xrightarrow{\\beta} L' \Rightarrow L \xrightarrow{\* \beta} L'$.

3. The smallest preorder containing $\xrightarrow{\beta}$ is $\xrightarrow{\* \beta} = \xrightarrow{\* \beta}$.

4. If $L \xrightarrow{\beta} L'$ and $M \xrightarrow{\\beta} M'$ then ${M/x}L \xrightarrow{\beta} {M'/x}L'$.

Proof: By induction on the structure of terms or by induction on the number of steps in any proof. QED

Theorem 12.3 $\xrightarrow{\beta}$ has the diamond property.
Proof of theorem 12.3

Proof: We need to prove for all $L$

$$N \xrightarrow{\beta} L \xrightarrow{\beta} M \Rightarrow \exists P : N \xrightarrow{\beta} P \xrightarrow{\beta} M$$

We prove this by induction on the structure of $L$ and a case analysis of the rule applied in definition 12.1.

Case $L \equiv x \in V$. Then $L \equiv M \equiv N \equiv P$.

Before dealing with the other inductive cases we dispose of some trivial sub-cases that arise in some or all of them.

Case $L \equiv_{\alpha} M$. Choose $P \equiv_{\alpha} N$ to complete the diamond.

Case $L \equiv_{\alpha} N$. Then choose $P \equiv_{\alpha} M$.

Case $M \equiv_{\alpha} N$. Then there is nothing to prove.

In the sequel we assume $N \not\equiv_{\alpha} L \not\equiv_{\alpha} M \not\equiv_{\alpha} N$ and proceed by induction on the structure of $L$.

Case $L \equiv \lambda x[L_1]$. Then clearly $M$ and $N$ were both obtained in proofs whose last step was an application of rule $\beta_1 Abs$ and so $M \equiv \lambda x[M_1]$ and $N \equiv \lambda x[N_1]$ for some $M_1$ and $N_1$ respectively and hence $N_1 \xrightarrow{\beta} L_1 \xrightarrow{\beta} M_1$. By the induction hypothesis we have

$$\exists P_1 : N_1 \xrightarrow{\beta} P_1 \xrightarrow{\beta} M_1$$

Hence by choosing $P \equiv \lambda x[P_1]$ we obtain the required result.

Case $L \equiv (L_1 L_2)$ and $L_1$ is not an abstraction.

The rule $\beta_1 App$ is the only rule that must have been applicable in the last step of the proofs of $N \xrightarrow{\beta} L \xrightarrow{\beta} M$. Clearly then there exist $M_1, M_2, N_1, N_2$ such that $N_1 \xrightarrow{\beta} L_1 \xrightarrow{\beta} M_1$ and $N_2 \xrightarrow{\beta} L_2 \xrightarrow{\beta} M_2$. Again by the induction hypothesis, we have

$$\exists P_1 : N_1 \xrightarrow{\beta} P_1 \xrightarrow{\beta} M_1$$

and

$$\exists P_2 : N_2 \xrightarrow{\beta} P_2 \xrightarrow{\beta} M_2$$

By choosing $P \equiv (P_1 P_2)$ we obtain the desired result.

Case $L \equiv (\lambda x[L_1] L_2)$.

Here we have four sub-cases depending upon whether each of $M$ and $N$ were obtained by an application of $\beta_1 App$ or $\beta_1 Abs$. Of these
the sub-case when both \( M \) and \( N \) were obtained by applying \( \parallel \beta_1 App \) is easy and similar to the previous case. That leaves us with three subcases.

**Sub-case: Both \( M \) and \( N \) were obtained by applying rule \( \parallel \beta_1 Abs \).**

Then we have

\[
\{N_2/x\}N_1 \equiv N \xrightarrow{1}_\beta L \equiv (\lambda x[L_1] L_2) \rightarrow^{1}_\beta M \equiv \{M_2/x\}M_1
\]

for some \( M_1, M_2, N_1, N_2 \) such that

\[
N_1 \xrightarrow{1}_\beta L_1 \rightarrow^{1}_\beta M_1
\]

and

\[
N_2 \xrightarrow{1}_\beta L_2 \rightarrow^{1}_\beta M_2
\]

By the induction hypothesis

\[
\exists P_1 : N_1 \rightarrow^{1}_\beta P_1 \xrightarrow{1}_\beta M_1
\]

and

\[
\exists P_2 : N_2 \rightarrow^{1}_\beta P_2 \xrightarrow{1}_\beta M_2
\]

and the last part of lemma 12.2 we have

\[
\exists P \equiv \{P_2/x\}P_1 : N \rightarrow^{1}_\beta P \xrightarrow{1}_\beta M
\]

completing the proof.

**Sub-case: \( M \) was obtained by applying rule \( \parallel \beta_1 Abs \) and \( N \) by \( \parallel \beta_1 App \).**

Then we have the form

\[
(\lambda x[N_1] N_2) \equiv N \xrightarrow{1}_\beta L \equiv (\lambda x[L_1] L_2) \rightarrow^{1}_\beta M \equiv \{M_2/x\}M_1
\]

where again

\[
N_1 \xrightarrow{1}_\beta L_1 \rightarrow^{1}_\beta M_1
\]

and

\[
N_2 \xrightarrow{1}_\beta L_2 \rightarrow^{1}_\beta M_2
\]

By the induction hypothesis

\[
\exists P_1 : N_1 \rightarrow^{1}_\beta P_1 \xrightarrow{1}_\beta M_1
\]

and

\[
\exists P_2 : N_2 \rightarrow^{1}_\beta P_2 \xrightarrow{1}_\beta M_2
\]
and finally we have

\[ \exists P \equiv \{P_2/x\}P_1 : N \rightarrow^1_{\beta} P \rightarrow^1_{\beta} M \]

completing the proof.

*Sub-case: M was obtained by applying rule \( ||_{\beta} App \) and N by \( ||_{\beta} Abs2. \)
Similar to the previous sub-case.

QED

**Theorem 12.4** \( \rightarrow^1_{\beta} \) is confluent.

**Proof:** We need to show that for all \( L, M, N, \)

\[ N \rightarrow^*_{\beta} L \rightarrow^*_{\beta} M \Rightarrow \exists P : N \rightarrow^*_{\beta} P \rightarrow^*_{\beta} M \]

We prove this by induction on the length of the sequences

\[ L \rightarrow^1_{\beta} M_1 \rightarrow^1_{\beta} M_2 \rightarrow^1_{\beta} \cdots \rightarrow^1_{\beta} M_m \equiv M \]

and

\[ L \rightarrow^1_{\beta} N_1 \rightarrow^1_{\beta} N_2 \rightarrow^1_{\beta} \cdots \rightarrow^1_{\beta} N_n \equiv N \]

where \( m, n \geq 0. \) More specifically we prove this by induction on the pairs of integers \((j, i)\) bounded by \((n, m)\), where \((j, i) < (j', i')\) if and only if either \( j < j' \) or \( (j = j') \) and \( i < i' \). The interesting cases are those where both \( m, n > 0. \) So we repeatedly apply theorem 12.3 to complete the rectangle

\[
\begin{array}{ccccccc}
L & \rightarrow^1_{\beta} & M_1 & \rightarrow^1_{\beta} & M_2 & \rightarrow^1_{\beta} & \cdots & \rightarrow^1_{\beta} & M_m \equiv M \\
N_1 & \rightarrow^1_{\beta} & P_{11} & \rightarrow^1_{\beta} & P_{12} & \rightarrow^1_{\beta} & \cdots & \rightarrow^1_{\beta} & P_{1m} \\
N_2 & \rightarrow^1_{\beta} & P_{21} & \rightarrow^1_{\beta} & P_{22} & \rightarrow^1_{\beta} & \cdots & \rightarrow^1_{\beta} & P_{2m} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
N_n & \rightarrow^1_{\beta} & P_{n1} & \rightarrow^1_{\beta} & P_{n2} & \rightarrow^1_{\beta} & \cdots & \rightarrow^1_{\beta} & P_{nm} \equiv P \\
\end{array}
\]

QED

**Corollary 12.5** \( \rightarrow^1_{\beta} \) is confluent.
Proof: Since \( \rightarrow^* \beta = \rightarrow^{\|}\beta \) it follows from theorem 12.4 that \( \rightarrow^1 \beta \) is confluent.

QED

Corollary 12.6 If a term reduces to a \( \beta \)-normal form then the normal form is unique (upto \( \equiv \alpha \)).

Proof: If \( N_1 \xrightarrow{\beta} L \xrightarrow{\beta} N_2 \) and both \( N_1 \) and \( N_2 \) are \( \beta \)-nfs, then by the corollary 12.5 they must both be \( \beta \)-reducible to a third element \( N_3 \) which is impossible if both \( N_1 \) and \( N_2 \) are \( \beta \)-nfs. Hence \( \beta \)-nfs are unique whenever they exist.

QED

Corollary 12.7 \( \rightarrow^1 \beta \) is Church-Rosser.

Proof: Follows from corollary 12.5 and theorem 13.2.

QED

13. Confluence Characterizations
The Church-Rosser Property
Confluence and Church-Rosser

Lemma 13.1 *Every confluent relation is also semi-confluent*

Theorem 13.2 *The following statements are equivalent for any reduction relation \( \rightarrow \).

1. \( \rightarrow \) is Church-Rosser.
2. \( \rightarrow \) is confluent.
Proof of theorem 13.2

Proof: (1 ⇒ 2) Assume → is Church-Rosser and let

\[ N \xleftarrow{*} L \rightarrow{*} M \]

Clearly then \( N \xleftarrow{*} M \). If → is Church-Rosser then

\[ \exists P : N \rightarrow{*} P \xleftarrow{*} M \]

which implies that it is confluent.

(2 ⇒ 1) Assume → is confluent and let \( L \xleftarrow{*} M \). We proceed by induction on the length of the chain (4).

\[ L \equiv_a M_0 \rightarrow / \leftarrow M_1 \rightarrow / \leftarrow \cdots \rightarrow / \leftarrow M_m \equiv_a M \]

**Basis.** \( m = 0 \). This case is trivial since for any \( P \), \( L \rightarrow{*} P \) iff \( M \rightarrow{*} P \)

**Induction Hypothesis (IH).**

The claim is true for all chains of length \( k \), \( 0 \leq k < m \).

**Induction Step.** Assume the chain is of length \( m = k + 1 \). i.e.

\[ L \equiv_a M_0 \rightarrow / \leftarrow M_1 \rightarrow / \leftarrow \cdots \rightarrow / \leftarrow M_k \rightarrow / \leftarrow M_{k+1} \equiv_a M \]

**Case** \( M_k \rightarrow M \). Then by the induction hypothesis and semi-confluence we have

\[ L \xleftarrow{*} \quad M_k \]

\[ \exists Q \quad \exists P \]

which proves the claim.

**Case** \( M_k \xleftarrow{} M \). Then the claim follows from the induction hypothesis and the following diagram

\[ L \xleftarrow{*} \quad M_k \leftarrow M \]

\[ \exists P \]
Lemma 13.3 If a terminating relation is locally confluent then it is semi-confluent.

Proof: Assume \( L \rightarrow M \) and \( L \rightarrow^* N \). We need to show that there exists \( P \) such that \( M \rightarrow^* P \) and \( N \rightarrow^* P \). We prove this by induction on the length of \( L \rightarrow^* N \). If \( L \equiv_\alpha N \) then \( P \equiv_\alpha M \), otherwise assume \( L \rightarrow N_1 \rightarrow \cdots \rightarrow N_n = N \) for some \( n > 0 \). By the local confluence we have there exists \( P_1 \) such that \( M \rightarrow^* P_1 \). By successively applying the induction hypothesis we get terms \( P_2, \ldots, P_n \) such that \( P_{j-1} \rightarrow^* P_j \) and \( N_j \rightarrow^* P_j \) for each \( j, 1 \leq j \leq m \). In effect we complete the following rectangle

\[
\begin{array}{cccc}
L & \rightarrow & N_1 & \rightarrow & \cdots & \rightarrow & N_n = M \\
\downarrow & & \downarrow & & \cdots & & \downarrow \\
M & \rightarrow & P_1 & \rightarrow & \cdots & \rightarrow & P_n \\
\end{array}
\]

QED

From lemma 13.3 and theorem 13.2 we have the following theorem.

Theorem 13.4 If a terminating relation is locally confluent then it is confluent.

Proof:
\( \rightarrow \) on \( \Lambda \) is given to be terminating and locally confluent. We need to show that it is confluent. That is for any \( L \), we are given that

1. there is no infinite sequence of reductions of \( L \), i.e. every maximal sequence of reductions of \( L \) is of length \( n \) for some \( n \geq 0 \).
2. \( N_1 \leftarrow^1 L \rightarrow 1 M_1 \Rightarrow \exists P : M_1 \rightarrow^* P \leftarrow N_1 \) (5)

We need to show for any term \( L \) that

\[ N \leftarrow^* L \rightarrow^* M \Rightarrow \exists S : M \rightarrow^* S \leftarrow N \] (6)

Let \( L \) be any term. Consider the graph \( G(L) = (\Gamma(L), \rightarrow^1) \) such that \( \Gamma(L) = \{ M \mid L \rightarrow^* M \} \). Since \( \rightarrow \) is a terminating reduction

Fact 13.5 The graph \( G(L) \) is acyclic for any term \( L \).
If \( G(L) \) is not acyclic, there must be a cycle of length \( k > 0 \) such that \( M_0 \rightarrow^1 M_1 \rightarrow^1 \cdots \rightarrow^1 M_{k-1} \rightarrow^1 M_0 \) which implies there is also an infinite reduction sequence of the form \( L \rightarrow^* M_0 \rightarrow^k M_0 \rightarrow^k \cdots \) which is impossible.

Since there are only a finite number of sub-terms of \( L \) that may be reduced under \( \rightarrow \), for each \( L \) there is a maximum number \( p \geq 0 \), which is the length of the longest reduction sequence.

**Fact 13.6** For every \( M \in \Gamma(L) \),

1. \( G(M) \) is a sub-graph of \( G(L) \) and
2. For every \( M \in \Gamma(L) - \{L\} \), the length of the longest reduction sequence of \( M \) is less than \( p \).

We proceed by induction on \( p \).

**Basis.** \( p = 0 \). Then \( \Gamma(L) = \{L\} \) and there are no reductions possible, so it is trivially confluent.

**Induction Hypothesis (IH).**

For any \( L \) whose longest reduction sequence is of length \( k \), \( 0 \leq k < p \), property (6) holds.

**Induction Step.** Assume \( L \) is a term whose longest reduction sequence is of length \( p > 0 \). Also assume \( N \rightarrow^* L \rightarrow^* M \) i.e. \( \exists m, n \geq 0 : N \rightarrow^m L \rightarrow^m M \).

**Case** \( m = 0 \). If \( m = 0 \) then \( M \equiv_\alpha L \) and hence \( S \equiv_\alpha N \).

**Case** \( n = 0 \). Then \( N \equiv_\alpha L \) and we have \( S \equiv_\alpha M \).

**Case** \( m, n > 0 \). Then consider \( M_1 \) and \( N_1 \) such that

\[
N \rightarrow^* N_1 \rightarrow^1 L \rightarrow^1 M_1 \rightarrow^* M
\]

(7)

See figure (1). By (5), \( \exists P : M_1 \rightarrow^* P \rightarrow^* N_1 \). Clearly \( M_1, N_1, P \in \Gamma(L) - \{L\} \). Hence by fact 13.6, \( G(M_1), G(N_1) \) and \( G(P) \) are all sub-graphs of \( G(L) \) and all their reduction sequences are of length smaller than \( p \). Hence by induction hypothesis, we get

\[
P \rightarrow^* M \rightarrow^* \exists Q : M \rightarrow^* Q \rightarrow^* P
\]

(8)

and

\[
N \rightarrow^* N_1 \rightarrow^* P \Rightarrow \exists R : P \rightarrow^* R \rightarrow^* N
\]

(9)
But by (8) and (9) and the induction hypothesis we have

\[ R \leftarrow^* P \rightarrow^* Q \Rightarrow \exists S : Q \rightarrow^* S \leftarrow^* R \quad (10) \]

Combining (10) with (7), (8) and (9) we get

\[ N \leftarrow^* L \rightarrow^* M \Rightarrow \exists S : M \rightarrow^* S \leftarrow^* N \quad (11) \]

**Theorem 13.7** If a terminating relation is locally confluent then it is Church-Rosser.
Proof: Follows from theorem 13.4 and theorem 13.2 QED
14.1. FL with recursion

An Applied Lambda-Calculus
A Simple Language of Terms: FL

Let $X$ be an infinite collection of variables (names). Consider the language (actually a collection of abstract syntax trees) of terms $T_{\Sigma}(X)$ defined by the following constructors (along with their intended meanings).

<table>
<thead>
<tr>
<th>Construct</th>
<th>Arity</th>
<th>Informal Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$</td>
<td>0</td>
<td>The number 0</td>
</tr>
<tr>
<td>$T$</td>
<td>0</td>
<td>The truth value true</td>
</tr>
<tr>
<td>$F$</td>
<td>0</td>
<td>The truth value false</td>
</tr>
<tr>
<td>$P$</td>
<td>1</td>
<td>The predecessor function on numbers</td>
</tr>
<tr>
<td>$S$</td>
<td>1</td>
<td>The successor function on numbers</td>
</tr>
<tr>
<td>ITE</td>
<td>3</td>
<td>The if-then-else construct (on numbers and truth values)</td>
</tr>
<tr>
<td>IZ</td>
<td>1</td>
<td>The is-zero predicate on numbers</td>
</tr>
<tr>
<td>GTZ</td>
<td>1</td>
<td>The greater-than-zero predicate on numbers</td>
</tr>
</tbody>
</table>
FL: Language, Datatype of Instruction Set?

The set of terms $T_{\Sigma}(X)$ may be alternatively defined by the BNF:

$$t ::= x \in X \mid Z \mid (P \ t) \mid (S \ t) \mid T \mid F \mid (ITE \langle t, t_1, t_0 \rangle) \mid (IZ \ t) \mid (GTZ \ t)$$

- It could be thought of as a user-defined data-type
- It could be thought of as the instruction-set of a particularly simple hardware machine.
- It could be thought of as a simple functional programming language without recursion.
- It is a language with two simple types of data: integers and booleans
Extending the language

To make this simple language safe we require

**Type-checking** : to ensure that arbitrary expressions are not mixed in ways they are not “intended” to be used. For example

- $t$ cannot be a boolean expression in $S(t)$, $P(t)$, $IZ(t)$ and $GTZ(t)$
- $ITE(t, t_1, t_0)$ may be used as a conditional expression for both integers and booleans, but $t$ needs to be a boolean and either both $t_1$ and $t_0$ are integer expressions or both are boolean expressions.

**Functions** : To be a useful programming language we need to be able to define functions.

**Recursion** : to be able to define complex functions in a well-typed fashion. Recursion should also be well-typed
FL: The Power of Functions

To make the language powerful we require the ability to define functions, both non-recursive and recursive. We define an applied lambda-calculus of lambda terms \( \Lambda_\Sigma(X) \) over this set of terms as follows:

\[
L, M ::= t \in T_\Sigma(X) \mid \lambda x[L] \mid (L \ M)
\]
The Normal forms for Integers

We need reduction rules to simplify (non-recursive) expressions.

Zero. \(Z\) is the unique representation of the number 0 and every integer expression that is equal to 0 must be reducible to \(Z\).

Positive integers. Each positive integer \(k\) is uniquely represented by the expression \(S^k(Z)\) where the super-script \(k\) denotes a \(k\)-fold application of \(S\).

Negative integers. Each negative integer \(-k\) is uniquely represented by the expression \(P^k(Z)\) where the super-script \(k\) denotes a \(k\)-fold application of \(P\).

\(\delta\)-rules

\[
\begin{align*}
(P \ (S \ x)) & \longrightarrow_{\delta} x \quad (12) \\
(S \ (P \ x)) & \longrightarrow_{\delta} x \quad (13)
\end{align*}
\]
Reduction Rules for Boolean Expressions

Pure Boolean Reductions. The constructs $T$ and $F$ are the normal forms for boolean values.

\[
\begin{align*}
(\text{ITE} \langle T, x, y \rangle) & \rightarrow_\delta x \\
(\text{ITE} \langle F, x, y \rangle) & \rightarrow_\delta y
\end{align*}
\] (14) (15)

Testing for zero.

\[
\begin{align*}
(\text{IZ Z}) & \rightarrow_\delta T \\
(\text{IZ} (S^k Z)) & \rightarrow_\delta F, k > 0 \\
(\text{IZ} (P^k Z)) & \rightarrow_\delta F, k > 0
\end{align*}
\] (16) (17) (18)
Testing for Positivity

\[(\text{GTZ } Z) \rightarrow \delta F \quad (19)\]
\[(\text{GTZ } (S^k Z)) \rightarrow \delta T, \ k > 0 \quad (20)\]
\[(\text{GTZ } (P^k Z)) \rightarrow \delta F, \ k > 0 \quad (21)\]
Other Non-recursive Operators

We may “program” the other boolean operations as follows:

\[
\text{NOT} \overset{df}{=} \lambda x[\text{ITE} \langle x, F, T \rangle]
\]
\[
\text{AND} \overset{df}{=} \lambda \langle x, y \rangle[\text{ITE} \langle x, y, F \rangle]
\]
\[
\text{OR} \overset{df}{=} \lambda \langle x, y \rangle[\text{ITE} \langle x, T, y \rangle]
\]

We may also “program” the other integer comparison operations as follows:

\[
\text{GEZ} \overset{df}{=} \lambda x[\text{OR} \langle (\text{IZ} x), (\text{GTZ} x) \rangle]
\]
\[
\text{LTZ} \overset{df}{=} \lambda x[\text{NOT} \ (\text{GEZ} x)]
\]
\[
\text{LEZ} \overset{df}{=} \lambda x[\text{OR} \langle (\text{IZ} x), (\text{LTZ} x) \rangle]
\]
Recursion in the Untyped Lambda-calculus

The full power of a programming language will not be realised without a recursion mechanism.
The untyped lambda-calculus has “paradoxical combinators” which behave like recursion operators upto $=\beta$.

**Definition 14.1** A combinator $Y$ is called a fixed-point combinator if for every lambda term $L$, $(Y\ L) = \beta (L\ (Y\ L))$

**Curry’s $Y_C$ combinator**

$$Y_C \overset{df}{=} \lambda f[(C\ C)]$$

where

$$C \overset{df}{=} \lambda x[(f\ (x\ x))]$$

**Turing’s $Y_T$ combinator**

$$Y_T \overset{df}{=} (T\ T)$$

where

$$T \overset{df}{=} \lambda y\ x[(x\ (y\ y\ x))]$$
FL: Adding Recursion

But the various $Y$ combinators unfortunately will not satisfy any typing rules that we may define for the language. Instead it is more convenient to use the fixed-point property and define a new constructor with a $\delta$-rule which satisfies the fixed-point property (definition 14.1).

We extend the language FL with a new constructor

$$L ::= \ldots | (\text{REC } L)$$

and add the fixed point property as a $\delta$-rule

$$(\text{REC } L) \longrightarrow^\delta (L \ (\text{REC } L)) \quad (22)$$
Recursion Example: Addition

Consider addition on integers as a binary operation to be defined in this language. We use the following properties of addition on the integers to define it inductively.

\[
x + y = \begin{cases} 
y & \text{if } x = 0 \\ 
(x - 1) + (y + 1) & \text{if } x > 0 \\ 
(x + 1) + (y - 1) & \text{if } x < 0
\end{cases}
\] (23)
Using the constructors of FL we require that any (curried) definition of addition on numbers should be a solution to the following equation in FL for all (integer) expression values of $x$ and $y$.

$$\text{plusc}\ x\ y =_{\beta\delta} \text{ITE} \langle \langle \text{IZ}\ x\rangle,\ y,\ \text{ITE} \langle \langle \text{GTZ}\ x\rangle,\ \text{plusc}\ (P\ x)\ (S\ y),\ \text{plusc}\ (S\ x)\ (P\ y)\rangle\rangle$$  (24)

Equation (24) may be rewritten using abstraction as follows:

$$\text{plusc} =_{\beta\delta} \lambda x[y[\text{ITE} \langle \langle \text{IZ}\ x\rangle,\ y,\ \text{ITE} \langle \langle \text{GTZ}\ x\rangle,\ \text{plusc}\ (P\ x)\ (S\ y),\ \text{plusc}\ (S\ x)\ (P\ y)\rangle\rangle]]$$  (25)

We may think of equation (25) as an equation to be solved in the unknown variable $\text{plusc}$.

Consider the (applied) $\lambda$-term obtained from the right-hand-side of equation (25) by simply abstracting the unknown $\text{plusc}$.

$$\text{addc} \overset{df}{=} \lambda f[y[\text{ITE} \langle \langle \text{IZ}\ x\rangle,\ y,\ \text{ITE} \langle \langle \text{GTZ}\ x\rangle,\ f\ (P\ x)\ (S\ y),\ f\ (S\ x)\ (P\ y)\rangle\rangle]]$$  (26)

Claim 14.2

$$(\text{REC}\ \text{addc}) \rightarrow_{\delta} (\text{addc}\ (\text{REC}\ \text{addc}))$$  (27)

and hence

$$(\text{REC}\ \text{addc}) =_{\beta\delta} (\text{addc}\ (\text{REC}\ \text{addc}))$$  (28)

Claim 14.3 $(\text{REC}\ \text{addc})$ satisfies exactly the equation (25). That is

$$\langle \langle \text{REC}\ \text{addc}\ x\ y\rangle\rangle =_{\beta\delta} \text{ITE} \langle \langle \text{IZ}\ x\rangle,\ y,\ \text{ITE} \langle \langle \text{GTZ}\ x\rangle,\ \text{addc}\ (P\ x)\ (S\ y),\ \text{addc}\ (S\ x)\ (P\ y)\rangle\rangle\rangle$$  (29)

Hence we may regard $(\text{REC}\ \text{addc})$ where addc is defined by right-hand-side of definition (26) as the required solution to the equation (24) in which $\text{plusc}$ is an unknown.

The abstraction shown in (26) and the claims (14.2) and (14.3) simply go to show that $M =_{\alpha} \lambda f[\{f/z\}L]$ is a solution to the equation $z =_{\beta\delta} L$, whenever such a solution does exist. Further, the claims also show that we may
“unfold” the recursion (on demand) by simply performing the substitution \( \{L/z\}L \) for each free occurrence of \( z \) within \( L \).
Exercise 14.1

1. Prove that the relation $\rightarrow_\delta$ is confluent.

2. The language FL does not have any operators that take boolean arguments and yields integer values. Define a standard conversion function $B2I$ which maps the value $F$ to $Z$ and $T$ to $S(Z)$.

3. Prove that $Y_C$ and $Y_T$ are both fixed-point combinators.

4. Using the combinator add and the other constructs of $\Lambda_\Sigma(X)$ to
   (a) define the equation for products of numbers in the language.
   (b) define the multiplication operation $\text{mult}$ on integers and prove that it satisfies the equation(s) for products.

5. The equation (23) is defined conditionally. However the following is equally valid for all integer values $x$ and $y$.
   $$ x + y = (x - 1) + (y + 1) $$

   (a) Follow the steps used in the construction of $\text{addc}$ to define a new applied $\text{addc'}$ that instead uses equation (30).
   (b) Is $\text{(REC addc')} =_\beta\delta (\text{addc'} (\text{REC addc'}))$?
   (c) Is $\text{addc} =_\beta\delta \text{addc'}$?
   (d) Is $\text{(REC addc)} =_\beta\delta (\text{REC addc'})$?

6. The function $\text{addc}$ was defined in curried form. Use the pairing function in the untyped $\lambda$-calculus, to define
   (a) addition and multiplication as binary functions independently of the existing functions.
   (b) the binary 'curry' function which takes a binary function and its arguments and creates a curried version of the binary function.
Typing FL expressions

We have already seen that the simple language FL has

- A types is an important *attribute* of any variable, constant or expression.

- two kinds of expressions: integer expressions and boolean expressions,

- there are also constructors which take integer expressions as arguments and yield boolean values

- there are also function types which allow various kinds of functions to be defined on boolean expressions and integer expressions.
The Need for typing in FL

Besides the need for type-checking rules on $T_\Sigma(X)$ to prevent illegal constructor operations,

- rules are necessary to ensure that $\lambda$-applications occur only between terms of appropriate types in order to remain meaningful.

- rules are necessary to ensure that all terms have clearly defined types at compile-time so that there are no run-time type violations.
TL: A Simple Language of Types

Consider the following language of types (in fully parenthesized form) defined over an infinite collection \( 'a \in TV \) of type variables. We also have two type constants \texttt{int} and \texttt{bool}.

\[
\sigma, \tau ::= \texttt{int} \mid \texttt{bool} \mid 'a \in TV \mid (\sigma \ast \tau) \mid (\sigma \rightarrow \tau)
\]

Notes.

- \texttt{int} and \texttt{bool} are \textit{type constants}.
- \( \ast \) is the product operation on types and \( \rightarrow \) is function operator on types.
- We require \( \ast \) because of the possibility of defining functions of various kinds of arities in \( \Lambda^\Sigma(X) \).
- \textbf{Precedence.} We assume \( \ast \) has a higher precedence than \( \rightarrow \).
- \textbf{Associativity.} \( \rightarrow \) is \textit{right} associative whereas \( \ast \) is \textit{left} associative.
- In any type expression \( \tau \), \( TV\operatorname{ar}(\tau) \) is the set of type variables
Type-inference Rules: Infrastructure

The question of assigning types to complicated expressions which may have variables in them still remains to be addressed.

Type inferencing. Can be done using type assignment rules, by a recursive travel of the abstract syntax tree.

Free variables (names) are already present in the environment (symbol table).

Constants and Constructors. May have their types either pre-defined or ther may be axioms assigning them types.

Bound variables. May be necessary to introduce “fresh” type variables in the environment.
Type Assignment: Infrastructure

- Assume $\Gamma$ is the environment (an association list) which may be looked up to determine the types of individual names. For each variable $x \in X$, $\Gamma(x)$ yields the type of $x$ i.e. $\Gamma(x) = \sigma$ if $x : \sigma \in \Gamma$.

- For each (sub-)expression in FL we define a set $C$ of type constraints of the form $\sigma = \tau$, where $T$ is the set of type variables used in $C$.

- The type constraints are defined by induction on the structure of the expressions in the language FL.

- The expressions of FL could have free variables. The type of the expression would then depend on the types assigned to the free variables.
Constraint Typing Relation

Definition 14.4 For each term \( L \in \Lambda_{\Sigma}(X) \) the constraint typing relation is of the form

\[
\Gamma \vdash L : \tau \uparrow_T C
\]

where

- \( \Gamma \) is called the context and defines the stack of assumptions that may be needed to assign a type (expression) to the (sub-)expression \( L \).
- \( \tau \) is the type(-expression) assigned to \( L \)
- \( C \) is the set of constraints
- \( T \) is the set of “fresh” type variables used in the (sub-)derivations
Typing axioms: Basic

The following axioms may be applied during the scanning and parsing phases of the compiler to assign types to the individual tokens.

\[
\begin{align*}
\Gamma \vdash Z : \text{int} & \quad \Gamma \vdash T : \text{bool} & \quad \Gamma \vdash F : \text{bool} \\
\Gamma \vdash S : \text{int} \rightarrow \text{int} & \quad \Gamma \vdash P : \text{int} \rightarrow \text{int} \\
\Gamma \vdash IZ : \text{int} \rightarrow \text{bool} & \quad \Gamma \vdash GTZ : \text{int} \rightarrow \text{bool} \\
\Gamma \vdash ITE : \text{bool} \star \text{int} \star \text{int} \rightarrow \text{int} & \quad \Gamma \vdash ITEB : \text{bool} \star \text{bool} \star \text{bool} \rightarrow \text{bool}
\end{align*}
\]

Notice that the constructor ITE is *overloaded* and actually is two constructors ITEI and ITEB. Which constructor is actually used will depend on the context and the type-inferencing mechanism.
Type Rules for FL: 2

\[
\begin{array}{c}
\text{Var} \quad \frac{}{\Gamma \vdash x : \Gamma(x) \triangleright_{\emptyset} \emptyset} \\
\text{Abs} \quad \frac{\Gamma, x : \sigma \vdash L : \tau \triangleright_\Gamma C}{\Gamma \vdash \lambda x[L] : \sigma \rightarrow \tau \triangleright_\Gamma C} \\
\text{App} \quad \frac{\Gamma \vdash L : \sigma \triangleright_{T_1} C_1 \\
\Gamma \vdash M : \tau \triangleright_{T_2} C_2}{\Gamma \vdash (L M) : \arrow{a} \triangleright_{T'} C'}
\end{array}
\]

(Conditions 1. and 2.)

where

- **Condition 1.** \(T_1 \cap T_2 = T_1 \cap TVar(\tau) = T_2 \cap TVar(\sigma) = \emptyset\)

- **Condition 2.** \(\arrow{a} \not\in T_1 \cup T_2 \cup TVar(\sigma) \cup TVar(\tau) \cup TVar(C_1) \cup TVar(C_2)\).

- \(T' = T_1 \cup T_2 \cup \{\arrow{a}\}\)

- \(C' = C_1 \cup C_2 \cup \{\sigma = \tau \rightarrow \arrow{a}\}\)
Example 14.5 Consider the following simple combinator \( C \equiv \lambda x[\lambda y[\lambda z[(x \ (y \ z))))] \) which defines the function composition operator. Since there are three bound variables \( x, y \) and \( z \) we begin with an initial assumption \( \Gamma = 'a, y : 'b, z : 'c \) which assigns arbitrary types to the bound variables, represented by the type variables \( 'a, 'b \) and \( 'c \) respectively. Our inference for the type of \( C \) then proceeds as follows. Note, that since \( C \) has no free variables, its type does not depend on the types of any variables. We expect that at the end of the proof there would be no assumptions.

1. \( x : 'a, y : 'b, z : 'c \vdash x : 'a \triangleright_0 \emptyset \)  
   (Var)

2. \( x : 'a, y : 'b, z : 'c \vdash y : 'b \triangleright_0 \emptyset \)  
   (Var)

3. \( x : 'a, y : 'b, z : 'c \vdash z : 'c \triangleright_0 \emptyset \)  
   (Var)

4. \( x : 'a, y : 'b, z : 'c \vdash (y \ z) : 'd \triangleright \{d\} \ {'b = 'c \rightarrow 'd} \)  
   (App)

5. \( x : 'a, y : 'b, z : 'c \vdash (x \ (y \ z)) : 'e \triangleright \{d, e\} \ {'b = 'c \rightarrow 'd, 'a = 'd \rightarrow 'e} \)  
   (App)

6. \( x : 'a, y : 'b \vdash \lambda z[(x \ (y \ z))] : 'c \rightarrow 'e \triangleright \{d, e\} \ {'b = 'c \rightarrow 'd, 'a = 'd \rightarrow 'e} \)  
   (Abs)

7. \( x : 'a \vdash \lambda x[\lambda y[\lambda z[(x \ (y \ z))]]] : 'b \rightarrow 'c \rightarrow 'e \triangleright \{d, e\} \ {'b = 'c \rightarrow 'd, 'a = 'd \rightarrow 'e} \)  
   (Abs)

8. \( \vdash \lambda x[\lambda y[\lambda z[(x \ (y \ z))]]] : 'a \rightarrow 'b \rightarrow 'c \rightarrow 'e \triangleright \{d, e\} \ {'b = 'c \rightarrow 'd, 'a = 'd \rightarrow 'e} \)  
   (Abs)

Hence \( \lambda x[\lambda y[\lambda z[(x \ (y \ z))]]] : 'a \rightarrow 'b \rightarrow 'c \rightarrow 'e \) subject to the constraints given by \( \{'b = 'c \rightarrow 'd, 'a = 'd \rightarrow 'e\} \) which yields \( \lambda x[\lambda y[\lambda z[(x \ (y \ z))]]] : ('d \rightarrow 'e) \rightarrow ('c \rightarrow 'd) \rightarrow 'c \rightarrow 'e \)
Principal Type Schemes

Definition 14.6 A solution for $\Gamma \vdash L : \tau \searrow^T C$ is a pair $\langle S, \sigma \rangle$ where $S$ is a substitution of type variables in $\tau$ such that $S(\tau) = \sigma$.

- The rules yield a principal type scheme for each well-typed applied $\lambda$-term.
- The term is ill-typed if there is no solution that satisfies the constraints.
- Any substitution of the type variables which satisfies the constraints $C$ is an instance of the most general polymorphic type that may be assigned to the term.
Exercise 14.2

1. The language has several constructors which behave like functions. Derive the following rules for terms in $T_{\Sigma}(X)$ from the basic typing axioms and the rule App.

   \[ \text{Sx} \quad \frac{\Gamma \vdash t : \tau \triangleright_T C}{\Gamma \vdash (St) : \text{int} \triangleright_T C \cup \{\tau = \text{int}\}} \]

   \[ \text{Px} \quad \frac{\Gamma \vdash t : \tau \triangleright_T C}{\Gamma \vdash (Pt) : \text{int} \triangleright_T C \cup \{\tau = \text{int}\}} \]

   \[ \text{IZx} \quad \frac{\Gamma \vdash t : \tau \triangleright_T C}{\Gamma \vdash (IZt) : \text{bool} \triangleright_T C \cup \{\tau = \text{int}\}} \]

   \[ \text{GTZx} \quad \frac{\Gamma \vdash t : \tau \triangleright_T C}{\Gamma \vdash (GTZt) : \text{bool} \triangleright_T C \cup \{\tau = \text{int}\}} \]

   \[ \text{ITEx} \quad \frac{\Gamma \vdash t : \sigma \triangleright_T C \quad \Gamma \vdash t_1 : \tau \triangleright_{T_1} C_1 \quad \Gamma \vdash t_0 : \nu \triangleright_{T_0} C_0}{\Gamma \vdash (\text{ITE}(t, t_1, t_0)) : \tau \triangleright_{T'} C'} \]

   where $T' = T \cup T_1 \cup T_0$ and $C' = C \cup C_1 \cup C_0 \cup \{\sigma = \text{bool}, \tau = \nu\}$

2. Use the rules to define the type of $S$?
3. How would you define a type assignment for the recursive function \texttt{addc} defined by equation (26).

4. Prove that the terms, $\omega = \lambda x[ (x \ x) ]$ and $\Omega = ( \omega \ \omega )$ are ill-typed.

5. Are the following well-typed or ill-typed? Prove your answer.

   (a) $(K \ S)$
   (b) $((K \ S) \ \omega )$
   (c) $(((S \ K) \ K) \ \omega )$
   (d) $(\text{ITE} \ (\text{ITE} \ (x), \ T, \ (K \ x)))$
15. Logic Programming and Prolog

Prolog: Abstract Interpreter

**Algorithm 5** A simple abstract interpreter for Prolog

**Require:** A Prolog program $P$ and ground goal $G$

**Ensure:** *yes* if $P \vdash G$ else *no*

1: resolvent := \{ $G$ \}

2: while $\neg$ empty(resolvent) do

3: Choose goal $A$ from resolvent

4: Choose a *ground* instance of some clause $A' \leftarrow B_1, \ldots, B_k$ from $P$ such that $A \equiv A'$

5: if $A'$ does not exist then

6: exit loop

7: end if

8: resolvent := (resolvent $\setminus \{ A \}) \cup \{ B_1, \ldots, B_k \}$

9: end while

10: if empty(resolvent) then

11: return *yes*

12: else

13: return *no*

14: end if