Scanning and Parsing

Stephen A. Edwards

Columbia University

Summer 2016
The First Question

How do you represent one of many things?

*Compilers should accept many programs; how do we describe which one we want?*
Use continuously varying values?

Very efficient, but has serious noise issues

Edison Model B Home Cylinder phonograph, 1906
The ENIAC: Programming with Spaghetti
Have one symbol per thing?

Works nicely when there are only a few things

Sholes and Glidden Typewriter, E. Remington and Sons, 1874
Have one symbol per thing?

Not so good when there are many, many things

Nippon Typewriter SH-280, 2268 keys
Solution: Use a Discrete Combinatorial System

Use *combinations* of a small number of things to represent (exponentially) many different things.
Every Human Writing System Does This

Hieroglyphics (24+)

Cuneiform (1000 – 300)

Sanskrit (36)

Chinese (214 – 4000)

IBM Selectric (88–96)

Mayan (100)

Roman (21–26)
The Second Question

How do you describe only certain combinations?

Compilers should only accept correct programs; how should a compiler check that its input is correct?
Just List Them?

Gets annoying for large numbers of combinations
Just List Them?

Can be really redundant
Choices: CS Research Jargon Generator

Pick one from each column

<table>
<thead>
<tr>
<th>an integrated</th>
<th>mobile</th>
<th>network</th>
</tr>
</thead>
<tbody>
<tr>
<td>a parallel</td>
<td>functional</td>
<td>preprocessor</td>
</tr>
<tr>
<td>a virtual</td>
<td>programmable</td>
<td>compiler</td>
</tr>
<tr>
<td>an interactive</td>
<td>distributed</td>
<td>system</td>
</tr>
<tr>
<td>a responsive</td>
<td>logical</td>
<td>interface</td>
</tr>
<tr>
<td>a synchronized</td>
<td>digital</td>
<td>protocol</td>
</tr>
<tr>
<td>a balanced</td>
<td>concurrent</td>
<td>architecture</td>
</tr>
<tr>
<td>a virtual</td>
<td>knowledge-based</td>
<td>database</td>
</tr>
<tr>
<td>a meta-level</td>
<td>multimedia</td>
<td>algorithm</td>
</tr>
</tbody>
</table>

E.g., “a responsive knowledge-based preprocessor.”

http://www.cs.purdue.edu/homes/dec/essay.topic.generator.html
SClgen: An Automatic CS Paper Generator

Rooter: A Methodology for the Typical Unification of Access Points and Redundancy

Jeremy Stribling, Daniel Aguayo and Maxwell Krohn

ABSTRACT

Many physicists would agree that, had it not been for congestion control, the evaluation of web browsers might never have occurred. In fact, few hackers worldwide would disagree with the essential unification of voice-over-IP and public-private key pair. In order to solve this riddle, we confirm that SMPs can be made stochastic, cacheable, and interposable.

I. INTRODUCTION

Many scholars would agree that, had it not been for active networks, the simulation of Lamport clocks might never have occurred. The notion that end-users synchronize with the investigation of Markov models is rarely outdated. A theoretical grand challenge in theory is the important unification

The rest of this paper is organized as follows. For starters, we motivate the need for fiber-optic cables work in context with the prior work in this area. To address this obstacle, we disprove that even the tauted autonomous algorithm for the construction of digital-to-analog converters by Jones [10] is NP-complete. Oriented languages can be made signed, decentralized, and compatible. Along these same lines, to accomplish this task, we concentrate our efforts on showing that the famous ubiquitous algorithm for the refinement of read-ahead logging by Sato [14] is linear-time. This may or may not actually hold in reality.

II. ARCHITECTURE

Our research is principled. Consider the early methodology by Martin and Smith; our model is similar, but will actually achieve this goal. Despite the fact that such work in context with the prior work in this area. To address this obstacle, we disprove that even the tauted autonomous algorithm for the construction of digital-to-analog converters by Jones [10] is NP-complete. Oriented languages can be made signed, decentralized, and compatible. Along these same lines, to accomplish this task, we concentrate our efforts on showing that the famous ubiquitous algorithm for the refinement of read-ahead logging by Sato [14] is linear-time. This may or may not actually hold in reality.
hey Jude

don't
- make it bad
- be afraid
- let me down

remember to
- let her into your heart
- let her under your skin

then you
- can start
- begin
- to make it better

better better better better better better waaaaa

na
How about more structured collections of things?

The boy eats hot dogs.
The dog eats ice cream.
Every happy girl eats candy.
A dog eats candy.
The happy happy dog eats hot dogs.

Pinker, *The Language Instinct*
Lexical Analysis
Lexical Analysis (Scanning)

Translate a stream of characters to a stream of tokens

```
foo = a + bar ( 0 , 42 , q ) ;
```

<table>
<thead>
<tr>
<th>Token</th>
<th>Lexemes</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQUALS</td>
<td>=</td>
<td>an equals sign</td>
</tr>
<tr>
<td>PLUS</td>
<td>+</td>
<td>a plus sign</td>
</tr>
<tr>
<td>ID</td>
<td>a foo bar</td>
<td>letter followed by letters or digits</td>
</tr>
<tr>
<td>NUM</td>
<td>0 42</td>
<td>one or more digits</td>
</tr>
</tbody>
</table>
Lexical Analysis

Goal: simplify the job of the parser and reject some wrong programs, e.g.,

```plaintext
%#@$^# !@#$%$
```

is not a C program†

Scanners are usually much faster than parsers.

Discard as many irrelevant details as possible (e.g., whitespace, comments).

Parser does not care that the identifier is “supercalifragilisticexpialidocious.”

Parser rules are only concerned with tokens.

† It is what you type when your head hits the keyboard
Describing Tokens

**Alphabet:** A finite set of symbols
Examples: \{ 0, 1 \}, \{ A, B, C, \ldots, Z \}, ASCII, Unicode

**String:** A finite sequence of symbols from an alphabet
Examples: \( \epsilon \) (the empty string), Stephen, \( \alpha\beta\gamma \)

**Language:** A set of strings over an alphabet
Examples: \( \emptyset \) (the empty language), \{ 1, 11, 111, 1111 \}, all English words, strings that start with a letter followed by any sequence of letters and digits
Operations on Languages

Let \( L = \{ \epsilon, \text{wo} \} \), \( M = \{ \text{man}, \text{men} \} \)

Concatenation: Strings from one followed by the other

\( LM = \{ \text{man}, \text{men}, \text{woman}, \text{women} \} \)

Union: All strings from each language

\( L \cup M = \{ \epsilon, \text{wo}, \text{man}, \text{men} \} \)

Kleene Closure: Zero or more concatenations

\( M^* = \{ \epsilon \} \cup M \cup MM \cup MMM \cdots = \{ \epsilon, \text{man}, \text{men}, \text{manman}, \text{manmen}, \text{menman}, \text{menmen}, \text{manmanman}, \text{manmanmen}, \text{manmenman}, \text{manmenmen}, \ldots \} \)
“*” is named after Stephen Cole Kleene, the inventor of regular expressions, who pronounced his last name “CLAY-nee.”

His son Ken writes “As far as I am aware this pronunciation is incorrect in all known languages. I believe that this novel pronunciation was invented by my father.”
Regular Expressions over an Alphabet $\Sigma$

A standard way to express languages for tokens.

1. $\epsilon$ is a regular expression that denotes $\{\epsilon\}$
2. If $a \in \Sigma$, $a$ is an RE that denotes $\{a\}$
3. If $r$ and $s$ denote languages $L(r)$ and $L(s)$,
   \[
   \begin{align*}
   (r) \mid (s) & \text{ denotes } L(r) \cup L(s) \\
   (r)(s) & \{tu : t \in L(r), u \in L(s)\} \\
   (r)^* & \cup_{i=0}^{\infty} L(r)^i \\
   \text{where} & \quad L(r)^0 = \{\epsilon\} \\
   \text{and} & \quad L(r)^i = L(r)L(r)^{i-1}
   \end{align*}
   \]
### Regular Expression Examples

\[ \Sigma = \{a, b\} \]

<table>
<thead>
<tr>
<th>Regexp.</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a \mid b)</td>
<td>{a, b}</td>
</tr>
<tr>
<td>((a \mid b)(a \mid b))</td>
<td>{aa, ab, ba, bb}</td>
</tr>
<tr>
<td>(a^*)</td>
<td>{(\epsilon), a, aa, aaa, aaaa, \ldots}</td>
</tr>
<tr>
<td>((a \mid b)^*)</td>
<td>{(\epsilon), a, b, aa, ab, ba, bb, aaa, aab, aba, abb, \ldots}</td>
</tr>
<tr>
<td>(a \mid a^*b)</td>
<td>{a, b, ab, aab, aaab, aaaaab, \ldots}</td>
</tr>
</tbody>
</table>
Typical choice: $\Sigma = \text{ASCII characters, i.e., } \{\_!, ",", ",\#, ",$, ",…, ",0, ",1, ",…, ",9, ",…, ",A, ",…, ",Z, ",…, \~\}\\\\
\textbf{letters: } A \mid B \mid \cdots \mid Z \mid a \mid \cdots \mid z \\
\textbf{digits: } 0 \mid 1 \mid \cdots \mid 9 \\
\textbf{identifier: } \text{letter(} \text{letter} \mid \text{digit} )^*
Implementing Scanners Automatically

Regular Expressions (Rules)
→
Nondeterministic Finite Automata
→
Subset Construction
→
Deterministic Finite Automata
→
Tables
Nondeterministic Finite Automata

“All strings containing an even number of 0’s and 1’s”

1. Set of states
   \[ S : \{ \text{A}, \text{B}, \text{C}, \text{D} \} \]

2. Set of input symbols \( \Sigma : \{0, 1\} \)

3. Transition function \( \sigma : S \times \Sigma \epsilon \rightarrow 2^S \)

<table>
<thead>
<tr>
<th>state</th>
<th>( \epsilon )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( \emptyset )</td>
<td>{B}</td>
<td>{C}</td>
</tr>
<tr>
<td>B</td>
<td>( \emptyset )</td>
<td>{A}</td>
<td>{D}</td>
</tr>
<tr>
<td>C</td>
<td>( \emptyset )</td>
<td>{D}</td>
<td>{A}</td>
</tr>
<tr>
<td>D</td>
<td>( \emptyset )</td>
<td>{C}</td>
<td>{B}</td>
</tr>
</tbody>
</table>

4. Start state \( s_0 : \text{A} \)

5. Set of accepting states
   \[ F : \{ \text{A} \} \]
The Language induced by an NFA

An NFA accepts an input string $x$ iff there is a path from the start state to an accepting state that “spells out” $x$.

Show that the string “010010” is accepted.
Translating REs into NFAs (Thompson’s algorithm)

- **Symbol**: $a$
- **Sequence**: $r_1 r_2$
- **Choice**: $r_1 \mid r_2$
- **Kleene Closure**: $(r)^*$
Why So Many Extra States and Transitions?

Invariant: Single start state; single end state; at most two outgoing arcs from any state: helpful for simulation.

What if we used this simpler rule for Kleene Closure?

Now consider $a^* b^*$ with this rule:

Is this right?
Translating REs into NFAs

Example: Translate \((a \mid b)^* abb\) into an NFA. Answer:

![NFA Diagram]

Show that the string "aabb" is accepted. Answer:

![Acceptance Diagram]
Simulating NFAs

Problem: you must follow the “right” arcs to show that a string is accepted. How do you know which arc is right?

Solution: follow them all and sort it out later.

“Two-stack” NFA simulation algorithm:

1. Initial states: the $\epsilon$-closure of the start state
2. For each character $c$,
   - New states: follow all transitions labeled $c$
   - Form the $\epsilon$-closure of the current states
3. Accept if any final state is accepting
Simulating an NFA: \(aabb\), Start
Simulating an NFA: \( aabb, \varepsilon \)-closure
Simulating an NFA: $a \cdot abb$
Simulating an NFA: $a \cdot abb$, $\epsilon$-closure
Simulating an NFA: $aa \cdot bb$
Simulating an NFA: $aa \cdot bb$, $\epsilon$-closure
Simulating an NFA: $aab \cdot b$
Simulating an NFA: $aab \cdot b$, $\epsilon$-closure
Simulating an NFA: $aabb$.
Simulating an NFA: $aabb\cdot$, Done
Deterministic Finite Automata

Restricted form of NFAs:

- No state has a transition on $\epsilon$
- For each state $s$ and symbol $a$, there is at most one edge labeled $a$ leaving $s$.

Differs subtly from the definition used in COMS W3261 (Sipser, *Introduction to the Theory of Computation*)

Very easy to check acceptance: simulate by maintaining current state. Accept if you end up on an accepting state. Reject if you end on a non-accepting state or if there is no transition from the current state for the next symbol.
Deterministic Finite Automata

{  
  type token = ELSE | ELSEIF  
}

rule token =
  parse "else" { ELSE }
  | "elseif" { ELSEIF }

Deterministic Finite Automata

\[
\begin{align*}
\text{rule } \text{token} &= \text{parse } "if" \\text{as } \text{lit} \{ \text{ID} (\text{lit}) \} \\
&\quad \text{as } \text{num} \{ \text{NUM} (\text{num}) \}
\end{align*}
\]
Building a DFA from an NFA

Subset construction algorithm

Simulate the NFA for all possible inputs and track the states that appear.

Each unique state during simulation becomes a state in the DFA.
Subset construction for \((a \mid b)^* abb\)
Subset construction for $(a \mid b)^* abb$
Subset construction for \((a | b)^* abb\)
Subset construction for \((a | b)^* abb\)
Subset construction for \((a \mid b)^* abb\)
Result of subset construction for \((a | b)^* abb\)

Is this minimal?
Minimized result for $(a | b)^* abb$
Problem: Translate \((a \mid b)^* abb\) into an NFA and perform subset construction to produce a DFA.

Solution:

<table>
<thead>
<tr>
<th>NFA State</th>
<th>DFA State</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>{0,1,2,4,7}</td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>{1,2,3,4,6,7,8}</td>
<td>B</td>
<td>B</td>
<td>D</td>
</tr>
<tr>
<td>{1,2,4,5,6,7}</td>
<td>C</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>{1,2,4,5,6,7,9}</td>
<td>D</td>
<td>B</td>
<td>E</td>
</tr>
<tr>
<td>{1,2,4,5,6,7,10}</td>
<td>E</td>
<td>B</td>
<td>C</td>
</tr>
</tbody>
</table>
An DFA can be exponentially larger than the corresponding NFA.

$n$ states versus $2^n$

Tools often try to strike a balance between the two representations.
Lexical Analysis with Ocamllex
Constructing Scanners with Ocamlllex

An example:

```ocaml
{ open Parser }

rule token =
parse [' ' '	' '' '
'] { token lexbuf }
| '+' { PLUS }
| '-' { MINUS }
| '*' { TIMES }
| '/' { DIVIDE }
| ['0'-'9']+ as lit { LITERAL(int_of_string lit) }
| eof { EOF }
```
Ocamlllex Specifications

```ocaml
{ (* Header: verbatim OCaml code; mandatory *) }

(* Definitions: optional *)
let ident = regexp
let ...

(* Rules: mandatory *)
rule entrypoint1 [arg1 ... argn] =
  parse pattern1 { action (* OCaml code *) }
  | ...
  | patternn { action }
and entrypoint2 [arg1 ... argn]} =
  ...
and ...

{ (* Trailer: verbatim OCaml code; optional *) }
}```
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>'c'</td>
<td>A single character</td>
</tr>
<tr>
<td>_</td>
<td>Any character (underline)</td>
</tr>
<tr>
<td>eof</td>
<td>The end-of-file</td>
</tr>
<tr>
<td>&quot;foo&quot;</td>
<td>A literal string</td>
</tr>
<tr>
<td>['1' '5' 'a'–'z']</td>
<td>“1,” “5,” or any lowercase letter</td>
</tr>
<tr>
<td>[^ '0'–'9']</td>
<td>Any character except a digit</td>
</tr>
<tr>
<td>( pattern )</td>
<td>Grouping</td>
</tr>
<tr>
<td>identifier</td>
<td>A pattern defined in the <code>let</code> section</td>
</tr>
<tr>
<td>pattern *</td>
<td>Zero or more patterns</td>
</tr>
<tr>
<td>pattern +</td>
<td>One or more patterns</td>
</tr>
<tr>
<td>pattern ?</td>
<td>Zero or one patterns</td>
</tr>
<tr>
<td>pattern₁ pattern₂</td>
<td><code>pattern₁</code> followed by <code>pattern₂</code></td>
</tr>
<tr>
<td>pattern₁</td>
<td>pattern₂</td>
</tr>
<tr>
<td>pattern as id</td>
<td>Bind the matched pattern to variable <code>id</code></td>
</tr>
</tbody>
</table>
An Example

{ type token = PLUS | IF | ID of string | NUM of int }

let letter = ['a'..'z' 'A'..'Z']
let digit = ['0'..'9']

rule token =
  parse [' ' '
' '	'] { token lexbuf } (* Ignore whitespace *)
  | '+' { PLUS } (* A symbol *)
  | "if" { IF } (* A keyword *)
  | letter (letter | digit | '_')* as id { ID(id) } (* Identifiers *)
  | digit+ as lit { NUM(int_of_string lit) } (* Numeric literals *)
  | "/*" { comment lexbuf } (* C-style comments *)

and comment =
  parse "*/" { token lexbuf } (* Return to normal scanning *)
  | _ { comment lexbuf } (* Ignore other characters *)
Free-Format Languages

Typical style arising from scanner/parser division

Program text is a series of tokens possibly separated by whitespace and comments, which are both ignored.

- keywords (if while)
- punctuation (, ( +)
- identifiers (foo bar)
- numbers (10 -3.14159e+32)
- strings ("A String")
Free-Format Languages

Java    C    C++    C#    Algol    Pascal

Some deviate a little (e.g., C and C++ have a separate preprocessor)

But not all languages are free-format.
FORTRAN 77 is not free-format. 72-character lines:

```
100 IF(IN .EQ. 'Y' .OR. IN .EQ. 'y' .OR. $ IN .EQ. 'T' .OR. IN .EQ. 't') THEN
```

When column 6 is not a space, line is considered part of the previous.

Fixed-length line works well with a one-line buffer.

Makes sense on punch cards.
The Python scripting language groups with indentation

```python
i = 0
while i < 10:
    i = i + 1
    print i      # Prints 1, 2, ..., 10

i = 0
while i < 10:
    i = i + 1
print i      # Just prints 10
```

This is succinct, but can be error-prone.

How do you wrap a conditional around instructions?
Does syntax matter? Yes and no
More important is a language’s semantics—its meaning.
The syntax is aesthetic, but can be a religious issue.
But aesthetics matter to people, and can be critical.
Verbosity does matter: smaller is usually better.
Too small can be problematic: APL is a succinct language with its own character set.
There are no APL programs, only puzzles.
Some syntax is error-prone. Classic fortran example:

```
DO 5 I = 1,25 ! Loop header (for i = 1 to 25)
DO 5 I = 1.25 ! Assignment to variable DO5I
```

Trying too hard to reuse existing syntax in C++:

```
vector< vector<int> > foo;
vector<vector<int>> foo; // Syntax error
```

C distinguishes > and >> as different operators. Bjarne Stroustrup tells me they have finally fixed this.
Modeling Sentences
Simple Sentences Are Easy to Model

The boy eats hot dogs.
The dog eats ice cream.
Every happy girl eats candy.
A dog eats candy.
The happy happy dog eats hot dogs.

Pinker, *The Language Instinct*
If the boy eats hot dogs, then the girl eats ice cream.

Either the boy eats candy, or every dog eats candy.

Does this work?
Automata Have Poor Memories

Want to “remember” whether it is an “either-or” or “if-then” sentence. Only solution: duplicate states.
Automata in the form of Production Rules

Problem: automata do not remember where they’ve been

\[
\begin{align*}
S & \rightarrow \text{Either } A \\
S & \rightarrow \text{If } A \\
A & \rightarrow \text{the } B \\
A & \rightarrow \text{the } C \\
A & \rightarrow a \ B \\
A & \rightarrow a \ C \\
A & \rightarrow \text{every } B \\
A & \rightarrow \text{every } C \\
B & \rightarrow \text{happy } B \\
B & \rightarrow \text{happy } C \\
C & \rightarrow \text{boy } D \\
C & \rightarrow \text{girl } D \\
C & \rightarrow \text{dog } D \\
D & \rightarrow \text{eats } E \\
E & \rightarrow \text{hot dogs } F \\
E & \rightarrow \text{ice cream } F \\
E & \rightarrow \text{candy } F \\
F & \rightarrow \text{or } A \\
F & \rightarrow \text{then } A \\
F & \rightarrow \epsilon
\end{align*}
\]
Solution: Context-Free Grammars

Context-Free Grammars have the ability to “call subroutines:”

\[
\begin{align*}
S & \rightarrow \text{Either } P, \text{ or } P. & \text{Exactly two } Ps \\
S & \rightarrow \text{If } P, \text{ then } P. \\
P & \rightarrow A \ H \ N \ \text{eats } O & \text{One each of } A, H, N, \text{ and } O \\
A & \rightarrow \text{the} \\
A & \rightarrow a \\
A & \rightarrow \text{every} \\
H & \rightarrow \text{happy } H & H \text{ is “happy” zero or more times} \\
H & \rightarrow \varepsilon \\
N & \rightarrow \text{boy} \\
N & \rightarrow \text{girl} \\
N & \rightarrow \text{dog} \\
O & \rightarrow \text{hot dogs} \\
O & \rightarrow \text{ice cream} \\
O & \rightarrow \text{candy}
\end{align*}
\]
A Context-Free Grammar for a Simplified C

\[
\begin{align*}
program & \rightarrow \epsilon \mid program \ vdecl \mid program \ fdecl \\
fdecl & \rightarrow \text{id} ( \text{formals} ) \{ \text{vdecls stmts} \} \\
formals & \rightarrow \text{id} \mid formals , \text{id} \\
vdecls & \rightarrow \text{vdecl} \mid vdecls \ vdecl \\
vdecl & \rightarrow \text{int} \ \text{id} ; \\
stmts & \rightarrow \epsilon \mid stmts \ stmt \\
stmt & \rightarrow \text{expr} ; \mid \text{return} \ \text{expr} ; \mid \{ \ \text{stmts} \} \mid \text{if} ( \ \text{expr} ) \ \text{stmt} \mid \\
& \quad \text{if} ( \ \text{expr} ) \ \text{stmt} \ \text{else} \ \text{stmt} \mid \\
& \quad \text{for} ( \ \text{expr} ; \ \text{expr} ; \ \text{expr} ) \ \text{stmt} \mid \text{while} ( \ \text{expr} ) \ \text{stmt} \\
expr & \rightarrow \text{lit} \mid \text{id} \mid \text{id} ( \ \text{actuals} ) \mid ( \ \text{expr} ) \\
& \quad \text{expr} + \ \text{expr} \mid \text{expr} - \ \text{expr} \mid \text{expr} \ast \ \text{expr} \mid \text{expr} / \ \text{expr} \\
& \quad \text{expr} == \ \text{expr} \mid \text{expr} != \ \text{expr} \mid \text{expr} < \ \text{expr} \mid \text{expr} <= \ \text{expr} \\
& \quad \text{expr} > \ \text{expr} \mid \text{expr} >= \ \text{expr} \mid \text{expr} = \ \text{expr} \\
actuals & \rightarrow \text{expr} \mid actuals , \text{expr}
\end{align*}
\]
Constructing Grammars and Ocamllyacc
Parsing

Objective: build an abstract syntax tree (AST) for the token sequence from the scanner.

\[
2 \times 3 + 4 \quad \Rightarrow \quad \begin{array}{c}
+ \\
/ \\
\times \\
/ \\
2 & 3
\end{array}
\]

Goal: verify the syntax of the program, discard irrelevant information, and “understand” the structure of the program.

Parentheses and most other forms of punctuation removed.
Ambiguity

One morning I shot an elephant in my pajamas.
One morning I shot an elephant in my pajamas. How he got in my pajamas I don’t know. —Groucho Marx
Ambiguity in English

*I shot an elephant in my pajamas*

\[
\begin{align*}
S & \rightarrow NP \ VP \\
VP & \rightarrow V \ NP \\
VP & \rightarrow V \ NP \ PP \\
NP & \rightarrow NP \ PP \\
NP & \rightarrow Pro \\
NP & \rightarrow Det \ Noun \\
NP & \rightarrow Poss \ Noun \\
PP & \rightarrow P \ NP \\
V & \rightarrow shot \\
Noun & \rightarrow elephant \\
Noun & \rightarrow pajamas \\
Pro & \rightarrow I \\
Det & \rightarrow an \\
P & \rightarrow in \\
Poss & \rightarrow my
\end{align*}
\]

Jurafsky and Martin, *Speech and Language Processing*
The Dangling Else Problem

Who owns the else?

```
if (a) if (b) c(); else d();
```

Should this be
```
if (a) if (b) c(); d();
```

or
```
if (a) if (b) c() d();
```

Grammars are usually ambiguous; manuals give disambiguating rules such as C’s:

As usual the “else” is resolved by connecting an else with the last encountered elseless if.
The Dangling Else Problem

Problem comes after matching the first statement. Question is whether an “else” should be part of the current statement or a surrounding one since the second line tells us “stmt ELSE” is possible.
The Dangling Else Problem

Some languages resolve this problem by insisting on nesting everything.
E.g., Algol 68:

```
if a < b then a else b fi;
```

“fi” is “if” spelled backwards. The language also uses do–od and case–esac.
Another Solution to the Dangling Else Problem

Idea: break into two types of statements: those that have a dangling “then” (“dstmt”) and those that do not (“cstmt”). A statement may be either, but the statement just before an “else” must not have a dangling clause because if it did, the “else” would belong to it.

\[
\text{stmt : dstmt} \\
\quad | \quad \text{cstmt}
\]

\[
\text{dstmt : IF expr THEN stmt} \\
\quad | \quad \text{IF expr THEN cstmt ELSE dstmt}
\]

\[
\text{cstmt : IF expr THEN cstmt ELSE cstmt} \\
\quad | \quad \text{other statements...}
\]

We are effectively carrying an extra bit of information during parsing: whether there is an open “then” clause. Unfortunately, duplicating rules is the only way to do this in a context-free grammar.
Ambiguous Arithmetic

Ambiguity can be a problem in expressions. Consider parsing

\[ 3 - 4 \times 2 + 5 \]

with the grammar

\[ e \rightarrow e + e \mid e - e \mid e \times e \mid e / e \mid N \]
Operator Precedence and Associativity

Usually resolve ambiguity in arithmetic expressions
Like you were taught in elementary school:
“My Dear Aunt Sally”
Mnemonic for multiplication and division before addition and subtraction.
Operator Precedence

Defines how “sticky” an operator is.

\[ 1 \times 2 + 3 \times 4 \]

\* at higher precedence than +:
\[(1 \times 2) + (3 \times 4)\]

+ at higher precedence than \*:
\[1 \times (2 + 3) \times 4\]
Associativity

Whether to evaluate left-to-right or right-to-left

Most operators are left-associative

\[
1 - 2 - 3 - 4
\]

(left associative)

\[
((1 - 2) - 3) - 4
\]

(left associative)

(right associative)
Fixing Ambiguous Grammars

A grammar specification:

```
expr :
    expr PLUS expr
| expr MINUS expr
| expr TIMES expr
| expr DIVIDE expr
| NUMBER
```

Ambiguous: no precedence or associativity.

Ocamlyacc’s complaint: “16 shift/reduce conflicts.”
Assigning Precedence Levels

Split into multiple rules, one per level

```
expr : expr PLUS expr  
     | expr MINUS expr  
     | term

term : term TIMES term  
      | term DIVIDE term  
      | atom

atom : NUMBER
```

Still ambiguous: associativity not defined

Ocamlyacc’s complaint: “8 shift/reduce conflicts.”
Assigning Associativity

Make one side the next level of precedence

```
expr : expr PLUS term
    | expr MINUS term
    | term

term : term TIMES atom
    | term DIVIDE atom
    | atom

atom : NUMBER
```

This is left-associative.

No shift/reduce conflicts.
Statement separators/terminators

C uses ; as a statement terminator.

```c
if (a<b)
    printf("a less");
else {
    printf("b"); printf(" less");
}
```

Pascal uses ; as a statement separator.

```pascal
if a < b then
    writeln(’a less’)
else begin
    write(’a’); writeln(’ less’)
end
```

Pascal later made a final ; optional.
Ocamlyacc Specifications

{%
   (* Header: verbatim OCaml; optional *)
%

   /* Declarations: tokens, precedence, etc. */
%

   /* Rules: context-free rules */
%

   (* Trailer: verbatim OCaml; optional *)
%}
Declarations

- %token `symbol` ...
  Define symbol names (exported to .mli file)
- %token `< type > symbol` ...
  Define symbols with attached attribute (also exported)
- %start `symbol` ...
  Define start symbols (entry points)
- %type `< type > symbol` ...
  Define the type for a symbol (mandatory for start)
- %left `symbol` ...
- %right `symbol` ...
- %nonassoc `symbol` ...
  Define precedence and associativity for the given symbols, listed in order from lowest to highest precedence
Rules

- **nonterminal** is the name of a rule, e.g., "program," "expr"
- **symbol** is either a terminal (token) or another rule
- **semantic-action** is OCaml code evaluated when the rule is matched
- In a **semantic-action**, $1$, $2$, ... returns the value of the first, second, ... symbol matched
- A rule may include "\%prec symbol" to override its default precedence
An Example .mly File

```ml
%token <int> INT
%token PLUS MINUS TIMES DIV LPAREN RPAREN EOL

%left PLUS MINUS /* lowest precedence */
%left TIMES DIV
%nonassoc UMINUS /* highest precedence */

%start main /* the entry point */
%type <int> main

%%

main:
    expr EOL { $1 }

expr:
    INT { $1 }
    | LPAREN expr RPAREN { $2 }
    | expr PLUS expr { $1 + $3 }
    | expr MINUS expr { $1 - $3 }
    | expr TIMES expr { $1 * $3 }
    | expr DIV expr { $1 / $3 }
    | MINUS expr %prec UMINUS { - $2 }
```
Parsing Algorithms
There are $O(n^3)$ algorithms for parsing arbitrary CFGs, but most compilers demand $O(n)$ algorithms.

Fortunately, the LL and LR subclasses of CFGs have $O(n)$ parsing algorithms. People use these in practice.
Rightmost Derivation of $\text{Id} \ast \text{Id} + \text{Id}$

$e$

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

At each step, expand the *rightmost* nonterminal.

"handle": The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.
Rightmost Derivation of $\text{Id} \ast \text{Id} + \text{Id}$

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

At each step, expand the rightmost nonterminal.

"handle": The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.
Rightmost Derivation of $\text{Id} \ast \text{Id} + \text{Id}$

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

At each step, expand the rightmost nonterminal.

“handle”: The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.
Rightmost Derivation of $\text{Id} \ast \text{Id} + \text{Id}$

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

At each step, expand the rightmost nonterminal.

"handle": The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.
Rightmost Derivation of \( \text{Id} \ast \text{Id} + \text{Id} \)

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

At each step, expand the rightmost nonterminal.

“handle”: The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.
Rightmost Derivation of $\text{Id} \ast \text{Id} + \text{Id}$

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

At each step, expand the rightmost nonterminal.

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.
Rightmost Derivation of $\text{Id} \ast \text{Id} + \text{Id}$

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

At each step, expand the rightmost nonterminal.

nonterminal

"handle": The right side of a production

Dragon-book style: underline handles

$e \rightarrow t + e \rightarrow t + t \rightarrow t + \text{Id} \rightarrow \text{Id} \ast t + \text{Id} \rightarrow \text{Id} \ast \text{Id} + \text{Id}$
Rightmost Derivation: What to Expand

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} * t$
4: $t \rightarrow \text{Id}$

Expand here ↑

Terminals only
Reverse Rightmost Derivation

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

viable prefixes

<table>
<thead>
<tr>
<th>viable prefixes</th>
<th>terminals</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Id} \ast \text{Id} + \text{Id} )</td>
<td>( \text{Id} \ast \text{Id} + \text{Id} )</td>
</tr>
</tbody>
</table>
Reverse Rightmost Derivation

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

viable prefixes  
\[ \text{Id} \ast \text{Id} + \text{Id} \]

terminals  
\[ \text{Id} \]

\[ t \]
Reverse Rightmost Derivation

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow Id \ast t$
4: $t \rightarrow Id$

Viable prefixes

Terminals
Reverse Rightmost Derivation

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

viable prefixes

terminals
Reverse Rightmost Derivation

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

viable prefixes  \hspace{2cm} \text{terminals}
Reverse Rightmost Derivation

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

---

viable prefixes

---

terminals
Shift/Reduce Parsing Using an Oracle

1: \( e \rightarrow t + e \)

2: \( e \rightarrow t \)

3: \( t \rightarrow \text{Id} * t \)

4: \( t \rightarrow \text{Id} \)

- Stack: \( \text{Id} * \text{Id} + \text{Id} \)
- Input:

- Shift/Reduce: \( e \rightarrow t + e \)
Shift/Reduce Parsing Using an Oracle

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

**Stack**

- \( \text{Id} \ast \text{Id} + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)

**Input**

- \( e \rightarrow t + e \)
- \( t + e \)
- \( t + \text{Id} \)
- \( \text{Id} \ast t + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)

*Shift* 

- \( t + e \)
- \( t + t \)
- \( t + \text{Id} \)
- \( \text{Id} \ast t + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)

*Reduce 4* 

- \( \text{Id} \ast \text{Id} + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)

*Reduce 3* 

- \( \text{Id} \ast \text{Id} + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)

*Reduce 2* 

- \( \text{Id} \ast \text{Id} + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)

*Reduce 1* 

- \( \text{Id} \ast \text{Id} + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)

*Accept*
Shift/Reduce Parsing Using an Oracle

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

Stack

Input
Shift/Reduce Parsing Using an Oracle

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

Input:

```
Id * Id + Id
Id * Id + Id
Id * Id + Id
Id * Id + Id
```

Stack:

```
shift
shift
shift
reduce 4
```
Shift/Reduce Parsing Using an Oracle

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow Id * t$
4: $t \rightarrow Id$

stack

input
Shift/Reduce Parsing Using an Oracle

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

```
stack
input
```

```
Id \ast Id + Id
Id \ast Id + Id
Id \ast Id + Id
Id \ast t + Id
  t + Id
  
  shift
  shift
  shift
  reduce 4
  reduce 3
  shift
```
Shift/Reduce Parsing Using an Oracle

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

Stack:
- $e$
- $t + e$
- $t + t$
- $t + \text{Id}$
- $\text{Id} \ast t + \text{Id}$
- $\text{Id} \ast \text{Id} + \text{Id}$
- $\text{Id} \ast \text{Id} + \text{Id}$
- $\text{Id} \ast \text{Id} + \text{Id}$
- $\text{Id} \ast \text{Id} + \text{Id}$

Input:
- $e$
- $t + e$
- $t + t$
- $t + \text{Id}$
- $\text{Id} \ast t + \text{Id}$
- $\text{Id} \ast \text{Id} + \text{Id}$
- $\text{Id} \ast \text{Id} + \text{Id}$
- $\text{Id} \ast \text{Id} + \text{Id}$

Actions:
- Shift 1
- Shift 2
- Shift 3
- Reduce 4
- Shift 5
- Shift 6
- Shift 7
- Shift 8
Shift/Reduce Parsing Using an Oracle

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$
Shift/Reduce Parsing Using an Oracle

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

Stack:

- \( \text{Id} \ast \text{Id} + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)
- \( \text{Id} \ast t + \text{Id} \)
- \( t + \text{Id} \)
- \( t + \text{Id} \)
- \( t + t \)

Input:

- \( e \)
- \( t + e \)
- \( t + t \)
- \( t + \text{Id} \)
- \( \text{Id} \ast t + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)
- \( \text{Id} \ast \text{Id} + \text{Id} \)

Shift:

- Shift
- Shift
- Shift
- Shift
- Shift
- Shift
- Shift

Reduce:

- Reduce 4
- Reduce 3
- Reduce 4
- Reduce 2
Shift/Reduce Parsing Using an Oracle

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

stack

input

shift
shift
shift
reduce 4
reduce 3
shift
shift
reduce 4
reduce 2
reduce 1
Shift/Reduce Parsing Using an Oracle

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} * t \)
4: \( t \rightarrow \text{Id} \)

Stack:
- \( e \)
- \( t + e \)
- \( t + t \)
- \( t + \text{Id} \)
- \( \text{Id} * t + \text{Id} \)
- \( \text{Id} * \text{Id} + \text{Id} \)

Input:
- \( e \)
- \( t + e \)
- \( t + t \)
- \( t + \text{Id} \)
- \( \text{Id} * \text{Id} + \text{Id} \)

Actions:
- Shift 1
- Shift 2
- Shift 3
- Reduce 4
- Shift 4
- Reduce 3
- Shift 5
- Shift 6
- Reduce 4
- Shift 7
- Reduce 2
- Shift 8
- Reduce 1
- Accept
Handle Hunting

**Right Sentential Form:** any step in a rightmost derivation

**Handle:** in a sentential form, a RHS of a rule that, when rewritten, yields the previous step in a rightmost derivation.

The big question in shift/reduce parsing:

```
When is there a handle on the top of the stack?
```

Enumerate all the right-sentential forms and pattern-match against them? *Usually infinitely many; let’s try anyway.*
Some Right-Sentential Forms and Their Handles

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} * t$
4: $t \rightarrow \text{Id}$
Some Right-Sentential Forms and Their Handles

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

Diagram:

```
  e
 / \      /   \
 t   t+e  t+ t+e
 /     /     /     /
ld*ld*ld*ld*ld*ld
```

```
  t
 /  \
ld*ld*ld*ld*ld*ld
```

```
  ld*ld*ld*ld*ld*ld
```

```
  ld*ld*ld*ld*ld*ld
```

```
  ld*ld*ld*ld*ld*ld
```

```
  ld*ld*ld*ld*ld*ld
```

```
  ld*ld*ld*ld*ld*ld
```

```
  ld*ld*ld*ld*ld*ld
```
Some Right-Sentential Forms and Their Handles

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{ld} \ast t$
4: $t \rightarrow \text{ld}$

Patterns:

$\text{ld} \ast \text{ld} \ast \ldots \ast \text{ld} \ast t \ldots$

$t + t + \ldots + t + e$

$t + t + \ldots + t + \text{ld}$

$t + t + \ldots + t + \text{ld} \ast \text{ld} \ast \ldots \ast \text{ld} \ast t$

$t + t + \ldots + t$
The Handle-Identifying Automaton

Magical result, due to Knuth: An automaton suffices to locate a handle in a right-sentential form.

\[ \text{Id} \ast \text{Id} \ast \cdots \ast \text{Id} \ast t \cdots \]
\[ \text{Id} \ast \text{Id} \ast \cdots \ast \text{Id} \cdots \]
\[ t + t + \cdots + t + e \]
\[ t + t + \cdots + t + \text{Id} \]
\[ t + t + \cdots + t + \text{Id} \ast \text{Id} \ast \cdots \ast \text{Id} \ast t \]
\[ t + t + \cdots + t \]
\[ e \]
Building the Initial State of the LR(0) Automaton

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

Key idea: automata identify viable prefixes of right sentential forms. Each state is an equivalence class of possible places in productions.
At the beginning, any viable prefix must be at the beginning of a string expanded from \( e \). We write this condition as \( e' \rightarrow \epsilon e \)
Building the Initial State of the LR(0) Automaton

1: e → t + e
2: e → t
3: t → Id * t
4: t → Id

Key idea: automata identify viable prefixes of right sentential forms. Each state is an equivalence class of possible places in productions.

At the beginning, any viable prefix must be at the beginning of a string expanded from e. We write this condition “e' → e”

There are two choices for what an e may expand to: t + e and t. So when e' → e, e → t + e and e → t are also true, i.e., it must start with a string expanded from t.
Building the Initial State of the LR(0) Automaton

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

Key idea: automata identify viable prefixes of right sentential forms. Each state is an equivalence class of possible places in productions.

At the beginning, any viable prefix must be at the beginning of a string expanded from $e$. We write this condition “$e' \rightarrow \varepsilon e$”

There are two choices for what an $e$ may expand to: $t + e$ and $t$. So when $e' \rightarrow \varepsilon e$, $e \rightarrow \varepsilon t + e$ and $e \rightarrow \varepsilon t$ are also true, i.e., it must start with a string expanded from $t$.

Also, $t$ must be $\text{Id} \ast t$ or $\text{Id}$, so $t \rightarrow \varepsilon \text{Id} \ast t$ and $t \rightarrow \varepsilon \text{Id}$.

This is a closure, like $\varepsilon$-closure in subset construction.
Building the LR(0) Automaton

\[ \begin{align*}
S_0 : & \quad e' \to \varepsilon_e \\
& \quad e \to \varepsilon_t + e \\
S_1 : & \quad t \to \varepsilon_Id \ast t \\
& \quad t \to \varepsilon_Id
\end{align*} \]

The first state suggests a viable prefix can start as any string derived from \( e \), any string derived from \( t \), or \( \text{Id} \).
Building the LR(0) Automaton

"Just passed a string derived from e"

S7: \( e' \rightarrow e \epsilon \)

S0: \( e \rightarrow \epsilon t \)
\( t \rightarrow \epsilon Id \ast t \)
\( t \rightarrow \epsilon Id \)

"Just passed a prefix ending in a string derived from t"

S2: \( e \rightarrow t \epsilon + e \)
\( e \rightarrow t \epsilon \)

The first state suggests a viable prefix can start as any string derived from \( e \), any string derived from \( t \), or \( Id \).

The items for these three states come from advancing the \( \epsilon \) across each thing, then performing the closure operation (vacuous here).

"Just passed a prefix that ended in an Id"
Building the LR(0) Automaton

In S2, a + may be next. This gives $t + e$.

In S1, * may be next, giving $Id * e$. 

The first state suggests a viable prefix can start as any string derived from $e$, any string derived from $t$, or $Id$.

The items for these three states come from advancing across each thing, then performing the closure operation (vacuous here).
Building the LR(0) Automaton

In S2, a + may be next. This gives \( t + e \). Closure adds 4 more items.

In S1, * may be next, giving \( \text{Id} * t \) and two others.
Building the LR(0) Automaton

- **S0**: $e' \rightarrow e$
- **S1**: $t \rightarrow Id * t$
- **S2**: $e \rightarrow t + e$
- **S3**: $t \rightarrow t * t$
- **S4**: $e \rightarrow t + e$
- **S5**: $t \rightarrow Id * t$
- **S6**: $e \rightarrow t + e$

The first state suggests a viable prefix can start as any string derived from $e$, any string derived from $t$, or $Id$. The items for these three states come from advancing across each thing, then performing the closure operation (vacuous here). In S2, a $+$ may be next. This gives $t + e$. Closure adds 4 more items. In S1, $*$ may be next, giving $Id * t$ and two others.
What to do in each state?

### S1:
- **1:** $e ightarrow t + e$
- **2:** $e ightarrow t$
- **3:** $t ightarrow \text{Id} * t$
- **4:** $t ightarrow \text{Id}$

### Stack

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{Id} * \text{Id} * \cdots * \text{Id}</td>
<td>* \cdots</td>
<td>Shift</td>
</tr>
<tr>
<td>\text{Id} * \text{Id} * \cdots * \text{Id}</td>
<td>+ \cdots</td>
<td>Reduce 4</td>
</tr>
<tr>
<td>\text{Id} * \text{Id} * \cdots * \text{Id}</td>
<td></td>
<td>Reduce 4</td>
</tr>
<tr>
<td>\text{Id} * \text{Id} * \cdots * \text{Id}</td>
<td>\text{Id} \cdots</td>
<td>Syntax Error</td>
</tr>
</tbody>
</table>

### Reducible Strings
- $\text{Id} * \text{Id} * \cdots * \text{Id} * t \cdots$
- $\text{Id} * \text{Id} * \cdots * \text{Id} \cdots$
- $t + t + \cdots + t + e$
- $t + t + \cdots + t + \text{Id}$
- $t + t + \cdots + t + \text{Id} * \text{Id} * \cdots * \text{Id} * t$
- $t + t + \cdots + t$
- $e$
The first function

If you can derive a string that starts with terminal $t$ from some sequence of terminals and nonterminals $\alpha$, then $t \in \text{first}(\alpha)$.

1. Trivially, $\text{first}(X) = \{X\}$ if $X$ is a terminal.
2. If $X \rightarrow \epsilon$, then add $\epsilon$ to $\text{first}(X)$.
3. For each prod. $X \rightarrow Y \cdots$, add $\text{first}(Y) - \{\epsilon\}$ to $\text{first}(X)$.
   
   If $X$ can produce something, $X$ can start with whatever that starts with

4. For each prod. $X \rightarrow Y_1 \cdots Y_k Z \cdots$ where $\epsilon \in \text{first}(Y_i)$ for $i = 1, \ldots, k$, add $\text{first}(Z) - \{\epsilon\}$ to $\text{first}(X)$.

Skip all potential $\epsilon$'s at the beginning of whatever $X$ produces

$\text{first}(\text{Id}) = \{\text{Id}\}$

$\text{first}(t) = \{\text{Id}\}$ because $t \rightarrow \text{Id} * t$ and $t \rightarrow \text{Id}$

$\text{first}(e) = \{\text{Id}\}$ because $e \rightarrow t + e$, $e \rightarrow t$, and $\text{first}(t) = \{\text{Id}\}$.
The follow function

If \( t \) is a terminal, \( A \) is a nonterminal, and \( \cdots \ A t \cdots \) can be derived, then \( t \in \text{follow}(A) \).

1. Add \$ \( (\text{"end-of-input"}) \) to follow\((S)\) (start symbol).
   
   End-of-input comes after the start symbol

2. For each prod. \( \rightarrow \cdots A \alpha \), add first\((\alpha) - \{ \epsilon \} \) to follow\((A)\).

   \( A \) is followed by the first thing after it

3. For each prod. \( A \rightarrow \cdots B \) or \( A \rightarrow \cdots B \alpha \) where \( \epsilon \in \text{first} \( (\alpha) \) \), then add everything in follow\((A)\) to follow\((B)\).

   If \( B \) appears at the end of a production, it can be followed by whatever follows that production

1 : \( e \rightarrow t + e \)  
2 : \( e \rightarrow t \)  
3 : \( t \rightarrow \text{Id} \ast t \)  
4 : \( t \rightarrow \text{Id} \)

\( \text{follow}(e) = \{ \$ \} \)  
\( \text{follow}(t) = \{ \} \)  
\( \text{first}(t) = \{ \text{Id} \} \)  
\( \text{first}(e) = \{ \text{Id} \} \)  

1. Because \( e \) is the start symbol
The follow function

If \( t \) is a terminal, \( A \) is a nonterminal, and \( \ldots At\ldots \) can be derived, then \( t \in \text{follow}(A) \).

1. Add $ ("end-of-input") to \text{follow}(S) (\text{start symbol}).
   \text{End-of-input comes after the start symbol}

2. For each prod. \( \rightarrow \cdots A\alpha \), add \text{first}(\alpha) - \{\epsilon\} to \text{follow}(A).
   \text{A is followed by the first thing after it}

3. For each prod. \( A \rightarrow \cdots B \) or \( A \rightarrow \cdots B\alpha \) where \( \epsilon \in \text{first}(\alpha) \),
   then add everything in \text{follow}(A) to \text{follow}(B).
   \text{If \( B \) appears at the end of a production, it can be followed by whatever follows that production}

\begin{align*}
1 & : e \rightarrow t + e & \text{follow}(e) = \{\}$
2 & : e \rightarrow t & \text{follow}(t) = \{ + \}
3 & : t \rightarrow \text{Id} \ast t & 2. \text{Because } e \rightarrow t + e \text{ and } \text{first}(+) = \{+\}
4 & : t \rightarrow \text{Id} & \text{first}(t) = \{\text{Id}\}
\end{align*}

first(\text{e}) = \{\text{Id}\}
The follow function

If \( t \) is a terminal, \( A \) is a nonterminal, and \( \cdots At\cdots \) can be derived, then \( t \in \text{follow}(A) \).

1. Add $ ("end-of-input") to follow(\( S \)) (start symbol).  
   \text{End-of-input comes after the start symbol}
2. For each prod. \( \rightarrow \cdots A\alpha \), add first(\( \alpha \)) − \{\( \epsilon \}\} to follow(\( A \)).  
   \text{A is followed by the first thing after it}
3. For each prod. \( A \rightarrow \cdots B \) or \( A \rightarrow \cdots B\alpha \) where \( \epsilon \in \text{first}(\( \alpha \)) \),  
   then add everything in follow(\( A \)) to follow(\( B \)). 
   \text{If \( B \) appears at the end of a production, it can be followed by whatever follows that production}

\[
\begin{align*}
1 : e & \rightarrow t + e & \text{follow}(e) = \{$$
2 : e & \rightarrow t & \text{follow}(t) = \{+, \}$$
3 : t & \rightarrow \text{Id} \ast t
4 : t & \rightarrow \text{Id}
\end{align*}
\]

first(\( t \)) = \{\text{Id}\}  
first(\( e \)) = \{\text{Id}\}

3. \text{Because } e \rightarrow t \text{ and } $ \in \text{follow}(e)$
The follow function

If \( t \) is a terminal, \( A \) is a nonterminal, and \( \cdots At\cdots \) can be derived, then \( t \in \text{follow}(A) \).

1. Add $ (“end-of-input”) to follow(\( S \)) (start symbol).
   \textit{End-of-input comes after the start symbol}

2. For each prod. \( \rightarrow \cdots A\alpha \), add first(\( \alpha \)) – \( \{\epsilon\} \) to follow(\( A \)).
   \textit{A is followed by the first thing after it}

3. For each prod. \( A \rightarrow \cdots B \) or \( A \rightarrow \cdots B\alpha \) where \( \epsilon \in \text{first}(\alpha) \), then add everything in follow(\( A \)) to follow(\( B \)).
   \textit{If \( B \) appears at the end of a production, it can be followed by whatever follows that production}

\begin{align*}
1 : e & \rightarrow t + e & \text{follow}(e) = \{\$\} \\
2 : e & \rightarrow t & \text{follow}(t) = \{+,\$\} \\
3 : t & \rightarrow \text{Id} * t \\
4 : t & \rightarrow \text{Id} \\
\text{first}(t) &= \{\text{Id}\} \\
\text{first}(e) &= \{\text{Id}\}
\end{align*}
Converting the LR(0) Automaton to an SLR Table

**State Action Goto**

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Goto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>+ * $</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>s1</td>
<td>7 2</td>
</tr>
</tbody>
</table>

From S0, shift an Id and go to S1; or cross a $t$ and go to S2; or cross an $e$ and go to S7.

follow($e$) = {$\$$}  
follow($t$) = {+, $\$$}
Converting the LR(0) Automaton to an SLR Table

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

follow($e$) = {$}$
follow($t$) = {+, $}$

From S1, shift a * and go to S3; or, if the next input $\in$ follow($t$), reduce by rule 4.
Converting the LR(0) Automaton to an SLR Table

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

follow(\( e \)) = \{\$\}
follow(\( t \)) = \{+,\$\}

From S2, shift a + and go to S4; or, if the next input \( \in \) follow(\( e \)), reduce by rule 2.
Converting the LR(0) Automaton to an SLR Table

State Action Goto

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Goto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>+</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>$</td>
</tr>
<tr>
<td>0</td>
<td>s1</td>
<td>7 2</td>
</tr>
<tr>
<td>1</td>
<td>r4 s3 r4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>s4 r2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>s1</td>
<td></td>
</tr>
</tbody>
</table>

follow(e) = {$}
follow(t) = {+, $}

From S3, shift an Id and go to S1; or cross a t and go to S5.
Converting the LR(0) Automaton to an SLR Table

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

**Follow Set Calculation**

- \( \text{follow}(e) = \{\$\} \)
- \( \text{follow}(t) = \{+, \$\} \)

From S4, shift an \text{Id} and go to S1; or cross an \( e \) or a \( t \).
Converting the LR(0) Automaton to an SLR Table

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \cdot t \)
4: \( t \rightarrow \text{Id} \)

follow \((e) = \{\$\} \)
follow \((t) = \{+, \$\} \)

State | Action | Goto |
--- | --- | --- |
Id | + | $\mid e \ t$
Id | $ | t
Id | $ | t
Id | $ | t
Id | $ | t
Id | $ | t
Id | $ | t

From S5, reduce using rule 3 if the next symbol \( \in \) follow(t).
Converting the LR(0) Automaton to an SLR Table

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

follow($e$) = {$}$
follow($t$) = {+, $}$

From S6, reduce using rule 1 if the next symbol $\in$ follow($e$).
If, in S7, we just crossed an e, accept if we are at the end of the input.
### Shift/Reduce Parsing with an SLR Table

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Goto</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>s1</td>
<td>7 2</td>
</tr>
<tr>
<td>1</td>
<td>r4 s3 r4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>s4 r2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>s1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>s1</td>
<td>6 2</td>
</tr>
<tr>
<td>5</td>
<td>r3 r3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>r1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Look at the state on top of the stack and the next input token. Find the action (shift, reduce, or error) in the table. In this case, shift the token onto the stack and mark it with state 1.
Shift/Reduce Parsing with an SLR Table

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{ld} \ast t \)
4: \( t \rightarrow \text{ld} \)

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Goto</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>s1</td>
<td>7 2</td>
</tr>
<tr>
<td>1</td>
<td>r4 s3 r4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>s4 r2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>s1</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>s1</td>
<td>6 2</td>
</tr>
<tr>
<td>5</td>
<td>r3 r3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>r1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Id</td>
<td>* Id + Id $</td>
<td>Shift, goto 3</td>
</tr>
<tr>
<td>0 Id</td>
<td>* Id + Id $</td>
<td>Shift, goto 3</td>
</tr>
</tbody>
</table>

Here, the state is 1, the next symbol is \( \ast \), so shift and mark it with state 3.
1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow \text{Id} \ast t \)
4: \( t \rightarrow \text{Id} \)

Here, the state is 1, the next symbol is +. The table says reduce using rule 4.
### Shift/Reduce Parsing with an SLR Table

1. \( e \rightarrow t + e \)
2. \( e \rightarrow t \)
3. \( t \rightarrow \text{Id} \ast t \)
4. \( t \rightarrow \text{Id} \)

#### State Action Goto

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>( \text{Id} \ast \text{Id} + \text{Id} $ ) Shift, goto 1</td>
</tr>
<tr>
<td></td>
<td>0 ( \text{Id} )</td>
<td>( * \text{Id} + \text{Id} $ ) Shift, goto 3</td>
</tr>
<tr>
<td></td>
<td>0 ( \text{Id} \ast )</td>
<td>( \text{Id} + \text{Id} $ ) Shift, goto 1</td>
</tr>
<tr>
<td></td>
<td>0 ( \text{Id} \ast \text{Id} \ast \text{Id} )</td>
<td>( + \text{Id} $ ) Reduce 4</td>
</tr>
<tr>
<td></td>
<td>0 ( \text{Id} \ast )</td>
<td>( + \text{Id} $ )</td>
</tr>
</tbody>
</table>

Remove the RHS of the rule (here, just \( \text{Id} \)), observe the state on the top of the stack, and consult the “goto” portion of the table.
Shift/Reduce Parsing with an SLR Table

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Goto</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Id</td>
<td>+</td>
<td>*</td>
<td>$</td>
</tr>
<tr>
<td>0</td>
<td>s1</td>
<td></td>
<td></td>
<td>Shift, goto 1</td>
</tr>
<tr>
<td>1</td>
<td>r4 s3 r4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>s4 r2</td>
<td></td>
<td></td>
<td>Shift, goto 3</td>
</tr>
<tr>
<td>3</td>
<td>s1</td>
<td></td>
<td></td>
<td>Id + Id$</td>
</tr>
<tr>
<td>4</td>
<td>s1</td>
<td></td>
<td></td>
<td>Shift, goto 1</td>
</tr>
<tr>
<td>5</td>
<td>r3 r3</td>
<td></td>
<td></td>
<td>+ Id$</td>
</tr>
<tr>
<td>6</td>
<td>r1</td>
<td></td>
<td></td>
<td>Reduce 4</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>✓</td>
<td></td>
<td>Accept</td>
</tr>
</tbody>
</table>

Here, we push a $t$ with state 5. This effectively “backs up” the LR(0) automaton and runs it over the newly added nonterminal.

In state 5 with an upcoming $+$, the action is “reduce 3.”
### Shift/Reduce Parsing with an SLR Table

1: $e \rightarrow t + e$
2: $e \rightarrow t$
3: $t \rightarrow \text{Id} \ast t$
4: $t \rightarrow \text{Id}$

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Goto</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>s1</td>
<td>e</td>
</tr>
<tr>
<td>1</td>
<td>r4</td>
<td>s3</td>
</tr>
<tr>
<td>2</td>
<td>s4</td>
<td>r2</td>
</tr>
<tr>
<td>3</td>
<td>s1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>s1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>r3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>r1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

This time, we strip off the RHS for rule 3, $\text{Id} \ast t$, exposing state 0, so we push a $t$ with state 2.
**Shift/Reduce Parsing with an SLR Table**

1: \(e \rightarrow t + e\)
2: \(e \rightarrow t\)
3: \(t \rightarrow \text{Id} \ast t\)
4: \(t \rightarrow \text{Id}\)

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Goto</th>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>s1</td>
<td></td>
<td>(\text{Id} + \ast \text{Id} + \text{Id} \ast \text{Id} + \text{Id}) 0</td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 1</td>
<td>Shift, goto 1</td>
</tr>
<tr>
<td>1</td>
<td>r4</td>
<td></td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 0</td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 1</td>
<td>Shift, goto 3</td>
</tr>
<tr>
<td>2</td>
<td>s4</td>
<td></td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 0</td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 1</td>
<td>Shift, goto 1</td>
</tr>
<tr>
<td>3</td>
<td>s1</td>
<td></td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 0</td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 1</td>
<td>Reduce 4</td>
</tr>
<tr>
<td>4</td>
<td>s1</td>
<td></td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 0</td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 1</td>
<td>Reduce 3</td>
</tr>
<tr>
<td>5</td>
<td>r3</td>
<td></td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 0</td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 1</td>
<td>Shift, goto 4</td>
</tr>
<tr>
<td>6</td>
<td>r1</td>
<td></td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 0</td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 1</td>
<td>Reduce 4</td>
</tr>
<tr>
<td>7</td>
<td>√</td>
<td></td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 0</td>
<td>(\text{Id} + \text{Id} + \text{Id} + \text{Id}) 1</td>
<td>Accept</td>
</tr>
</tbody>
</table>

**Notes:**
- **Shift:** Move to the next symbol in the input.
- **Reduce:** Replace a derivation with a single symbol.
- **Accept:** The parser has successfully parsed the input.
L, R, and all that

LR parser: “Bottom-up parser”: 
L = Left-to-right scan, R = (reverse) Rightmost derivation

RR parser: R = Right-to-left scan (from end)
I called them “Australian style”; nobody uses these

LL parser: “Top-down parser”:
L = Left-to-right scan: L = (reverse) Leftmost derivation

LR(1): LR parser that considers next token (lookahead of 1)
LR(0): Only considers stack to decide shift/reduce

SLR(1): Simple LR: lookahead from first/follow rules
Derived from LR(0) automaton

LALR(1): Lookahead LR(1): fancier lookahead analysis
Uses same LR(0) automaton as SLR(1)

Ocamlyacc builds LALR(1) tables.
This is a tricky, but mechanical procedure. The Ocamlyacc parser generator uses a modified version of this technique to generate fast bottom-up parsers.

You need to understand it to comprehend error messages:

**Shift/reduce conflicts are caused by a state like**

\[ t \rightarrow \cdot \text{Else} \cdot s \]
\[ t \rightarrow \cdot \]

If the next token is *Else*, do you reduce it since *Else* may follow a \( t \), or shift it?

**Reduce/reduce conflicts are caused by a state like**

\[ t \rightarrow \text{Id} \ast t \cdot \]
\[ e \rightarrow t + e \cdot \]

Do you reduce by “\( t \rightarrow \text{Id} \ast t \)” or by “\( e \rightarrow t + e \)”?