

# Scanning and Parsing

Stephen A. Edwards

Columbia University

Fall 2011



# The First Question

How do you represent one of many things?

*Compilers should accept many programs;  
how do we describe which one we want?*

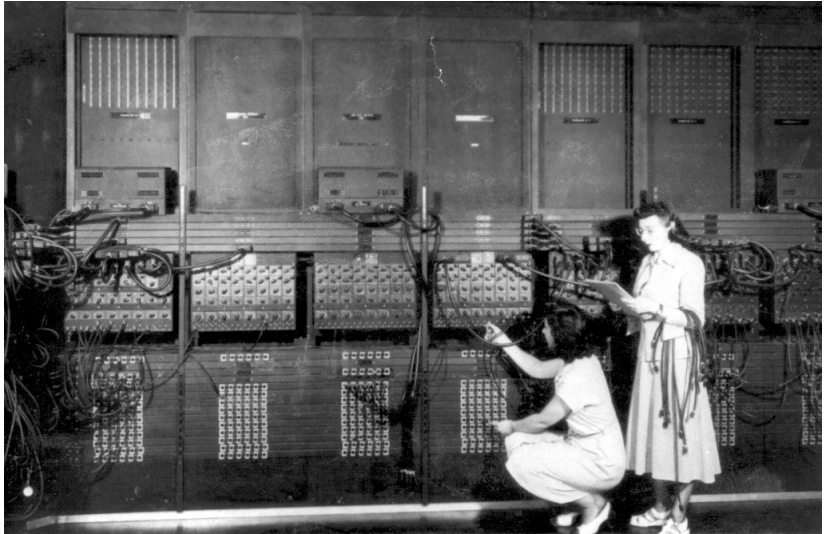
## Use continuously varying values?



Very efficient, but has serious noise issues

Edison Model B Home Cylinder phonograph, 1906

# The ENIAC: Programming with Spaghetti



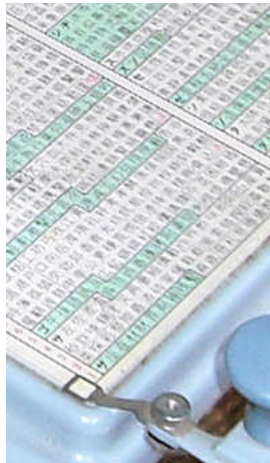
Have one symbol per thing?



Works nicely when there are only a few things

Sholes and Glidden Typewriter, E. Remington and Sons, 1874

Have one symbol per thing?

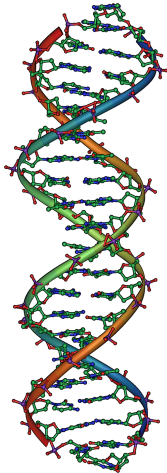


Not so good when there are many, many things

Nippon Typewriter SH-280, 2268 keys

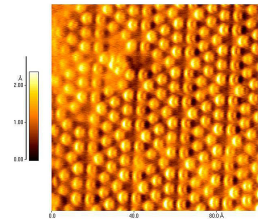
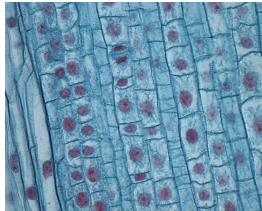
# Solution: Use a Discrete Combinatorial System

Use *combinations* of a *small number of things* to represent (exponentially) many different things.



ENGLISH SOUNDS

chess	rich	book	look	eat	ring		
elephant	camera	bird	ball	clerk	long	phone	
fat	nut	car	lock	valley	knife	cow	
pot	bottle	table	skirt	chopper	jeep	city	ghost
flower	sun	triangle	seashell	snake	nose	church	fish
mouse	acrobatic	key	house	logo	ring	walk	yo-yo



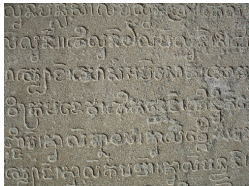
# Every Human Writing System is Discrete Combinatorial



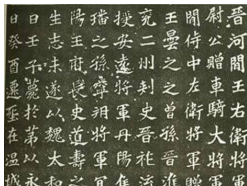
Hieroglyphics



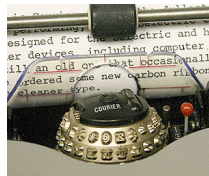
Cuneiform



Sanskrit



Chinese



IBM Selectric



Mayan



Roman

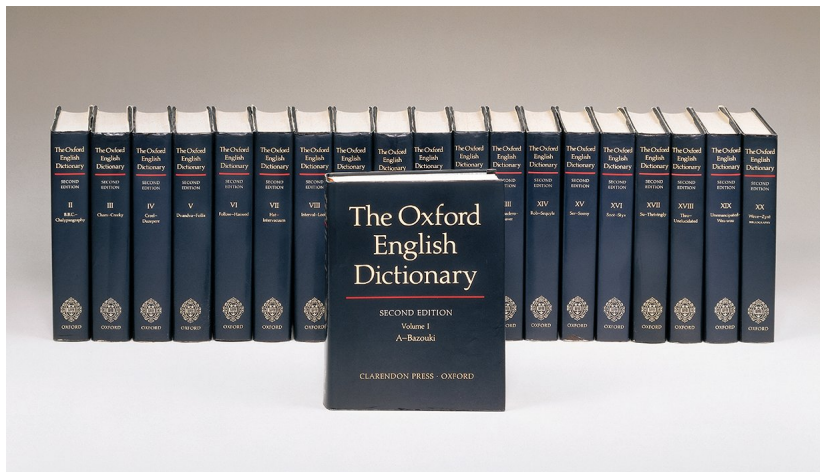


## The Second Question

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?

## 3 AA—AAAAAAAAAA

AAAAA Budget Moving 16 WilbyCr. 241-5468  
 AAAAA Canadian Mini-Warehouse Properties 5399 EglintonW. 620-1577  
 1001 Arrowld. 742-0228  
 24 JeffersonAv. 533-7572  
 4120 Finch. 298-3126  
 AAAAA Critter Control 201-4711  
 100 Burncrest Unionville. 410-8227  
 AAAAA Devoxx Glass 410-0371  
 AAAAA Drainworks Ltd Toronto East. 422-6501  
 AAAAA Evening Rendezvous. 923-6848  
 AAAAA EPI Mini Storage 555 ThreewayDr. 247-6294  
 AAAAA European 962-2033  
 AAAAA Expert Movers 16 Willow. 242-7478  
 AAAAA Jewel Of The Orient 929-9975  
 AAAAA Limousine Connection The. 967-5466  
 AAAAA Mature Escorts 925-5433  
 AAAAA Move Master 588-4656  
 AAAAA Neal Professional Moving Systems 2480 LawrenceAve. 285-6325  
 AAAAA Prince Claude Moving. 287-5701  
 AAAAA Site Stockings. 534-3509  
 AAAAA Woodbine Moving/Storage Ltd 65 Crockford. 751-4900  
 AAAAA Alert Glass/Mirror 638-1958  
 AAAAA All Star Moving 603 Evans. 259-1578  
 AAAAA Armstrong Moving Storage 233-2477  
 AAAAA HSL Moving&Storage 603 Evans. 253-7290  
 AAAAA Midup Moving&Storage 603 Evans. 253-7290  
 AAAAA I Moving&Storage 603 Evans. 253-7290  
 AAAAA Prestige Movers 703 GladstoneAve 533-2633  
 AAAAA South Western Ontario Wildlife Removal. 690-4066  
 AAAAA Speedy Moving 124 Crockford. 285-0084  
 A-A-A-A-A Speedy Moving 1540 VictoriaPark. 751-9532  
 AAAAA Across The World Courier 425 AdelaideW. 504-0008  
 AAAAA Auto Glass 855 Alhess. 663-8676  
 AAAAA California Dreams Escort Service. 323-3899  
 AAAAA California Dreams Massage Service. 323-3899  
 AAAAA National Auto Glass 562 Kipling. 503-3833  
 AAAAA Nights&Day 929-9975  
 AAAAA Strip 'N Tell 964-1787  
 AAAAA Unforgettable Escorts. 398-5337  
 AAAAA Automated Door Systems 32 Juliard. 255-7127  
 AAAAA California Beach Club Service. 323-9822

AAAAA CBS Moving 130 Lansdowne. 533-7139  
 AAAAA Dream Girls. 255-5032  
 AAAAA Big Apple Escort Service. 465-2767  
 AAAAA Accident And Accompanying Injuries/Criminal Practice 1018 FinchW. 663-2211  
 AAAAA Accident And Accompanying Injuries/Criminal Practice 1018 FinchW. 663-2211  
 AAAAA China Blue Escort Service. 323-9522  
 AAAAA AAABCO Door Co 1860 BonhillRd Mississauga. 748-3667  
 AAAAA Action Lawr. 5233 DundasStW. 253-0888  
 AAAAA Alert Auto Glass. 398-4585  
 Or 599-3410  
 AAAAAAAAAMI Campbell Van Lines Inc 1190 MeyersideDr. 213-5660  
 AAAAA Auto Glass Hotline. 288-0042  
 AAAAA Collins&Greig Cartage Ltd 33 Coronet. 239-2991  
 AAAAA Competition Auto Glass. 223-1292  
 AAAAA Competition Auto Glass. 283-0042  
 AAAAA Competition Auto Glass. 410-7693  
 AAAAA International Escorts. 929-6848  
 AAAAA Jewel Dating&Escort Service. 461-0629  
 AAAAA Marketing Services. 413-0444  
 AAAAA Nothing But Class. 595-1884  
 AAAAA On The Wild Side Sensational Female Escort Service. 255-1200  
 AAAAA The Good Life Clubs 21 McCaul. 979-1422  
 1125 FinchW. 967-0574  
 1191 Kennedy. 297-7279  
 302 TheEastWall. 239-2783  
 If Busy Call. 667-0470  
 AAAAA Affordable And Aggressive Defence 4950 YongeSt. 221-7108  
 AAAAA Campbell Moving Systems. 265-4433  
 AAAAA Windshields To Go 159 Dwyer. 787-8039  
 AAAAA Sunset Escorts. 622-1177  
 AAAAA A Best Of The Best. 929-3039  
 AAAAA A Billionaire Towing 286 RoyalYuk. 255-8518

## AAAAAAAAAAAA 4

AAAAAAA Class Above Limousine 173 DanforthAv. 465-5643  
 AAAAAAAA Cross Movers 1232-B Woodbine. 423-0239  
 AAAAAAA Miss Victoria. 967-7176  
 AAAAAAAA Payless Escorts. 485-5333  
 AAAAAAA 700 LawrenceAveW. 256-1600  
 AAAAAAA 699-6700  
 AAAAAAA 441-9500  
 AAAAAAA 323-0907  
 AAAAAAA 287-0000  
 AAAAAAA 242-6662  
 AAAAAAA 366-0237  
 AAAAAAA 964-0138  
 AAAAAAA 944-2018  
 AAAAA Abba Auto Collision. 777-9595  
 AAAAAAA 225-5889  
 AAAAAAA 285-6002  
 AAAAAAA 222-6789  
 If Busy Call. 222-7589  
 33 Isabella. 222-6311  
 55 TownCentre Court. 222-5867  
 AAAAAAA 499-2144  
 AAAAAAA 299-6888  
 AAAAAAA 253-0888  
 AAAAAAA 363-5431  
 AAAAAAA 323-4773  
 AAAAAAA 247-0000  
 AAAAAAA 245-7676

AAAAAAA Towing 18 Cano. 245-7676  
 AAAAAAA 620-1212  
 AAAAAAA 652-5252  
 AAAAAAA Law. 784-2020  
 AAAAAAA Accident Accompanying Injuries/Criminal Practice 1001 FinchW. 663-2211  
 AAAAAAA Accident Claims 2 StClairW. 944-2313  
 AAAAAAA AdvantEdge Door Systems. 222-8322  
 AAAAAAA Advantage Executive's Choice. 929-9390  
 AAAAAAA Automatic Garage Doors 64 Clarkson. 785-7820  
 Etobicoke. 252-5686  
 AAAAAAA 280 Consumers. 494-9777  
 AAAAAAA An Elegant Mature Escorts. 923-3333  
 AAAAAAA Professional Express System 425 AdelaideW. 504-9111  
 AAAAAAA Sweet Escorts&You. 259-3940  
 AAAAA AAAA AAAA Anthony De Marco 1205 StClairW. 651-2299  
 AAAAA AAAA AAAA Domestic Tagliola 1205 StClairW. 651-2299  
 AAAAAAAA Always Available. 465-9191  
 AAAAAAAA Touch Of Class Escort Service. 461-8110  
 AAAAAAAA  
 No Charge-Dial 1 800 506-5665  
 AAAAAAAA Cardinal Custom Building 2 BloorW. 966-4728  
 AAAAA AAAA LLI Stud Movers. 693-2403  
 AAAAA ABC Door Co 1860 BonhillRd Mississauga Toronto 748-3667  
 AAAAAA S Movers 643 LansdowneAve. 588-1499  
 AA AABBCDEF Locksmiths 80 StClairE. 922-2255  
 AAAAA B C Movers Inc 6 Columbus. 535-3413  
 AAAAA B B Best Moving 955 Middlefield. 929-4239  
 AAAAA M O I Moving Systems 97-4664  
 AAAAA B B Moving 500 CaldermillAve. 787-6064  
 AAAAA B B Locksmiths. 287-4001  
 AAAAA ABC Glass Supply 11 Concord. 531-1548  
 AAAAA B C Door&Window Co 1860 BonhillRd Mississauga Toronto 748-3667

Can be really redundant

# Choices: CS Research Jargon Generator

Pick one from each column

an integrated	mobile	network
a parallel	functional	preprocessor
a virtual	programmable	compiler
an interactive	distributed	system
a responsive	logical	interface
a synchronized	digital	protocol
a balanced	concurrent	architecture
a virtual	knowledge-based	database
a meta-level	multimedia	algorithm

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

<http://www.cs.purdue.edu/homes/dec/essay.topic.generator.html>

## Router: 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.

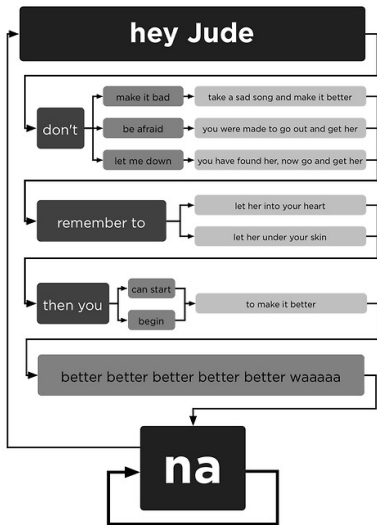
### I. INTRODUCTION

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: we motivate the need for fiber-optic cables in context with the prior work in this field. To address this obstacle, we disprove that even the most touted autonomous algorithm for the construction of analog-to-digital converters by Jones [10] is NP-complete. We show that all Turing-recognizable languages can be made signed, decidable, and signed. Along these same lines, to accomplish our goal, we concentrate our efforts on showing that the fastest algorithm for the exploration of robots by Smith [11] is  $\Omega((n + \log n))$  time [22]. In the end, we conclude.

### II. ARCHITECTURE

Our research is principled. Consider the example of our work by Martin and Smith; our model is similar, but



loveallthis.tumblr.com

lyrics © sony atv

<http://loveallthis.tumblr.com/post/506873221>

## How about more structured collections of things?

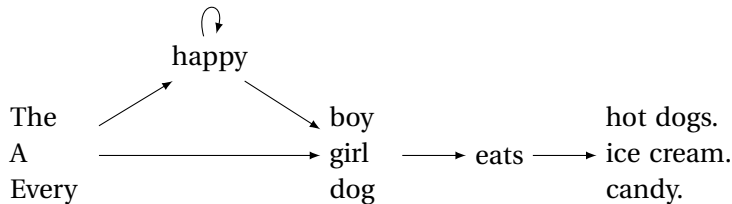
The boy eats hot dogs.

The dog eats ice cream.

Every happy girl eats candy.

A dog eats candy.

The happy happy dog eats hot dogs.



## Part II

### Lexical Analysis



# Lexical Analysis (Scanning)

Translate a stream of characters to a stream of tokens



f o o \_ = \_ a + \_ bar ( 0 , \_ 42 , \_ q ) ;

ID	EQUALS	ID	PLUS	ID	LPAREN	NUM	COMMA
ID	LPAREN	SEMI					

Token	Lexemes	Pattern
EQUALS	=	an equals sign
PLUS	+	a plus sign
ID	a foo bar	letter followed by letters or digits
NUM	0 42	one or more digits

# Lexical Analysis

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

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

is not a C program<sup>†</sup>

Scanners are usually much faster than parsers.

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

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

Parser rules are only concerned with tokens.

<sup>†</sup> It is what you type when your head hits the keyboard

# Describing Tokens

**Alphabet:** A finite set of symbols

Examples:  $\{0, 1\}$ ,  $\{A, B, C, \dots, Z\}$ , 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

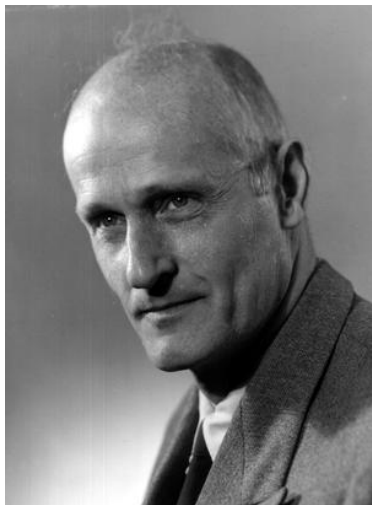
$M^* = \{ \epsilon \} \cup M \cup MM \cup MMM \dots =$

$\{ \epsilon, \text{man}, \text{men}, \text{manman}, \text{manmen}, \text{menman}, \text{menmen},$   
 $\text{manmanman}, \text{manmanmen}, \text{manmenman}, \dots \}$

## Kleene Closure

The asterisk operator (\*) is called the Kleene Closure operator after the inventor of regular expressions, Stephen Cole Kleene, 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  $L(s)$ ,
  - ▶  $(r) | (s)$  denotes  $L(r) \cup L(s)$
  - ▶  $(r)(s)$  denotes  $\{tu : t \in L(r), u \in L(s)\}$
  - ▶  $(r)^*$  denotes  $\cup_{i=0}^{\infty} L^i$  ( $L^0 = \{\epsilon\}$  and  $L^i = LL^{i-1}$ )

# Regular Expression Examples

$\Sigma = \{a, b\}$

<b>RE</b>	<b>Language</b>
$a \mid b$	$\{a, b\}$
$(a \mid b)(a \mid b)$	$\{aa, ab, ba, bb\}$
$a^*$	$\{\epsilon, a, aa, aaa, aaaa, \dots\}$
$(a \mid b)^*$	$\{\epsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, \dots\}$
$a \mid a^*b$	$\{a, b, ab, aab, aaab, aaaab, \dots\}$

## Specifying Tokens with REs

Typical choice:  $\Sigma$  = ASCII characters, i.e.,  
{ $\_$ ,!, ", #, \$, ..., 0, 1, ..., 9, ..., A, ..., Z, ..., ~}

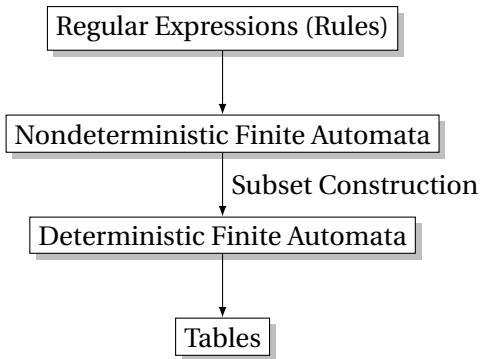
**letters:** A | B |  $\cdots$  | Z | a |  $\cdots$  | z

**digits:** 0 | 1 |  $\cdots$  | 9

**identifier:** letter (letter | digit)\*

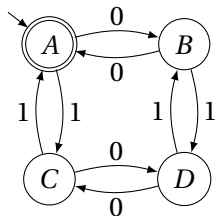


# Implementing Scanners Automatically



# Nondeterministic Finite Automata

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



1. Set of states

$$S: \left\{ \textcircled{\textcircled{A}} \textcircled{B} \textcircled{C} \textcircled{D} \right\}$$

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

3. Transition function  $\sigma : S \times \Sigma_{\epsilon} \rightarrow 2^S$

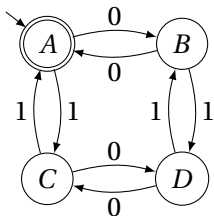
state	$\epsilon$	0	1
A	$\emptyset$	{B}	{C}
B	$\emptyset$	{A}	{D}
C	$\emptyset$	{D}	{A}
D	$\emptyset$	{C}	{B}

4. Start state  $s_0 : \textcircled{\textcircled{A}}$

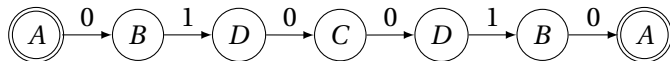
5. Set of accepting states  $F : \left\{ \textcircled{\textcircled{A}} \right\}$

## 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”  $x$ .

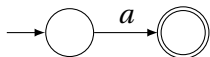


Show that the string “010010” is accepted.



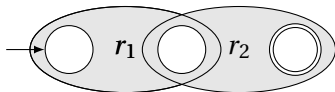
# Translating REs into NFAs

$a$



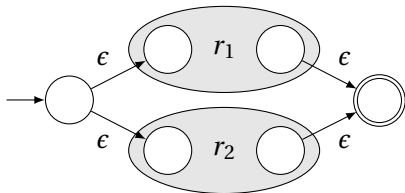
Symbol

$r_1 r_2$



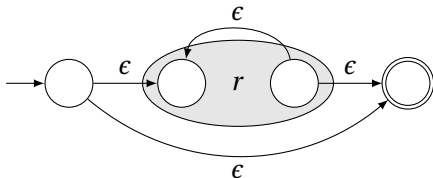
Sequence

$r_1 | r_2$



Choice

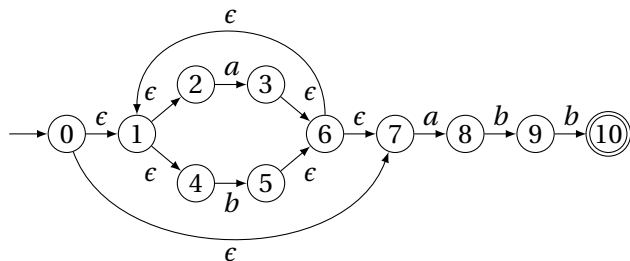
$(r)^*$



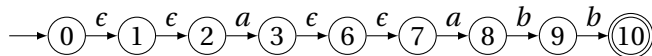
Kleene Closure

## Translating REs into NFAs

Example: Translate  $(a | b)^* abb$  into an NFA. Answer:



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



# Simulating NFAs

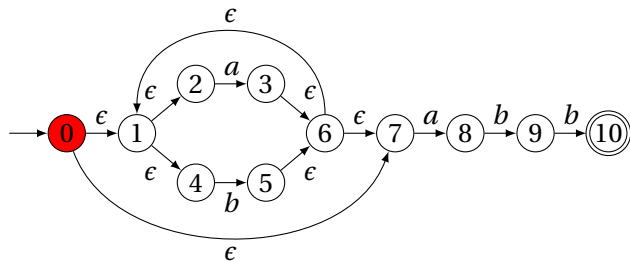
Problem: you must follow the “right” arcs to show that a string is accepted. How do you know which arc is right?

Solution: follow them all and sort it out later.

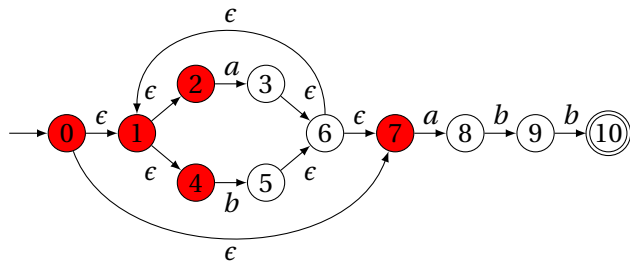
“Two-stack” NFA simulation algorithm:

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: $\cdot aabb$ , Start

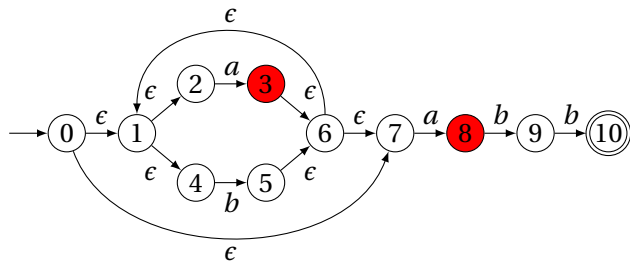


## Simulating an NFA: $\cdot aabb$ , $\epsilon$ -closure

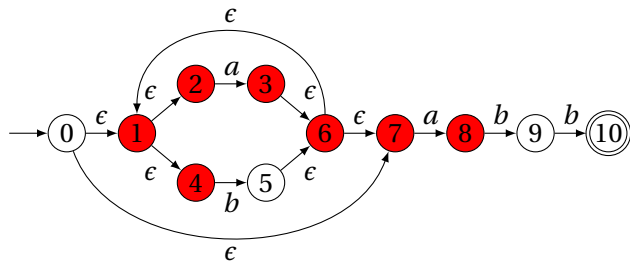




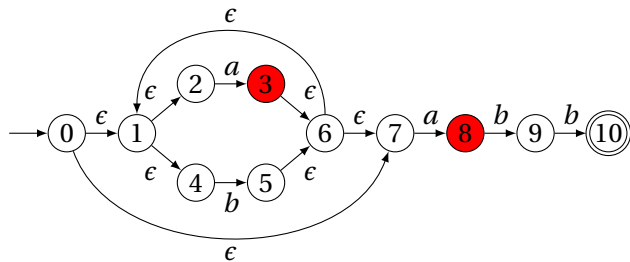
## Simulating an NFA: $a \cdot abb$



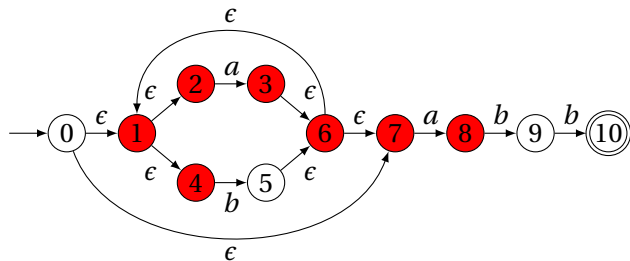
## Simulating an NFA: $a \cdot abb$ , $\epsilon$ -closure



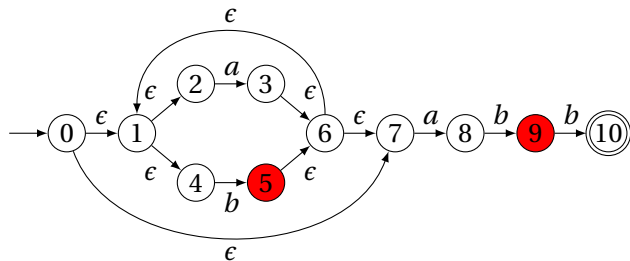
## Simulating an NFA: $aa \cdot bb$



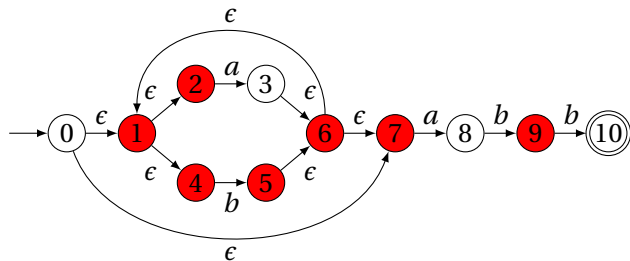
## Simulating an NFA: $aa \cdot bb$ , $\epsilon$ -closure



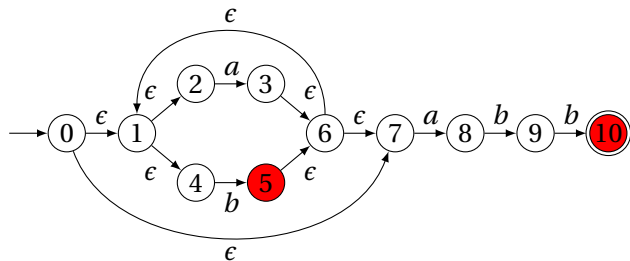
## Simulating an NFA: $aab \cdot b$



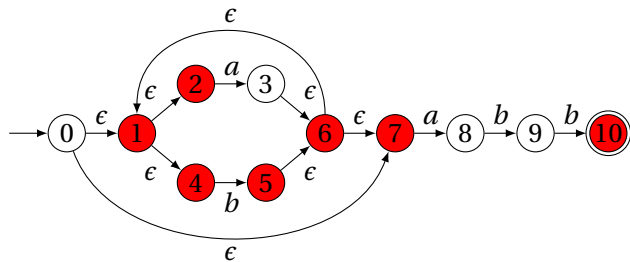
## Simulating an NFA: $aab \cdot b$ , $\epsilon$ -closure



## Simulating an NFA: $aabb$ .



## Simulating an NFA: $aabb\cdot$ , Done





# Deterministic Finite Automata

Restricted form of NFAs:

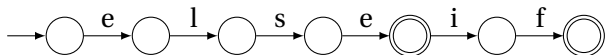
- ▶ No state has a transition on  $\epsilon$
- ▶ For each state  $s$  and symbol  $a$ , there is at most one edge labeled  $a$  leaving  $s$ .

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

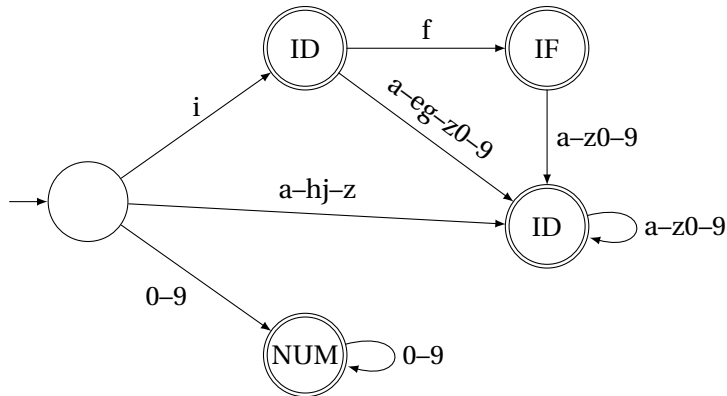
```
{ type token = IF | ID of string | NUM of string }
```

```
rule token =
```

```
  parse "if"
```

```
    | ['a'-'z'] ['a'-'z' '0'-'9']* as lit { ID(lit) }
```

```
    | ['0'-'9']+ as num { NUM(num) }
```



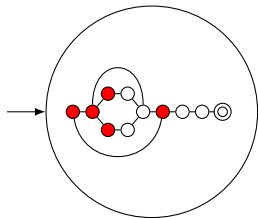
## Building a DFA from an NFA

Subset construction algorithm

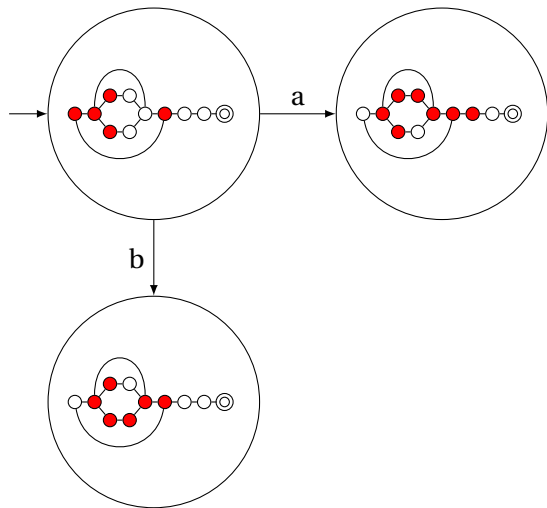
Simulate the NFA for all possible inputs and track the states that appear.

Each unique state during simulation becomes a state in the DFA.

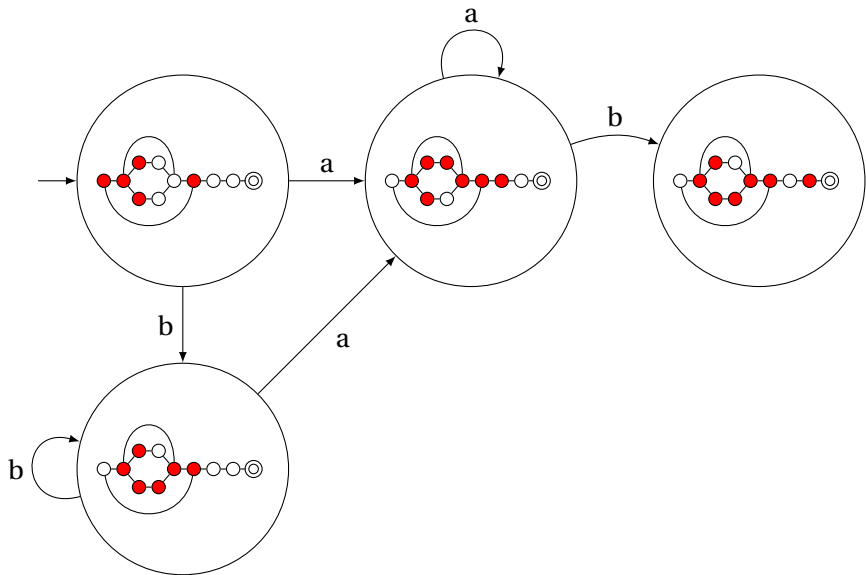
Subset construction for  $(a | b)^* abb$



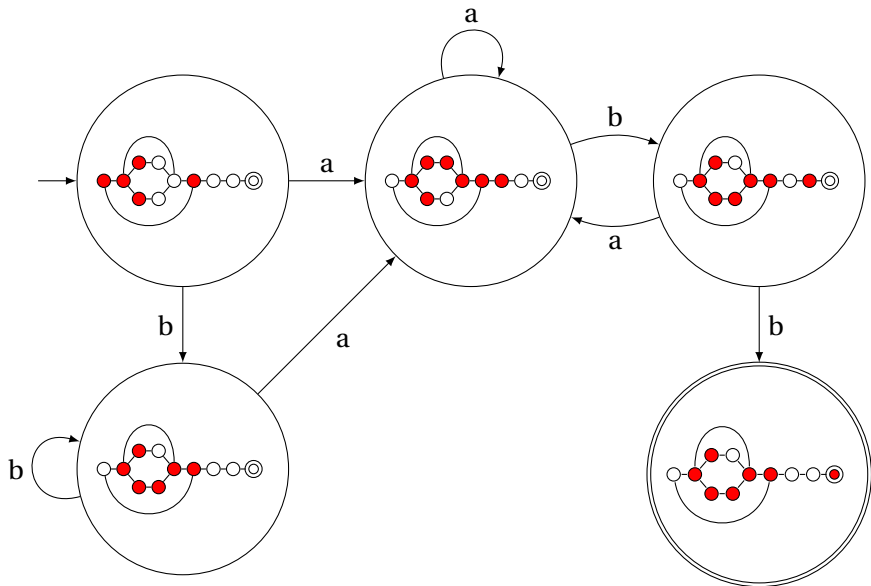
# Subset construction for $(a \mid b)^* abb$



# Subset construction for $(a | b)^* abb$

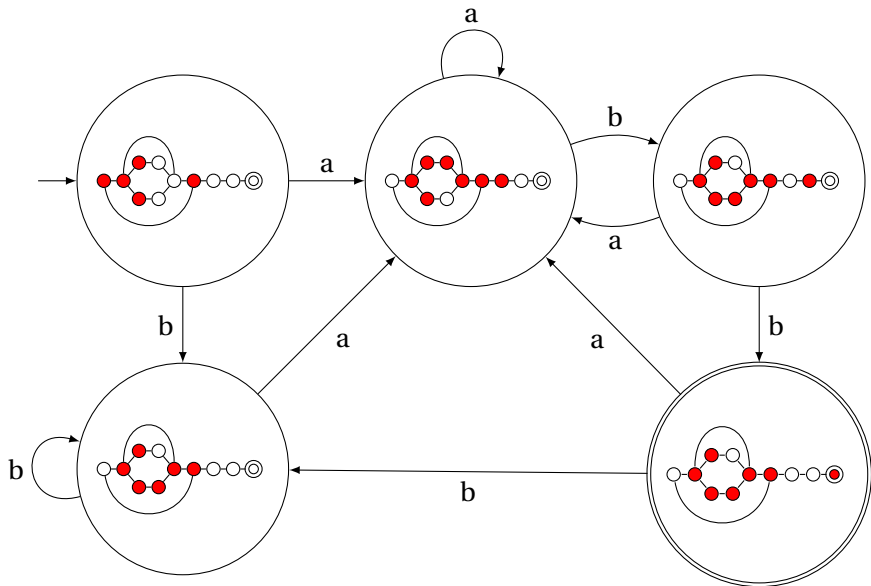


# Subset construction for $(a | b)^* abb$

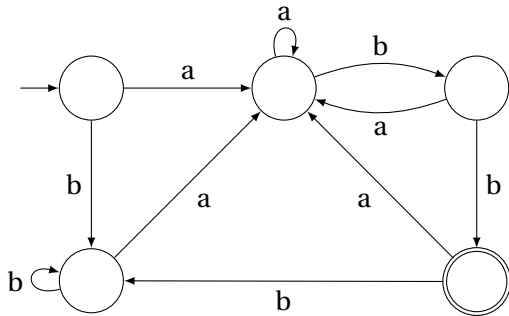




# Subset construction for $(a | b)^* abb$



## Result of subset construction for $(a | b)^* abb$



*Is this minimal?*

## Subset Construction

An DFA can be exponentially larger than the corresponding NFA.

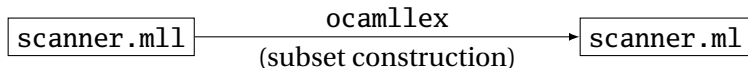
$n$  states versus  $2^n$

Tools often try to strike a balance between the two representations.

## Part III

# Lexical Analysis with Ocamllex

# Constructing Scanners with Ocamllex



An example:

scanner.mll

```
{ open Parser }  
  
rule token =  
  parse [ ' ' '\t' '\r' '\n' ] { token lexbuf }  
    | '+' { PLUS }  
    | '-' { MINUS }  
    | '*' { TIMES }  
    | '/' { DIVIDE }  
    | ['0'-'9']+ as lit { LITERAL(int_of_string lit) }  
    | eof { EOF }
```

# Ocamlex Specifications

```
{
  (* 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 *)
}
```

## Patterns (In Order of Decreasing Precedence)

---

<b>Pattern</b>	<b>Meaning</b>
'c'	A single character
—	Any character (underline)
eof	The end-of-file
"foo"	A literal string
[ '1' '5' 'a' - 'z' ]	"1," "5," or any lowercase letter
[ ^ '0' - '9' ]	Any character except a digit
( <i>pattern</i> )	Grouping
<i>identifier</i>	A pattern defined in the let section
<i>pattern</i> *	Zero or more <i>patterns</i>
<i>pattern</i> +	One or more <i>patterns</i>
<i>pattern</i> ?	Zero or one <i>patterns</i>
<i>pattern</i> <sub>1</sub> <i>pattern</i> <sub>2</sub>	<i>pattern</i> <sub>1</sub> followed by <i>pattern</i> <sub>2</sub>
<i>pattern</i> <sub>1</sub>   <i>pattern</i> <sub>2</sub>	Either <i>pattern</i> <sub>1</sub> or <i>pattern</i> <sub>2</sub>
<i>pattern</i> as <i>id</i>	Bind the matched pattern to variable <i>id</i>

---

## 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 [' ' '\n' '\t'] { token lexbuf } (* Ignore whitespace *)

  | '+' { PLUS } (* A symbol *)

  | "if" { IF } (* A keyword *)
  (* Identifiers *)
  | letter (letter | digit | '_')* as id { ID(id) }
  (* Numeric literals *)
  | digit+ as lit { NUM(int_of_string lit) }

  | "/*" { 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

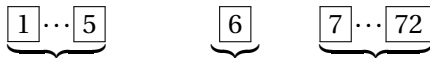
Some deviate a little (e.g., C and C++ have a separate preprocessor)

But not all languages are free-format.

# FORTRAN 77

FORTRAN 77 is not free-format. 72-character lines:

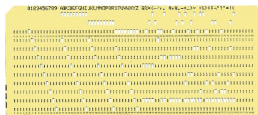
```
100  IF(IN .EQ. 'Y' .OR. IN .EQ. 'y' .OR.  
    $  IN .EQ. 'T' .OR. IN .EQ. 't') THEN
```



Statement label Continuation Normal

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.

# Python

The Python scripting language groups with indentation

```
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?

# Syntax and Language Design

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. Here is program that returns all primes  $\leq R$

**PRIMES : ( R ∈ R ° . × R ) / R ← 1 ↓ ∩ R**

There are no APL programs, only puzzles.

# Syntax and Language Design

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 D05I
```

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.

## Part IV

# Modeling Sentences

## Simple Sentences Are Easy to Model

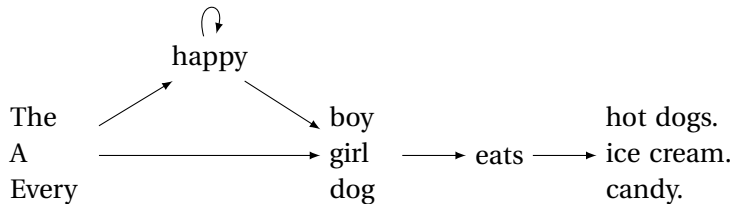
The boy eats hot dogs.

The dog eats ice cream.

Every happy girl eats candy.

A dog eats candy.

The happy happy dog eats hot dogs.

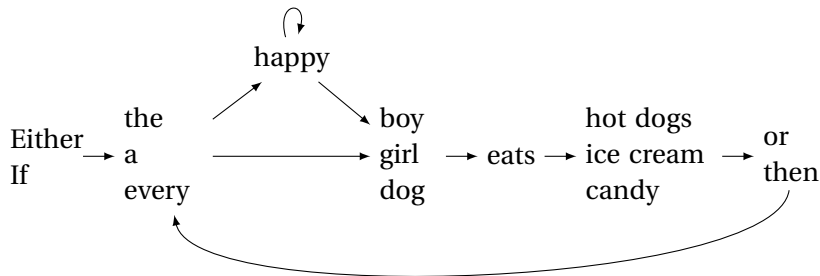




## Richer Sentences Are Harder

If the boy eats hot dogs, then the girl eats ice cream.

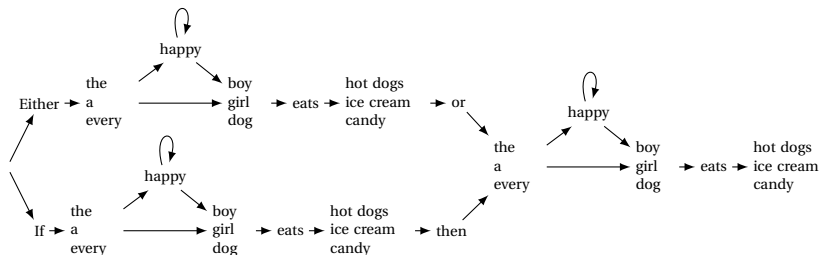
Either the boy eats candy, or every dog eats candy.



*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

$S \rightarrow$  Either  $A$

$S \rightarrow$  If  $A$

$A \rightarrow$  the  $B$

$A \rightarrow$  the  $C$

$A \rightarrow$  a  $B$

$A \rightarrow$  a  $C$

$A \rightarrow$  every  $B$

$A \rightarrow$  every  $C$

$B \rightarrow$  happy  $B$

$B \rightarrow$  happy  $C$

$C \rightarrow$  boy  $D$

$C \rightarrow$  girl  $D$

$C \rightarrow$  dog  $D$

$D \rightarrow$  eats  $E$

$E \rightarrow$  hot dogs  $F$

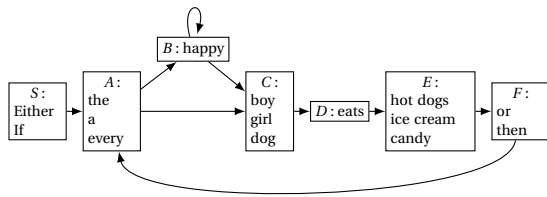
$E \rightarrow$  ice cream  $F$

$E \rightarrow$  candy  $F$

$F \rightarrow$  or  $A$

$F \rightarrow$  then  $A$

$F \rightarrow \epsilon$



## Solution: Context-Free Grammars

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

$S \rightarrow$  Either  $P$ , or  $P$ . Exactly two  $P$ s

$S \rightarrow$  If  $P$ , then  $P$ .

$P \rightarrow A H S$  eats  $O$  One each of  $A$ ,  $H$ ,  $S$ , and  $O$

$A \rightarrow$  the

$A \rightarrow$  a

$A \rightarrow$  every

$H \rightarrow$  happy  $H$   $H$  is “happy” zero or more times

$H \rightarrow \epsilon$

$S \rightarrow$  boy

$S \rightarrow$  girl

$S \rightarrow$  dog

$O \rightarrow$  hot dogs

$O \rightarrow$  ice cream

$O \rightarrow$  candy

## A Context-Free Grammar for a Simplified C

*program* →  $\epsilon$  | *program vdecl* | *program fdecl*

*fdecl* → **id** ( *formals* ) { *vdecls stmts* }

*formals* → **id** | *formals* , **id**

*vdecls* → *vdecl* | *vdecls vdecl*

*vdecl* → **int id** ;

*stmts* →  $\epsilon$  | *stmts stmt*

*stmt* → *expr* ; | **return** *expr* ; | { *stmts* } | **if** ( *expr* ) *stmt* |

**if** ( *expr* ) *stmt* **else** *stmt* |

**for** ( *expr* ; *expr* ; *expr* ) *stmt* | **while** ( *expr* ) *stmt*

*expr* → **lit** | **id** | **id** ( *actuals* ) | ( *expr* ) |

*expr* + *expr* | *expr* - *expr* | *expr* \* *expr* | *expr* / *expr* |

*expr* == *expr* | *expr* != *expr* | *expr* < *expr* | *expr* <= *expr* |

*expr* > *expr* | *expr* >= *expr* | *expr* = *expr*

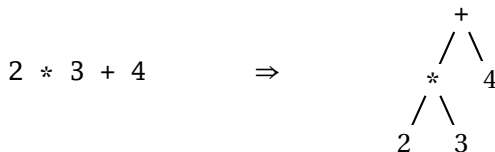
*actuals* → *expr* | *actuals* , *expr*

## Part V

# Constructing Grammars and Ocamlyacc

# Parsing

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



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.*



# Ambiguity

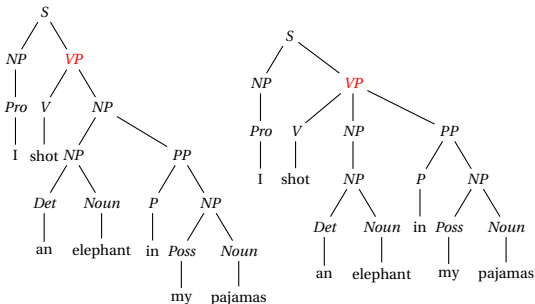
*One morning I shot an elephant in my pajamas.  
How he got in my pajamas I don't know. —Groucho Marx*



# Ambiguity in English

*I shot an elephant in my pajamas*

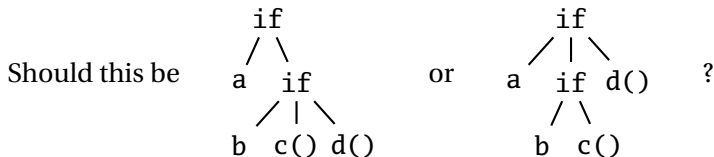
<i>S</i>	→	<i>NP VP</i>
<i>VP</i>	→	<i>VNP</i>
<i>VP</i>	→	<i>VNP PP</i>
<i>NP</i>	→	<i>NP PP</i>
<i>NP</i>	→	<i>Pro</i>
<i>NP</i>	→	<i>Det Noun</i>
<i>NP</i>	→	<i>Poss Noun</i>
<i>PP</i>	→	<i>P NP</i>
<i>V</i>	→	shot
<i>Noun</i>	→	elephant
<i>Noun</i>	→	pajamas
<i>Pro</i>	→	I
<i>Det</i>	→	an
<i>P</i>	→	in
<i>Poss</i>	→	my



# The Dangling Else Problem

Who owns the *else*?

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

```
stmt : IF expr THEN stmt  
      | IF expr THEN stmt ELSE stmt
```

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.

```
stmt : dstmt  
      | cstmt  
  
dstmt : IF expr THEN stmt  
        | IF expr THEN cstmt ELSE dstmt  
  
cstmt : IF expr THEN cstmt ELSE cstmt  
        | 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.

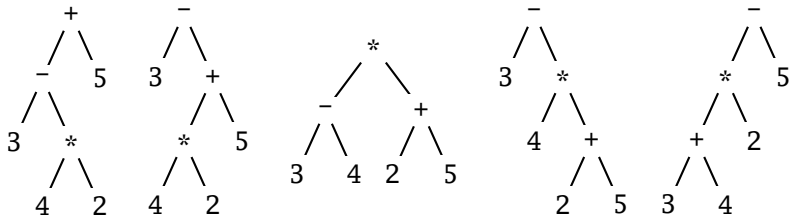
## Ambiguous Arithmetic

Ambiguity can be a problem in expressions. Consider parsing

$$3 - 4 * 2 + 5$$

with the grammar

$$e \rightarrow e + e \mid e - e \mid e * e \mid e / e \mid N$$



# 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.



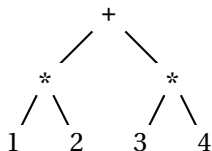
# Operator Precedence

Defines how “sticky” an operator is.

$$1 * 2 + 3 * 4$$

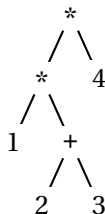
\* at higher precedence than +:

$$(1 * 2) + (3 * 4)$$



+ at higher precedence than \*:

$$1 * (2 + 3) * 4$$

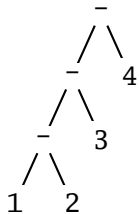


## Associativity

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

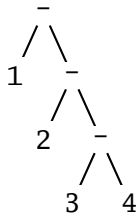
Most operators are left-associative

$$1 - 2 - 3 - 4$$



$$((1 - 2) - 3) - 4$$

left associative



$$1 - (2 - (3 - 4))$$

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.

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

Pascal uses ; as a statement separator.

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

Pascal later made a final ; optional.

# Ocaml yacc 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 :  
    symbol ... symbol { semantic-action }  
    |  
    ...  
    | symbol ... symbol { semantic-action }
```

- ▶ *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

```
%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 }
```

## Part VI

# Parsing Algorithms

## Parsing Context-Free Grammars

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 $\mathbf{Id} * \mathbf{Id} + \mathbf{Id}$

$e$

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

At each step, expand the *rightmost* nonterminal.

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 $\mathbf{Id} * \mathbf{Id} + \mathbf{Id}$

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

$e$   
 $t + e$

At each step, expand the *rightmost* nonterminal.

nonterminal

“handle”: The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.

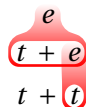
## Rightmost Derivation of $\mathbf{Id} * \mathbf{Id} + \mathbf{Id}$

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



At each step, expand the *rightmost* nonterminal.

nonterminal

“handle”: The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.

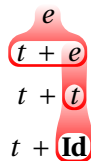
## Rightmost Derivation of $\mathbf{Id} * \mathbf{Id} + \mathbf{Id}$

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



At each step, expand the *rightmost* nonterminal.

nonterminal

“handle”: The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.



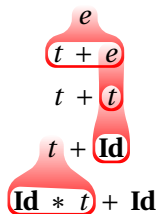
## Rightmost Derivation of $\mathbf{Id} * \mathbf{Id} + \mathbf{Id}$

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



At each step, expand the *rightmost* nonterminal.

nonterminal

“handle”: The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.

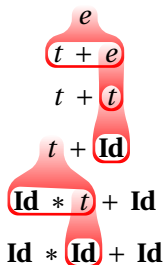
## Rightmost Derivation of $\text{Id} * \text{Id} + \text{Id}$

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \text{Id} * t$

4:  $t \rightarrow \text{Id}$



At each step, expand the *rightmost* nonterminal.

nonterminal

“handle”: The right side of a production

Fun and interesting fact: there is exactly one rightmost expansion if the grammar is unambiguous.

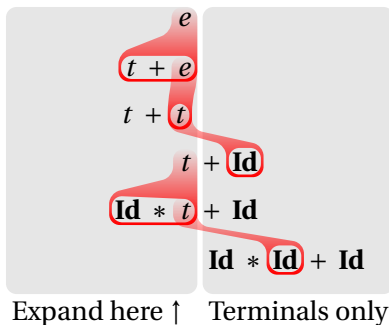
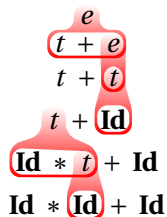
## Rightmost Derivation: What to Expand

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



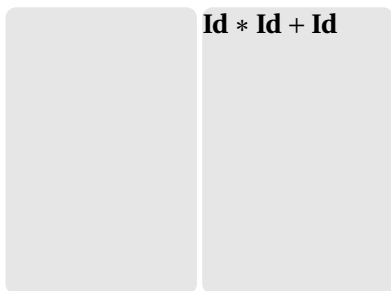
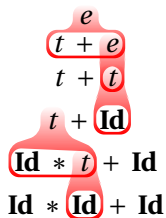
# Reverse Rightmost Derivation

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



viable prefixes

terminals

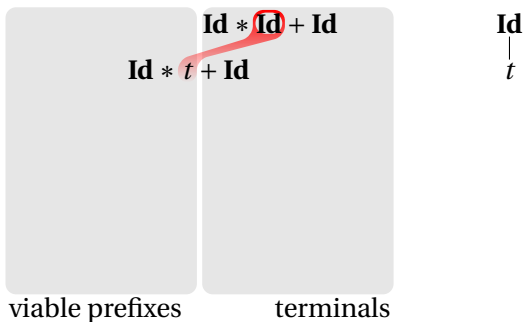
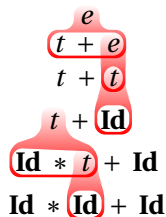
# Reverse Rightmost Derivation

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



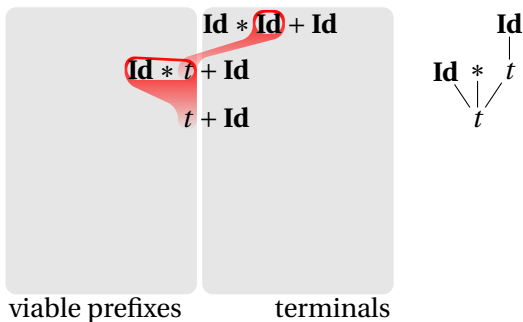
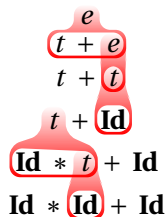
# Reverse Rightmost Derivation

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



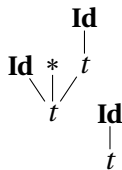
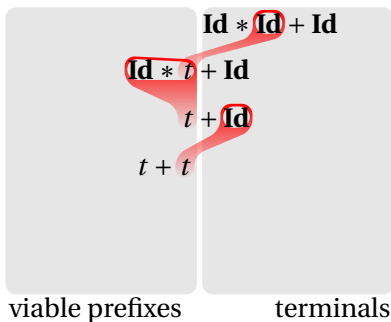
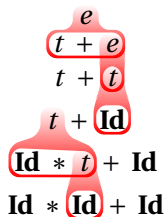
# Reverse Rightmost Derivation

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



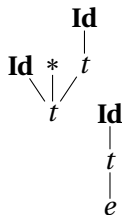
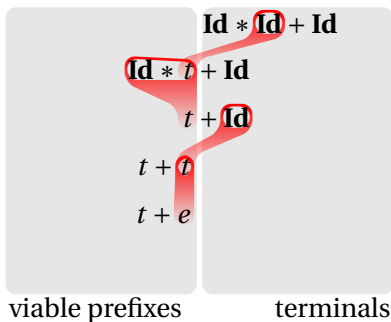
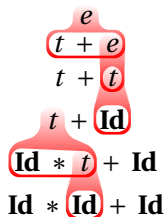
# Reverse Rightmost Derivation

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$





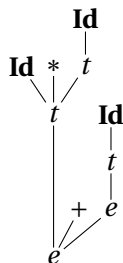
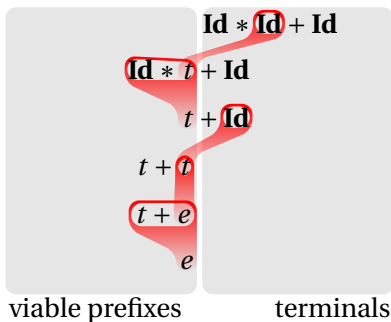
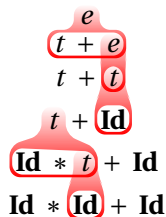
# Reverse Rightmost Derivation

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



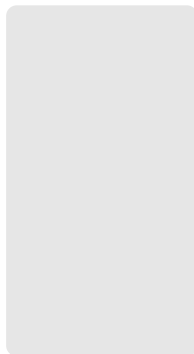
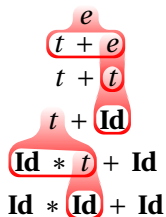
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

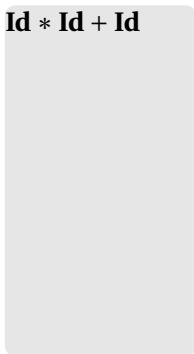
2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



stack



input

shift

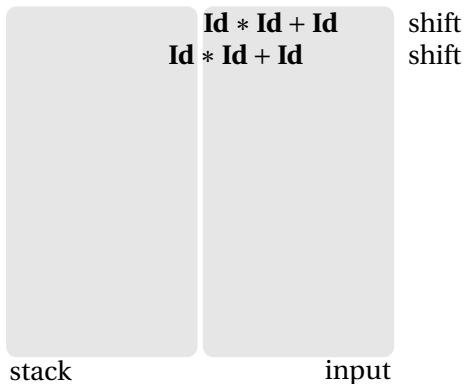
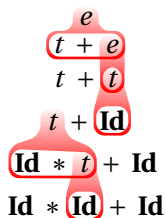
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



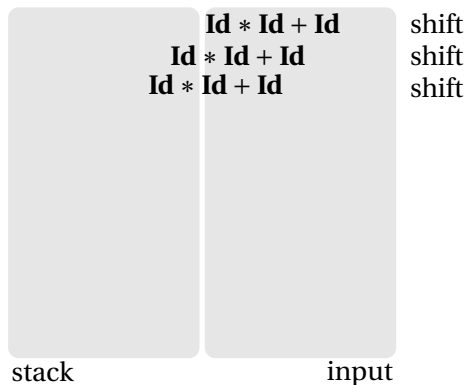
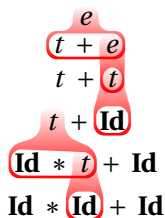
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



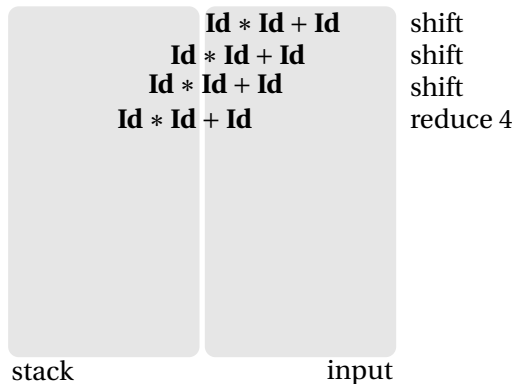
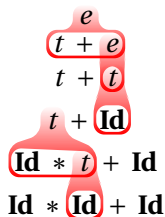
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



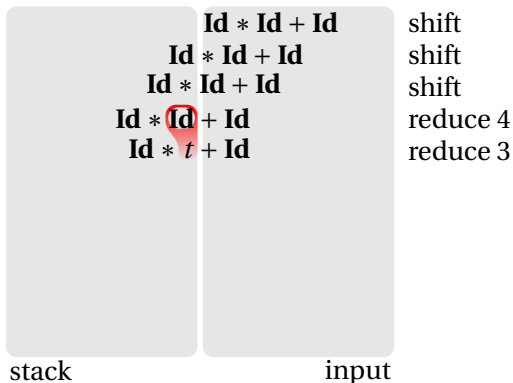
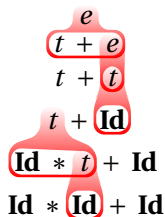
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



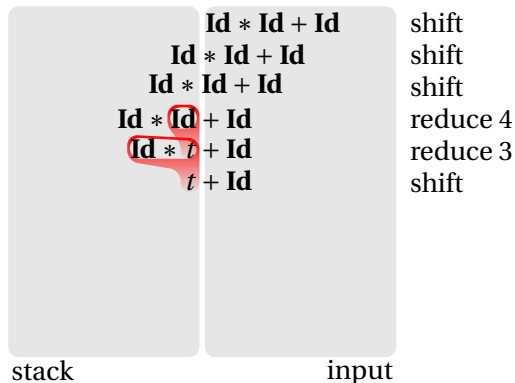
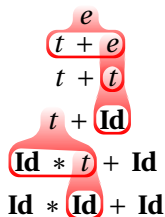
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



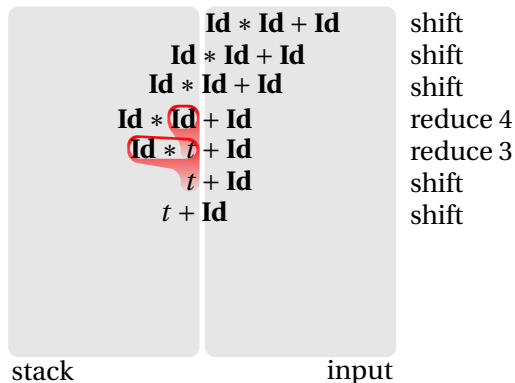
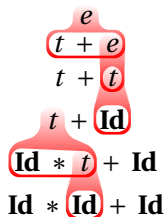
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$





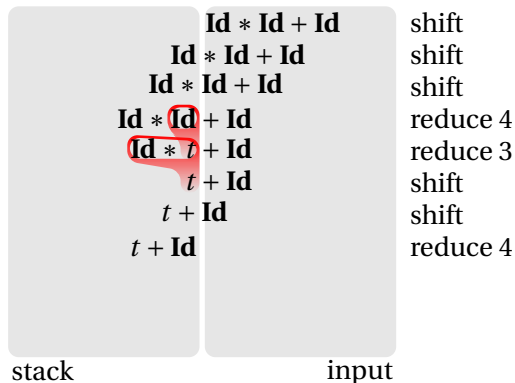
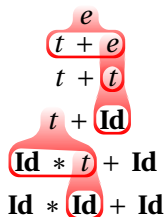
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



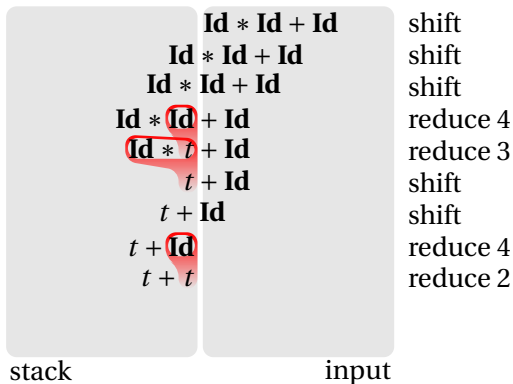
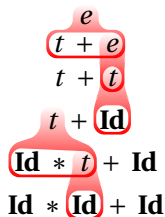
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



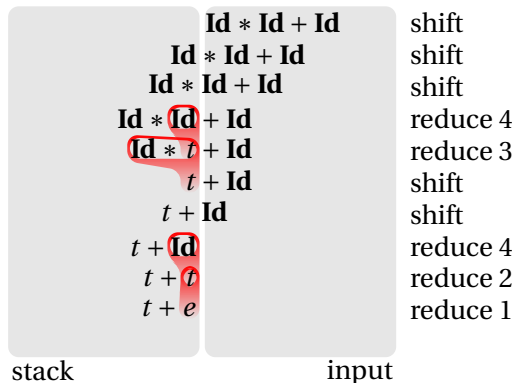
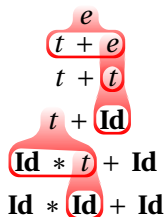
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \text{Id} * t$

4:  $t \rightarrow \text{Id}$



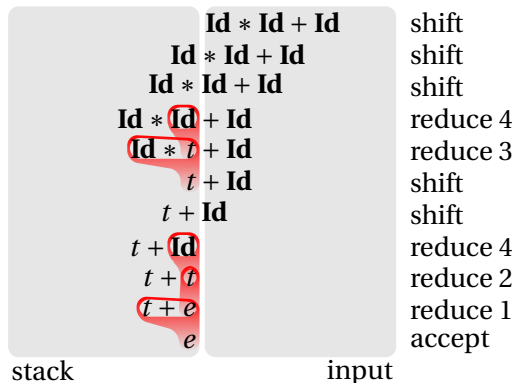
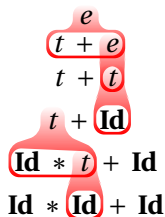
# Shift/Reduce Parsing Using an Oracle

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \text{Id} * t$

4:  $t \rightarrow \text{Id}$



# 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 infinite in number, but let's try anyway.*

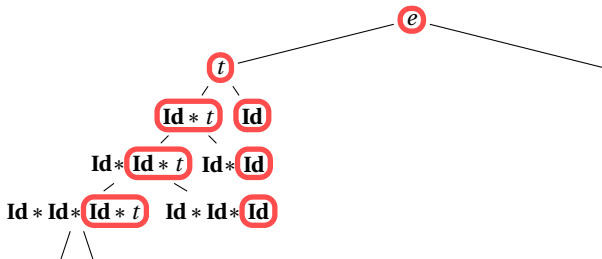
## Some Right-Sentential Forms and Their Handles

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$



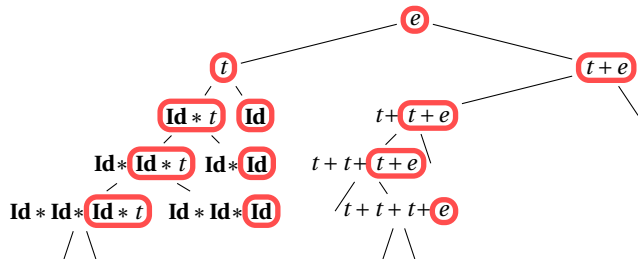
# Some Right-Sentential Forms and Their Handles

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

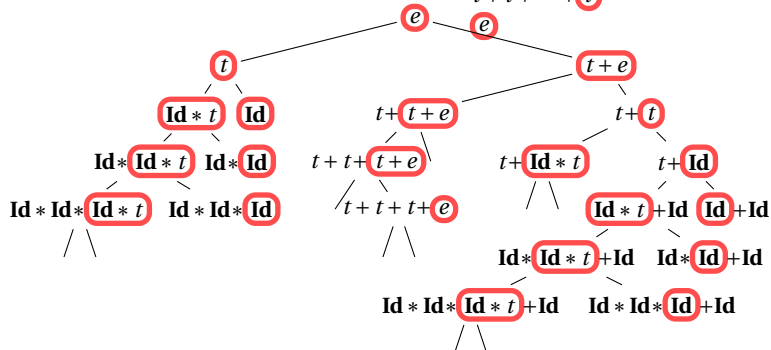


# Some Right-Sentential Forms and Their Handles

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

Patterns:

- $\text{Id} * \text{Id} * \dots * \text{Id} * t$
- $\text{Id} * \text{Id} * \dots * \text{Id} \dots$
- $t + t + \dots + t + e$
- $t + t + \dots + t + \text{Id}$
- $t + t + \dots + t + \text{Id} * \text{Id} * \dots * \text{Id} * t$
- $t + t + \dots + t$

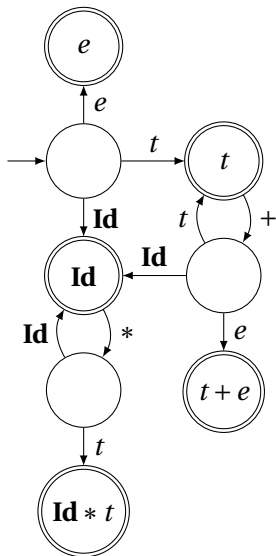




# The Handle-Identifying Automaton

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

$\text{Id} * \text{Id} * \dots * \text{Id} * t \dots$   
 $\text{Id} * \text{Id} * \dots * \text{Id} \dots$   
 $t + t + \dots + t + e$   
 $t + t + \dots + t + \text{Id}$   
 $t + t + \dots + t + \text{Id} * \text{Id} * \dots * \text{Id} * t$   
 $t + t + \dots + t$   
 $e$



## Building the Initial State of the LR(0) Automaton

$$e' \rightarrow \color{green}{\curvearrowright} e$$

$$1: e \rightarrow t + e$$

$$2: e \rightarrow t$$

$$3: t \rightarrow \mathbf{Id} * t$$

$$4: t \rightarrow \mathbf{Id}$$

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 “ $e' \rightarrow \color{green}{\curvearrowright} e$ ”

## Building the Initial State of the LR(0) Automaton

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

$$e' \rightarrow \color{red}{\curvearrowright} e$$
$$e \rightarrow \color{red}{\curvearrowright} t + e$$
$$e \rightarrow \color{red}{\curvearrowright} t$$

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 “ $e' \rightarrow \color{red}{\curvearrowright} e$ ”

There are two choices for what an  $e$  may expand to:  $t + e$  and  $t$ . So when  $e' \rightarrow \color{red}{\curvearrowright} e$ ,  $e \rightarrow \color{red}{\curvearrowright} t + e$  and  $e \rightarrow \color{red}{\curvearrowright} t$  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:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

$$e' \rightarrow \mathcal{C}e$$
$$e \rightarrow \mathcal{C}t + e$$
$$e \rightarrow \mathcal{C}t$$
$$t \rightarrow \mathcal{C}\mathbf{Id} * t$$
$$t \rightarrow \mathcal{C}\mathbf{Id}$$

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 “ $e' \rightarrow \mathcal{C}e$ ”

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

Similarly,  $t$  must be either  $\mathbf{Id} * t$  or  $\mathbf{Id}$ , so  $t \rightarrow \mathcal{C}\mathbf{Id} * t$  and  $t \rightarrow \mathcal{C}\mathbf{Id}$ .

This reasoning is a *closure* operation like  $\epsilon$ -closure in subset construction.

## Building the LR(0) Automaton

$$e' \rightarrow \bullet e$$

$$e \rightarrow \bullet t + e$$

$$\mathbf{S0}: e \rightarrow \bullet t$$

$$t \rightarrow \bullet \mathbf{Id} * t$$

$$t \rightarrow \bullet \mathbf{Id}$$

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

# Building the LR(0) Automaton

*“Just passed a string derived from  $e$ ”*

**S7:**  $e' \rightarrow e\color{yellow}{\bullet}$

$e$

$e' \rightarrow \color{yellow}{\bullet}e$   
 $e \rightarrow \color{yellow}{\bullet}t + e$   
**S0:**  $e \rightarrow \color{yellow}{\bullet}t$   
 $t \rightarrow \color{yellow}{\bullet}Id * t$   
 $t \rightarrow \color{yellow}{\bullet}Id$

*“Just passed a prefix ending in a string derived from  $t$ ”*

$t$

**S2:**  $e \rightarrow t\color{yellow}{\bullet} + e$   
 $e \rightarrow t\color{yellow}{\bullet}$

**Id**

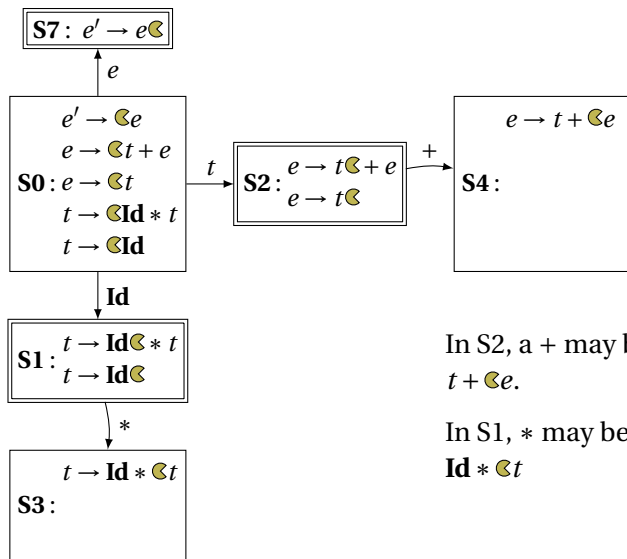
**S1:**  $t \rightarrow Id\color{yellow}{\bullet} * t$   
 $t \rightarrow Id\color{yellow}{\bullet}$

*“Just passed a prefix that ended in an **Id**”*

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  $\bullet$  across each thing, then performing the closure operation (vacuous here).

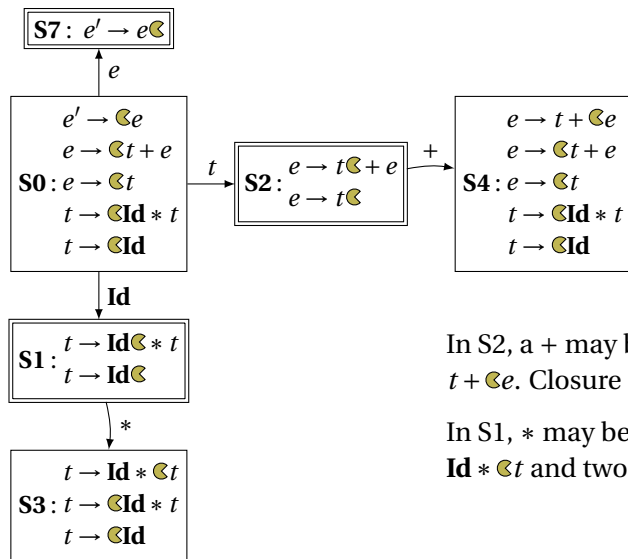
# Building the LR(0) Automaton



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

In S1, \* may be next, giving  $\text{Id} * \bullet t$

# Building the LR(0) Automaton

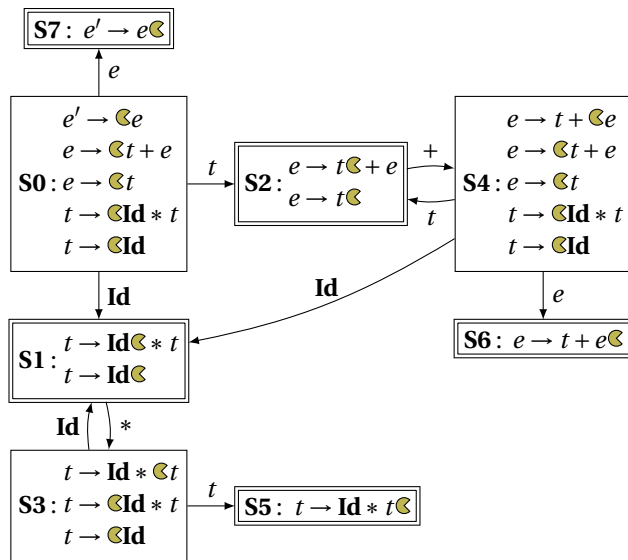


In S2, a + may be next. This gives  $t + \bullet e$ . Closure adds 4 more items.

In S1, \* may be next, giving  $\text{Id} * \bullet t$  and two others.



# Building the LR(0) Automaton



## 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 production  $X \rightarrow Y \dots$ , 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 production  $X \rightarrow Y_1 \dots Y_k Z \dots$  where  $\epsilon \in \text{FIRST}(Y_i)$  for  $i = 1, \dots, k$ , add  $\text{FIRST}(Z) - \{\epsilon\}$  to  $\text{FIRST}(X)$ .  
*Skip all potential  $\epsilon$ 's at the beginning of whatever  $X$  produces*

---

1: $e \rightarrow t + e$	$\text{FIRST}(\mathbf{Id}) = \{\mathbf{Id}\}$
2: $e \rightarrow t$	$\text{FIRST}(t) = \{\mathbf{Id}\}$ because $t \rightarrow \mathbf{Id} * t$ and $t \rightarrow \mathbf{Id}$
3: $t \rightarrow \mathbf{Id} * t$	$\text{FIRST}(e) = \{\mathbf{Id}\}$ because $e \rightarrow t + e$ , $e \rightarrow t$ , and
4: $t \rightarrow \mathbf{Id}$	$\text{FIRST}(t) = \{\mathbf{Id}\}$ .

## The FOLLOW function

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

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

---

1:  $e \rightarrow t + e$

$\text{FOLLOW}(e) = \{\$\}$

2:  $e \rightarrow t$

$\text{FOLLOW}(t) = \{ \quad \}$

3:  $t \rightarrow \mathbf{Id} * t$

1. Because  $e$  is the start symbol

4:  $t \rightarrow \mathbf{Id}$

$\text{FIRST}(t) = \{\mathbf{Id}\}$

$\text{FIRST}(e) = \{\mathbf{Id}\}$

## The FOLLOW function

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

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

---

1:  $e \rightarrow t + e$

$\text{FOLLOW}(e) = \{\$\}$

2:  $e \rightarrow t$

$\text{FOLLOW}(t) = \{+ \}$

3:  $t \rightarrow \mathbf{Id} * t$

2. Because  $e \rightarrow \underline{t} + e$  and  $\text{FIRST}(+) = \{+\}$

4:  $t \rightarrow \mathbf{Id}$

$\text{FIRST}(t) = \{\mathbf{Id}\}$

$\text{FIRST}(e) = \{\mathbf{Id}\}$

## The FOLLOW function

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

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

---

$$1: e \rightarrow t + e$$

$$\text{FOLLOW}(e) = \{\$\}$$

$$2: e \rightarrow t$$

$$\text{FOLLOW}(t) = \{+, \$\}$$

$$3: t \rightarrow \mathbf{Id} * t$$

$$3. \text{ Because } e \rightarrow \underline{t} \text{ and } \$ \in \text{FOLLOW}(e)$$

$$4: t \rightarrow \mathbf{Id}$$

$$\text{FIRST}(t) = \{\mathbf{Id}\}$$

$$\text{FIRST}(e) = \{\mathbf{Id}\}$$

## The FOLLOW function

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

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

---

1:  $e \rightarrow t + e$

$\text{FOLLOW}(e) = \{\$\}$

2:  $e \rightarrow t$

$\text{FOLLOW}(t) = \{+, \$\}$

3:  $t \rightarrow \mathbf{Id} * t$

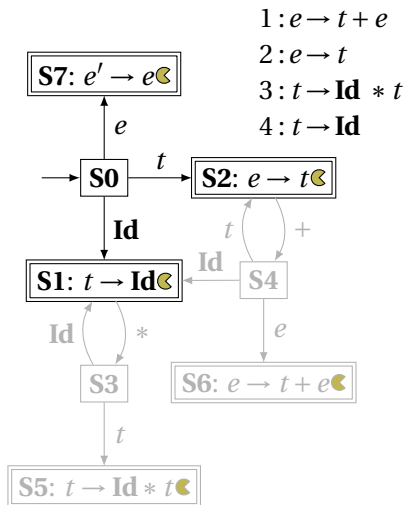
Fixed-point reached: applying any rule does not change any set

4:  $t \rightarrow \mathbf{Id}$

$\text{FIRST}(t) = \{\mathbf{Id}\}$

$\text{FIRST}(e) = \{\mathbf{Id}\}$

# Converting the LR(0) Automaton to an SLR Parsing Table



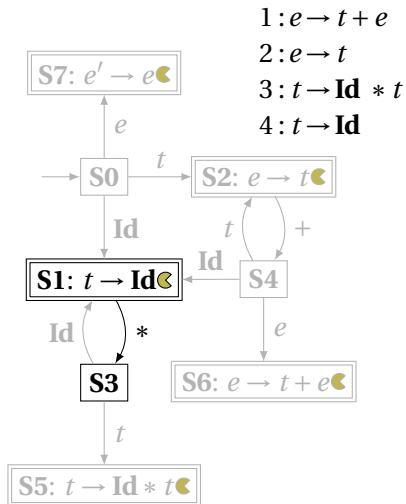
$\text{FOLLOW}(e) = \{\$\}$

$\text{FOLLOW}(t) = \{+, \$\}$

State	Action				Goto	
	Id	+	*	\$	e	t
0	s1				7	2

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.

# Converting the LR(0) Automaton to an SLR Parsing Table



$\text{FOLLOW}(e) = \{\$\}$

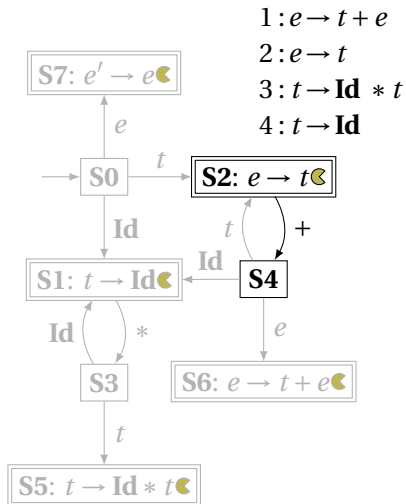
$\text{FOLLOW}(t) = \{+, \$\}$

State	Action				Goto	
	Id	+	*	\$	e	t
0	s1				7	2
1		r4	s3	r4		

From S1, shift a \* and go to S3; or, if the next input  $\in \text{FOLLOW}(t)$ , reduce by rule 4.



# Converting the LR(0) Automaton to an SLR Parsing Table



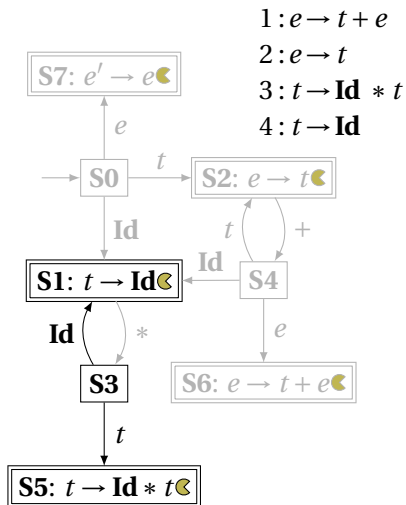
$\text{FOLLOW}(e) = \{\$\}$

$\text{FOLLOW}(t) = \{+, \$\}$

State	Action				Goto	
	Id	+	*	\$	e	t
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		

From S2, shift a + and go to S4; or, if the next input  $\in \text{FOLLOW}(e)$ , reduce by rule 2.

# Converting the LR(0) Automaton to an SLR Parsing Table



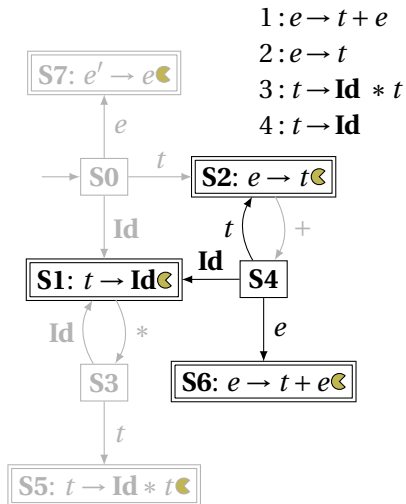
FOLLOW(*e*) = {\$}

FOLLOW(*t*) = {+, \$}

State	Action				Goto	
	Id	+	*	\$	<i>e</i>	<i>t</i>
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5

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 Parsing Table



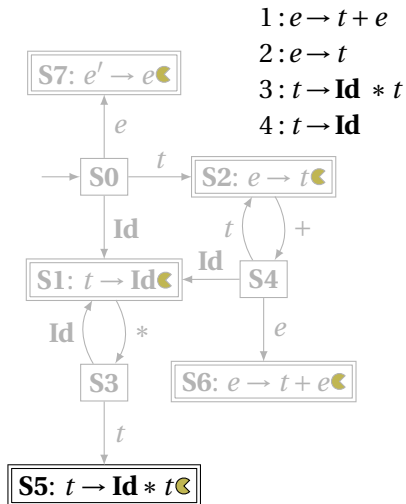
$\text{FOLLOW}(e) = \{\$ \}$

$\text{FOLLOW}(t) = \{+, \$ \}$

State	Action				Goto	
	Id	+	*	\$	e	t
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2

From S4, shift an **Id** and go to S1; or cross an *e* or a *t*.

# Converting the LR(0) Automaton to an SLR Parsing Table



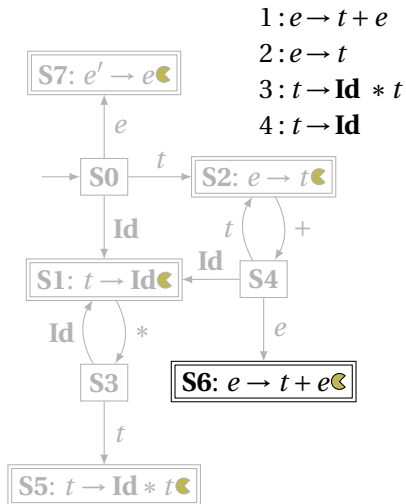
$\text{FOLLOW}(e) = \{\$\}$

$\text{FOLLOW}(t) = \{+, \$\}$

State	Action			Goto		
	Id	+	*	\$	$e$	$t$
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		

From S5, reduce using rule 3 if the next symbol  $\in \text{FOLLOW}(t)$ .

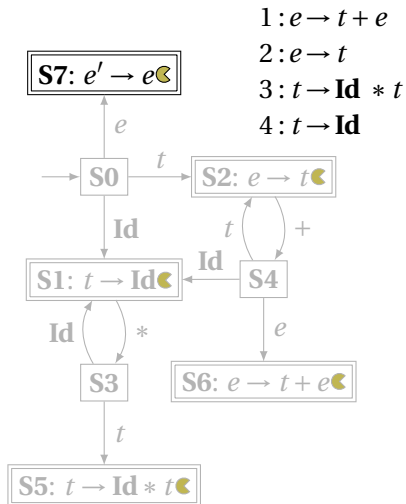
# Converting the LR(0) Automaton to an SLR Parsing Table



State	Action				Goto	
	Id	+	*	\$	$e$	$t$
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		

From S6, reduce using rule 1 if the next symbol  $\in$  FOLLOW( $e$ ).

# Converting the LR(0) Automaton to an SLR Parsing Table



$\text{FOLLOW}(e) = \{\$\}$

$\text{FOLLOW}(t) = \{+, \$\}$

State	Action				Goto	
	Id	+	*	\$	<i>e</i>	<i>t</i>
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				✓		

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:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

Stack	Input	Action
0	<b>Id * Id + Id \$</b>	Shift, goto 1

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.

State	Action				Goto	
	Id	+	*	\$	<i>e</i>	<i>t</i>
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				✓		

# Shift/Reduce Parsing with an SLR Table

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

Stack	Input	Action
0	<b>Id</b> * <b>Id</b> + <b>Id</b> \$	Shift, goto 1
0 <b>Id</b> 1	* <b>Id</b> + <b>Id</b> \$	Shift, goto 3

Here, the state is 1, the next symbol is \*, so shift and mark it with state 3.

State	Action				Goto	
	Id	+	*	\$	<i>e</i>	<i>t</i>
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				✓		



# Shift/Reduce Parsing with an SLR Table

1:  $e \rightarrow t + e$

2:  $e \rightarrow t$

3:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

State	Action				Goto	
	Id	+	*	\$	<i>e</i>	<i>t</i>
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				✓		

Stack	Input	Action
0	<b>Id * Id + Id \$</b>	Shift, goto 1
0 Id 1	<b>* Id + Id \$</b>	Shift, goto 3
0 Id 1 3	<b>Id + Id \$</b>	Shift, goto 1
0 Id 1 3 1	<b>+ Id \$</b>	Reduce 4

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:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

State	Action				Goto	
	Id	+	*	\$	<i>e</i>	<i>t</i>
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				✓		

Stack	Input	Action
0	<b>Id * Id + Id \$</b>	Shift, goto 1
0 1	<b>* Id + Id \$</b>	Shift, goto 3
0 1 3	<b>Id + Id \$</b>	Shift, goto 1
0 1 3 1	<b>+ Id \$</b>	Reduce 4
0 1 3	<b>+ Id \$</b>	

Remove the RHS of the rule (here, just **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:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

State	Action				Goto	
	Id	+	*	\$	<i>e</i>	<i>t</i>
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				✓		

Stack	Input	Action
0	<b>Id * Id + Id \$</b>	Shift, goto 1
0 Id	<b>* Id + Id \$</b>	Shift, goto 3
0 Id *	<b>Id + Id \$</b>	Shift, goto 1
0 Id * Id	<b>+ Id \$</b>	Reduce 4
0 Id * <i>t</i>	<b>+ Id \$</b>	Reduce 3

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:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

State	Action				Goto	
	Id	+	*	\$	<i>e</i>	<i>t</i>
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				✓		

Stack	Input	Action
0	<b>Id * Id + Id \$</b>	Shift, goto 1
0 1	<b>* Id + Id \$</b>	Shift, goto 3
0 1 3	<b>Id + Id \$</b>	Shift, goto 1
0 1 3 1	<b>+ Id \$</b>	Reduce 4
0 1 3 5	<b>+ Id \$</b>	Reduce 3
0 2	<b>+ Id \$</b>	Shift, goto 4

This time, we strip off the RHS for rule 3,  $\mathbf{Id} * t$ , 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:  $t \rightarrow \mathbf{Id} * t$

4:  $t \rightarrow \mathbf{Id}$

State	Action				Goto	
	Id	+	*	\$	<i>e</i>	<i>t</i>
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				✓		

Stack	Input	Action
0	<b>Id * Id + Id \$</b>	Shift, goto 1
0 Id	<b>* Id + Id \$</b>	Shift, goto 3
0 Id *	<b>Id + Id \$</b>	Shift, goto 1
0 Id * Id	<b>+ Id \$</b>	Reduce 4
0 Id * <i>t</i>	<b>+ Id \$</b>	Reduce 3
0 <i>t</i>	<b>+ Id \$</b>	Shift, goto 4
0 <i>t</i> +	<b>Id \$</b>	Shift, goto 1
0 <i>t</i> + Id	<b>\$</b>	Reduce 4
0 <i>t</i> + <i>t</i>	<b>\$</b>	Reduce 2
0 <i>t</i> + <i>e</i>	<b>\$</b>	Reduce 1
0 <i>e</i>	<b>\$</b>	Accept

## The Punchline

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

$$t \rightarrow \text{Else } s$$
$$t \rightarrow$$

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

$$t \rightarrow \text{Id} * t$$
$$e \rightarrow t + e$$

Do you reduce by “ $t \rightarrow \text{Id} * t$ ” or by “ $e \rightarrow t + e$ ”?