Lexical Analysis (Scanning)

Lexical Analysis

Goal: simplify the job of the parser.
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.

Lexical Analysis (Scanning)

Translates a stream of characters to a stream of tokens

<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>

Describing Tokens

**Alphabet:** A finite set of symbols
Examples: \{ 0, 1 \}, \{ A, B, C, ..., 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

Lexical Analysis

- Translates a stream of characters to a stream of tokens
- f o o = a + bar(2, q);
- ID EQUALS ID PLUS ID LPAREN NUM
- COMMA ID LPAREN SEMI

Operations on Languages

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

**Concatenation:** Strings from one followed by the other
\( LM = \{ \text{man, men, woman, women} \} \)

**Union:** All strings from each language
\( L \cup M = \{ \epsilon, wo, \text{man, men} \} \)

**Kleene Closure:** Zero or more concatenations
\( M^* = \{ \epsilon, M, MM, MMM, \ldots \} = \{ \epsilon, \text{man, men, manman, manmen, menman, menmen, manmanman, manmanmen, manmenman, manmenmen, ...} \} \)

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

\( \Sigma = \{ a, b \} \)

<table>
<thead>
<tr>
<th>RE</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>(a</td>
<td>b) (a</td>
</tr>
<tr>
<td>a*</td>
<td>{\epsilon, a, aa, aaa, aaaa,...}</td>
</tr>
<tr>
<td>(a</td>
<td>b)*</td>
</tr>
<tr>
<td>a</td>
<td>a*b</td>
</tr>
</tbody>
</table>
Specifying Tokens with REs

Typical choice: \(\Sigma = \text{ASCII characters, i.e.,} \{, !, *, #, \ldots, 0, 1, \ldots, 9, \ldots, A, \ldots, Z, \ldots, \} \)

letters: A|B|\ldots|Z|a|\ldots|z

digits: 0|1|\ldots|9

identifier: letter (letter | digit)*

Implementing Scanners Automatically

Translation: Regular Expressions (Rules)  
\rightarrow\text{nondeterministic finite automata}  
\rightarrow\text{subset construction}  
\rightarrow\text{deterministic finite automata}  
\rightarrow\text{tables}

Translating REs into NFAs

Example: translate \((a|b)^* abb\) into an NFA

Simulating an NFA: \(-aabb\), Start

Translating REs into NFAs

Example: translate \((a|b)^* abb\) into an NFA

Simulating an NFA: \(-aabb\), \(\epsilon\)-closure

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” \(x\).

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:

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 the current state. Accept if you end 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

ELSE: "else" ;
ELSEIF: "elseif" ;

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
An DFA can be exponentially larger than the corresponding NFA.
\( n \text{ states versus } 2^n \)
Tools often try to strike a balance between the two representations.
ANTLR uses a different technique.

The ANTLR Compiler Generator
Language and compiler for writing compilers
Running ANTLR on an ANTLR file produces Java source files that can be compiled and run.
ANTLR can generate
- Scanners (lexical analyzers)
- Parsers
- Tree walkers
An ANTLR File for a Simple Scanner

```java
class CalcLexer extends Lexer;

LPAREN : '(' ;  // Rules for punctuation
RPAREN : ')';
STAR : '*' ;
PLUS : '+' ;
SEMI : ';' ;

protected
DIGIT : '0'..'9' ;  // Any character between 0 and 9
INT : (DIGIT)+;  // One or more digits

WS : (' ' | '	' | '
' | '') { $setType(Token.SKIP); } ;  // Action: ignore
```

ANTLR Specifications for Scanners

Rules are names starting with a capital letter.

A character in single quotes matches that character.

```java
LPAREN : '(' ;
```

A string in double quotes matches the string

```java
IF : "if" ;
```

A vertical bar indicates a choice:

```java
OP : '+' | '-' | '*' | '/' ;
```

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")

Python

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?

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 a problem: APL is a compact, cryptic
language with its own character set (!)

```python
E=A*TEST*B;L
E=0.5
E=((A*A)+B*B)+L
```

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
1 5 |{z}|
6 |{z}|
7 |{z}|
72 |{z}|
```

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.

Syntax and Language Design

Some syntax is error-prone. Classic Fortran example:

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

C distinguishes > and >> as different operators.
**Parsing**

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

\[ 2 \times 3 + 4 \Rightarrow \frac{4}{2} \frac{3}{1} \]

Goal: discard irrelevant information to make it easier for the next stage.
Parentheses and most other forms of punctuation removed.

**Grammars**

Most programming languages described using a context-free grammar.

Compared to regular languages, context-free languages add one important thing: recursion.
Recursion allows you to count, e.g., to match pairs of nested parentheses.

Which languages do humans speak? I’d say it’s regular: I do not not not not not not not not not not understand this sentence.

**Languages**

Regular languages (\( t \) is a terminal):

\[ A \rightarrow t_1 \ldots t_n B \]

\[ A \rightarrow t_1 \ldots t_n \]

Context-free languages (\( P \) is terminal or a variable):

\[ A \rightarrow P_1 \ldots P_n \]

Context-sensitive languages:

\[ \alpha_1 \alpha_2 \rightarrow \alpha_1 B \alpha_2 \]

“\( B \rightarrow A \) only in the ‘context’ of \( \alpha_1 \ldots \alpha_2 \)”

**Issues**

Ambiguous grammars
Precedence of operators
Left- versus right-recursive
Top-down vs. bottom-up parsers
Parse Tree vs. Abstract Syntax Tree

**Ambiguous Grammars**

A grammar can easily be ambiguous. Consider parsing

\[ 3 - 4 \,* \, 2 + 5 \]

with the grammar

\[ e \rightarrow e + e | e - e | e * e | e / e | N \]

**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 +:

\[ (1 \:* \, 2) + (3 \:* \, 4) \]

\[ 1 \,* \, (2 \,*\, 3) \,*\, 4 \]

+ at higher precedence than *:

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

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

left associative
right associative
Fixing Ambiguous Grammars

Original ANTLR grammar specification

```
expr : expr '+' expr |
     | expr '-' expr |
     | expr '*' expr |
     | expr '/' expr |
     | NUMBER ;
```

Ambiguous: no precedence or associativity.

Assigning Precedence Levels

Split into multiple rules, one per level

```
expr : expr '+' term |
     | expr '-' term |
     | term ;

term : term '*' term |
     | term '/' term |
     | atom ;

atom : NUMBER ;
```

Still ambiguous: associativity not defined

Assigning Associativity

Make one side or the other the next level of precedence

```
expr : expr '+' term |
     | expr '-' term |
     | term ;

term : term '*' atom |
     | term '/' atom |
     | atom ;

atom : NUMBER ;
```

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.

Parsing LL(k) Grammars

LL: Left-to-right, Left-most derivation

k: number of tokens to look ahead

Parsed by top-down, predictive, recursive parsers

Basic idea: look at the next token to predict which production to use

ANTLR builds recursive LL(k) parsers

Almost a direct translation from the grammar.

Implementing a Top-Down Parser

```
stmt : 'if' expr 'then' expr |
     | 'while' expr 'do' expr |
     | expr :=' expr ;

stmt() {
    switch (next-token) {
        case IF: match(IF); expr(); match(THEN); expr(); break;
        case WHILE: match(WHILE); expr(); match(DO); expr(); break;
        case NUMBER or LPAREN: expr(); match(COLEQ); expr();
    }
}
```

Writing LL(k) Grammars

Cannot have left-recursion

```
expr : expr '+' term | term ;
```

becomes

```
AST expr() {
    switch (next-token) {
        case NUMBER : expr(); /* Infinite Recursion */
    }
}
```

Writing LL(1) Grammars

Cannot have common prefixes

```
expr : ID ('(' expr ')') |
     | ID '=' expr

becomes

```
```
```
```
```
Eliminating Left Recursion

Understand the recursion and add tail rules

\[
\text{expr} : \text{expr} \ ('+' \ \text{term} \ | \ '-' \ \text{term} ) \\
| \ \text{term} \\
\]

becomes

\[
\text{expr} : \ \text{term} \ \text{exprt} ; \\
\text{exprt} : \ '+' \ \text{term} \ \text{exprt} \\
| \ '-' \ \text{term} \ \text{exprt} \\
| \ /* \text{nothing} */ \\
\]

Using ANTLR’s EBNF

ANTLR makes this easier since it supports * and -:

\[
\text{expr} : \text{expr} \ '+' \ \text{term} \\
| \text{expr} \ '-' \ \text{term} \\
| \ \text{term} ; \\
\]

becomes

\[
\text{expr} : \ \text{term} \ ('+' \ \text{term} \ | \ '-' \ \text{term})*; \\
\]

The Dangling Else Problem

Who owns the else?

\[
\text{if} \ (a) \ \text{if} \ (b) \ c(); \ \text{else} \ d(); \\
\]

or if

\[
\text{if} \ a \ \text{if} \ b \ c(); \ \text{d()}; \\
\]

The Dangling Else Problem

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

ANTLR can resolve this problem by making certain rules “greedy.” If a conditional is marked as greedy, it will take that option even if the “nothing” option would also match:

\[
\text{stmt} : \ "\text{if}\) \ \text{expr} \ "\text{then}\) \ \text{stmt} \ \text{iftail} \\
| \ \text{other-statements} ; \\
\text{iftail} : \ "\text{else}\) \ \text{stmt} \\
| \ /* \text{nothing} */ \\
\]

Problem comes when matching “iftail.”

Normally, an empty choice is taken if the next token is in the “follow set” of the rule. But since “else” can follow an iftail, the decision is ambiguous.

The Dangling Else Problem

Some languages resolve this problem by insisting on nesting everything.

E.g., Algol 68:

\[
\text{if} \ a < b \ \text{then} \ a \ \text{else} \ b \ \text{fi}; \\
\]

“fi” is “if” spelled backwards. The language also uses do–od and case–esac.

Statement separators/terminators

C uses ; as a statement terminator.

\[
\text{if} \ (a<b) \ \text{printf}("a less"); \\
\text{else} \{ \\
\quad \text{printf}("b"); \ \text{printf}(" less"); \\
\} \\
\]

Pascal uses ; as a statement separator.

\[
\text{if} \ a < b \ \text{then} \ \text{writeln}('a less') \ \text{else} \text{begin} \\
\quad \text{write}('a'); \ \text{writeln}(' less') \end \\
\]

Pascal later made a final ; optional.

Rightmost Derivation

A rightmost derivation for \(\text{ld} \rightarrow \text{ld} + \text{ld}\):

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

Basic idea of bottom-up parsing:

\[
\begin{align*}
\text{ld} & \rightarrow \text{ld} + \text{ld} \\
\text{t} & \rightarrow \text{ld} \\
\text{e} & \rightarrow \text{t} \\
\end{align*}
\]

Here, I’ve drawn a box around each symbol to expand.

Bottom-up Parsing
Handles

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow Id + t \)
4: \( t \rightarrow Id \)

This is a reverse rightmost derivation for \( Id + Id + Id \).

Each highlighted region is a handle.

A handle is the right side of a rule that turns into a left.

In order, the handles build the tree leaves-to-root.

Shift-reduce Parsing

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow Id + t \)
4: \( t \rightarrow Id \)

Scan input left-to-right, looking for handles.

An oracle tells what to do.

LR Parsing

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow Id + t \)
4: \( t \rightarrow Id \)

1. Look at state on top of stack
2. and the next input token
3. to find the next action
4. In this case, shift the token onto the stack and go to state 1.

LR Parsing

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow Id + t \)
4: \( t \rightarrow Id \)

The states are places we could be in a reverse-rightmost derivation. Let’s represent such a place with a dot.

1: \( e \rightarrow t + e \)
2: \( e \rightarrow t \)
3: \( t \rightarrow Id + t \)
4: \( t \rightarrow Id \)

Say we were at the beginning \( (e) \). This corresponds to:

The first is a placeholder. The second are the two possibilities when we’re just before \( e \). The last two are the two possibilities when we’re just before \( t \).

Constructing the SLR Parse Table

This is a tricky, but mechanical procedure. The parser generators YACC, Bison, Cup, and others (but not ANTLR) use 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

\( t \rightarrow Id \cdot t \)
\( t \rightarrow Id + t \)

Reduce/reduce conflicts are caused by a state like

\( t \rightarrow Id \cdot t \)
\( e \rightarrow t + e \)

The Punchline

Each highlighted region is a handle.

In order, the handles build the tree leaves-to-root.