CS321 Languages and Compiler Design I
Fall 2010
Lecture 4

Lexical Analysis

Convert source file characters into **token stream**.

Remove content-free characters (comments, whitespace, ...)

Detect lexical errors (badly-formed literals, illegal characters, ...)

Output of lexical analysis is input to syntax analysis.

Could just do lexical analysis as part of syntax analysis.

But choose to handle separately for better modularity and portability, and to allow make syntax analysis easier.

Idea: Look for **patterns** in input character sequence, convert to **tokens** with **attributes**, and pass them to parser in **stream**.

Lexical Analysis Example

Pattern	Token	Attribute
`if`	IF	
`else`	ELSE	
`print`	PRINT	
`then`	THEN	
`:=`	ASSIGN	
= or < or >	RELOP	enum
letter followed by letters or digits	ID	symbol
digits	NUM	int
chars between double quotes	STRING	string

Source code:

```
if x>17 then count:= 2
          else (* oops !*) print "bad!"
```

Lexeme	Token	Attribute
`if`	IF	
`x`	ID	`"x"`
`>`	RELOP	`GT`
`17`	NUM	`17`
`then`	THEN	
`count`	ID	`"count"`
`:=`	ASSIGN	
`2`	NUM	`2`
`else`	ELSE	
`print`	PRINT	
`"bad!"`	STRING	`"bad!"`

More Details

A **token** describes a **class** of character strings with some distinguished meaning in language.

- May describe **unique** string (e.g., IF, ASSIGN)
- or set of possible strings, in which case an **attribute** is needed to indicate which.

(Tokens are typically represented as elements of an **enumeration**.)

A **lexeme** is the string in the input that actually matched the pattern for some token.

Attributes represent lexemes converted to a more useful form, e.g.,:

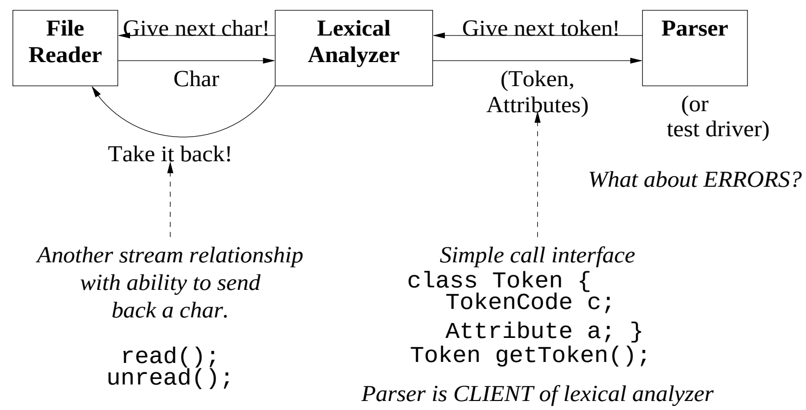
- strings
- symbols (like strings, but perhaps handled separately)
- numbers (integers, reals, ...)
- enumerations

Whitespace (spaces, tabs, new lines, ...) and **comments** usually just disappear (unless they affect program meaning).

Stream Interface

Could convert entire input file to list of tokens/attributes.

But parser needs only one token at a time, so use **stream** instead:



Hand-coded Scanner (in Pseudo-Java)

```
Token getToken() {
  while (true) {
     char c = read();
     if (c is whitespace)
       ignore it;
     else if (c is digit) {
        int n = 0;
        do {n = n * 10 + (c- '0');
            c = read(); }
        until (c not a digit);
        unread(c);
        return new Token(NUM,n);
     } else if (c is alpha) {
        String s = "";
        do { s = s + c;
             c = read();
        } until (c is not an alphanumeric);
        unread(c);
        return new Token(ID,S);
    } else ...   } }
```


Pros and Cons of Hand-coded Scanners

Efficient!

But easy to get wrong!

Note intermixed code for input, output, patterns, conversion.

Hard to specify! (esp. **patterns**).

Formalizing Pattern Descriptions

Ex.: "An **identifier** is a letter followed by any number of letters or digits."

- Exactly what is a letter?

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

- Exactly what is a digit?

```
DIGIT →  0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
```

- How can we express "letters or digits" ?

```
LORD → LETTER | DIGIT
```

- How can we express "any number of" ?

LORDS $\longrightarrow$ LORD*

- How can we express "followed by" ?

```
IDENTIFIER → LETTER LORDS
```


Regular Expressions

A **regular expression (R.E.)** is a concise formal characterization of a **regular language** (or **regular set**).

Example: The regular language containing all IDENTs is described by the regular expression

letter (letter | digit)*

where "|" means "or" and "e^*" means "zero or more copies of e."

Regular languages are one particular kind of **formal** languages.

Languages: Some preliminary definitions

- An **alphabet** is a set of symbols (e.g., the ASCII character set).
- A **language** over an alphabet is a set of strings of symbols from that alphabet.
- We write ϵ for the **empty string** (containing zero characters); some authors use λ instead.
- If x and y are strings, then the **concatenation** xy is the string consisting of the the characters of x followed by the characters of y.
- If L and M are languages, then their **concatenation** $LM = \{xy \mid x \in L, y \in M\}$.
- The **exponentiation** of a language L is defined thus: $L^0 = \{\epsilon\}$, the language containing just the empty string, and $L^i = L^{i-1}L$ for $i > 0$.

Regular Expressions and Languages

Each **R.E.** over an alphabet Σ denotes a **regular language** over Σ, according to the following **inductive definition**:

Base rules:

- The R.E. ϵ denotes $\{\epsilon\}$.
- For each $a \in \Sigma$, the R.E. a denotes $\{a\}$, the language containing the single string containing just a.

Inductive rules: If the R.E. R denotes L_R and the R.E. S denotes L_S, then

- $R \mid S$ denotes $L_R \cup L_S$.
- $R \cdot S$ (or just RS) denotes $L_R L_S$.
- R^* denotes $L_R^* = \bigcup_{i=0}^{\infty} L^i$, the **"Kleene closure"** (the concatenation of zero or more strings from L_R).

Also: (R) denotes L_R.

Precedence rules: $()$ before * before $\cdot$ before $|$.

REGULAR EXPRESSIONS

Examples (over alphabet $\{\mathtt{a},\mathtt{b}\}$)

$\mathtt{a}^*$	zero or more a's
$(\mathtt{a} \mid \mathtt{b})^*$	all strings of a's and b's of length ≥ 0
$(\mathtt{a}^*\mathtt{b}^*)^*$	ditto
$(\mathtt{aa} \mid \mathtt{ab} \mid \mathtt{ba} \mid \mathtt{bb})^*$	all strings of a's and b's of even length

Counterexamples (Not every language is regular!)

- $\{\mathtt{a}^n\mathtt{b}^n \mid n \geq 0\}$
- Set of strings over $\{(,)\}$ such that parentheses are properly matched.

Implication: regular languages can't be used to describe arithmetic expressions.

R.E.'s are everywhere in command-line programming tools

grep, Perl, shell commands, etc.

REGULAR DEFINITIONS

Give names to R.E.'s and then use these as a shorthand.

- Must avoid recursive definitions!
- Example of "**syntactic sugar**"

Examples:

id	$\rightarrow$	letter (letter $\mid$ digit)*	
num	$\rightarrow$	digit (digit)*	*or this shorthand:* digit$^+$
if	$\rightarrow$	`if`	*(not too useful!)*
then	$\rightarrow$	`then`	
relop	$\rightarrow$	< $\mid$ > $\mid$ <= $\mid$ >= $\mid$ =	*or:* <(ϵ $\mid$ =) $\mid$ >(ϵ $\mid$ =) $\mid$ =
assgn	$\rightarrow$	`:=`	
string	$\rightarrow$	"(nonquote)*"	
letter	$\rightarrow$	a $\mid$ b $\mid$ ... $\mid$ z $\mid$ A $\mid$ ... $\mid$ Z	*or this shorthand:* [a-zA-Z]
digit	$\rightarrow$	0 $\mid$ 1 $\mid$ 2 $\mid$ 3 $\mid$ 4 $\mid$ 5 $\mid$ 6 $\mid$ 7 $\mid$ 8 $\mid$ 9	*or this shorthand:* [0-9]
nonquote	$\rightarrow$	letter $\mid$ digit $\mid$! $\mid$ $ $\mid$ % $\mid$ ...	

Note that id and keywords have overlapping patterns.

SPECIFYING LEXICAL ANALYZERS

Can define lexical analyzer via list of pairs:

(**regular expression**, **action**)

where **regular expression** describes token pattern (maybe using auxiliary regular definitions),

and **action** is a piece of code, parameterized by the matching lexeme, that returns a (token,attribute) pair.

Example

```
(digit+,                {return new Token(NUM,parseInt(lexeme));})
(alpha(alpha|digit)*, {return new Token(ID,lexeme);})
(space|tab|newline,    {})
(.,.)
(.,.)
(.,.)
```

So R.E'.s can help us **specify** scanners.

But can they help us **generate** running code that does pattern matching?