A parser is a program that, given an input sentence, “recognizes” whether or not that sentence is in the language of a given grammar. Works by reconstructing a derivation for the sentence.

Parser “constructs” parse tree:

- **explicitly** – actual data structure is built; or
- **implicitly** – “semantic actions” are invoked at points corresponding to nodes in the tree, but no tree is actually built.

All parsers read input **left-to-right**, but they differ in how tree is constructed: **top-down** vs. **bottom-up**.

Any context-free grammar can be parsed by a (nondeterministic) pushdown automaton (NFA + stack), but not necessarily by a deterministic one (much less an efficient one).

### Top-Down Parsing

Idea: construct parse tree by starting at start symbol and “guessing” each derivation until we reach a string that matches input.

**Example Grammar:**

\[
S \rightarrow \text{if } E \text{ then } S \text{ else } S | \text{while } E \text{ do } S | \text{print} | \epsilon
\]

\[
E \rightarrow \text{true} | \text{false} | \text{id}
\]

**Token string:**

if id_b then while true do else print

**Tree:**

\[
S \\
if \quad E \quad \text{then} \quad S \quad \text{else} \quad S
\]

**Input:** if id_b then while true do else print

**Action:** Guess for \( S \)

**Tree:**

\[
S \\
if \quad E \quad \text{then} \quad S \quad \text{else} \quad S
\]

**Input:** if id_b then while true do else print

**Action:** if matches; guess for \( E \)
Tree:

if E then S else S

id

Input: id\textsubscript{b} then while true do else print
Action:

id\textsubscript{b} matches; then matches; guess for S

Tree:

if E then S else S

id\textsubscript{b} while E do S

Input: while true do else print
Action:

while matches; guess for E

Tree:

if E then S else S

id\textsubscript{b} while E do S

true

Input: true do else print
Action:

true matches; do matches; guess for S

Tree:

if E then S else S

id\textsubscript{b} while E do S

\texttt{\textepsilon}

Input: else print
Action:

\texttt{\textepsilon} matches; else matches; guess for S
Implementation of top-down parser using a recursive procedure for each non-terminal.

For many languages, can make perfect guesses (avoid back-tracking) by using 1-symbol lookahead. I.e., if 

\[ A \rightarrow \alpha_1 | \alpha_2 | \ldots | \alpha_n \]

choose correct \( \alpha_i \) by looking at first symbol it derives.

(If \( \epsilon \) is an alternative, choose it last.)

This approach is also called predictive parsing

Recursive-descent parsers are easy to write by hand and reasonably efficient.

Often must massage grammar into suitable form (more later).

Not all languages can be parsed this way.

void e() {
    if (tok == TRUE || tok == FALSE || tok == ID) {
        tok = lex(); /* get next input token */
        e();
    } else if (tok == THEN) {
        tok = lex();
        s(); /* recursive call! */
        if (tok == ELSE) {
            tok = lex();
            s();
            } else error(); /* issue error message */
        } else if (tok == ELSE) {
            tok = lex();
            s();
            } else error(); /* issue error message */
        } else if (tok == WHILE) {
            tok = lex();
            e();
            if (tok == DO) {
                tok = lex();
                s();
                } else error();
            } else/* epsilon case falls out */
        }

For the same grammar as before:

\[ S \rightarrow \text{if } E \text{ then } S \text{ else } S | \text{while } E \text{ do } S | \text{print } \epsilon \]

\[ E \rightarrow \text{true} | \text{false} | \text{id} \]

We write one function to parse expressions, and another to parse statements.

Each function returns normally if it successfully parsed; otherwise it calls error(), which we assume issues an error message and terminates the parser.

For simplicity in this example, the parser functions don’t return anything.

We assume that lex() returns the next token and advances the lexical analyzer’s token stream.

void s() {
    if (tok == IF) {
        tok = lex(); /* get next input token */
        e();
        if (tok == THEN) {
            tok = lex();
            s(); /* recursive call! */
            if (tok == ELSE) {
                tok = lex();
                s();
                } else error(); /* issue error message */
            } else error(); /* issue error message */
        } else if (tok == WHILE) {
            tok = lex();
            e();
            if (tok == DO) {
                tok = lex();
                s();
                } else error();
            } else if (tok == PRINT) {
                tok = lex();
            } else/* epsilon case falls out */
        }

For the same grammar as before:

\[ S \rightarrow \text{if } E \text{ then } S \text{ else } S | \text{while } E \text{ do } S | \text{print } \epsilon \]

\[ E \rightarrow \text{true} | \text{false} | \text{id} \]

We write one function to parse expressions, and another to parse statements.

Each function returns normally if it successfully parsed; otherwise it calls error(), which we assume issues an error message and terminates the parser.

For simplicity in this example, the parser functions don’t return anything.

We assume that lex() returns the next token and advances the lexical analyzer’s token stream.

void e() {
    if (tok == TRUE || tok == FALSE || tok == ID) {
        tok = lex(); /* get next input token */
        e();
    } else if (tok == THEN) {
        tok = lex();
        s(); /* recursive call! */
        if (tok == ELSE) {
            tok = lex();
            s();
            } else error(); /* issue error message */
        } else if (tok == ELSE) {
            tok = lex();
            s();
            } else error(); /* issue error message */
        } else if (tok == WHILE) {
            tok = lex();
            e();
            if (tok == DO) {
                tok = lex();
                s();
                } else error();
            } else/* epsilon case falls out */
        }

Recursive-Descent Parsing

Recursive-Descent Example

For the same grammar as before:

\[ S \rightarrow \text{if } E \text{ then } S \text{ else } S | \text{while } E \text{ do } S | \text{print } \epsilon \]

\[ E \rightarrow \text{true} | \text{false} | \text{id} \]

We write one function to parse expressions, and another to parse statements.

Each function returns normally if it successfully parsed; otherwise it calls error(), which we assume issues an error message and terminates the parser.

For simplicity in this example, the parser functions don’t return anything.

We assume that lex() returns the next token and advances the lexical analyzer’s token stream.
PROBLEMS FOR RECURSIVE-DESCRIPT PARSING

• Left recursion: a derivation
  \[ A \Rightarrow A\alpha \]
causes parser to loop!

Solution: **Remove** left recursion from grammar.

• Need to backtrack (inefficient) because one-symbol lookahead can’t “guess” correctly, e.g.:
  \[ S \rightarrow V := \text{int} \]
  \[ V \rightarrow \text{alpha ‘[’ int ‘]’ | alpha} \]
Possible inputs: \( x := 77 \) or \( x[2] := 17 \).

Which alternative should we choose for \( V \)?

Solution: **Left-factor** the grammar.

• These problems arise naturally in expression grammars. (Can usually prevent them in statement grammars by careful language design.)

---

ELIMINATING IMMEDIATE LEFT RECURSION

Replace left-recursive productions of the form
  \[ A \rightarrow A\alpha | \beta \]
which generate sentences of the form
  \( \beta, \beta\alpha, \beta\alpha\alpha, \ldots \)
by the **right-recursive** productions
  \[ A \rightarrow \beta A' \]
  \[ A' \rightarrow \alpha A' | \epsilon \]

Yields different parse trees but same language:

---

ELIMINATING ALL LEFT-RECURSION

Consider
  \[ S \rightarrow Aa | b \]
  \[ A \rightarrow Sc | d \]

Non-terminal \( S \) is left-recursive in two steps:
  \( S \Rightarrow Aa \Rightarrow Sca \Rightarrow Aaca \Rightarrow Scaca \Rightarrow \ldots \)

**Fairly General Algorithm**

(Works unless \( A \Rightarrow A \) or \( A \Rightarrow \epsilon \). Fully general algorithm exists, but is complicated!)

• arrange non-terminals in some order \( A_1, \ldots, A_n \).
  • for \( i = 1 \) to \( n \) do
    • for \( j = 1 \) to \( i-1 \) do
      • for any production \( A_i \rightarrow A_j\alpha \)
        replace it by substituting definition of \( A_j \) into r.h.s.,
        i.e., by changing it to
        \[ A_i \rightarrow \beta_1\alpha | \ldots | \beta_m\alpha \]
        where current productions for \( A_j \) are
        \[ A_j \rightarrow \beta_1 | \ldots | \beta_m \]
      • eliminate any immediate left-recursion in \( A_i \)

For a simplified expression grammar:
  \[ E \rightarrow E \ast T | E \ast T | T \]
  \[ T \rightarrow T \ast F | T / F | F \]
  \[ F \rightarrow (E) | \text{id} \]
becomes
  \[ E \rightarrow T E' \]
  \[ E' \rightarrow \ast T E' | \ast T E' | \epsilon \]
  \[ T \rightarrow F T' \]
  \[ T' \rightarrow \ast F T' | F T' | \epsilon \]
  \[ F \rightarrow (E) | \text{id} \]

But note that the desired left-associativity has been lost!
Example: Consider a grammar with the non-terminals ordered as follows:

(1) \( S \rightarrow P \ a | b \)
(2) \( P \rightarrow Sc | Rd \)
(3) \( R \rightarrow Sc | f \)

Step i=2, j=1: In-line (1) into (2), giving

(2) \( P \rightarrow P ac | bc \)
and then remove immediate left-recursion, giving

(2) \( P \rightarrow bcP' | RdP' \)
(2') \( P' \rightarrow acP' | \epsilon \)

Step i=3, j=1: Inline (1) into (3), giving

(3) \( R \rightarrow P ae | be | f \)

Step i=3, j=2: Inline new (2) into new (3), giving

(3) \( R \rightarrow bcP' ae | RdP' ae | be | f R' \)
and then remove immediate left-recursion, giving

(3) \( R \rightarrow bcP' aeR' | beR' | f R' \)
(3') \( R' \rightarrow dP' aeR' | \epsilon \)

Expression Parsing Using Recursive Descent (II)

We can simplify this code (and improve its performance) by turning the recursions into iterations, e.g.:

```c
e1() {
  if (tok == '+') {
    tok = lex(); t(); e1();
  } /* epsilon case falls through */
}
```

becomes:

```c
e1() {
  while (tok == '+') {
    tok = lex(); t();
  }
}
```
We then inline functions that are (now) called only once, e.g.:

```c
e() { t(); e1(); }
```

becomes:

```c
e() {
    t();
    while (tok == '+') {
        tok = lex();
        t();
    }
}
```

Performing the same transformation on `t` and `t1`, and adding the usual recursive descent code for `f`, we get...

```c
e() {
    t();
    while (tok == '+') { lex(); t(); }
}
```

```c
t() {
    f();
    while (tok == '*') { lex(); f(); }
}
```

```c
f() {
    if (tok == ID)
        lex();
    else if (tok == '(') {
        lex();
        e();
        if (tok == ')
            lex();
        else error();
        else error();
    }
```