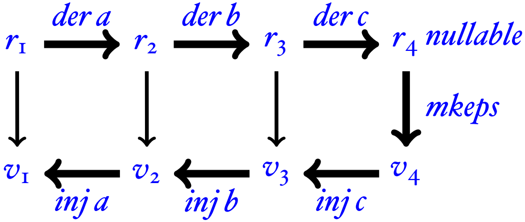
Compilers and Formal Languages (6)

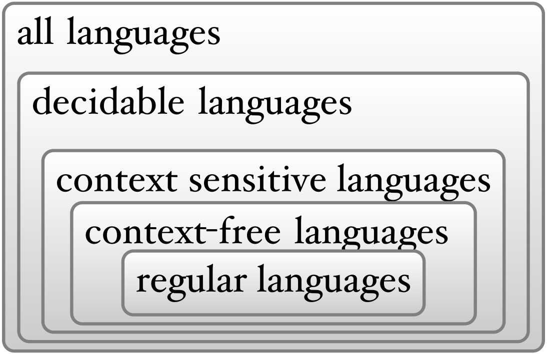
Email:	christian.urban at kcl.ac.uk
Office:	S1.27 (1st floor Strand Building)
Slides:	KEATS (also home work is there)



$$
\begin{aligned}
inj\ (c)\ c\ Empty &\stackrel{\text{def}}{=} Char\ c\\
inj\ (r_1 + r_2)\ c\ Left(v) &\stackrel{\text{def}}{=} Left(inj\ r_1\ c\ v)\\
inj\ (r_1 + r_2)\ c\ Right(v) &\stackrel{\text{def}}{=} Right(inj\ r_2\ c\ v)\\
inj\ (r_1 \cdot r_2)\ c\ Seq(v_1, v_2) &\stackrel{\text{def}}{=} Seq(inj\ r_1\ c\ v_1, v_2)\\
inj\ (r_1 \cdot r_2)\ c\ Left(Seq(v_1, v_2)) &\stackrel{\text{def}}{=} Seq(inj\ r_1\ c\ v_1, v_2)\\
inj\ (r_1 \cdot r_2)\ c\ Right(v) &\stackrel{\text{def}}{=} Seq(mkeps(r_1), inj\ r_2\ c\ v)\\
inj\ (r^*)\ c\ Seq(v, vs) &\stackrel{\text{def}}{=} inj\ r\ c\ v\ ::\ vs
\end{aligned}
$$

Hierarchy of Languages

Recall that languages are sets of strings.



Two Grammars

Which languages are recognised by the following two grammars?

$$
\begin{array}{rcl}
S & \rightarrow & 1 \cdot S \cdot S \\
  & | & \epsilon
\end{array}
$$

$$
\begin{array}{rcl}
U & \rightarrow & 1 \cdot U \\
  & | & \epsilon
\end{array}
$$

Atomic parsers, for example

$$1 :: rest \;\Rightarrow\; \{(1, rest)\}$$

- you consume one or more token from the input (stream)
- also works for characters and strings

Alternative parser (code $p \;||\; q$)

- apply p and also q; then combine the outputs

$$p(\text{input}) \cup q(\text{input})$$

Sequence parser (code $p \sim q$)

- apply first p producing a set of pairs
- then apply q to the unparsed parts
- then combine the results:

$$((\text{output}_1, \text{output}_2), \text{unparsed part})$$

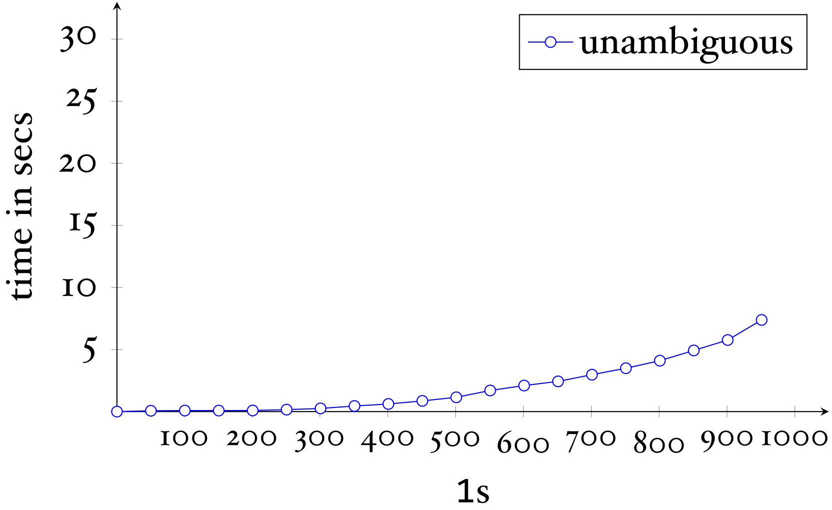
$$\{((o_1, o_2), u_2) \mid (o_1, u_1) \in p(\text{input}) \wedge (o_2, u_2) \in q(u_1)\}$$

Function parser (code $p \Rightarrow f$)

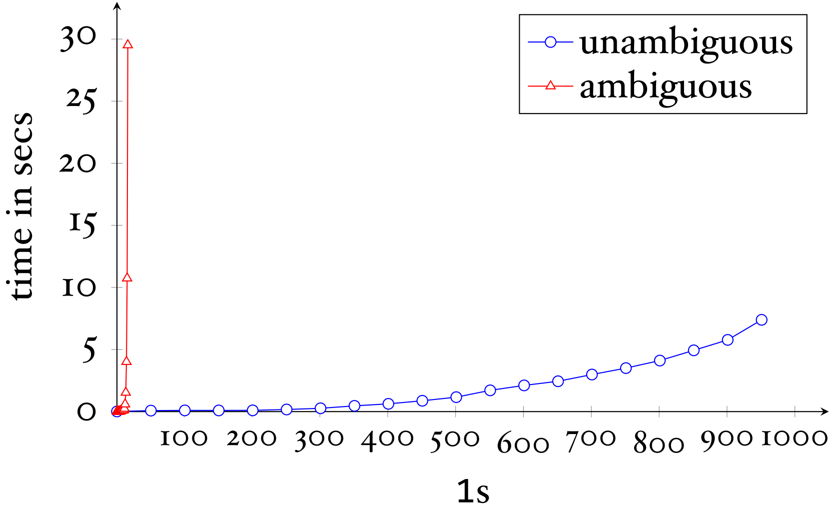
- apply p producing a set of pairs
- then apply the function f to each first component

$$\{(f(o_1), u_1) \mid (o_1, u_1) \in p(\text{input})\}$$

Ambiguous Grammars



Ambiguous Grammars



Arithmetic Expressions

A grammar for arithmetic expressions and numbers:

$$
\begin{array}{lcl}
E & \rightarrow & E \cdot + \cdot E \mid E \cdot * \cdot E \mid (\cdot E \cdot) \mid N \\
N & \rightarrow & N \cdot N \mid 0 \mid 1 \mid \ldots \mid 9
\end{array}
$$

Unfortunately it is left-recursive (and ambiguous).

A problem for recursive descent parsers
(e.g. parser combinators).

Arithmetic Expressions

A grammar for arithmetic expressions and numbers:

$$\begin{array}{lcl} E & \rightarrow & E \cdot + \cdot E \mid E \cdot * \cdot E \mid (\cdot E \cdot) \mid N \\ N & \rightarrow & N \cdot N \mid 0 \mid 1 \mid \ldots \mid 9 \end{array}$$

Unfortunately it is left-recursive (and ambiguous).

A problem for recursive descent parsers (e.g. parser combinators).

Numbers

$$N \;\rightarrow\; N \cdot N \mid 0 \mid 1 \mid \ldots \mid 9$$

A non-left-recursive, non-ambiguous grammar for numbers:

$$N \;\rightarrow\; 0 \cdot N \mid 1 \cdot N \mid \ldots \mid 0 \mid 1 \mid \ldots \mid 9$$

Removing Left-Recursion

The rule for numbers is directly left-recursive:

$$N \rightarrow N \cdot N \mid 0 \mid 1 \quad (\ldots)$$

Translate

$$\begin{array}{rcl} N & \rightarrow & N \cdot \alpha \\ & \mid & \beta \end{array} \quad \Rightarrow \quad \begin{array}{rcl} N & \rightarrow & \beta \cdot N' \\ N' & \rightarrow & \alpha \cdot N' \\ & \mid & \epsilon \end{array}$$

Removing Left-Recursion

The rule for numbers is directly left-recursive:

$$N \;\rightarrow\; N \cdot N \mid 0 \mid 1 \quad (\ldots)$$

Translate

$$\begin{array}{rcl} N & \rightarrow & N \cdot \alpha \\ & \mid & \beta \end{array} \quad\Rightarrow\quad \begin{array}{rcl} N & \rightarrow & \beta \cdot N' \\ N' & \rightarrow & \alpha \cdot N' \\ & \mid & \epsilon \end{array}$$

Which means

$$\begin{array}{rcl} N & \rightarrow & 0 \cdot N' \mid 1 \cdot N' \\ N' & \rightarrow & N \cdot N' \mid \epsilon \end{array}$$

Operator Precedences

To disambiguate

$$E \;\rightarrow\; E \cdot + \cdot E \mid E \cdot * \cdot E \mid (\cdot E \cdot) \mid N$$

Decide on how many precedence levels, say

highest for $()$, medium for $*$, lowest for $+$

$$\begin{array}{lcl} E_{low} & \rightarrow & E_{med} \cdot + \cdot E_{low} \mid E_{med} \\ E_{med} & \rightarrow & E_{hi} \cdot * \cdot E_{med} \mid E_{hi} \\ E_{hi} & \rightarrow & (\cdot E_{low} \cdot) \mid N \end{array}$$

Operator Precedences

To disambiguate

$$E \rightarrow E \cdot + \cdot E \mid E \cdot * \cdot E \mid (\cdot E \cdot) \mid N$$

Decide on how many precedence levels, say

highest for $()$, medium for $*$, lowest for $+$

$$\begin{array}{lcl} E_{low} & \rightarrow & E_{med} \cdot + \cdot E_{low} \mid E_{med} \\ E_{med} & \rightarrow & E_{hi} \cdot * \cdot E_{med} \mid E_{hi} \\ E_{hi} & \rightarrow & (\cdot E_{low} \cdot) \mid N \end{array}$$

What happens with $1 + 3 + 4$?

Chomsky Normal Form

All rules must be of the form

$$A \rightarrow a$$

or

$$A \rightarrow B \cdot C$$

No rule can contain ϵ.

ϵ-Removal

1. If $A \rightarrow \alpha \cdot B \cdot \beta$ and $B \rightarrow \epsilon$ are in the grammar, then add $A \rightarrow \alpha \cdot \beta$ (iterate if necessary).
2. Throw out all $B \rightarrow \epsilon$.

$$\begin{aligned} N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \\ N' &\rightarrow N \cdot N' \mid \epsilon \end{aligned}$$

$$\begin{aligned} N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \mid 0 \mid 1 \\ N' &\rightarrow N \cdot N' \mid N \mid \epsilon \end{aligned}$$

$$\begin{aligned} N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \mid 0 \mid 1 \\ N' &\rightarrow N \cdot N' \mid N \end{aligned}$$

ϵ-Removal

1. If $A \rightarrow \alpha \cdot B \cdot \beta$ and $B \rightarrow \epsilon$ are in the grammar, then add $A \rightarrow \alpha \cdot \beta$ (iterate if necessary).
2. Throw out all $B \rightarrow \epsilon$.

$$
\begin{aligned}
N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \\
N' &\rightarrow N \cdot N' \mid \epsilon
\end{aligned}
$$

$$
\begin{aligned}
N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \mid 0 \mid 1 \\
N' &\rightarrow N \cdot N' \mid N \mid \epsilon
\end{aligned}
$$

$$
\begin{aligned}
N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \mid 0 \mid 1 \\
N' &\rightarrow N \cdot N' \mid N
\end{aligned}
$$

$$N \rightarrow 0 \cdot N \mid 1 \cdot N \mid 0 \mid 1$$

CYK Algorithm

If grammar is in Chomsky normalform ...

$$
\begin{array}{lcl}
S & \rightarrow & N \cdot P \\
P & \rightarrow & V \cdot N \\
N & \rightarrow & N \cdot N \\
N & \rightarrow & \texttt{students} \mid \texttt{Jeff} \mid \texttt{geometry} \mid \texttt{trains} \\
V & \rightarrow & \texttt{trains}
\end{array}
$$

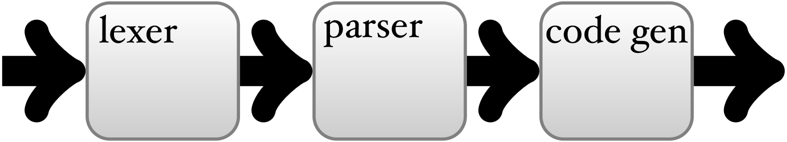
`Jeff trains geometry students`

CYK Algorithm

- fastest possible algorithm for recognition problem
- runtime is $O(n^3)$
- grammars need to be transferred into CNF

The Goal of this Course

Write a Compiler



We have lexer and parser.

$$
\begin{array}{lcl}
\mathit{Stmt} & \rightarrow & \texttt{skip}\\
& | & \mathit{Id}\ \texttt{:=}\ \mathit{AExp}\\
& | & \texttt{if}\ \mathit{BExp}\ \texttt{then}\ \mathit{Block}\ \texttt{else}\ \mathit{Block}\\
& | & \texttt{while}\ \mathit{BExp}\ \texttt{do}\ \mathit{Block}\\
& | & \texttt{read}\ \mathit{Id}\\
& | & \texttt{write}\ \mathit{Id}\\
& | & \texttt{write}\ \mathit{String}\\
\mathit{Stmts} & \rightarrow & \mathit{Stmt}\ \texttt{;}\ \mathit{Stmts}\\
& | & \mathit{Stmt}\\
\mathit{Block} & \rightarrow & \{\ \mathit{Stmts}\ \}\\
& | & \mathit{Stmt}\\
\mathit{AExp} & \rightarrow & \ldots\\
\mathit{BExp} & \rightarrow & \ldots
\end{array}
$$

```
write "Fib";
read n;
minus1 := 0;
minus2 := 1;
while n > 0 do {
       temp := minus2;
       minus2 := minus1 + minus2;
       minus1 := temp;
       n := n - 1
};
write "Result";
write minus2
```

An Interpreter

$$\{\; x := 5;\; y := x * 3;\; y := x * 4;\; x := u * 3 \;\}$$

- the interpreter has to record the value of x before assigning a value to y

An Interpreter

$$
\begin{array}{l}
\{ \\
\quad x := 5; \\
\quad y := x * 3; \\
\quad y := x * 4; \\
\quad x := u * 3 \\
\}
\end{array}
$$

- the interpreter has to record the value of x before assigning a value to y
- eval(stmt, env)

Interpreter

$$
\begin{array}{lcl}
\text{eval}(n, E) & \stackrel{\text{def}}{=} & n \\
\text{eval}(x, E) & \stackrel{\text{def}}{=} & E(x) \quad \text{lookup } x \text{ in } E \\
\text{eval}(a_1 + a_2, E) & \stackrel{\text{def}}{=} & \text{eval}(a_1, E) + \text{eval}(a_2, E) \\
\text{eval}(a_1 - a_2, E) & \stackrel{\text{def}}{=} & \text{eval}(a_1, E) - \text{eval}(a_2, E) \\
\text{eval}(a_1 * a_2, E) & \stackrel{\text{def}}{=} & \text{eval}(a_1, E) * \text{eval}(a_2, E) \\
\\
\text{eval}(a_1 = a_2, E) & \stackrel{\text{def}}{=} & \text{eval}(a_1, E) = \text{eval}(a_2, E) \\
\text{eval}(a_1 \mathbin{!=} a_2, E) & \stackrel{\text{def}}{=} & \neg(\text{eval}(a_1, E) = \text{eval}(a_2, E)) \\
\text{eval}(a_1 < a_2, E) & \stackrel{\text{def}}{=} & \text{eval}(a_1, E) < \text{eval}(a_2, E)
\end{array}
$$

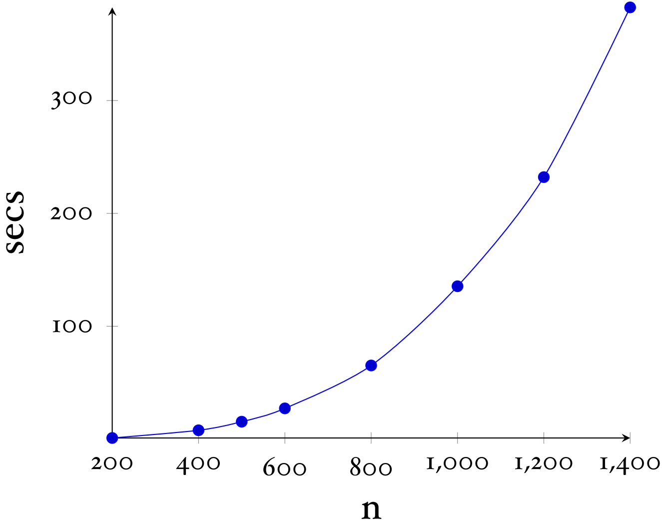
Interpreter (2)

$$
\begin{array}{l}
\text{eval}(\text{skip}, E) \quad \stackrel{\text{def}}{=} \quad E \\
\text{eval}(x := a, E) \quad \stackrel{\text{def}}{=} \quad E(x \mapsto \text{eval}(a, E)) \\
\text{eval}(\text{if } b \text{ then } cs_1 \text{ else } cs_2, E) \stackrel{\text{def}}{=} \\
\qquad\qquad \text{if eval}(b, E) \text{ then eval}(cs_1, E) \\
\qquad\qquad\qquad\qquad\qquad\quad \text{else eval}(cs_2, E) \\
\text{eval}(\text{while } b \text{ do } cs, E) \stackrel{\text{def}}{=} \\
\qquad\qquad \text{if eval}(b, E) \\
\qquad\qquad \text{then eval}(\text{while } b \text{ do } cs, \text{eval}(cs, E)) \\
\qquad\qquad \text{else } E \\
\text{eval}(\text{write } x, E) \quad \stackrel{\text{def}}{=} \quad \{ \text{ println}(E(x)) \;;\; E \text{ } \}
\end{array}
$$

Test Program

```
start := 1000;
x := start;
y := start;
z := start;
while 0 < x do {
 while 0 < y do {
  while 0 < z do { z := z - 1 };
  z := start;
  y := y - 1
 };
 y := start;
 x := x - 1
}
```

Interpreted Code



Java Virtual Machine

- introduced in 1995
- is a stack-based VM (like Postscript, CLR of .Net)
- contains a JIT compiler
- many languages take advantage of JVM's infrastructure (JRE)
- is garbage collected $\Rightarrow$ no buffer overflows
- some languages compile to the JVM: Scala, Clojure...