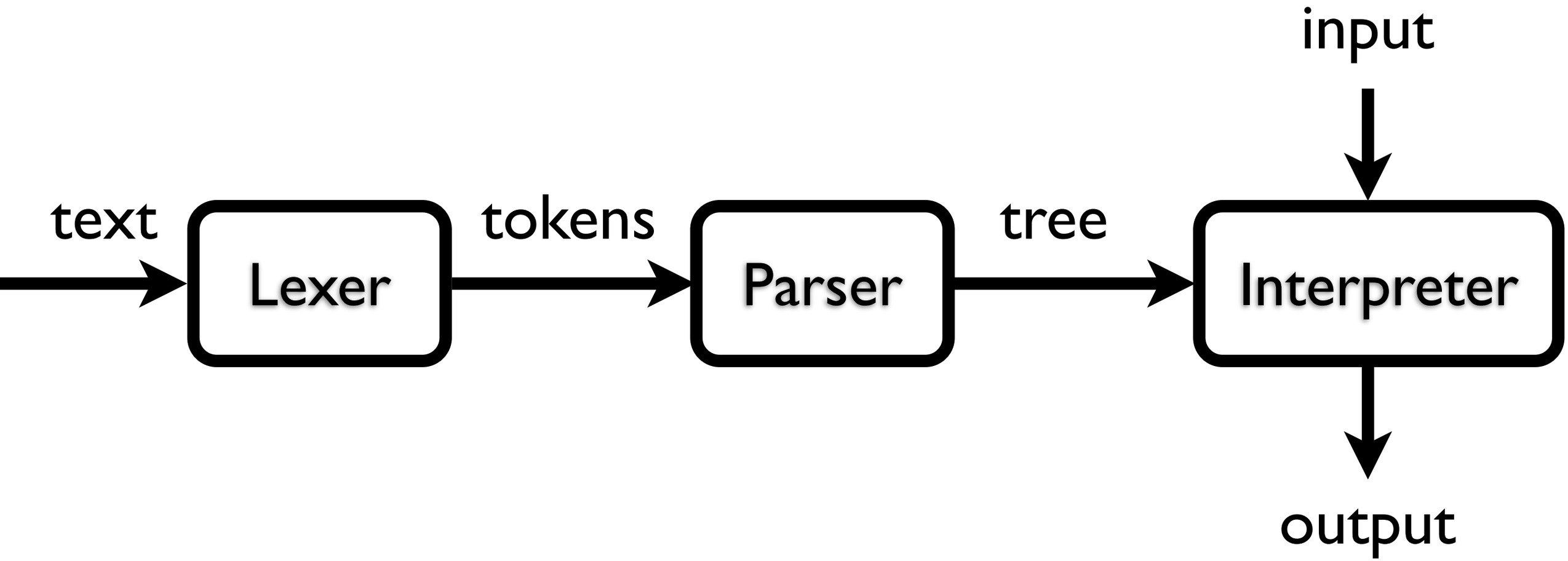
Parsing: Tokens to Trees

Matt Might
University of Utah
matt.might.net

Today

- Context-free languages
- Context-free grammars
- Recursive-descent parsers
- LALR parser-generators

Interpreters


input
text
Lexer
tokens
Parser
tree
Interpreter
output

What is parsing?

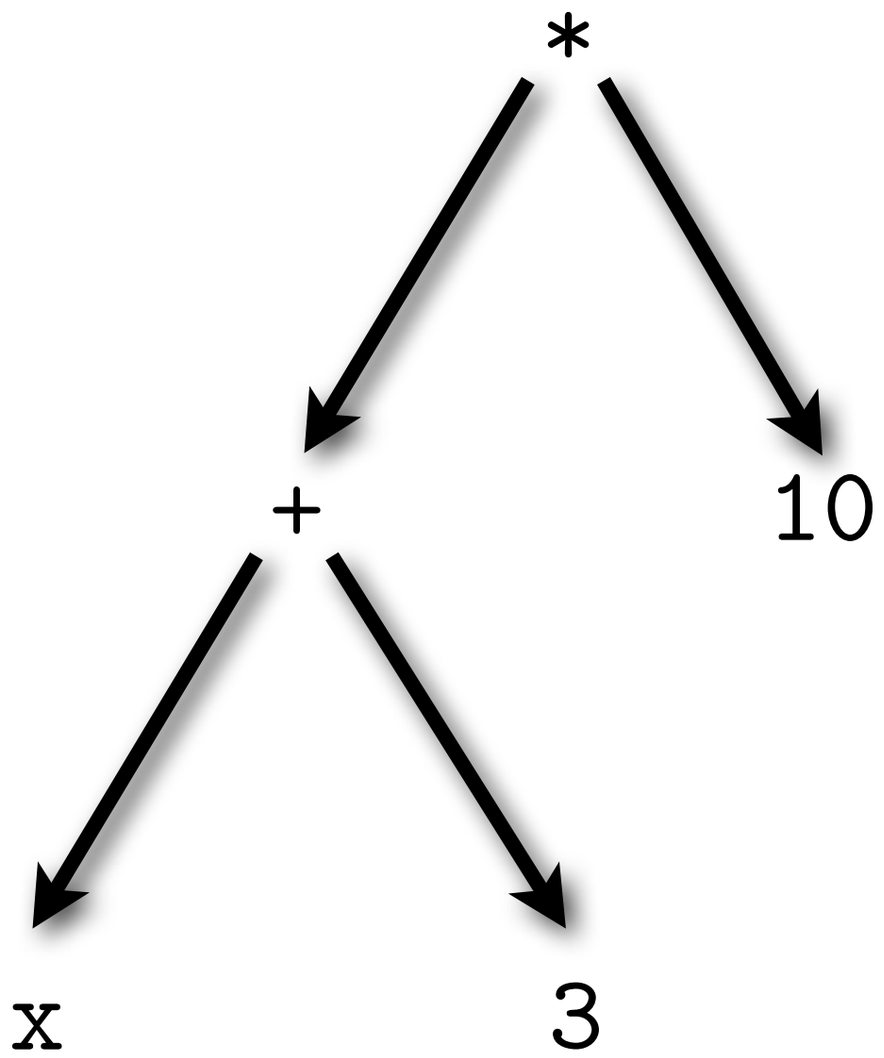
Unstructure to structure.

A transformation from a sequence to a tree.

$$(x + 3) * 10$$

```
(x + 3) * 10
```

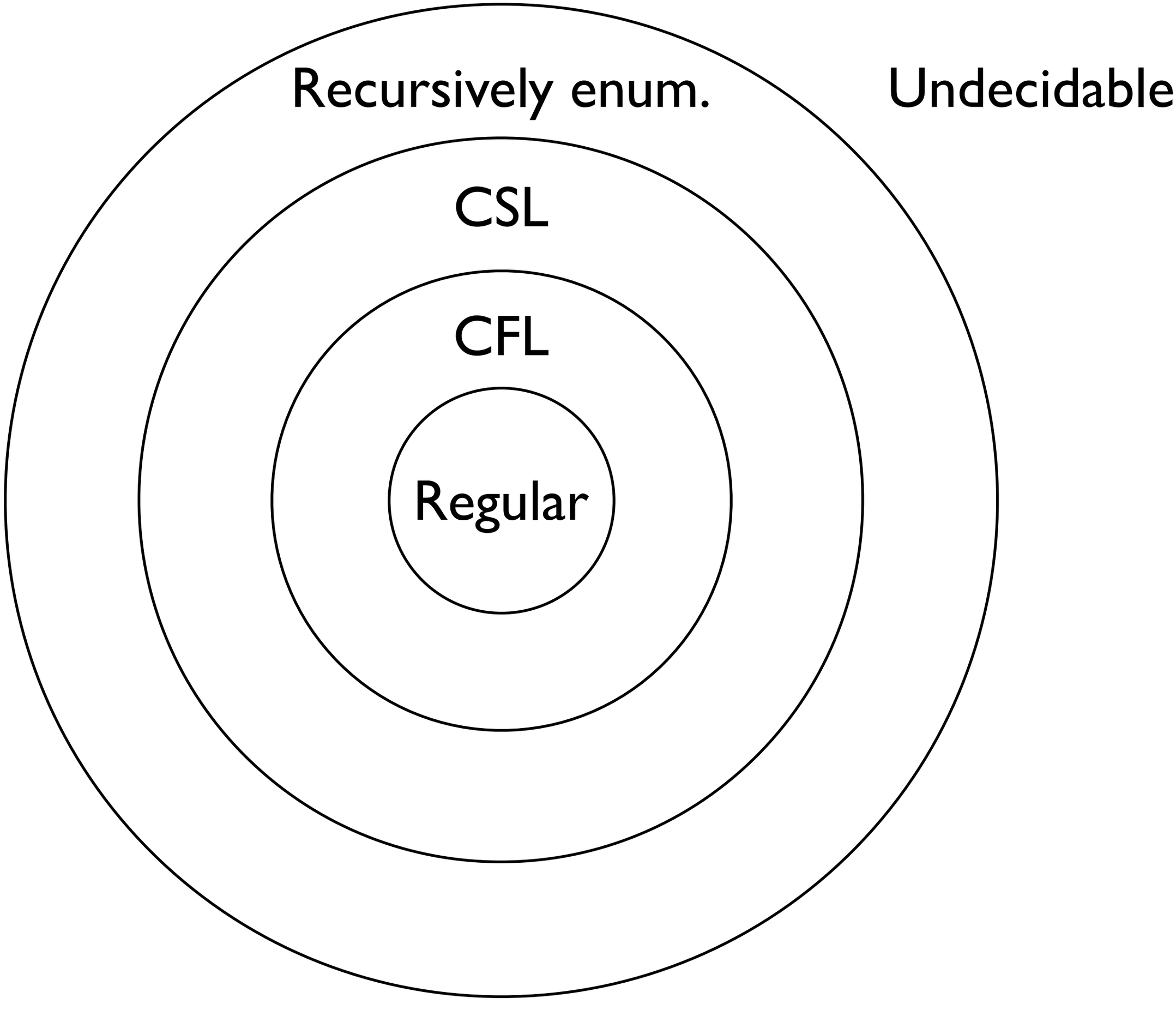
```
(x + 3) * 10
```

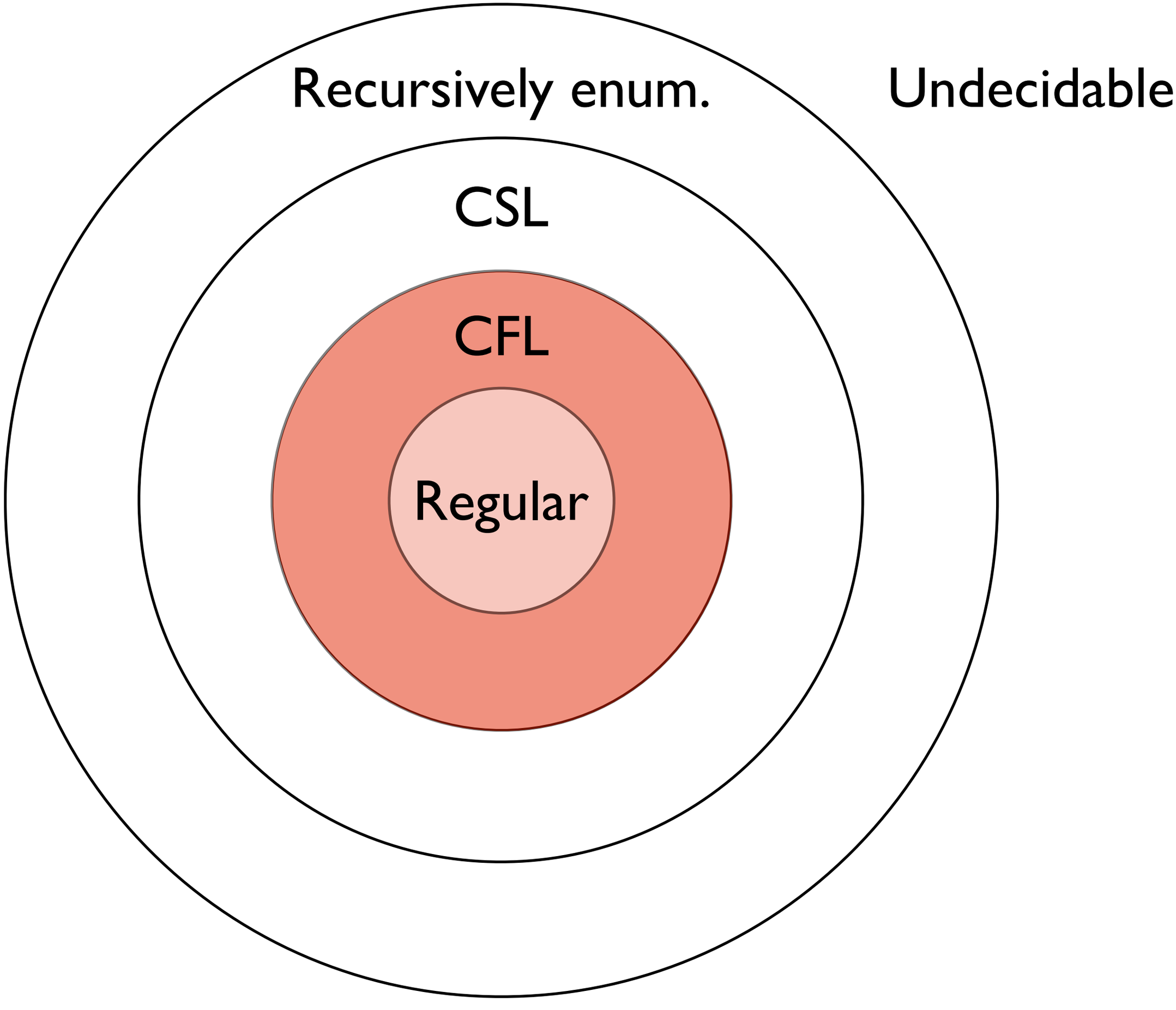

*
+
10
x
3

grammars :: tree structure

What's a context-free grammar?

Recursive regular expressions.

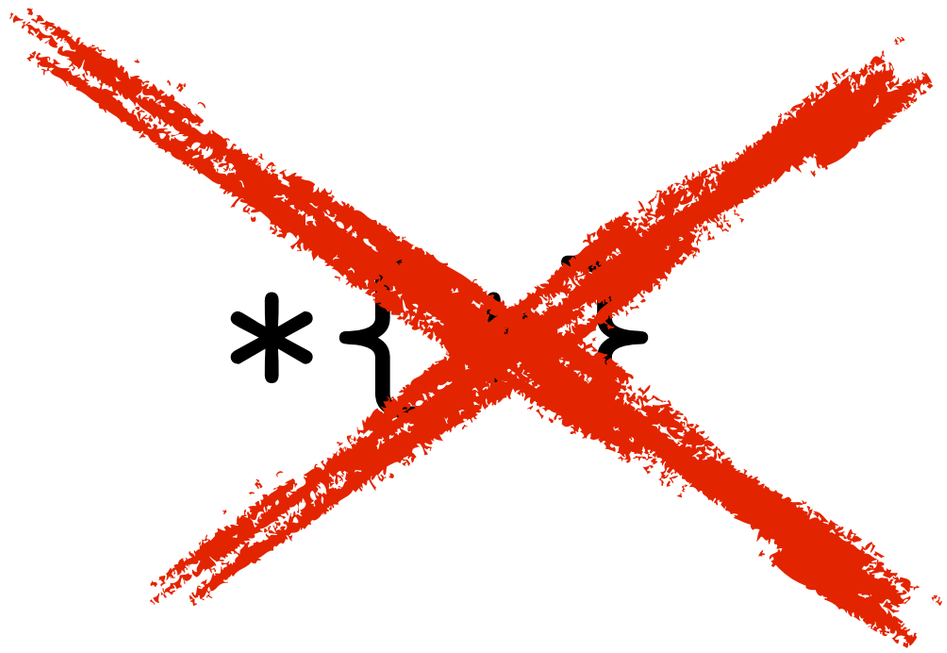

Recursively enum.
Undecidable
CSL
CFL
Regular


Recursively enum.
Undecidable
CSL
CFL
Regular

Regex for balanced braces?

$$|\{\}|\{\{\}\}|\{\{\{\}\}\}|\{\{\{\{\}\}\}\}|\ldots$$

{*}*



$$S ::= \{S\} | \varepsilon$$

CFG ingredients

- Sequence
- Alternation
- Recursion

Kleene star?

$$\ldots A^{*}\ldots$$

$$\ldots B \ldots$$

$$B ::= AB$$

$$\mid \varepsilon$$

Context-free grammar

- **Terminals**: Set of atomic symbols
- **Nonterminals**: Phrase types
- **Productions**: Phrase structure rules
- **Start symbol**: Top-level phrase

Notations

- Backus-Naur Form (BNF)
- Extended Backus-Naur Form (EBNF)
- Phrase structure rules
- Set-based language construction

Backus-Naur Form

$$\langle nonterm \rangle \texttt{ ::= } exp_1 \texttt{ ... } exp_2$$

Backus-Naur Form

exp is <*nonterm*> or *term*

EBNF

exp is <*nonterm*>
or *term*
or *exp** or { *exp* }
or *exp*+
or *exp*? or [*exp*]
or (*exp*)
or *exp* | *exp*

Phrase structure rules

$$NT \rightarrow S_1 \ldots S_n$$

Phrase structure rules

S is nonterminal or terminal

Set-based languages

$$L' = \epsilon \text{ or } \{\langle\rangle\}$$

$$L' = c \text{ or } \{c\}$$

$$L' = \emptyset \text{ or } \{\}$$

$$L' = L_1 \cup L_2$$

$$L' = L_1 \circ L_2$$

Examples

S-Expression

```
<S>  ::= "(" <S>* ")" | <A>

<A>  ::= SYMBOL
       |  NUMBER
```

Terminal symbols

```
<S>  ::= "(" <S>* ")" | <A>

<A>  ::= SYMBOL
       | NUMBER
```

Non-terminal symbols

```
<S> ::= "(" <S>* ")" | <A>

<A> ::= SYMBOL
      |  NUMBER
```

Non-terminal symbols

```
<S> ::= "(" <S>* ")" | <A>

<A> ::= SYMBOL
      | NUMBER
```

Productions

```
<S>  ::= "(" <S>* ")" | <A>
```

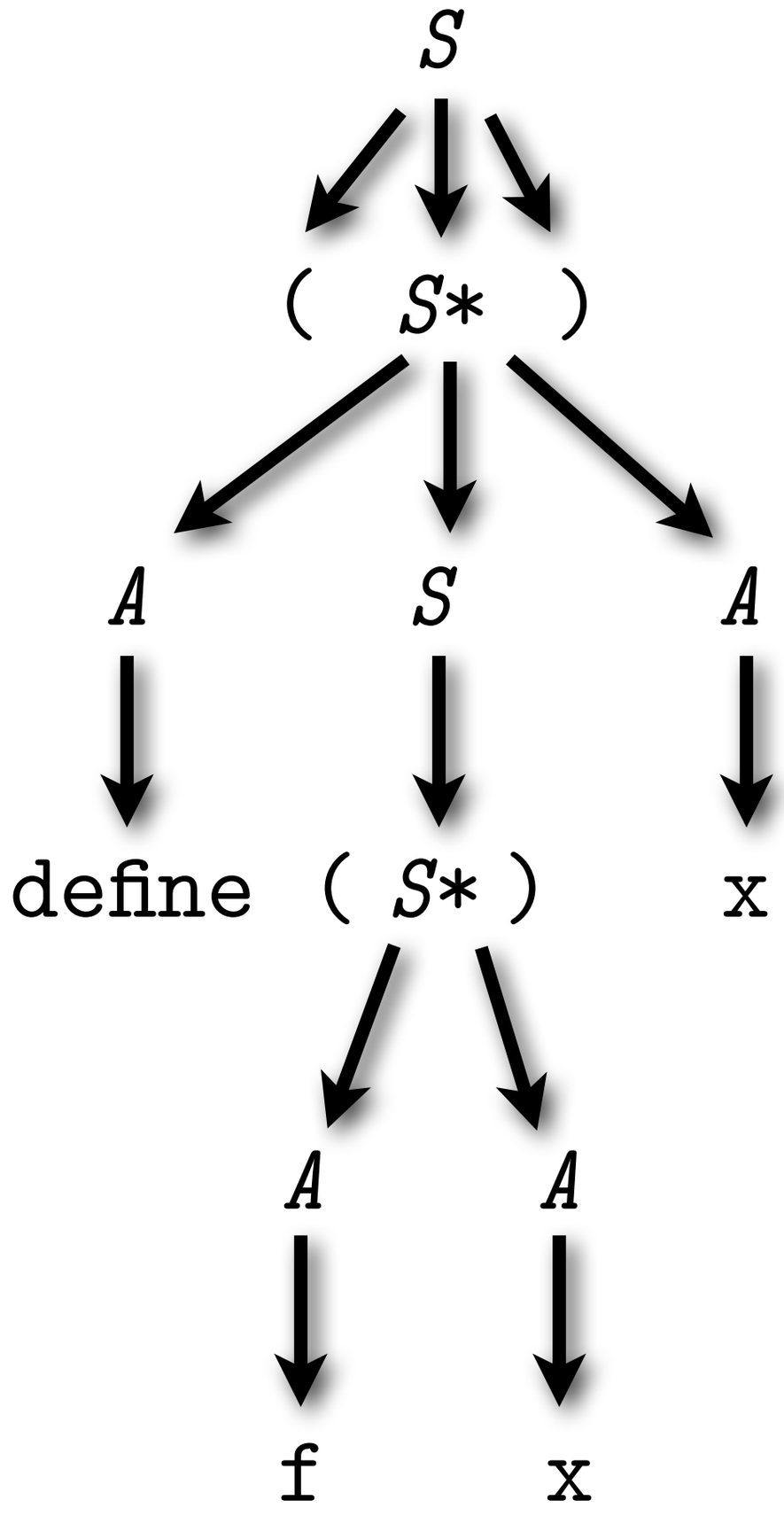
```
<A>  ::= SYMBOL
       |  NUMBER
```

Start symbol

```
<S> ::= "(" <S>* ")" | <A>

<A> ::= SYMBOL
      | NUMBER
```

```
(define (id x) x)
```


S
(S*)
A
S
A
define (S*) x
A
A
f
x

Terms

$$E \to T + E$$

$$E \to T$$

$$T \to F \times T$$

$$T \to F$$

$$F \to (\ E \ )$$

$$F \to n$$

Terminal symbols

$$E \rightarrow T + E$$

$$E \rightarrow T$$

$$T \rightarrow F \times T$$

$$T \rightarrow F$$

$$F \rightarrow (E)$$

$$F \rightarrow n$$

Nonterminal symbols

$$E \rightarrow T + E$$

$$E \rightarrow T$$

$$T \rightarrow F \times T$$

$$T \rightarrow F$$

$$F \rightarrow (\ E \ )$$

$$F \rightarrow n$$

Nonterminal symbols

$$E \to T + E$$
$$E \to T$$
$$T \to F \times T$$
$$T \to F$$
$$F \to (E)$$
$$F \to n$$

Productions

$$E \to T + E$$

$$E \to T$$

$$T \to F \times T$$

$$T \to F$$

$$F \to (\ E \ )$$

$$F \to n$$

Start symbol

$$\boxed{E} \rightarrow T + E$$

$$\boxed{E} \rightarrow T$$

$$T \rightarrow F \times T$$

$$T \rightarrow F$$

$$F \rightarrow (\ E \ )$$

$$F \rightarrow n$$

Balanced parens

$$P = (\circ P \circ) \circ P \cup \epsilon$$

Terminals

$$P = \boxed{(} \circ P \circ \boxed{)} \circ P \cup \epsilon$$

Nonterminals

$$P = (\circ P \circ) \circ P \cup \epsilon$$

Productions

$$P = (\circ P \circ) \circ P \cup \epsilon$$

Start symbol

$$P = (\circ P \circ) \circ P \cup \epsilon$$

How to parse (the hard way)

Recursive descent

- One function per nonterminal
- Allowed to look one token ahead

Recursive descent

- **peek()** : preview next token
- **next()** : consume & return next token
- **eat(token)** : match next token

Recursive descent examples

Expressions

$$E \to T + E$$

$$E \to T$$

$$T \to F \times T$$

$$T \to F$$

$$F \to (\ E \ )$$

$$F \to n$$

S-Expressions

```
<S>  ::= "(" <SL> ")"
       |  <A>
<SL> ::= <S> <SL>
       |
<A>  ::= SYMBOL
       |  INTEGER
       |  #t
       |  #f
```