Lexical Error Recovery

A character sequence that can't be scanned into any valid token is a *lexical error*.

Lexical errors are uncommon, but they still must be handled by a scanner. We won't stop compilation because of so minor an error.

Approaches to lexical error handling include:

- Delete the characters read so far and restart scanning at the next unread character.
- Delete the first character read by the scanner and resume scanning at the character following it.

Both of these approaches are reasonable.

The first is easy to do. We just reset the scanner and begin scanning anew.

The second is a bit harder but also is a bit safer (less is immediately deleted). It can be implemented using scanner backup.

Usually, a lexical error is caused by the appearance of some illegal character, mostly at the beginning of a token.

(Why at the beginning?)

In these case, the two approaches are equivalent.

The effects of lexical error recovery might well create a later *syntax error*, handled by the parser.

Consider

```
...for$tnight...
```

The **`$`** terminates scanning of **`for`**. Since no valid token begins with **`$`**, it is deleted. Then **`tnight`** is scanned as an identifier. In effect we get

```
...for tnight...
```

which will cause a syntax error. Such "false errors" are unavoidable, though a syntactic error-repair may help.

Error Tokens

Certain lexical errors require special care. In particular, runaway strings and runaway comments ought to receive special error messages.

In Java strings may not cross line boundaries, so a runaway string is detected when an end of a line is read within the string body. Ordinary recovery rules are inappropriate for this error. In particular, deleting the first character (the double quote character) and restarting scanning is a *bad* decision.

It will almost certainly lead to a cascade of "false" errors as the string text is inappropriately scanned as ordinary input.

One way to handle runaway strings is to define an *error token*.

An error token is *not* a valid token; it is never returned to the parser. Rather, it is a *pattern* for an error condition that needs special handling. We can define an error token that represents a string terminated by an end of line rather than a double quote character.

For a valid string, in which internal double quotes and back slashes are escaped (and no other escaped characters are allowed), we can use

" (Not(" | Eol | \ ) | \" | \\ )* "

For a runaway string we use

" (Not(" | Eol | \ ) | \" | \\ )* Eol

(**Eol** is the end of line character.)

When a runaway string token is recognized, a special error message should be issued.

Further, the string may be "repaired" into a correct string by returning an ordinary string token with the closing Eol replaced by a double quote.

This repair may or may not be "correct." If the closing double quote is truly missing, the repair will be good; if it is present on a succeeding line, a cascade of inappropriate lexical and syntactic errors will follow.

Still, we have told the programmer exactly what is wrong, and that is our primary goal.

In languages like C, C++, Java and CSX, which allow multiline comments, improperly terminated (runaway) comments present a similar problem.

A runaway comment is not detected until the scanner finds a close comment symbol (possibly belonging to some other comment) or until the end of file is reached. Clearly a special, detailed error message is required.

Let's look at Pascal-style comments that begin with a { and end with a }. Comments that begin and end with a pair of characters, like /* and */ in Java, C and C++, are a bit trickier.

Correct Pascal comments are defined quite simply:

{ Not(})* }

To handle comments terminated by **Eof**, this error token can be used:

{ Not(})* Eof

We want to handle comments unexpectedly closed by a close comment belonging to another comment:

```
{... missing close comment
 ... { normal comment }...
```

We will issue a *warning* (this form of comment is lexically legal).

Any comment containing an open comment symbol in its body is most probably a missing } error.

We split our legal comment definition into two token definitions.

The definition that accepts an open comment in its body causes a warning message ("Possible unclosed comment") to be printed.

We now use:

{ Not({ | })* } and
{ (Not({ | })* { Not({ | })*)+ }

The first definition matches correct comments that do not contain an open comment in their body.

The second definition matches correct, but suspect, comments that contain at least one open comment in their body.

Single line comments, found in Java, CSX and C++, are terminated by Eol.

They can fall prey to a more subtle error—what if the last line has no Eol at its end?

The solution?

Another error token for single line comments:

// Not(Eol)*

This rule will only be used for comments that don't end with an Eol, since scanners always match the longest rule possible.

Regular Expressions and Finite Automata

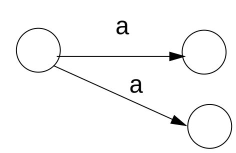
Regular expressions are fully equivalent to finite automata.

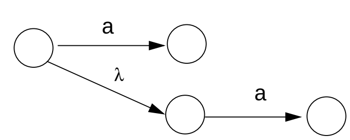
The main job of a scanner generator like JLex is to transform a regular expression definition into an equivalent finite automaton.

First it transforms a regular expression into a *nondeterministic finite automaton* (NFA).

Unlike an ordinary deterministic finite automaton, an NFA need not make a unique (deterministic) choice of a successor state to visit. For example, as shown below, an NFA is allowed to have a state that has two transitions (arrows) coming out of it, labeled by

the same symbol. An NFA may also have transitions labeled with λ.





Transitions are normally labeled with individual characters in Σ, and although λ is a string (the string with no characters in it), it is definitely *not* a character. In the above example, when the automaton is in the state at the left and the next input character is **a**, it may choose to use the transition labeled **a** *or* first follow the λ transition (you can always find λ wherever you look for it) and *then* follow an **a** transition. FAs that contain no λ transitions and that always have unique successor states for any symbol are *deterministic.*

Building Finite Automata From Regular Expressions

We make an FA from a regular expression in two steps:

- Transform the regular expression into an NFA.
- Transform the NFA into a deterministic FA.

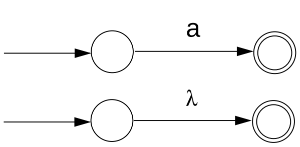
The first step is easy.

Regular expressions are all built out of the *atomic* regular expressions **a** (where **a** is a character in Σ) and λ by using the three operations

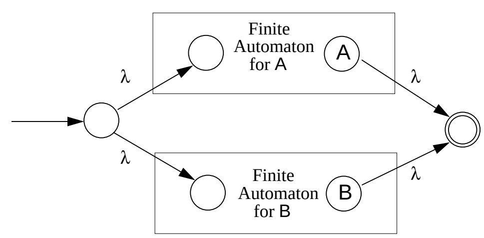
A B and **A | B** and $\mathbf{A}^*$.

Other operations (like $\mathbf{A}^+$) are just abbreviations for combinations of these.

NFAs for **a** and λ are trivial:



Suppose we have NFAs for **A** and **B** and want one for **A | B**. We construct the NFA shown below:

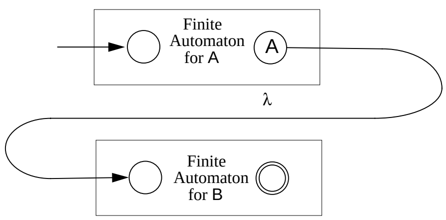


The states labeled **A** and **B** were the accepting states of the automata for **A** and **B**; we create a new accepting state for the combined automaton.

A path through the top automaton accepts strings in **A**, and a path through the bottom automation accepts strings in **B**, so the whole automaton matches **A | B**.

As shown below, the construction for **A B** is even easier. The accepting state of the combined automaton is the same state that was the accepting state of **B**. We must follow a path through **A**'s automaton, then through **B**'s automaton, so overall **A B** is matched.

We could also just merge the accepting state of **A** with the initial state of **B**. We chose not to only because the picture would be more difficult to draw.



Finally, let's look at the NFA for $\mathbf{A}^*$. The start state reaches an accepting state via λ, so λ is accepted. Alternatively, we can follow a path through the FA for **A** one or more times, so zero or more strings that belong to **A** are matched.