AXIOM	DESCRIPTION
$r\|s = s\|r$	$\|$ is commutative
$r\|(s\|t) = (r\|s)\|t$	$\|$ is associative
$(rs)t = r(st)$	concatenation is associative
$r(s\|t) = rs\|rt$ $(s\|t)r = sr\|tr$	concatenation distributes over $\|$
$\epsilon r = r$ $r\epsilon = r$	ϵ is the identity element for concatenation
$r^* = (r\|\epsilon)^*$	relation between * and ϵ
$r^{**} = r^*$	* is idempotent

Fig. 3.9. Algebraic properties of regular expressions.

Regular Definitions

For notational convenience, we may wish to give names to regular expressions and to define regular expressions using these names as if they were symbols. If Σ is an alphabet of basic symbols, then a *regular definition* is a sequence of definitions of the form

$$\begin{aligned} &d_1 \rightarrow r_1 \\ &d_2 \rightarrow r_2 \\ &\cdots \\ &d_n \rightarrow r_n \end{aligned}$$

where each d_i is a distinct name, and each r_i is a regular expression over the symbols in $\Sigma \cup \{d_1, d_2, \ldots, d_{i-1}\}$, i.e., the basic symbols and the previously defined names. By restricting each r_i to symbols of Σ and the previously defined names, we can construct a regular expression over Σ for any r_i by repeatedly replacing regular-expression names by the expressions they denote. If r_i used d_j for some $j \geq i$, then r_i might be recursively defined, and this substitution process would not terminate.

To distinguish names from symbols, we print the names in regular definitions in boldface.

Example 3.4. As we have stated, the set of Pascal identifiers is the set of strings of letters and digits beginning with a letter. Here is a regular definition for this set.

$$\begin{aligned} \textbf{letter} &\rightarrow \texttt{A} \mid \texttt{B} \mid \cdots \mid \texttt{Z} \mid \texttt{a} \mid \texttt{b} \mid \cdots \mid \texttt{z} \\ \textbf{digit} &\rightarrow \texttt{0} \mid \texttt{1} \mid \cdots \mid \texttt{9} \\ \textbf{id} &\rightarrow \textbf{letter} \; (\; \textbf{letter} \mid \textbf{digit} \;)^* \end{aligned}$$

□

Example 3.5. Unsigned numbers in Pascal are strings such as `5280`, `39.37`, 6.336E4, or 1.894E-4. The following regular definition provides a precise specification for this class of strings:

$$
\begin{aligned}
\textbf{digit} &\to 0 \mid 1 \mid \cdots \mid 9 \\
\textbf{digits} &\to \textbf{digit}\ \textbf{digit}^* \\
\textbf{optional\_fraction} &\to .\ \textbf{digits} \mid \epsilon \\
\textbf{optional\_exponent} &\to (\ \mathtt{E}\ (\ +\mid - \mid \epsilon\ )\ \textbf{digits}\ ) \mid \epsilon \\
\textbf{num} &\to \textbf{digits}\ \ \textbf{optional\_fraction}\ \ \textbf{optional\_exponent}
\end{aligned}
$$

This definition says that an **optional_fraction** is either a decimal point followed by one or more digits, or it is missing (the empty string). An **optional_exponent**, if it is not missing, is an E followed by an optional + or - sign, followed by one or more digits. Note that at least one digit must follow the period, so **num** does not match 1. but it does match 1.0. □

Notational Shorthands

Certain constructs occur so frequently in regular expressions that it is convenient to introduce notational shorthands for them.

1. *One or more instances.* The unary postfix operator $^+$ means "one or more instances of." If r is a regular expression that denotes the language $L(r)$, then $(r)^+$ is a regular expression that denotes the language $(L(r))^+$. Thus, the regular expression a^+ denotes the set of all strings of one or more a's. The operator $^+$ has the same precedence and associativity as the operator *. The two algebraic identities $r^* = r^+|\epsilon$ and $r^+ = rr^*$ relate the Kleene and positive closure operators.

2. *Zero or one instance.* The unary postfix operator ? means "zero or one instance of." The notation $r?$ is a shorthand for $r|\epsilon$. If r is a regular expression, then $(r)?$ is a regular expression that denotes the language $L(r) \cup \{\epsilon\}$. For example, using the $^+$ and ? operators, we can rewrite the regular definition for **num** in Example 3.5 as

$$
\begin{aligned}
\textbf{digit} &\to 0 \mid 1 \mid \cdots \mid 9 \\
\textbf{digits} &\to \textbf{digit}^+ \\
\textbf{optional\_fraction} &\to (\ .\ \textbf{digits}\ )? \\
\textbf{optional\_exponent} &\to (\ \mathtt{E}\ (\ + \mid -\ )?\ \textbf{digits}\ )? \\
\textbf{num} &\to \textbf{digits}\ \ \textbf{optional\_fraction}\ \ \textbf{optional\_exponent}
\end{aligned}
$$

3. *Character classes.* The notation [abc] where a, b, and c are alphabet symbols denotes the regular expression a | b | c. An abbreviated character class such as [a−z] denotes the regular expression a | b | · · · | z. Using character classes, we can describe identifiers as being strings generated by the regular expression

   [A−Za−z][A−Za−z0−9]*