It is now time for us to provide a rigorous definition for the expressions we have been playing with.

We have all the parts we need in order to define regular expressions recursively. The symbols that appear in regular expressions are: the letters of the alphabet Σ, the symbol for the null string Λ, parentheses, the star operator, and the plus sign.

DEFINITION

The set of regular expressions is defined by the following rules:

Rule 1 Every letter of Σ can be made into a regular expression by writing it in boldface; $\mathbf{\Lambda}$ is a regular expression.

Rule 2 If $\mathbf{r}_1$ and $\mathbf{r}_2$ are regular expressions, then so are

$$(\mathbf{r}_1) \quad \mathbf{r}_1\mathbf{r}_2 \quad \mathbf{r}_1 + \mathbf{r}_2 \quad \mathbf{r}_1{}^*$$

Rule 3 Nothing else is a regular expression. ■

We could have included the plus sign as a superscript $\mathbf{r}_1{}^+$ as part of the definition, but since we know that $\mathbf{r}_1{}^+ = \mathbf{r}_1\mathbf{r}_1{}^*$, this would add nothing valuable.

This is a language of language definers. It is analogous to a book that lists all the books in print. Every word in this book is a book-definer. The same confusion occurs in everyday speech. The string "French" is both a word (an adjective) and a language-defining name (a noun). However difficult Computer Theory may seem, English is much harder.

Because of Rule 1 we may have trouble in distinguishing when we write an *a* whether we mean *a* the letter in Σ, *a* the word in Σ^*, $\{a\}$ the one word language, or **a** the regular expression for that language. Context and typography will guide us.

As with the recursive definition of arithmetic expressions, we have included the use of parentheses as an option, not a requirement. Let us emphasize again the implicit parentheses in $\mathbf{r}_1{}^*$. If $\mathbf{r}_1 = \mathbf{aa} + \mathbf{b}$ then the expression $\mathbf{r}_1{}^*$ technically refers to the expression

$$\mathbf{r}_1{}^* = \mathbf{aa} + \mathbf{b}^*$$

which is the formal concatenation of the symbols for $\mathbf{r}_1$ with the symbol *, but what we generally mean when we write $\mathbf{r}_1{}^*$ is actually $(\mathbf{r}_1)^*$

$$(\mathbf{r}_1)^* = (\mathbf{aa} + \mathbf{b})^*$$

which is different. Both are regular expressions and both can be generated

from the rules. Care should always be taken to produce the expression we actually want, but this much care is too much to ask of mortals, and when we write $\mathbf{r}_1$* in the rest of the book we really mean $(\mathbf{r}_1)$*.

Another example of excessive care is the worry about the language that contains no words at all. The set of words in this language is the null set, not the null word. The null word is a word, so the language that contains no words cannot contain it. The language of no words cannot technically be defined by a regular expression since Rule 1 starts by putting something into the language. We finesse this point by saying that the language of no words is defined by the regular expression of no symbols.

To make the identification between the regular expressions and their associated languages more explicit, we need to define the operation of multiplication of sets of words.

DEFINITION

If S and T are sets of strings of letters (whether they are finite or infinite sets), we define the product set of strings of letters to be

ST = {all combinations of a string from S concatenated with a string from T } ■

EXAMPLE

If

$$S = \{a \quad aa \quad aaa\} \qquad T = \{bb \quad bbb\}$$

then

$$ST = \{abb \quad abbb \quad aabb \quad aabbb \quad aaabb \quad aaabbb\}$$

Note that these words are not in proper order. ■

EXAMPLE

If

$$S = \{a \quad bb \quad bab\} \qquad T = \{a \quad ab\}$$

then

$$ST = \{aa \quad aab \quad bba \quad bbab \quad baba \quad babab\}$$ ■

PROBLEMS

1. Let $\mathbf{r}_1$, $\mathbf{r}_2$, and $\mathbf{r}_3$ be three regular expressions. Show that the language associated with $(\mathbf{r}_1 + \mathbf{r}_2)\mathbf{r}_3$ is the same as the language associated with $\mathbf{r}_1\mathbf{r}_3 + \mathbf{r}_2\mathbf{r}_3$. Show that $\mathbf{r}_1(\mathbf{r}_2 + \mathbf{r}_3)$ is equivalent to $\mathbf{r}_1\mathbf{r}_2 + \mathbf{r}_1\mathbf{r}_3$. This will be the same as "proving a distributive law" for regular expressions.

   Construct a regular expression defining each of the following languages over the alphabet $\Sigma = \{a, b\}$.

2. All words in which a appears tripled, if at all. This means that every clump of a's contains 3 or 6 or 9 or 12 . . . a's.

3. All words that contain at least one of the strings s_1 s_2 s_3 or s_4.

4. All words that contain exactly three b's in total.

5. All words that contain exactly two b's or exactly three b's, not more.

6. (i) All strings that end in a double letter.
   (ii) All strings that have exactly one double letter in them.

7. All strings in which the letter b is *never* tripled. This means that no word contains the substring *bbb*.

8. All words in which a is tripled or b is tripled, but not both. This means each word contains the substring *aaa* or the substring *bbb* but not both.

9. (i) All strings that do not have the substring *ab*.
   (ii) All strings that do not have both the substrings *bba* and *abb*.

10. All strings in which the *total* number of a's is divisible by three, such as *aabaabbaba*.

11. (i) All strings in which any b's that occur are found in clumps of an odd number at a time, such as *abaabbbab*.
    (ii) All strings that have an even number of a's and an odd number of b's.
    (iii) All strings that have an odd number of a's and an odd number of b's.

12. Let us reconsider the regular expression

$$\mathbf{(a + b)^*a(a + b)^*b(a + b)^*}$$

(i) Show that this is equivalent to

$$\mathbf{(a + b)^*ab(a + b)^*}$$

in the sense that they define the same language.

(ii) Show that

$$\mathbf{(a + b)^*ab(a + b)^* + b^*a^* = (a + b)^*}$$

(iii) Show that

$$\mathbf{(a + b)^* ab[(a + b)^*ab(a + b)^* + b^*a^*] + b^*a^* = (a + b)^*}$$

(iv) Is (iii) the last variation of this theme or are there more beasts left in this cave?

13. We have defined the product of two sets of strings in general. If we apply this to the case where both factors are the same set, $S = T$, we obtain squares, S^2. Similarly we can define S^3, S^4,. . . . Show that

(i) $S^* = \Lambda + S + S^1 + S^2 + S^3 + S^4 + \ldots$

(ii) $S^+ = S + S^1 + S^2 + S^3 + S^4 + \ldots$

Show that the following pairs of regular expressions define the same language over the alphabet $\Sigma = \{a, b\}$.

14. (i) $\mathbf{(ab)^*a}$ and $\mathbf{a(ba)^*}$
(ii) $\mathbf{(a^* + b)^*}$ and $\mathbf{(a + b)^*}$
(iii) $\mathbf{(a^* + b^*)^*}$ and $\mathbf{(a + b)^*}$

15. (i) $\mathbf{\Lambda^*}$ and $\mathbf{\Lambda}$
(ii) $\mathbf{(a^*b)^*a^*}$ and $\mathbf{a^*(ba^*)^*}$
(iii) $\mathbf{(a^*bbb)^*a^*}$ and $\mathbf{a^*(bbba^*)^*}$

16. (i) $\mathbf{((a + bb)^*aa)^*}$ and $\mathbf{\Lambda + (a + bb)^*aa}$
(ii) $\mathbf{(aa)^*(\Lambda + a)}$ and $\mathbf{a^*}$
(iii) $\mathbf{a(aa)^*(\Lambda + a)b + b}$ and $\mathbf{a^*b}$

17. (i) **a(ba + a)*b** and **aa*b(aa*b)***
(ii) **Λ + a(a + b)* + (a + b)* aa(a + b)*** and **((b*a)*ab*)***

Describe (in English phrases) the languages associated with the following regular expressions.

18. (i) **(a + b)* a(Λ + bbbb)**
(ii) **(a(a + bb)*)***
(iii) **(a(aa)*b(bb)*)***
(iv) **(b(bb)*)*(a(aa)*b(bb)*)***
(v) **(b(bb)*)*(a(aa)*b(bb)*)*(a(aa)*)***
(vi) **((a + b)a)***

19. (D.N. Arden) Let R, S, and T be three languages and assume that Λ is not in S. Prove the following statements.
(i) From the premise that $R = SR + T$, we can conclude that $R = S^*T$.
(ii) From the premise that $R = S^*T$, we can conclude that $R = SR + T$.

20. Explain why we can take any pair of equivalent regular expressions and replace the letter a in both with any regular expression **R** and the letter b with any regular expression **S** and the resulting regular expressions will have the same language. For example, 15.(ii)

$$\mathbf{(a^*b)^*a^* = a^*(ba^*)^*}$$

becomes the identity

$$\mathbf{(R^*S)^*R^* = R^*(SR^*)^*}$$

which is true for all regular expressions **R** and **S**. In particular **R = a + bb**, **S = ba*** results in the complicated identity

$$\mathbf{((a + bb)^*(ba^*))^*(a + bb)^* = (a + bb)^* ((ba^*)(a + bb)^*)^*}$$

What is the deeper meaning of this transformation?
What identity would result from using

$$\mathbf{R = (ba^*)^* \quad S = (\Lambda + b)}$$