MCS-265: The Theory of Computation Handout #F1
San Skulrattanakulchai
Gustavus Adolphus College February 11, 2014

Formal Languages

Sipser Ch 0: p13–14; Ch 1: p44–45

Languages are our model for the data manipulated by computers.

Definitions. The term *symbol* (or *letter*) is undefined. An *alphabet* is a nonempty, finite set of *symbols*, e.g., letting $\Sigma = \{\mathtt{a}, \mathtt{b}\}$ we have Σ is the alphabet while a and b are symbols.

A *string* (or *word* or *sentence*) is a finite sequence (list) of symbols chosen from an alphabet, e.g., $\langle \mathtt{a}, \mathtt{b}, \mathtt{a} \rangle$, usually written as aba. (In the general formal language theory, infinite strings are allowed.) The length of a string w, denoted $|w|$, is the length of the sequence. The *empty string*, denoted ε, has length 0. If $|w| = n$, we usually write w as $w_1 w_2 \dots w_n$. For example, letting $w = \mathtt{aba}$, we have $w_1 = w_3 = \mathtt{a}$, and $w_2 = \mathtt{b}$. For any string w and symbol a, we write $|w|_a$ to denote the number of times the symbol a occurs in string w. For example, $|\mathtt{aba}|_\mathtt{a} = 2$, $|\mathtt{aba}|_\mathtt{b} = 1$, and $|\mathtt{aba}|_\mathtt{c} = 0$.

Let $x = x_1 x_2 \dots x_m$ and $y = y_1 y_2 \dots y_n$ be strings of length m and n respectively. The *concatenation* of x and y, written xy, is the string $x_1 x_2 \dots x_m y_1 y_2 \dots y_n$ of length $m + n$ that results from appending y to the end of x, e.g., concatenating back and bone gives backbone.

String concatenation is associative, and ε is the identity element. Therefore, strings form a monoid under concatenation.

If w is a string and n is a positive integer, we write w^n to mean the concatenation of n copies of w. The notation w^0 is defined to be ε.

A string y is a *substring* (or *subword*) of string w if there exist strings x, z such that $w = xyz$. A string x is a *prefix* of string w if there exists a string y such that $w = xy$. A string y is a *suffix* of string w if there exists a string x such that $w = xy$. So by definition an empty string is a substring, prefix, and suffix of any string; and any string is a substring, prefix, and suffix of itself. String x is a *subsequence* of string y if x can be obtained by striking out 0 or more symbols from y. For example `bat` is a subsequence of of `habitat`.

Let $w = w_1 w_2 \dots w_n$ be a string of length n. By the *reversal of* w, notated w^R, we mean the string $w_n w_{n-1} \dots w_1$. For example, $\texttt{star}^R = \texttt{rats}$. A string w is a *palindrome* if $w^R = w$. Examples of palindromes are `eve`, `madam`, `racecar`, `deified`, `rotator`.

Given an alphabet Σ, we define Σ^* to be the set of all strings over Σ, e.g., for $\Sigma = \{\texttt{a}, \texttt{b}\}$ we have $\Sigma^* = \{\varepsilon, \texttt{a}, \texttt{b}, \texttt{aa}, \texttt{ab}, \texttt{ba}, \texttt{bb}, \texttt{aaa}, \dots\}$. The listing of strings here is in *shortlex order* (or *string order* or *radix order*), i.e., ordered like in a dictionary, except that a shorter string always precede a longer one.

Exercises.

1. Define $<$ for dictionary order (or lexicographic order) precisely.
2. Define $<$ for shortlex order (or string order, or radix order) precisely.
3. What is the position of the string `ab`, when the strings of $\{\texttt{a}, \texttt{b}\}^*$ are arranged in dictionary order? in shortlex order?

A *language over the alphabet* Σ is any subset of Σ^*. Here are some example languages.

1. The set of all strings with an odd number of `a`'s.
2. The set of all palindromes.
3. The set of all strings of "balanced" left and right parentheses.
4. The set of all strings containing equal numbers of `a`'s, `b`'s, and `c`'s.
5. The set of all binary strings that represent prime numbers.
6. The set of all graphs with a Hamiltonian cycle, where the graph is encoded as a string.
7. $\emptyset$ and $\{\varepsilon\}$ are (different) languages

Remarks.

1. The subject matter of this course concerns languages and machines that recognize/compute them!

2. Finite languages are trivial.

3. A lone letter like `a` is ambiguous. It either represents a symbol or a string of length 1. Context decides which meaning is intended.

4. The concept of "string", "concatenation", "length of of string", "reversal of a string", etc., can be defined recursively as well.

Operations on Languages

Set Operations $\cup$, $\cap$, $\setminus$, $\triangle$, complement $\overline{A}$ of a language A, etc.

Concatenation The concatenation of two languages A and B is AB, i.e., the set of all strings xy where $x \in A$ and $y \in B$.

When precision is desired, concatenation is denoted by $\circ$, e.g., $x \circ y$, $A \circ B$.

Examples. Let $O = \{$all strings of odd length$\}$, $E = \{$all strings of even length$\}$, and $N = \{\mathtt{a}\}$. Then $ON = \{$ all strings of even length ending in $\mathtt{a}$ $\}$, $OE = O$, and $E^2 = E$.

For any language A, language A^0 denotes $\{\varepsilon\}$; languages A^i denotes AA^{i-1} whenever $i > 0$.

Kleene Closure $A^* = \bigcup_{i=0}^{\infty} A^i$. For example, $\emptyset^* = \{\varepsilon\}$. Note how this definition of * agrees nicely with our previous definition of * in Σ^* if we identify a string of length one with the symbol contained in it.

Positive Closure $A^+ = \bigcup_{i=1}^{\infty} A^i$.

Exercises.

1. Is it true that $A^+ = A^* \setminus \{\varepsilon\}$ for every language A?.

2. Which of the 7 example languages A above have $A = A^*$?

3. Characterize languages A that satisfy $A^* = A^+$?

4. Describe these languages $A\emptyset$, $A\{\varepsilon\}$, $A \cup \emptyset$, $A \cup \{\varepsilon\}$.

5. The $\cup$ and the $\circ$ operators for languages are comparable to the $+$ and the $*$ operators for numbers, respectively. What is the identity element for $\cup$? for $\circ$? What rules governing $+$ and $*$ are also obeyed by $\cup$ and $\circ$?