MCS-265: The Theory of Computation Handout #R1
San Skulrattanakulchai
Gustavus Adolphus College February 13, 2013

Formal Languages

Sipser Ch 0: p13–14; Ch 1: p44–45

Languages are our model for the data manipulated by computers.

Definitions. An *alphabet* is a nonempty finite set of *symbols* (or *letters*), e.g., $\{a, b\}$.

A *string* (or *word*) is a finite sequence of symbols from the alphabet, e.g., $\langle a, b, a\rangle$, usually written as *aba*. 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 $w_1 w_2 \dots w_n$.

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* 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.

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

Exercises.

1. Define $<$ for dictionary (or lexicographic) order precisely.
2. Define $<$ for shortlex (or string) order precisely.

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, where a *palindrome* is a string w that equals its reverse w^R, e.g., *madam* or *eve*.
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 circuit, 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.

Operations on Languages

Set Operations $\cup$, $\cap$, 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 $O = \{\text{all strings of odd length}\}$, $E = \{\text{all strings of even length}\}$, and $N = \{a\}$. Then $ON = \{\text{all strings of even length ending in } a\}$, $OE = O$, 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. Which of the 7 example languages have $A = A^*$?
2. What languages A satisfy $A^* = A^+$?
3. Describe these languages $A\emptyset$, $A\{\varepsilon\}$, $A \cup \emptyset$, $A \cup \{\varepsilon\}$.
4. 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 $\circ$?