CSCI/MATH 2113 Discrete Structures

6.1 Language: The Set Theory of Strings

Alyssa Motas

March 3, 2021

Contents

1 Definition of an alphabet

An *alphabet* is a finite nonempty set. We write Σ for an alphabet and we sometimes call the elements of Σ *letters*. For example, we may have $\Sigma = \{0, 1\}$ or $\Sigma = \{a, b, c, d, e\}$.

2 Powers of an alphabet

If Σ is an alphabet and $n \in \mathbb{Z}^+$, we define the *powers* of Σ recursively as follows:

1. $\Sigma^1 = \Sigma$; and
2. $\Sigma^{n+1} = \{xy \mid x \in \Sigma, y \in \Sigma^n\}$, where xy denotes the juxtaposition of x and y.

2.1 Example

Let Σ be an alphabet. With $\Sigma = \{0, 1\}$, we find that

$$\Sigma^2 = \{00, 01, 10, 11\} \text{ and } |\Sigma^2| = |\Sigma|^2 = 2^2 \text{ two-symbol strings.}$$

In general, we have $|\Sigma^n| = |\Sigma|^n$.

3 Empty string

For an alphabet Σ, we define $\Sigma^0 = \{\lambda\}$, where λ denotes the *empty string*. That is, the string consisting of *no* symbols taken from Σ. Note that even though $\lambda \notin \Sigma$, we do have $\varnothing \subseteq \Sigma$. Also, $\{\lambda\} \neq \varnothing$ because $|\{\lambda\}| = 1 \neq 0 = |\varnothing|$.

4 Union of alphabets

If Σ is an alphabet, then

(a) $\Sigma^+ = \bigcup_{n=1}^{\infty} \Sigma^n = \bigcup_{n \in \mathbb{Z}^+} \Sigma^n$;

(b) $\Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n$.

The difference between (a) and (b) is that $\lambda \in \Sigma^n$ only when $n = 0$. Also, $\Sigma^* = \Sigma^+ \cup \Sigma^0$.

We shall also refer to the elements of Σ^+ or Σ^* as *words* and sometimes as *sentences*. Finally, we note that even though the sets Σ^+ and Σ^* are *infinite*, the elements of these sets are *finite* strings of symbols.

4.1 Example

- For $\Sigma = \{0, 1\}$ the set Σ^* consists of all finite strings (binary words) of 0's and 1's together with the empty string.
- If $\Sigma = \{+, \times, 0, 1, \ldots, 9, (,), \quad \}$ we have

$$((14 + 12) \times 3) \times 1009 \in \Sigma^* \text{ or }) + (x)1 + (\times 3) \in \Sigma^*.$$

5 Equality of sets

If $w_1, w_2 \in \Sigma^+$, then we may write

$$w_1 = x_1 x_2 \ldots x_m \text{ and } w_2 = y_1 y_2 \ldots y_n$$

for $m, n \in \mathbb{Z}^+$ and $x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n \in \Sigma$. We say that the strings w_1 and w_2 are *equal*, and we write $w_1 = w_2$, if $m = n$, and $x_i = y_i$ for all $1 \leq i \leq m$.

6 Length

Let $w = x_1 x_2 \ldots x_n \in \Sigma^+$, where $x_i \in \Sigma$ for each $1 \leq i \leq n$. We define the *length* of w, which is denoted by $||w||$, as the value n. For the case of λ, we have $||\lambda|| = 0$.

7 Concatenation

Let $x, y \in \Sigma^+$ with $x = x_1 x_2 \ldots x_m$ and $y = y_1 y_2 \ldots y_n$, so that each x_i, for $1 \leq i \leq m$, and each y_j, for $1 \leq j \leq n$, is in Σ. The *concatenation* of x and y, which we write as xy, is the string

$$x_1 x_2 \ldots x_m y_1 y_2 \ldots y_n.$$

The concatenation of x and λ is $x\lambda = x_1 x_2 \ldots x_m \lambda = x_1 x_2 \ldots x_m = x$, and the concatenation of λ and x is $\lambda x = \lambda x_1 x_2 \ldots x_m = x_1 x_2 \ldots x_m = x$. Finally, the concatenation of λ and λ is $\lambda\lambda = \lambda$.

Here, we have defined a closed binary operation on Σ^* (and Σ^+). This operation is associative but not commutative unless $|\Sigma| = 1$. The λ is also the identity for the operation of concatenation. We also have

$$||xy|| = ||x|| + ||y||, \qquad \text{for all } x, y \in \Sigma^*.$$

8 Powers of a string

For each $x \in \Sigma^*$, we define the *powers* of x by $x^0 = \lambda, x^1 = x, x^2 = xx, x^3 = xx^2, \dots, x^{n+1} = xx^n, \dots,$ where $n \in \mathbb{N}$.

9 Proper prefix and suffix

If $x, y \in \Sigma^*$ and $w = xy$, then the string x is called a *prefix* of w, and if $y \neq \lambda$, then x is said to be a *proper prefix.* Similarly, the string y is called a *suffix* of w; it is a *proper suffix* when $x \neq \lambda$.

In general, for an alphabet Σ, if $n \in \mathbb{Z}^+$ and $x_i \in \Sigma$, for all $1 \leq i \leq n$, then each of $\lambda, x_1, x_1x_2, x_1x_2x_3, \dots,$ and $x_1x_2x_3 \dots x_n$ is a prefix of the string $x = x_1x_2x_3 \dots x_n$. And $\lambda, x_n, x_{n-1}x_n, x_{n-2}x_{n-1}x_n, \dots,$ and $x_1x_2x_3 \dots x_n$ are all suffixes of x. So, x has $n+1$ prefixes, n of which are proper, and the situation is the same for suffixes.

10 Substring

If $x, y, z \in \Sigma^*$ and $w = xyz$, then y is called a *substring* of w. When at least one of x and z is different from λ (so that y is different from w), we call y a *proper substring or subword.*

11 Language

For a given alphabet Σ, any subset of Σ^* is called a *language* over Σ. This includes the subset $\varnothing$, which we call the *empty language.*

11.1 Example

With $\Sigma = \{0, 1\}$, the sets

$$A = \{0, 01, 001\}$$

and

$$B = \{0, 01, 001, 0001, \dots\}$$

are examples of languages over Σ.

11.2 Concatenation

For an alphabet Σ and languages $A, B \subseteq \Sigma^*$, the *concatenation* of A and B, denoted AB is $\{ab \mid a \in A, b \in B\}$.

We might compare concaenation with the cross product. We shall see that just as $A \times B \neq B \times A$ in general, we also have $AB \neq BA$ in general. For A, B finite we did have $|A \times B| = |B \times A|$, but here $|AB| \neq |BA|$ is possible for finite languages.

Example. Let $\Sigma = \{x, y, z\}$, and let A, B be the finite languages $A = \{x, xy, z\}$, $B = \{\lambda, y\}$. Then $AB = \{x, xy, z, xyy, zy\}$ and $BA = \{x, xy, z, yx, yxy, yz\}$, so

1. $|AB| = 5 \neq 6 = |BA|$; and
2. $|AB| = 5 \neq 6 = 3 \cdot 2 = |A||B|$.

This suggests that for finite languages A and B, $|AB| \leq |A||B|$.

11.3 Properties

For an alphabet Σ, let $A, B, C \subseteq \Sigma^*$. then

(a) $A\{\lambda\} = \{\lambda\}A = A$

(b) $(AB)C = A(BC)$

(c) $A(B \cup C) = AB \cup AC$

(d) $(B \cup C)A = BA \cup CA$

(e) $A(B \cap C) \subseteq AB \cap AC$

(f) $(B \cap C)A \subseteq BA \cap CA$

Proof. Let us prove (d) and (f).

(d) Starting with $x \in \Sigma^*$ we find that

$$\begin{aligned}
x \in (B \cup C)A &\Rightarrow x = yz && \text{for } y \in B \cup C, z \in A \\
x &\Rightarrow yz && \text{for } y \in B \text{ or } y \in C,\ z \in A \\
&\Rightarrow x \in BA \cup x \in CA \\
&\Rightarrow (B \cup C)A \subseteq BA \cup CA.
\end{aligned}$$

Conversely, it follows that

$$\begin{aligned} x \in BA \cup CA &\Rightarrow x \in BA \text{ or } x \in CA \\ &\Rightarrow (x = ba_1, b \in B, a_1 \in A) \text{ or } (x = ca_2, c \in C, a_2 \in A). \end{aligned}$$

Assume that $x = ba_1$ for $b \in B, a_1 \in A$. Since $B \subseteq B \cup C$, we have $x = ba_1$, where $b \in B \cup C, a_1 \in A$. Then $x \in (B \cup C)A$, so $BA \cup CA \subseteq (B \cup C)A$. The argument is similar when $x = ca_2$. With both inclusions established, it follows that $(B \cup C)A = BA \cup CA$.

(f) For $x \in \Sigma^*$, we see that $x \in (B \cap C)A \Rightarrow x = yz$ where $y \in B \cap C$ and $z \in A \Rightarrow x = yz$ for $y \in B, z \in A$ and $x = yz$ for $y \in C$. This implies that $x \in BA$ and $x \in CA$, then $x \in BA \cap CA$. Thus, $(B \cap C)A \subseteq BA \cap CA$.

□

11.4 Positive and Kleene Closure

For a given language $A \subseteq \Sigma^*$ we can construct other languages as follows:

1. $A^0 = \{\lambda\}$, $A^1 = A$, and for all $n \in \mathbb{Z}^+$, $A^{n+1} = \{ab \mid a \in A, b \in A^n\}$.
2. $A^+ = \bigcup_{n \in \mathbb{Z}^+} A^n$, the *positive closure* of A.
3. $A^* = A^+ \cup \{\lambda\}$. The language A^* is called the *Kleene* closure of A, in honor of the American logician Stephen Cole Kleene (1909-1994).

Examples. For $\Sigma = \{x, y, z\}$, $A = \{x\}$ then

$$\begin{aligned} A^0 &= \{\lambda\} \qquad & A^+ &= \{x^n \mid n > 0\} \\ A^n &= \{x^n\} \qquad & A^* &= \{x^n \mid n \in \mathbb{N}\} \end{aligned}$$

For $\Sigma = \{x, y\}$ we have

(a) $A = \{xx, xy, yx, yy\}$

$$\begin{aligned} &A^* \subseteq \Sigma^* \\ &A^* = \{w \in \Sigma^* \mid ||w|| \text{ is even}\} \end{aligned}$$

(b) $B = \{x, y\}$

$$B(A^*) = \{w \in \Sigma^* \mid ||w|| \text{ is odd}\}$$

$$\text{Here } BA^* = A^*B \text{ and}$$

$$\Sigma^* = A^* \cup A^*B$$

(c)

$$\{x\}\{x,y\}^* = \{w \in \Sigma^* \mid \text{ w has } x \text{ as a prefix}\}$$
$$\{x\}\{x,y\}^+ = \{w \in \Sigma^* \mid \text{ w has } x \text{ as a proper prefix}\}$$

There are languages A and B for which $AB \neq BA$. Conversely, there are languages A and B for which $AB = BA$.

Example. For $\Sigma = \{x, y\}$ we have

$$\begin{aligned} A &= \{\lambda, x, x^3, x^4, \dots\} = \{x^n \mid n \in \mathbb{N}\} \setminus \{x^2\} \\ B &= \{x^n \mid n \in \mathbb{N}\} \\ &\Rightarrow A^2 = B^2 (= B) \text{ but } A \neq B. \end{aligned}$$

11.5 Theorem

Let Σ be an alphabet, with languages $A, B \subseteq \Sigma^*$.

(a) $A \subseteq \Rightarrow A^n \subseteq B^n, \forall n \in \mathbb{N}$

(b) $A \subseteq AB^*$

(c) $A \subseteq B^* A$

(d) $A \subseteq B \Rightarrow A^+ \subseteq B^+$ and $A^* \subseteq B^*$

(e) $A^+ = AA^* = A^* A$

(f) $A^* A^* = A^* = (A^*)^* = (A^*)^+ = (A^+)^*$

(g) $(A \cup B)^* = (A^* \cup B^*)^* = (A^* B^*)^*$