Strings: Basic notation

PREREQUISITES

- sets (basic notation)

Strings play a very prominent role in computational linguistics. A string is a sequence of symbols, like *nfm*, *wendigo*, or *105§/*. In contrast to sets, strings are ordered and can contain duplicates.

EXAMPLE 1.
The sets $\{m, a, d\}$, $\{d, a, m\}$, and $\{a, d, a, m\}$ are equivalent, but for strings $mad \neq dam \neq adam$.

EXERCISE 1.
Fill in $=$ or $\neq$ as appropriate for each pair of strings below.

- $abba$ _ $ABBA$
- 10 _ $5 + 5$
- $\{m, a, d\}$ _ $\{d, a, m\}$

Caution: { and } can be symbols just like m, a, or d.

1 Alphabet

When talking about strings, one usually fixes a finite set of symbols over which the strings are built. This is called an **alphabet**. Alphabets are often given labels like Σ or Ω. A string over alphabet Σ is also called a Σ-string.

EXAMPLE 2.
The set of Latin characters (A-Z, a-z) is an alphabet that's familiar to all of you. Strings over it include:

- string
- alphabet
- aaaaaaa
- c

EXAMPLE 3.
The set of Arabic digits is an alphabet with symbols 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Every natural number (0, 1, 2, ...), when represented in decimal as usual, is

a string over this alphabet. But not every string over this alphabet is a number of the decimal system. For instance, 000134095 is not a valid number, although 134095 is.

Example 4.
The set $\mathbb{N}$ of all natural numbers is not a valid alphabet because it isn't finite.

Exercise 2.
For each one of the following, say whether it is a valid alphabet. Justify your answer.

- $\{a\}$
- $\{0, 1\}$
- the set of all English words that are spelled with at most 5 characters
- all natural numbers less than 1000
- the nucleobases of DNA: adenine, cytosine, guanine, thymine

2 String length

The length of a Σ-string s is indicated by $|s|$. For instance, $|\text{ant}| = 3$, $|0770001| = 7$, and $|\text{a}| = 1$. The set of all strings over Σ whose length is exactly n is denoted by Σ^n.

Example 5.
Let $\Sigma := \{a, b\}$. Then Σ^3 contains all of the following strings, and only those:

- aaa
- aab
- aba
- abb
- baa
- bab
- bba
- bbb

The size of Σ^n is always fixed. If Σ has m members, then Σ^n contains m^n strings.

Example 6.
In the previous example, Σ contains two symbols, so Σ^n should consist of $2^3 = 8$ distinct strings. That's exactly what we found.

Exercise 3.
Which one of the following are members of $\{a,b\}^4$, i.e. Σ^4 where Σ contains a, b, and nothing else?

- $aaab$
- aba
- $aaaaa$
- b
- $abca$

Exercise 4.
List all members of $\{k,o,z\}^2$.

Very often expressions like a^n are used as a shorthand for $\{a\}^n$.

Example 7.
The expression ba^5c^3d is a shorthand for $baaaaaccd$.

Exercise 5.
Write each one of the following in a more compact fashion using exponents.

- ABBA
- loool
- aardvark

3 Infinite string sets over Σ

Since alphabets must be finite, Σ^n is necessarily finite for any alphabet Σ and $n \geq 0$. But the set of all strings over Σ is infinite.

Example 8.
Let $\Sigma := \{a\}$. Then a is a string over Σ, and so are aa, aaa, $aaaa$, and so on. This enumeration continues indefinitely, so there must be infinitely many distinct strings over Σ.

Two infinite string sets are commonly defined over Σ. They are Σ^* and Σ^+, respectively. The former contains all strings over Σ, whereas the latter contains all strings whose length is at least 1. The only difference between the two is that Σ^* also contains the **empty string** ε. The empty string is the string counterpart of the number 0: it represents nothing. In fact, ε is the only string whose length is 0.

Example 9.
Let $\Sigma = \{a, b\}$. Then Σ^* contains

- ε,
- a,
- b,
- aa,
- ab,
- ba,
- bb,
- aaa,
- aab,
- aba,
- abb,
- and so on.

All these strings are also members of Σ^+, except ε.

Σ^* is also called the **Kleene closure**, named after Stephen C. Kleene.

Here's a little bit of background to make it easier for you to remember the difference between Σ^* and Σ^+. As you might know from search engines, the Kleene star $*$ is sometimes used as a wildcard that matches everything. So Σ^* can be translated as "every string built over Σ". On the other hand Σ^+ only contains those strings whose length is at least 1, or in other words, whose length is positive. And $+$ is a common abbreviation for positive (just think of batteries).

Exercise 6.
Enumerate the five shortest members of $\{a\}^*$.

4 Concatenation

Given two Σ-strings u and v, their **concatenation** $u \cdot v$ is the result of "glueing" the left end of v to the right end of u.

Example 10.
Here are a few examples of concatenation:

- $math \cdot ematics = mathematics$,
- $2000 \cdot 18 = 200018$,
- Thomas · Graf = ThomasGraf.

Just like addition, concatenation is **associative**. This means that if we carry out multiple concatenations, it does not matter which concatenation step we resolve first: $u \cdot (v \cdot w) = (u \cdot v) \cdot w = u \cdot v \cdot w$.

Example 11.
It does not matter in which order we combine *is* with *concatenation* and *associative* below:

- $(\textit{concatenation} \cdot \textit{is}) \cdot \textit{associative} = \textit{concatenation is associvative}$
- $\textit{concatenation} \cdot (\textit{is} \cdot \textit{associative}) = \textit{concatenation is associvative}$

Even though concatenation is associative, it is not **commutative**. That is to say, $u \cdot v$ and $v \cdot u$ are not necessarily the same.

Example 12.
Let $u :=$ house and $v :=$ boat. Then $u \cdot v$ is *houseboat*, whereas $v \cdot u$ is *boathouse*. Those are not the same strings (and they also happen to mean completely different things).

Note the special behavior of the empty string: $u \cdot \varepsilon = \varepsilon \cdot u = u$. This makes sense because adding nothing to u does not change u, just like adding 0 to a number does not change that number.

Sometimes concatenation is not explicitly indicated, so that instead of $u \cdot v$ one may simply write uv.

Exercise 7.
Give an example of distinct u and v such that $u \cdot v = v \cdot u$ and neither u nor v is the empty string.

Exercise 8.
Is the following true or false? If $u \neq v$, then $u \cdot v \neq v \cdot u$?

5 Recap

- A string is a sequence of symbols drawn from some alphabet.
- A Σ-string is a string over alphabet Σ.
- The length of string s is denoted by $|s|$.
- The empty string ε is the unique string of length 0.
- Σ^n is the set of all Σ-strings s such that $|s| = n$.
- a^n is a shorthand for $\{a\}^n$.
- The Kleene closure Σ^* is the set of all Σ-strings (including ε).
- The positive closure Σ^+ contains all Σ-strings except ε.

- Concatenation of strings u and v is denoted by $u \cdot v$.