

Strings: Basic notation

PREREQUISITES

- sets (basic notation)

Strings play a very prominent role in linguistics and language technology. 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**. It is common but not necessary to require alphabets to contain at least one symbol. 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 $(0, 1, 2, \dots)$, 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 $(0, 1, 2, \text{ and so on})$ 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
- the set of all natural numbers less than 1000
- the set of 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 $|a| = 1$. 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^3 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 $baaaaaaccccd$.

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 set Σ^* contains all strings over Σ , whereas Σ^+ 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 (e.g. with 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:

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

Just like addition, concatenation is **associative**. This means that if we carry out

multiple concatenations, it does not matter in what order we resolve the concatenation steps: $u \cdot (v \cdot w) = (u \cdot v) \cdot w = u \cdot v \cdot w$.

EXAMPLE 11.

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

- $(\text{concatenation} \cdot \text{is}) \cdot \text{associative} = \text{concatenationis} \cdot \text{associative} = \text{concatenationisassociative}$
- $\text{concatenation} \cdot (\text{is} \cdot \text{associative}) = \text{concatenation} \cdot \text{isassociative} = \text{concatenationisassociative}$

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. They might be, but it's not guaranteed.

EXAMPLE 12.

Let $u := \text{house}$ and $v := \text{boat}$. 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 is fairly intuitive because adding a string of length 0 to u should not change the length of u , which means that u does not change at all — 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 strings u and v such that $uv = vu$ and neither u nor v is the empty string.

EXERCISE 8.

Is the following true or false? If $u \neq v$, then $uv \neq vu$?

EXERCISE 9.

Is the following true or false? If $uv \neq vu$, then $u \neq v$?

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 $u \cdot v$ or simply uv .