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, …), 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, 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 $|\text{a}| = 1$. 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 $baaaaacccd$.

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 · 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:

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

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 $u :=$ 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 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 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.

Strings: Basic notation

- 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.

Solutions

Solution to Exercise 1.

1. $abba \neq ABBA$
2. $10 \neq 5 + 5$
3. $\{m, a, d\} \neq \{d, a, m\}$

Explanation.

1. The symbols a and A are distinct (and so are b and B). Hence $abba$ and $ABBA$ already contain different symbols in their very first position, and consequently the strings are distinct.
2. While it is true that the sum of 5 and 5 is 10, the exercise asks you to compare the string $5+5$ to the string 10. These are very different strings. For example, only the former contains the symbol $+$, and only the latter contains 1 and 0.
3. This one is tricky. If we were talking about sets, then it would be the case that $\{m, a, d\}$ is equivalent to $\{d, a, m\}$ because sets are unordered. But here you are asked to consider strings, not sets. These strings just so happen to start with { and end with }, but they are still strings. And since the second symbol of the left string is m, whereas the right string has d in its second position, these two strings are distinct.

Solution to Exercise 2.
An alphabet must be a finite set of symbols. Therefore:

- The set $\{a\}$ is a valid alphabet because it contains exactly one symbol and thus is finite.
- The set $\{0, 1\}$ is a valid alphabet because it contains exactly two symbols and thus is finite.
- The set of all English words that are spelled with at most 5 characters is also a finite set: The English alphabet has 26 letters, which means that there are at most $26^5 = 11,881,376$ words that fit these requirements. While that is a large number, it is still finite. The fact that each symbol is itself a string of characters does not matter, that is an artefact of our writing system. If instead we had a unique symbol to refer to each one of these words, e.g. pictograms, the issue wouldn't even arise.
- The set of all natural numbers less than 1000 is also a possible alphabet because this set is finite. Again you might be concerned that we write natural numbers as strings of digits and it may seem weird to you to have symbols that are internally complex. But just like with the English words, this is just an artefact of how we write numbers. In hexademical notation, for instance, the natural number 11 would be a single symbol b.
- The nucleobases of DNA once again form a finite set and hence are a valid

alphabet. Again these are symbols with internal structure, in this case their molecular structure. When we treat the nucleobases as the symbols of our alphabet, we are making a decision to ignore their internal structure and treat them as the smallest unit of analysis (just like, say, a phonologist may treat sounds as the smallest unit of analysis and ignore the sound waves that actually give rise to the sounds).

Solution to Exercise 3.
Among the listed strings, *aaab* is the only member of $\{a, b\}^4$.

Explanation.
The $\{a, b\}^4$ contains all strings that meet both of the following criteria:

1. The only symbols that may occur in the string are a and b.
2. The string must contain exactly four symbols.

The only strings above with exactly four symbols are *aaab* and *abca*, and only the former contains no symbols besides a and b. Note that *aaaa*, which isn't listed above, is also a member $\{a, b\}^4$ — the requirement that the strings may contain no symbols other than a and b does not entail that the strings must contain both a and b.

Solution to Exercise 4.
The members are *kk*, *ko*, *kz*, *ok*, *oo*, *oz*, *zk*, *zo*, *zz*.

Explanation.
The set $\{k, o, z\}^2$ contains all strings of length 2 that can be built using only the symbols k, o, and z. This gives us three options starting with k (*kk*, *ko*, *kz*), three options starting with o (*ok*, *oo*, *oz*), and three options starting with z (*zk*, *zo*, *zz*).

Solution to Exercise 5.
1. $\mathrm{AB^2A}$
2. $\mathrm{lo^3l}$
3. $\mathrm{a^2rdvark}$

Explanation.
The general strategy is to find parts of the string where the same symbol x is iterated n times and to then replace that part with x^n. For example, in *ABBA* we have *A* followed by two *B*s and then another *A*, i.e. AB^2A.

Solution to Exercise 6.

1. ε
2. a
3. aa
4. aaa
5. $aaaa$

Explanation.

The expression $\{a\}^*$ is short for "the set of all strings of length 0 or more that can be built over the alphabet $\{a\}$". With only one symbol to choose from, there can never be two distinct strings of the same length. Hence we just start with the shortest possible string, which is the empty string ε, and then keep adding as until we have enumerated five strings.

Solution to Exercise 7.

There are infinitely many solutions. For example, if $u := aba$ and $v := abaaba$, then $u \cdot v = aba \cdot abaaba = abaabaaba = abaaba \cdot aba = v \cdot u$. The simplest solution is actually $u := a$ and $v := aa$, where we have $u \cdot v = a \cdot aa = aaa = aa \cdot a = v \cdot u$. Quite generally, we can pick any non-empty string u, set $v = uu$, and it will always be the case that $u \neq v$ yet $u \cdot v = v \cdot u = uuu$.

Note that $u := a$ and $v := ab$ is not a valid solution. In this case, $u \cdot v = aab$, whereas $v \cdot u = aba$, which is distinct from aab. Similarly, $u := a$ and $v := a$ is not a valid solution because the exercise requires u and v to be distinct strings.

Solution to Exercise 8.

False.

Explanation.

This follows immediately from the previous exercise, where you had to pick distinct u and v such that $u \cdot v \neq v \cdot u$.

Solution to Exercise 9.

True.

Explanation.

Assume towards a contradiction that $uv \neq vu$ yet $u = v$. If $u = v$, then $uv = uu = vv = vu$, which contradicts our initial assumption that $uv \neq vu$.