Parts of strings

Prerequisites

- strings (basic notation)

It is often important to refer to specific substructures of a string. The most important notion is that of **substring**, with the special cases of **prefix** and **suffix**. But for some applications, **subsequences** are also relevant.

1 Substrings

A **substring** is a continuous part of a string.

Example 1.
The string *abcd* has 11 substrings:

- ε
- a
- b
- c
- d
- ab
- bc
- cd
- abc
- bcd
- $abcd$

Some authors like to write $u \sqsubseteq v$ to indicate that u is a substring of v.
Note that

1. the empty string is a substring of every string, and
2. every string is a substring of itself.

A substring u of v is a **proper** substring iff $u \neq v$.

Example 2.
All the strings listed above are proper substrings of *abcd*, except *abcd* itself.

If u is substring that spans from the very beginning of v, we call it a **prefix**. And if u is a substring that spans to the end of v, we call it a **suffix**. Make sure not to confuse these with the linguistic notions of prefix and suffix, which work very differently.

Example 3.
Among the strings listed above, all of the following are prefixes of $abcd$:

- ε (if you find this confusing, check the formal definition below)
- a
- ab
- abc
- $abcd$

And the all of the following are suffixes of $abcd$:

- ε (if you find this confusing, check the formal definition below)
- d
- cd
- bcd
- $abcd$

Substrings, prefixes, and suffixes are formally defined via concatenation.

Definition 1. Given Σ-strings u and v, u is a **substring** of v ($u \sqsubseteq v$) iff there are $x, y \in \Sigma^*$ such that $v = x \cdot u \cdot y$. We furthermore call u

- a **proper substring** iff $u \neq v$,
- a **prefix** iff $y = \varepsilon$,
- a **suffix** iff $x = \varepsilon$.

Exercise 1.
For each one of the string pairs below, indicate whether the first string is a substring of the second string, a proper substring, or neither:

- a & $aaaa$
- a & b
- ε & b
- ε & ε
- aa & $abbbca$
- bc & $abbbca$
- cb & $abbbca$

Exercise 2.
For every string u, there are two substrings that are both prefixes and suffixes of u. What are they? For which string are these two substrings not distinct?

2 Subsequence

Whereas substrings must be continuous, **subsequences** are allowed to also be discontinuous. However, a subsequence need not be discontinuous.

Example 4.
The string $abcd$ has 16 subsequences:

- ε
- a
- b
- c
- d
- ab
- ac
- ad
- bc
- bd
- cd
- abc
- abd
- acd
- bcd
- $abcd$

Note that ca is not a subsequence of $abcd$, but it is a subsequence of $abcda$.

Exercise 3.
List all distinct subsequences of the string $aaaa$ (without duplicates).

Just like substrings, a subsequence u of v is **proper** iff $u \neq v$.

The formal definition of subsequences is quite a bit more verbose than that of substrings. This is because the option of discontinuity requires the use of additional string variables that can be interleaved with the subsequence in order to obtain the original string.

Definition 2. Let v be a Σ-string and $u := u_1 u_2 \cdots u_n$ a member of Σ^n. Then u is a **subsequence** of v iff there are strings $x_0, x_1, \ldots, x_{n+1} \in \Sigma^*$ such that

$$v = x_0 \cdot u_1 \cdot x_1 \cdot u_2 \cdot x_2 \ldots \cdot u_n \cdot x_{n+1}$$

A subsequence u of v is **proper** iff $u \neq v$.

Exercise 4.
For each one of the string pairs below, indicate whether the first string is a subsequence of the second string, a proper subsequence, or neither:

- a & $aaaa$
- a & b
- ε & b
- ε & ε
- aa & $abbbca$
- bc & $abbbca$
- cb & $abbbca$

Exercise 5.
Say whether the following is True or False: Every substring of some string s is also a subsequence of s, but not the other way round. Justify your answer.

3 Recap

- A **substring** is a continuous part of a string. Initial substrings are called **prefixes**, and final ones are called **suffixes**.
- A **subsequence** is a discontinuous part of a string.
- The empty string is a substring, a prefix, a suffix, and a subsequence of every string.
- Every string s is both a substring and a subsequence of itself. The substrings and subsequences of s that are distinct from s are **proper**.

Solutions

Solution to Exercise 1.

- a is a proper substring of $aaaa$
- a is not a substring of b
- ε is a proper substring b
- ε is a substring of ε
- aa is not a substring of $abbbca$
- bc is a proper substring of $abbbca$
- cb is not a substring of $abbbca$

Solution to Exercise 2.

The two substrings are ε and u itself. The empty string is the only case where the two are not distinct as $u = \varepsilon$.

Explanation.

No matter what u looks like, it is the case that $u = \varepsilon \cdot u \cdot \varepsilon$. But by the definition of prefixes, this means that u is both a prefix ($y = \varepsilon$) and a suffix ($x = \varepsilon$) of u. But note that we can also decompose u into $x \cdot \varepsilon \cdot u$ such that $x = \varepsilon$, which means that ε is a prefix of u. Similarly, $u = u \cdot \varepsilon y$ with $y = \varepsilon$, and thus ε is also a suffix of u.

Solution to Exercise 3.

- ε
- a
- aa
- aaa
- aaaa

Solution to Exercise 4.

- a is a proper subsequence of $aaaa$
- a is not a subsequence of b
- ε is a proper subsequence of b
- ε is a subsequence of ε
- aa is a proper subsequence of $abbbca$
- bc is a proper subsequence of $abbbca$
- cb is not a subsequence of $abbbca$

Solution to Exercise 5.

This is correct. A substring is a subsequence where all symbols happen to be adjacent. For example, ab is both a substring and a subsequence of abc.

In the other direction, not every subsequence is a substring. For example, ac is

a subsequence of *abc* but not a substring.