What is a Formal Language?

Arno Pauly

January 29, 2021

Words over an alphabet

Definition (Words of fixed length)

Let Σ be a (finite) set (which we will call the *alphabet*) and $n \in \mathbb{N}$ a natural number. With Σ^n we denote the set of functions from $\{0, 1, \ldots, n-1\}$ to Σ, also called *words* of length n.

Example

Let $\Sigma = \{a, b\}$. Then
$\Sigma^3 = \{aaa, aab, aba, abb, baa, bab, bba, bbb\}$, where e.g. aba denotes the function returning a for inputs 0 and 2, and b for input 1.

Remarks

- Calling Σ an alphabet is expressing what we use Σ for, it is not constraining what kind of set Σ could be.
- We treat elements of Σ as atomic. If $\Sigma = \{10\}$, then $\Sigma^2 = \{(10)(10)\}$. If $\Sigma = \{0, 1\}$, then $101 \in \Sigma^3$ denotes the string “101”, not the numbers 101 or 9.
- While Σ and Σ^1 are formally distinct things, we can often identify them (like implicit type casting in programming).
- Regardless of Σ, there is a single element in Σ^0, the *empty word* ε.

Formal languages

Definition

Let $\Sigma^* = \bigcup_{n \in \mathbb{N}} \Sigma^n$ be the union of the set of words of all lengths over Σ.

Definition

For $w \in \Sigma^*$, we write $|w| = n$ to express that $w \in \Sigma^n$.

Definition

A (formal) language over the alphabet Σ is a subset L of Σ^*.

Examples of formal languages

- Let Σ be the set of ASCII-characters, and L the set of valid Java programs.
- Let Σ be the set of ASCII-characters, and L the set of valid Java programs that don't erase the disk.
- Let Σ be the words listed in the Oxford dictionary, and L be the set of grammatically correct English sentences using these words.
- Let $\Sigma = \{0, 1\}$, and L be the set of binary encodings for graphs having a Hamiltonian cycle.

Composition

Definition

We define the *composition* of words $\circ : \Sigma^* \times \Sigma^* \to \Sigma^*$ via $|w \circ u| = |w| + |u|$, $(w \circ u)(i) = w(i)$ if $i < |w|$, and $(w \circ u)(i) = u(i - |w|)$ else.

- It holds that $(u \circ v) \circ w = u \circ (v \circ w)$.
- $\varepsilon \circ u = u \circ \varepsilon = u$.
- We usually drop the $\circ$, and just write e.g. *uvw* for $(u \circ v) \circ w$.