language

Canonical name	Language
Date of creation	2013-03-22 12:17:10
Last modified on	2013-03-22 12:17:10
Owner	mps (409)
Last modified by	mps (409)
Numerical id	28
Author	mps (409)
Entry type	Definition
Classification	msc 03C07
Classification	msc 68Q45
Synonym	λ-free
Related topic	Alphabet
Related topic	ContextFreeLanguage
Related topic	RegularLanguage
Related topic	DeterministicFiniteAutomaton
Related topic	NonDeterministicFiniteAutomaton
Related topic	KleeneStar
Related topic	FormalGrammar
Related topic	FirstOrderLanguage
Related topic	TermsAndFormulas
Related topic	Word
Defines	string
Defines	empty language
Defines	substring
Defines	proper language
Defines	improper language
Defines	alphabet of a language
Defines	alphabet of a word
Defines	finite language
Defines	atomic language

Let Σ be an alphabet. We then define the following using the powers of an alphabet and infinite union, where $n \in \mathbb{Z}$.

$$\begin{aligned}\Sigma^+ &= \bigcup_{n=1}^{\infty} \Sigma^n \\ \Sigma^* &= \bigcup_{n=0}^{\infty} \Sigma^n = \Sigma^+ \cup \{\lambda\}\end{aligned}$$

where λ is the element called *empty string*. A *string* is an element of Σ^*, meaning that it is a grouping of symbols from Σ one after another (via concatenation). For example, $abbc$ is a string, and $cbba$ is a different string. A string is also commonly called a *word*. Σ^+, like Σ^*, contains all finite strings except that Σ^+ does not contain the empty string λ. Given a string $s \in \Sigma^*$, a string t is a *substring* of s if $s = utv$ for some strings $u, v \in \Sigma^*$. For example, $lp, al, ha, alpha$, and λ (the empty string) are all substrings of the string $alpha$.

Definition. A *language* over an alphabet Σ is a subset of Σ^*, meaning that it is a `http://planetmath.org/Set`set of strings made from the symbols in the alphabet Σ.

Take for example an alphabet $\Sigma = \{\clubsuit, \wp, 63, a, A\}$. The following are all languages over Σ:

- $\{aaa, \lambda, A\wp 63, 63\clubsuit, AaAaA\}$,
- $\{\wp a, \wp aa, \wp aaa, \wp aaaa, \cdots\}$,
- The empty set $\varnothing$. In the context of languages, $\varnothing$ is called the *empty language*.
- $\{63\}$
- $\{a^{2n} \mid n \geq 0\}$

A language L is said to be *proper* if the empty string does not belong to it: $\lambda \notin L$. A proper language is also said to be λ-*free*. Otherwise, it is *improper*. In the examples above, all but the first and the last examples are λ-free. L is a *finite language* if L is a finite set, and *atomic* if it is a singleton subset of Σ, such as the fourth example above. A language can be arbitrarily formed, or constructed via some set of rules called a formal grammar.

Given a language L over Σ, the *alphabet of* L is defined as the maximal subset $\Sigma(L)$ of Σ such that every symbol in $\Sigma(L)$ occurs in some word in L. Equivalently, define the alphabet of a word w to be the set $\Sigma(w)$ of all symbols that occur in w, then $\Sigma(L)$ is the union of all $\Sigma(w)$, where w ranges over L.

Remark. A language can also be described in terms of "infinite" alphabets. For example, in model theory, a language is built from a set of symbols, together with a set of variables. These sets are often infinite. Another way of generalizing the notion of a language is to allow the strings to have infinite lengths. The way to do this is to think of a string as a partial function f from some set X to the alphabet A such that $|\operatorname{dom}(f)| < |X|$. Then the length of a string $f : X \to A$ is just $|\operatorname{dom}(f)|$. This specializes to the finite case if we take X to be the set of all non-negative integers.