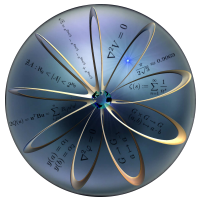

planetmath.org

Math for the people, by the people.

formal grammar

Canonical name	FormalGrammar
Date of creation	2013-03-22 16:27:10
Last modified on	2013-03-22 16:27:10
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	31
Author	CWoo (3771)
Entry type	Definition
Classification	msc 91F20
Classification	msc 68Q42
Classification	msc 68Q45
Classification	msc 03D05
Synonym	phrase-structure grammar
Synonym	unrestricted grammar
Synonym	grammar
Synonym	phrase-structure language
Synonym	terminal symbol
Synonym	non-terminal symbol
Synonym	initial-symbol
Synonym	initial symbol
Synonym	start symbol
Related topic	SemiThueSystem
Related topic	Language
Related topic	PostSystem
Defines	formal language
Defines	terminal
Defines	non-terminal
Defines	starting symbol
Defines	production
Defines	generable by a formal grammar
Defines	sentential form

Introduction

A grammar, loosely speaking, is a set of rules that can be applied to words to generate sentences in a language. For example, with the grammar of the English language, one can form syntactically correct sentences such as "The elephant drove his bicycle to the moon," regardless whether the sentence is meaningful or not.

The mathematical abstraction of a grammar is known as a *formal grammar*. Instead of generating sentences from words, a formal grammar generates words from symbols and other words. The following basic ingredients are necessary in a formal grammar:

- a collection of symbols called an alphabet,
- a collection of rules, called *rewriting rules*, specifying how one can *generate* new words from existing ones, and
- a collection of *initial* words that serve to initialize the generation of new words.

To see how these rewriting rules work, let us look at an example. Let $\{a, b\}$ be the alphabet as well as the set of initial words. With the rewriting rules given by: from a word x we can form the word ax, as well as the word xa, we would be able to generate words like

$$aa, \quad aaa, \quad ab, \quad baa$$

However, words such as

$$bb, \quad baba, \quad baaab$$

can not be produced.

Note that by adding an extra symbol σ to the above alphabet, and two additional rewriting rules given by "from σ form a" and "from σ form b", it is not hard to see that any word that can be generated by the first grammar can be generated by this new grammar.

Definition

Formalizing what we have discussed above, we say that a *formal grammar* G is a quadruple (Σ, N, P, σ), where

1. (Σ, P) is a rewriting system;

2. N is a subset of Σ whose elements are called *non-terminals*, and $T := \Sigma - N$ the set of *terminals*;

3. an element $\sigma \in N$ called the *starting symbol*.

Instead of writing $G = (\Sigma, N, P, \sigma)$, the quadruple (T, N, P, σ) is another way of representing G (as long as the conditions $\Sigma = T \cup N$ and $T \cap N = \varnothing$ are satisfied).

A formal grammar is variously known as a *phrase-structure grammar*, an *unrestricted grammar*, or simply a *grammar*. A formal grammar is sometimes also called a rewriting system in the literature, although the two notions are distinct on PlanetMath.

Given a formal grammar G, a word W over Σ such that $\sigma \stackrel{*}{\Rightarrow} W$ is called a *sentential form* of G. A sentential form over T is called a *word* generated by G. The set of all words generated by G is called the *formal language* generated by G, and is denoted by $L(G)$. In other words,

$$L(G) := \{w \in T^* \mid \sigma \stackrel{*}{\Rightarrow} w\},$$

where $\stackrel{*}{\Rightarrow}$ is the derivability relation in the rewriting system (Σ, P). A *formal language* is also known as a *phrase-structure language*.

A language L over an alphabet A is said to be *generable by a formal grammar* if there is a formal grammar G such that $L = L(G) \cap A^*$.

Example. Continuing from the example in the previous section, we can put $T = \{a, b\}$ and $N = \{\sigma\}$. For the set P of productions, we have four

1. $\sigma \to \sigma a$

2. $\sigma \to a\sigma$

3. $\sigma \to a$

4. $\sigma \to b$

Then $G = (\Sigma, N, P, \sigma)$ is a formal grammar. It is not hard to see that $\sigma \stackrel{*}{\Rightarrow} baa$, as $\sigma \to \sigma a \to \sigma aa \to baa$. In fact, $L(G)$ consists of all words such that a occurs at least once and b occurs at most once.

Remarks.

- Not every language can be generated by a formal grammar. Given a finite alphabet Σ, Σ^* is countably infinite, and therefore there are uncountably many languages over Σ. However, there are only a countably infinitely many languages that can be generated by formal grammars.
- Every language generated by a formal grammar is recursively enumerable.
- Every context-sensitive grammar is equivalent to a formal grammar, and under the Chomsky hierarchy, the class of formal languages is of class 0.

References

[1] H.R. Lewis, C.H. Papadimitriou *Elements of the Theory of Computation.* Prentice-Hall, Englewood Cliffs, New Jersey (1981).