Journal of Automata, Languages and Combinatorics **7** (2002) 3, 377–394

ON THE COMBINATORIAL ALPHABETS OF A LANGUAGE

Daniel-Claudian Voinescu
Faculty of Mathematics, University of Bucharest
2 Riul Dorna block F1 apt 28, R-773441 Bucharest 6, Romania
e-mail: d_c_voinescu@yahoo.com

ABSTRACT

The *combinatorial degree* (or *dimension*) of a language L over a finite alphabet Σ is the positive integer $d(L) = \min\{\mathrm{card}(A) \mid A \subset \Sigma^*,\ L \subset A^*\}$. A subset A of Σ^* for which $L \subset A^*$ and $\mathrm{card}(A) = d(L)$ will be called a *combinatorial alphabet of the language* L. The set of all the combinatorial alphabets of L will be denoted by $\mathcal{CA}(L)$. In this paper we shall establish the main properties and the structure of $\mathcal{CA}(L)$ and also we shall prove that for every language L there exists a finite part L_{f} – which will be called, by analogy with the famous notion of a test-set, a *finite combinatorial test-set* – such that $\mathcal{CA}(L) = \mathcal{CA}(L_{\mathrm{f}})$. We shall prove also that each language has a finite subset which is simultaneously a test-set and a combinatorial test-set, we shall highlight some relations between test-sets and combinatorial test-sets and we shall give some interesting examples.

Keywords: Languages, combinatorial dimension, combinatorial alphabet

1. Introduction

1.1. Notations and Definitions

Throughout the paper Σ will denote a finite nonempty set called *alphabet*, and Σ^+ and Σ^* will be the free semigroup and the free monoid generated by Σ. The elements of Σ^* will be called *words* and any set of words will be called a *language*. The empty word will be denoted by λ. For $L, L_1, L_2 \subset \Sigma^*$, let be $L_1L_2 = \{x_1x_2 \mid x_1 \in L_1,\ x_2 \in L_2\}$, $L^0 = \{\lambda\}$, and for any positive integer n, $L^n = LL^{n-1}$, $L^{\leq n} = \bigcup_{k=0}^{n} L^k$, $L^{\geq n} = \bigcup_{k=n}^{\infty} L^k$, and $L^* = \bigcup_{n\geq 0} L^n$. The *combinatorial degree* or *dimension* of a language L is the positive integer $d(L) = \min\{\mathrm{card}(A) \mid A \subset \Sigma^*,\ L \subset A^*\}$. It follows immediately from the definition that $d(L) \leq \mathrm{card}(\Sigma)$ and if $L_1 \subset L_2$ then $d(L_1) \leq d(L_2)$. A $C \subset \Sigma^*$ is called a *code* if any $w \in C^*$ has a unique factorization over C. If C is a code and $\alpha \in C^*$, the set of the words from C that appear in the unique factorization of α over C will be denoted by $\mathrm{alph}_C(\alpha)$. For $B \subset C^*$, the set $\bigcup_{\alpha \in B} \mathrm{alph}_C(\alpha)$ will be denoted by $\mathrm{alph}_C(B)$. In the particular case $C = \Sigma$ we shall write simply $\mathrm{alph}(\alpha)$ and $\mathrm{alph}(B)$ instead of $\mathrm{alph}_\Sigma(\alpha)$ and $\mathrm{alph}_\Sigma(B)$. An $A \subset \Sigma^*$ for which $d(A) = \mathrm{card}(A)$ is called an *elementary set*. We shall denote the set of all the subsets of Σ^* which