Journal of Automata, Languages and Combinatorics **12** (2007) 3, 407–426

AN EXAMINATION OF OHLEBUSCH AND UKKONEN'S CONJECTURE ON THE EQUIVALENCE PROBLEM FOR E-PATTERN LANGUAGES [1]

DANIEL REIDENBACH [2]
Department of Computer Science, Loughborough University
Loughborough, Leicestershire, LE11 3TU, United Kingdom
e-mail: D.Reidenbach@lboro.ac.uk

ABSTRACT

Our paper contributes new facets to the discussion on the equivalence problem for E-pattern languages (also referred to as extended or erasing pattern languages). This fundamental open question asks for the existence of a computable function which, given any pair of patterns, decides whether or not they generate the same language. Our main result disproves Ohlebusch and Ukkonen's Conjecture (*Theoretical Computer Science* 186, 1997) on the equivalence problem; the respective argumentation – that largely deals with the nondeterminism of pattern languages and, thus, yields new insights into combinatorics on morphisms in free monoids – is restricted to terminal alphabets with at most four distinct letters. Additionally and with regard to larger alphabets, we examine the standard proof technique which in previous works has successfully been applied to restricted variants of the equivalence problem, and we show that it has to be adapted in an unexpected manner if the full class of E-pattern languages is considered. This necessity of modifying the analysed method is caused by a strongly counter-intuitive phenomenon concerning the expressive power of injective morphisms.

Keywords: Pattern languages, decision problems, combinatorics on words, terminal-preserving morphisms, ambiguity

1. Introduction

Patterns – finite strings that consist of variables and terminal symbols – are compact and "natural" devices for the definition of numerous regular and nonregular formal languages. A pattern generates a word by a substitution of the variables with arbitrary strings of terminal symbols (where several occurrences of a variable in the pattern must be replaced with the same word each). Accordingly, its language is the set of all words that can be obtained by such substitutions, which, in fact, can equivalently be interpreted as a particular (namely "terminal-preserving") type of morphisms. For instance, the language generated by the pattern $\alpha = x_1\, x_1\, \mathtt{a}\, \mathtt{b}\, x_2$ (with variables x_1, x_2 and terminals a, b) includes all words where the prefix consists of two occurrences

[1] A preliminary version of this paper has been presented at DLT 2004 [15].
[2] The work was done while the author was employed at the Technische Universität Kaiserslautern.