Journal of Automata, Languages and Combinatorics **1** (1996) 4, 259–303

ON PREFIXES OF FORMAL LANGUAGES AND THEIR RELATION TO THE AVERAGE-CASE COMPLEXITY OF THE MEMBERSHIP PROBLEM

RAINER KEMP
Fachbereich Informatik (20), Johann Wolfgang Goethe-Universität Frankfurt am Main
D–60054 Frankfurt am Main, Germany
e-mail: kemp@sads.informatik.uni-frankfurt.de

ABSTRACT

Given a formal language $\mathcal{L}$ over an alphabet furnished with a probability distribution, we shall present a simple general approach to the computation of the average length of the shortest prefix which has to be read in order to decide whether or not a given word of length n belongs to $\mathcal{L}$. This approach covers a complete average-case analysis of that parameter, including higher moments about the origin and the cumulative distribution function. We shall demonstrate our results by discussing various concrete languages, such as regular sets, Dyck sets, permutations, well-known languages encoding trees, etc.

Keywords: formal languages, membership problem, average-case analysis.

1. Introduction and Basic Observations

A fundamental problem in formal language theory is the so-called *membership problem*, that is the question whether or not a given word w belongs to a given language $\mathcal{L}$. A simple theoretical strategy to solve this problem is as follows: Let $\ell(w)$ be the length of the word w. Scan w from left to right letter by letter until the last symbol of the shortest prefix v which has no extension rightwards to any word of length $\ell(w)$ of the language $\mathcal{L}$. If $w \in \mathcal{L}$, then we have to read $\ell(w)$ symbols; but, if $w \notin \mathcal{L}$, then we only have to read $\ell(v) \leq \ell(w)$ symbols. Naturally, such a recognition procedure presupposes information about the words which have an extension rightwards to a word of length $\ell(w)$ belonging to $\mathcal{L}$ and those ones not having such a continuation.

Now, suppose that we are able to classify the words with respect to these two properties. Furthermore, let us assume that the possible input words of a fixed length are distributed according to a given probability distribution. In respect to the efficiency of the recognition procedure informally described above, we are mainly confronted with the following questions:

- What is the average length of the shortest prefix which we have to read in order to decide whether or not an input word belongs to the given language?