Journal of Automata, Languages and Combinatorics **8** (2003) 1, 51–57

ON THE CONNECTION BETWEEN LEXICOGRAPHICAL GENERATION AND RANKING

Jens Liebehenschel

Fachbereich Biologie und Informatik, Johann Wolfgang Goethe-Universität
D-60054 Frankfurt am Main, Germany
e-mail: `jens@sads.informatik.uni-frankfurt.de`

ABSTRACT

Assuming that all words of a given length in a language are equally likely, we point out a connection between the average costs for the lexicographical generation and the average costs for the ranking of these words.

Keywords: Analysis of algorithms, formal languages, lexicographical generation, ranking

1. Introduction and Basic Definitions

In this paper we investigate the connection between the average costs for the lexicographical generation and for the ranking of formal languages. Ranking a word in a language means to determine its position according to the lexicographical order of all words in the language. The measure is the number of symbols to be changed (resp. read) by an optimal algorithm for the lexicographical generation (resp. ranking). For any *language* $\mathcal{L} \subseteq T^n$ over the *ordered alphabet* T, we compute tight bounds for the sum of the average costs for the lexicographical generation and for the ranking.

From the results in this paper we can draw the following conclusions.

- If the lexicographical generation (resp. the ranking) for a language $\mathcal{L}$ is cheap, i.e. $\mathcal{O}(1)$, then the ranking (resp. the lexicographical generation) is expensive, i.e. $n - \mathcal{O}(1)$.
- If the lexicographical generation (resp. the ranking) for a language $\mathcal{L}$ is expensive, i.e. $n - \mathcal{O}(1)$, then the ranking (resp. the lexicographical generation) requires average costs between $\mathcal{O}(1)$ and $\frac{n}{2} + \mathcal{O}(1)$, but not $n - \mathcal{O}(1)$. For no language the average costs for both, the lexicographical generation and the ranking, are nearly equal to the length of the words.
- Any formal language $\mathcal{L}$ has at most about $\frac{n-1}{2}$ prefixes without a unique continuation to a word in $\mathcal{L}$ per word.

In the sequel, $\ell(w)$ stands for the *length* of $w \in T^\star$. The existence of an ordering on T implies an ordering on the words in $\mathcal{L}$, called *lexicographical order*. By $w^{\mathcal{L}}_{\min}$