On the Size of Higman-Haines Sets

Hermann Gruber*and Markus Holzer†and Martin Kutrib‡

A not so well known theorem in formal language theory is that of Higman [5, Theorem 4.4], which reads as follows:

> If X is any set of words formed from a finite alphabet, it is possible to find a *finite* subset X_0 of X such that, given a word w in X, it is possible to find w_0 in X_0 such that the letters of w_0 occur in w in their right order, though not necessarily consecutively.

In fact, this statement is a corollary to a more general theorem on well-partially-ordered sets. Here a partially ordered set is called well-partially-ordered, if every non-empty subset has at least one, but no more than a finite number of minimal elements (finite basis property). For instance, the set A^*, where A is a finite alphabet, under the scattered subword relation $\leq$, i.e., $v \leq w$ if and only if $v = v_1 \dots v_k$ and $w = w_1 v_1 \dots w_k v_k w_{k+1}$, for some integer k, where v_i and w_j are in A^*, for $1 \leq i \leq k$ and $1 \leq j \leq k+1$, is a well-partially-ordered set. Interestingly, the concept of well-partially-orders has been frequently rediscovered, for example, see [4, 5, 6, 7, 8]. Moreover, although Higman's result appears to be only of theoretical interest, it has some nice applications in formal language theory; see, e.g., [2, 3, 7]. It seems that one of the first applications has been given by Haines in [4, Theorem 3], where it is shown that the set of all scattered subwords, i.e., the *Higman-Haines sets*

$$\textsc{Down}(L) = \{\, v \in A^* \mid \text{there exists } w \in L \text{ such that } v \leq w \,\}$$

and

$$\textsc{Up}(L) = \{\, v \in A^* \mid \text{there exists } w \in L \text{ such that } w \leq v \,\},$$

are both regular for *any* language $L \subseteq A^*$. As pointed out in [4] this is an exceptional property, which is quite unexpected. It is worth mentioning that the regular languages $\textsc{Down}(L)$ and $\textsc{Up}(L)$ cannot be obtained constructively in general (in terms of finite automata).

Although the basic results for Higman-Haines sets date back to the 1950s and 1960s, surprisingly less is known with respect to the (descriptional) size of these sets. To our knowledge the only paper dealing with effective constructibility issues is [8], where an open problem raised in [4] has been solved, i.e.,

*Institut für Informatik, Ludwig-Maximilians-Universität München, Oettingenstraße 67, D-80538 München, Germany. Email: gruberh@tcs.ifi.lmu.de

†Institut für Informatik, Technische Universität München, Boltzmannstraße 3, D-85748 Garching bei München, Germany. Email: holzer@in.tum.de

‡Institut für Informatik, Universität Giessen, Arndtstraße 2, D-35392 Giessen, Germany. Email: kutrib@informatik.uni-giessen.de

Down(L) can effectively be constructed for a given context-free grammar G with $L = L(G)$. Moreover, it was also shown that Up(L) can be obtained effectively, if L is a context-free language. This immediately raises the question whether a similar result holds for the family of Church-Rosser languages. This language family lies in between the regular languages and the growing context-sensitive languages, but is incomparable to the family of context-free languages [1]. In fact, we show that for Church-Rosser languages the size of the Higman-Haines sets cannot be bounded by any recursive function; hence we obtain a non-recursive trade-off result, which implies that the Higman-Haines sets cannot effectively be constructed for Church-Rosser languages and all of its supersets. Moreover, we consider the problem of computing the Higman-Haines sets induced by the families of regular, context-free, and linear context-free languages. For the size of the Higman-Haines sets generated by regular languages upper and lower bounds are presented. That is, we prove that an exponential blow-up is sufficient and necessary in the worst case for a deterministic finite automaton to accept the Higman-Haines set Down(L) or Up(L) generated by some language that is represented by another deterministic finite automaton. This nicely contrasts the result about nondeterministic finite automata, where a linear matching upper and lower bound on the size of Higman-Haines sets can be shown. Furthermore, we investigate the descriptional complexity of the Higman-Haines sets when the underlying device is a context-free or linear context-free grammar.

References

[1] Gerhard Buntrock and Friedrich Otto. Growing context-sensitive languages and Church-Rosser languages. *Inform. Comput.*, 141(1):1–36, 1998.

[2] H. Fernau and F. Stephan. Characterizations of recursively enumerable sets by programmed grammars with unconditional transfer. *J. Autom., Lang. Comb.*, 4(2):117–152, 1999.

[3] Robert H. Gilman. A shrinking lemma for indexed languages. *Theoret. Comput. Sci.*, 163:277–281, 1966.

[4] Leonard H. Haines. On free monoids partially ordered by embedding. *J. Comb. Theory*, 6:94–98, 1969.

[5] Graham Higman. Ordering by divisibility in abstract algebras. *Proc. London Math. Soc., Series 2*, 2:326–336, 1952.

[6] Joseph B. Kruskal. The theory of well-quasi-ordering: A frequently discovered concept. *J. Comb. Theory*, 13:297–305, 1972.

[7] Jan van Leeuwen. A regularity condition for parallel rewriting systems. *SIGACT News*, 8(4):24–27, 1976.

[8] Jan van Leeuwen. Effective constructions in well-partially-ordered free monoids. *Discrete Mathematics*, 21:237–252, 1978.