Language Operations with Regular Expressions of Polynomial Size*

Hermann Gruber and Markus Holzer
Institut für Informatik, Justus-Liebig-Universität Gießen,
Arndtstraße 2, D-35392 Gießen, Germany
email: {gruber,holzer}@informatik.uni-giessen.de

In the last 20 years, a large body of research on the descriptional complexity of finite automata has been developed. To the authors' knowledge, the first systematic attempt to start a parallel development for the descriptional complexity of regular expressions was presented by Ellul et al. [4] at the workshop "Descriptional Complexity of Formal Systems" (DCFS), in 2002. In particular, they raised the question of determining how basic language operations such as complementation and intersection affect the required regular expression size. For the intersection and shuffle operation, exponential lower bounds are known, and complementation can even incur a doubly-exponential blow-up [5, 6]. In [6] it was shown that the star height of a regular language is at most logarithmic in the minimum regular expression size, and lower bounds are proved by finding families of languages for which the respective language operations give rise to a dramatic increase in star height. In contrast, it is well known that taking language quotients does not increase the star height [3]. This and similar language operations appear to be a natural testing ground for deepening our understanding of the descriptional complexity of regular expressions: Either one has to find some new lower bound techniques, or one has to find a nontrivial implementation of these operations on regular expressions, or both—a straightforward procedure would be to convert the expression into a finite automaton, implement the operation on a finite automaton, and convert back to a regular expression using state elimination. Yet that last step can incur an exponential blow-up in general, even over binary alphabets [6].

*Most of the work was done while the first author was at Institut für Informatik, Ludwig-Maximilians-Universität München, Oettingenstraße 67, D-80538 München, Germany, and the second author was at Institut für Informatik, Technische Universität München, Boltzmannstraße 3, D-85748 Garching bei München, Germany.

Here, we give upper bounds for the required expression size resulting from taking language quotients and circular shift. The *(left) quotient* of L with respect to a set of words W, denoted by $W^{-1}L$, is defined as $\bigcup_{w \in W} w^{-1}L$, where $w^{-1}L = \{ x \mid wx \in L \}$, for some word w. Moreover, the circular (or cyclic) shift of a language, denoted by $\circlearrowleft(L)$, is given by $\{ xw \mid wx \in L \}$. Descriptional complexity aspects of these operations were already studied in [1, 11] for the circular shift and [8, 9] for language quotients—the latter two references consider deterministic finite automata with multiple start states, but the results easily translate to state complexity results for (left) quotients. The basic idea is to implement the operation for the special case of linear expressions [2] called single-occurrence regular expressions in [5]. These are expressions in which every alphabetic symbol occurs exactly once, which makes it easier to deal with as they can describe only local languages. To cover the general case, we study the interplay of the operations with length-preserving homomorphisms. The main result reads as follows—the size of a regular expression is defined as the total number of occurrences of letters from the underlying alphabet:

Theorem 1 *Let r be a regular expression of size n denoting the language $L \subseteq \Sigma^*$, and let $W \subseteq \Sigma^*$. Then there is a regular expression of size $O(n^2)$ denoting $W^{-1}L$ and a regular expression of size $O(n^3)$ denoting $\circlearrowleft(L)$.*

Currently, we do not know whether these upper bounds have the right order of magnitude. Nevertheless, it is worth mentioning that for the latter operation, i.e., the circular shift operation, at least an almost quadratic blow-up can be necessary in the worst case.

Theorem 2 *There exist infinitely many regular languages L_m over a binary alphabet such that L_m admits a regular expression of alphabetic width m, but every regular expression describing $\circlearrowleft(L_m)$ has size at least $\Omega\left(\frac{m^2}{\log^2 m}\right)$.*

One task for further research is to find other regularity preserving operations for which this or similar approaches might work. For instance, for the language of scattered substrings (superstrings, respectively) of the language described by a regular expression over Σ, we simply replace every position a with a subexpression $\lambda + a$ (with a subexpression describing $\Sigma^* a \Sigma^*$, respectively) to obtain a regular expression denoting that language. Both operations can be thus performed with only linear increase in expression size provided Σ is fixed. Issues on the state complexity of these operations were studied recently in [7] and [10].

References

[1] P. R. J. Asveld. Generating all circular shifts by context-free grammars in Greibach normal form. *International Journal of Foundations of Computer Science*, 18(6):1139–1149, 2007.

[2] G. Berry and R. Sethi. From regular expressions to deterministic automata. *Theoretical Computer Science*, 48:117–126, 1986.

[3] R. S. Cohen and J. A. Brzozowski. General properties of star height of regular events. *Journal of Computer and System Sciences*, 4(3):260–280, 1970.

[4] K. Ellul, B. Krawetz, J. Shallit, and M. Wang. Regular expressions: New results and open problems. *Journal of Automata, Languages and Combinatorics*, 10(4):407–437, 2005.

[5] W. Gelade and F. Neven. Succinctness of the complement and intersection of regular expressions. In S. Albers and P. Weil, editors, *Proceedings of the* 25*th Symposium on Theoretical Aspects of Computer Science*, volume 08001 of *Dagstuhl Seminar Proceedings*, pages 325–336, Bordeaux, France, February 2008. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

[6] H. Gruber and M. Holzer. Finite automata, digraph connectivity, and regular expression size. In L. Aceto, I. Damgaard, L. A. Goldberg, M. M. Halldórsson, A. Ingólfsdóttir, and I. Walkuwiewicz, editors, *Proceedings of the* 35*th International Colloquium on Automata, Languages and Programming*, volume 5126 of *LNCS*, pages 39–50, Reykjavik, Iceland, July 2008. Springer.

[7] H. Gruber, M. Holzer, and M. Kutrib. More on the size of Higman-Haines sets: Effective constructions. In J. O. Durand-Lose and M. Margenstern, editors, *Proceedings of the* 5*th International Conference Machines, Computations, and Universality*, volume 4664 of *LNCS*, pages 193–204, Orléans, France, September 2007. Springer.

[8] M. Holzer, K. Salomaa, and S. Yu. On the state complexity of k-entry deterministic finite automata. *Journal of Automata, Languages and Combinatorics*, 6(4):453–466, 2001.

[9] M. Kappes. Descriptional complexity of deterministic finite automata with multiple initial states. *Journal of Automata, Languages and Combinatorics*, 5(3):269–278, 2000.

[10] A. Okhotin. On the state complexity of scattered substrings and superstrings. Technical Report TUCS Technical Report No.849, University of Turku - Department of Mathematics and Turku Centre for Computer Science and Academy of Finland, October 2007.

[11] A. Okhotin and G. Jirásková. State complexity of cyclic shift. *RAIRO—Theoretical Informatics and Applications*, 42(2):335–360, 2008.