

Quotient Complexity of Bifix-, Factor-, and Subword-Free Regular Languages ^{*}

Janusz Brzozowski¹ Galina Jirásková² Baiyu Li¹ Joshua Smith¹

¹ David R. Cheriton School of Computer Science, University of Waterloo,
Waterloo, ON, Canada N2L 3G1

{brzozo@, b5li@, j45smith@student.math.}uwaterloo.ca

² Mathematical Institute, Slovak Academy of Sciences,
Grešákova 6, 040 01 Košice, Slovakia
jiraskov@saske.sk

Abstract

A language L is prefix-free if, whenever words u and v are in L and u is a prefix of v , then $u = v$. Suffix-, factor-, and subword-free languages are defined similarly, where “subword” means “subsequence”. A language is bifix-free if it is both prefix- and suffix-free. We study the quotient complexity, more commonly known as state complexity, of operations in the classes of bifix-, factor-, and subword-free regular languages. We find tight upper bounds on the quotient complexity of intersection, union, difference, symmetric difference, concatenation, star, and reversal in these three classes of languages.

1 Introduction

The state complexity of a regular language L is the number of states in the minimal deterministic finite automaton (dfa) accepting L [28]. This complexity is the same as the quotient complexity [5] of L , which is the number of distinct left quotients of L . We prefer quotient complexity since it is more closely related to properties of languages. The quotient complexity of an operation in a class \mathcal{C} of regular languages is the worst-case quotient complexity of the language resulting from the operation, taken as a function of the quotient complexities of the operands in class \mathcal{C} . For surveys on state and quotient complexity see [5, 28].

^{*}This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant no. OGP0000871 and by the Slovak Research and Development Agency under contract APVV-0035-10 “Algorithms, Automata, and Discrete Data Structures”.

One of the first results concerning the state complexity of an operation is the 1966 theorem by Mirkin [19], who showed that the bound 2^n for the reversal of an n -state dfa can be attained. In 1970 Maslov [18] stated without proof the bounds on the complexities of union, concatenation, star, and several other operations in the class of regular languages, and gave languages meeting these bounds. In 1994 these operations, along with intersection, reversal, and left and right quotients, were studied in detail by Yu, Zhuang and Salomaa [27].

Results exist also for proper subclasses of the class of regular languages: unary [21, 27], finite [9, 11, 28], cofinite [2], prefix-free [13, 14], suffix-free [10, 12, 15], ideal [6], and closed [8]. The bounds can vary considerably.

Free languages (with the exception of $\{\varepsilon\}$, where ε is the empty word) are codes, which constitute an important class of languages and have applications in such areas as cryptography, data compression, and information transmission. They have been studied extensively; see, for example, [3, 16]. In particular, *prefix codes* [3] are prefix-free and suffix-free languages, respectively, *infix codes* [22, 23] are factor-free, and *hypercodes* [22, 23] are subword-free, where by subword we mean subsequence. Moreover, free languages are special cases of convex languages [1, 24]. We are interested only in regular free languages.

The state complexities of intersection, union, concatenation, star, and reversal were first studied by Han, K. Salomaa, and Wood [13] for prefix-free languages, and by Han and K. Salomaa [12] for suffix-free languages. In the present paper, these results are extended to bifix-, factor- and subword-free languages. In particular, we obtain tight upper bounds on the complexities of intersection, union, difference, symmetric difference, star, concatenation, and reversal in these three classes of free languages. All the omitted proofs can be found in [7].

2 Preliminaries

It is assumed that the reader is familiar with finite automata and regular languages as treated in [20, 26], for example. If Σ is a finite non-empty alphabet, then Σ^* is the set of all words over this alphabet, with ε as the empty word. For $w \in \Sigma^*$, let $|w|$ be the length of w . A language is any subset of Σ^* .

The following set operations are defined on languages: *complement* ($\bar{L} = \Sigma^* \setminus L$), *union* ($K \cup L$), *intersection* ($K \cap L$), *difference* ($K \setminus L$), and *symmetric difference* ($K \oplus L$). A general *boolean operation* with two arguments is denoted by $K \circ L$.

We also define the *product*, usually called *concatenation* or *catenation*, ($KL = \{w \in \Sigma^* \mid w = uv, u \in K, v \in L\}$), (Kleene) *star* ($L^* = \bigcup_{i \geq 0} L^i$ with $L^0 = \{\varepsilon\}$), and *positive closure* ($L^+ = \bigcup_{i \geq 1} L^i$).

The reverse w^R of a word $w \in \Sigma^*$ is defined inductively as follows: $\varepsilon^R = \varepsilon$, and $(wa)^R = aw^R$ for every symbol a in Σ and every word w in Σ^* . The *reverse* of a language L is denoted by L^R and is defined as $L^R = \{w^R \mid w \in L\}$.

Regular languages over Σ are languages that can be obtained from the *set of basic languages* $\{\emptyset, \{\varepsilon\}\} \cup \{\{a\} \mid a \in \Sigma\}$, using a finite number of operations of union, product, and star. We use regular expressions to represent languages. If E is a regular expression, then $\mathcal{L}(E)$ is the language denoted by that expression. For example, the regular expression $E = (\varepsilon \cup a)^*b$ denotes language $L = \mathcal{L}(E) = (\{\varepsilon\} \cup \{a\})^*\{b\}$. We usually do not distinguish notationally between regular languages and regular expressions.

Whenever convenient, we derive upper bounds on the state complexity of operations on free languages following the approach of [5]. A *quotient* of a language L by a word w is defined as $L_w = \{x \in \Sigma^* \mid wx \in L\}$. The number of distinct quotients of a language is called its *quotient complexity* and is denoted by $\kappa(L)$.

Quotients of regular languages [4, 5] can be computed as follows: First, the ε -*function* L^ε of a regular language L is $L^\varepsilon = \emptyset$ if $\varepsilon \notin L$, and $L^\varepsilon = \varepsilon$ if $\varepsilon \in L$. The *quotient by a letter* a in Σ is computed by induction: $b_a = \emptyset$ if $b \in \{\emptyset, \varepsilon\}$ or $b \in \Sigma$ and $b \neq a$, and $b_a = \varepsilon$ if $b = a$; $(\overline{L})_a = \overline{L_a}$; $(K \circ L)_a = K_a \circ L_a$; $(KL)_a = K_a L \cup K^\varepsilon L_a$; $(L^*)_a = L_a L^*$. The quotient by a word w in Σ^* is computed by induction on the length of w : $L_\varepsilon = L$ and $L_{wa} = (L_w)_a$. A quotient L_w is *accepting* if $\varepsilon \in L_w$; otherwise it is *rejecting*.

A *deterministic finite automaton* (dfa) is a quintuple $\mathcal{D} = (Q, \Sigma, \delta, q_0, F)$, where Q is a finite set of *states*, Σ is a finite *alphabet*, $\delta : Q \times \Sigma \rightarrow Q$ is the *transition function*, q_0 is the *initial state*, and $F \subseteq Q$ is the set of *final* or *accepting states*. As usual, the transition function is extended to $Q \times \Sigma^*$. The dfa \mathcal{D} accepts a word w in Σ^* if $\delta(q_0, w) \in F$. The set of all words accepted by \mathcal{D} is $L(\mathcal{D})$. By the *language of a state* q of \mathcal{D} we mean the language L_q accepted by the automaton $(Q, \Sigma, \delta, q, F)$. A state is *empty* if its language is empty.

The *quotient automaton* of a regular language L is the dfa $\mathcal{D} = (Q, \Sigma, \delta, q_0, F)$, where $Q = \{L_w \mid w \in \Sigma^*\}$, $\delta(L_w, a) = L_{wa}$, $q_0 = L_\varepsilon$, $F = \{L_w \mid \varepsilon \in L_w\}$. This is the minimal dfa accepting L . Hence the quotient complexity of L is equal to the state complexity of L , and we call it simply *complexity*.

3 Free Languages

If $u, v, w, x \in \Sigma^*$ and $w = uxv$, then u is a *prefix* of w , x is a *factor* of w , and v is a *suffix* of w . Both u and v are also factors of w . If $w = u_0v_1u_1 \cdots v_nu_n$, where $u_i, v_i \in \Sigma^*$, then $v = v_1v_2 \cdots v_n$ is a *subword* of w . Every factor of w is also a subword of w .

A language L is *prefix-free* (respectively, *suffix-free*, *factor-free*, or *subword-free*) if, whenever words u and v are in L and u is a prefix (respectively, suffix, factor, or subword) of v , then $u = v$. Additionally, L is *bifix-free* if it is both prefix and suffix-free. All subword-free languages are factor-free, and all factor-free languages are bifix-free. For convenience, we refer to prefix-, suffix-, bifix-, factor-, and subword-free languages together as *free* languages.

If ε is a quotient of L , then L also has the empty quotient, since $\varepsilon_a = \emptyset$, for all a in Σ . We say that a quotient L_w is *uniquely reachable* if $L_w = L_x$ implies that $w = x$. We now restate two propositions from [12, 13] in our terminology.

Proposition 1. *A non-empty language is prefix-free if and only if it has exactly one accepting quotient and that quotient is ε .*

Proposition 2. *The quotient by ε of a non-empty suffix-free language is uniquely reachable, and the language has the empty quotient.*

Let L be any language. If $(L_u)_x = L_v$ for some words u, v and a non-empty word x , then L_v is *positively reachable* from L_u , and we denote this by $L_u \rightarrow L_v$. The relation \rightarrow is transitive. The next proposition uses this relation to characterize finite languages.

Proposition 3. *If L is any language with the set of quotients $\{L_1, L_2, \dots, L_n\}$, and $u, v \in \Sigma^*$, then the following are equivalent:*

1. L is finite.
2. $L_u \rightarrow L_v$ and $L_v \rightarrow L_u$ if and only if $L_u = L_v = \emptyset$.
3. There exists a total order \preceq on the set of quotients:

$$L = L_1 \preceq L_2 \preceq \dots \preceq L_{n-1} \preceq L_n = \emptyset$$

which satisfies the condition that $(L_i)_a = L_j$ implies $L_i \prec L_j$ or $L_i = L_j = L_n$.

Since every subword-free language is finite, we get the next lemma, which we use later to prove that upper bounds on the quotient complexity of some operations on subword-free languages cannot be reached if the alphabet of the language does not have sufficiently many letters.

Lemma 1. *Let L be a subword-free language with $\kappa(L) = n$, where $n \geq 4$. Let the distinct quotients $L = L_\varepsilon = L_1, L_2, \dots, L_{n-2}, L_{n-1} = \varepsilon, L_n = \emptyset$ of L be ordered as in Proposition 3. If $L_w = L_2$ for some word w , then $|w| = 1$.*

Finally, we describe a simple method of constructing free languages.

Proposition 4. *Let $L \subseteq \Sigma^*$ be any language, and let $a \notin \Sigma$. Then (1) aL is suffix-free, (2) La is prefix-free, (3) aLa is factor-free.*

4 Boolean Operations

The complexity of boolean operations, in the class of prefix- and suffix-free regular languages, except for the difference and symmetric difference of suffix-free languages, was studied in [12, 13, 14, 15]. It was shown that for prefix-free languages, the tight bounds for union, intersection, difference, and symmetric difference are $mn - 2$, $mn - 2(m + n - 3)$, $mn - (m + 2n - 4)$, and $mn - 2$, respectively. For union and intersection of suffix-free languages, the tight bounds are $mn - (m + n - 2)$ and $mn - 2(m + n - 3)$, respectively. The bounds for difference and symmetric difference are $mn - (m + 2n - 4)$ and $mn - (m + n - 2)$, respectively, and the bounds for all four boolean operations are met by binary suffix-free languages [10]. The next two theorems provide results for boolean operations on bifix-, factor-, and subword-free languages.

Theorem 1 (Boolean Operations: Bifix- and Factor-Free Languages). *Let K and L be bifix- or factor-free languages over an alphabet Σ with $\kappa(K) = m$ and $\kappa(L) = n$, where $m, n \geq 4$. Then*

1. $\kappa(K \cap L) \leq mn - 3(m + n - 4)$, and the bound is tight if $|\Sigma| \geq 2$;
2. $\kappa(K \setminus L) \leq mn - (2m + 3n - 9)$, and the bound is tight if $|\Sigma| \geq 2$;
3. $\kappa(K \cup L) \leq mn - (m + n)$, and the bound is tight if $|\Sigma| \geq 3$;
4. $\kappa(K \oplus L) \leq mn - (m + n)$, and the bound is tight if $|\Sigma| \geq 3$.

Proof. Since K and L are bifix-free, by unique reachability we get a reduction of $m + n - 2$ from the general bound mn . Moreover, both languages K and L have ε and \emptyset as quotients. For intersection, we have $\emptyset \cap L_w = K_w \cap \emptyset = \emptyset$, and the quotients $\varepsilon \cap L_w$ and $K_w \cap \varepsilon$ are either empty or equal to ε . This gives the upper bound. For difference, we eliminate $m + n - 2$ quotients by unique reachability, $n - 2$ quotients by the fact that $\emptyset \setminus L_w = \emptyset$ (keeping only one representative $\emptyset \setminus \emptyset$), $m - 2$ quotients by the fact that $K_w \setminus \emptyset = K_w \setminus \varepsilon$ (keeping $K_w \setminus \emptyset$ as a representative), and $n - 3$ more quotients by the rule $\varepsilon \setminus L_w = \varepsilon$, for a total reduction of $(2m + 3n - 9)$. For union, we have the unique reachability reduction of $m + n - 2$, and a further reduction of 2 by the rule $\varepsilon \cup \varepsilon = \varepsilon \cup \emptyset = \emptyset \cup \varepsilon = \varepsilon$. For symmetric difference, we note that $\varepsilon \oplus \varepsilon = \emptyset \oplus \emptyset = \emptyset$ and $\varepsilon \oplus \emptyset = \emptyset \oplus \varepsilon = \varepsilon$.

For tightness, consider $K = a(c^*(a \cup b))^{m-3}$, $L = a(b^*(a \cup c))^{n-3}$; see Figure 1. If $w \in K$, then $w = av$ for some word v containing $m - 3$ occurrences of symbols from $\{a, b\}$ and ending in a or b . This means that no proper factor of w is in K , and so K is factor-free. A similar proof applies to L .

In the cross-product automaton for the boolean operations on K and L , all the states are reached from the initial state $(1, 1)$ by a word in $ab^*c^* \cup ac^*b^*$, except for state $(m - 1, n - 1)$ which is reached from state $(m - 2, n - 2)$ by a .

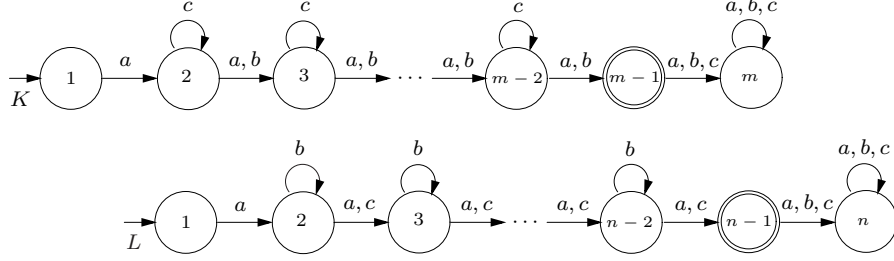


Figure 1: Factor-free languages meeting the upper bounds on boolean operations.

For intersection, the only accepting state is $(m - 1, n - 1)$. All the rejecting states in rows $m - 1$ and m and columns $n - 1$ and n are empty. The word a is accepted only from $(m - 2, n - 2)$, word $b^{m-2-i}c^{n-2-j}a$ ($2 \leq i \leq m - 2$, $2 \leq j \leq n - 2$) only from state (i, j) , and the word $ab^{n-4}c^{n-4}a$ only from state $(1, 1)$. This gives $mn - 3(m + n - 4)$ reachable and pairwise distinguishable states.

For difference, all the states of the cross-product automaton in row $m - 1$, except for $(m - 1, n - 1)$, are accepting and accept ε . All the states in row m , as well as state $(m - 1, n - 1)$ are empty. Moreover, states $(i, n - 1)$ and (i, n) with $2 \leq i \leq m - 2$ are equivalent. The word ab^{m-3} is accepted only from $(1, 1)$. Now let (i, j) and (k, ℓ) , where $2 \leq i \leq n - 1$, $2 \leq j \leq m - 2$, be two distinct states. If $i < k$, then $c^i b^{m-1-i}$ is accepted from (i, j) but not from (k, ℓ) . If $i = k$ and $j < \ell$, then $b^{m-2-i}c^{n-2-j}a$ is not accepted from (i, j) but is accepted from (k, ℓ) . This means that $mn - (2m + 3n - 9)$ states are pairwise distinguishable.

For union, all the states in row $m - 1$ and in column $n - 1$ are accepting, and moreover, the three states $(m, n - 1)$, $(m - 1, n - 1)$, and $(m - 1, n)$ are equivalent. The word ab^{m-3} is accepted only from $(1, 1)$. Consider two distinct rejecting states (i, j) and (k, ℓ) . If $i < k$, then $c^i b^{m-1-i}$ is accepted from (i, j) but not from (k, ℓ) . If $j < \ell$, then $b^m c^{n-1-j}$ is accepted from (i, j) but not from (k, ℓ) . Now consider two distinct accepting states different from $(m, n - 1)$ and $(m - 1, n)$. By c , the two states either go to two states one of which is accepting and the other rejecting, or to two distinct rejecting, and hence distinguishable, states. This proves distinguishability of $mn - (m + n)$ states.

The proof for symmetric difference is the same as for union, except that state $(m - 1, n - 1)$ is empty and states $(m, n - 1)$ and $(m - 1, n)$ are equivalent.

The bounds for intersection and difference are met by binary factor-free languages $K = a(ba^*b^{m-5}b)^*ba^*b^{m-5}a$ and $L = (ab^*)^{n-3}a$. We omit the proof due to space constraints. \square

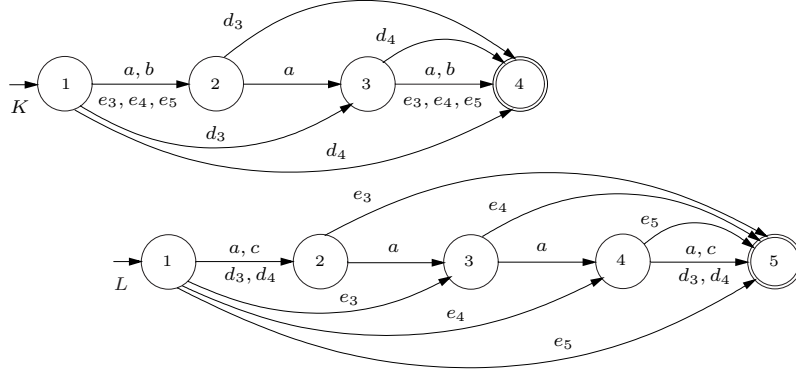


Figure 2: Subword-free witness languages for boolean operations; $m = 5$, $n = 6$.

Theorem 2 (Boolean Operations: Subword-Free Languages). *Let K and L be subword-free languages over an alphabet Σ with $\kappa(K) = m$ and $\kappa(L) = n$, where $m, n \geq 4$. Then*

1. $\kappa(K \cup L), \kappa(K \oplus L) \leq mn - (m + n)$, and the bound is tight if $|\Sigma| \geq m + n - 3$;
 2. $\kappa(K \cap L) \leq mn - 3(m + n - 4)$, and the bound is tight if $|\Sigma| \geq m + n - 7$;
 3. $\kappa(K \setminus L) \leq mn - (2m + 3n - 9)$, and the bound is tight if $|\Sigma| \geq m + n - 6$.
- Moreover, the bounds cannot be met for smaller alphabets.

Proof. Since subword-free languages are bifix-free, all the upper bounds apply. To prove tightness, let $\Sigma = \{a, b, c\} \cup \{d_i \mid 3 \leq i \leq m-1\} \cup \{e_j \mid 3 \leq j \leq n-1\}$. Consider the languages K and L defined by the following quotient equations:

$$\begin{aligned}
K_1 &= (a \cup b \cup e_3 \cup \dots \cup e_{n-1})K_2 \cup \bigcup_{i=3}^{m-1} d_i K_i, \\
K_i &= aK_{i+1} \cup d_{i+1}K_{m-1} && i = 2, 3, \dots, m-3, \\
K_{m-2} &= (a \cup b \cup d_{m-1} \cup e_3 \cup e_4 \cup \dots \cup e_{n-1})K_{m-1}, \\
K_{m-1} &= \varepsilon, \\
K_m &= \emptyset, \\
L_1 &= (a \cup c \cup d_3 \cup \dots \cup d_{m-1})L_2 \cup \bigcup_{j=3}^{n-1} e_j L_j, \\
L_j &= aL_{j+1} \cup e_{j+1}L_{n-1} && j = 2, 3, \dots, n-3, \\
L_{n-2} &= (a \cup c \cup e_{n-1} \cup d_3 \cup d_4 \cup \dots \cup d_{m-1})L_{n-1}, \\
L_{n-1} &= \varepsilon, \\
L_n &= \emptyset.
\end{aligned}$$

Figure 2 shows the quotient automata for languages K and L if $m = 5$ and $n = 6$. All the omitted transitions go to the empty states m and n .

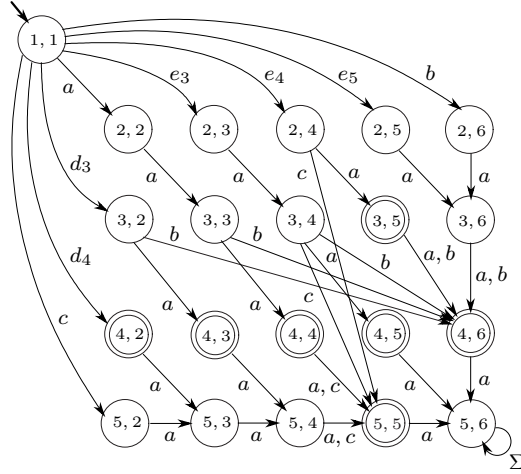


Figure 3: Reachability in the cross-product automaton for the union of languages from Figure 2 and transitions by b and c .

Let us show that languages K and L are subword-free. For this purpose, let $\Gamma = \{a, b, e_3, e_4, \dots, e_{n-1}\}$, and $\Delta = \{d_3, d_4, \dots, d_{m-1}\}$. Notice that no word in Γ^* of length less than $m - 2$ is in K . Now let w be a word in K . Then w either contains no letter from Δ , or contains at most two such letters. If w contains no letter from Δ , then w is a word in Γ^* of length $m - 2$, and so no its proper subword is in K . If w contains exactly one letter from Δ , then either $w = ud_i$ for some word u in Γ^* of length $i - 2$, or $w = d_iv$ for some word v in Γ^* of length $m - 1 - i$. In both cases, no proper subword of w is in K . Finally, if w contains two letters from Δ , then $w = d_ia^k d_{i+k+1}$ where $k \geq 0$ and $3 \leq i < i + k + 1 \leq m - 2$. No proper subword of such a word is in K . This means that K is subword-free. The proof for L is similar.

Figure 3 depicts the cross-product automaton of the dfa's for languages K and L defined in Figure 2, where we show only the transitions necessary to prove reachability and those caused by b and c . In the cross-product automaton, states in the first row and the first column, except for the initial state $(1, 1)$, are unreachable. Now consider the remaining states. All the states in the second row and the second column are reached from $(1, 1)$ by symbols in Σ . Each other state is reached from a state in the second row or second column by a word in a^* .

For union, all the states in row $m - 1$ and in column $n - 1$ are accepting, and the three states $(m, n - 1)$, $(m - 1, n - 1)$, and $(m - 1, n)$ are equivalent. These three states are distinguishable from all other accepting states, since each

of the other accepting states accepts at least one non-empty word. Now let (i, j) and (k, ℓ) be two distinct states other than the three states accepting only ε . First assume that $i < k$. If $i = m - 1$, then state (i, j) is accepting while state (k, ℓ) is rejecting. If $i \leq m - 2$, then $a^{m-2-i}b$ is accepted from state (i, j) , but not from state (k, ℓ) . Symmetrically, if $j < \ell$, then either ε or $a^{n-2-j}c$ distinguishes the two states. Therefore all the $mn - (m + n)$ states are pairwise distinguishable. For symmetric difference, $(m - 1, n - 1)$ is empty; the rest of the proof is the same as for union.

For intersection, the only accepting state is $(m - 1, n - 1)$, and all the rejecting states in the last two rows and last two columns are empty. Next, the word a is accepted only from state $(m - 2, n - 2)$, the word d_i ($3 \leq i \leq n - 2$) is accepted only from state $(i - 1, n - 2)$, while the word e_j ($3 \leq j \leq m - 2$), only from state $(m - 2, j - 1)$. This means that for each state (i, j) , there exists a word in $a^*(a \cup d_3 \cup \dots \cup d_{m-2} \cup e_3 \cup \dots \cup e_{n-2})$ that is accepted only from (i, j) . So we get $mn - 3(m + n - 4)$ pairwise distinguishable states. Notice, that here we do not use transitions by symbols b, c, d_{m-1}, e_{n-1} , and so we can simply omit these symbols to get witness languages over an alphabet of size $m + n - 7$.

For difference, all the states in row $m - 1$, except for state $(m - 1, n - 1)$, are accepting and accept ε . All the states in the last row, as well as state $(m - 1, n - 1)$, are empty, and states $(i, n - 1)$ and (i, n) with $2 \leq i \leq m - 2$ are equivalent. States in different rows (up to row $m - 1$) are distinguished by a word in a^*b . States in row $m - 2$ are distinguished by a word in $a \cup e_3 \cup e_4 \cup \dots \cup e_{n-2}$ because a distinguishes states $(m - 2, n - 2)$ and $(m - 2, n - 1)$, and if $2 \leq j < \ell \leq n - 1$ and $j \neq n - 2$, then word e_{j+1} is not accepted from $(m - 2, j)$ but is accepted from $(m - 2, \ell)$. Next, states $(i, n - 2)$ and $(i, n - 1)$ with $2 \leq i \leq m - 3$ are distinguished by d_{i+1} . Finally, if two distinct states are in the same row, then there is a word in a^* , by which the two states either go to two distinct states in row $m - 2$, or to two states $(i, n - 2)$ and $(i, n - 1)$ with $2 \leq i \leq m - 3$. In both cases the resulting states are distinguishable, which proves the distinguishability of $mn - (2m + 3n - 9)$ states. Notice that now we do not use transitions by c, d_{m-1}, e_{n-1} , and so the bound is met for an alphabet of size $m + n - 6$.

We now show that the upper bounds cannot be met using smaller alphabets. Let the quotients of K and L be $K = K_1, K_2, \dots, K_{m-2}, K_{m-1} = \varepsilon, K_m = \emptyset$, and $L = L_\varepsilon = L_1, L_2, \dots, L_{n-2}, L_{n-1} = \varepsilon, L_n = \emptyset$, ordered as in Proposition 3. By Lemma 1, all the quotients of the form $K_2 \cup L_i$ or $K_j \cup L_2$ must be reached by letters if the bound is to hold, and this is impossible if the size of the alphabet is smaller than the number of such quotients. \square

5 Product and Star

The complexity of product of prefix-free languages is $m + n - 2$ [13]. For suffix-free languages, the complexity is $(m - 1)2^{n-1} + 1$ [12]. Since bifix-free languages are prefix-free, and the witness prefix-free languages a^{m-2} and a^{n-2} are also subword-free, and we have the following result.

Theorem 3 (Product). *If K and L are bifix-free with $\kappa(K) = m$ and $\kappa(L) = n$, where $m, n \geq 2$, then $\kappa(KL) \leq m + n - 2$. Furthermore, there are unary subword-free languages that meet this bound.*

The complexity of star is n for prefix-free languages [13], and $2^{n-2} + 1$ for suffix-free languages [12]. We now extend these results to bifix-, factor-, and subword-free languages. The quotient of L^* by ε is $L^* = \varepsilon \cup LL^*$, and the following formula holds for a quotient of L^* by a non-empty word w [5]:

$$(L^*)_w = (L_w \cup \bigcup_{\substack{w=uv \\ u,v \in \Sigma^+}} (L^*)_u L_v) L^*.$$

Theorem 4 (Star). *If L is bifix-free with $\kappa(L) = n$, where $n \geq 3$, then $\kappa(L^*) \leq n - 1$. Furthermore, there are binary subword-free languages that meet this bound.*

Proof. Assume that L is bifix-free. Then it is prefix-free, has only one accepting quotient, namely ε , and has the empty quotient, by Proposition 1. Moreover, since L is suffix-free, the quotient L is uniquely reachable by ε , by Proposition 2.

Let L_w be a non-empty quotient of L by a non-empty word w . Let us show that $(L^*)_u^\varepsilon = \emptyset$ for every proper non-empty prefix u of w . Assume for contradiction that $\varepsilon \in (L^*)_u$, where $w = uv$ for some non-empty words u and v . Then $u \in L^*$, and so there exist words x in L and y in L^* such that $u = xy$. This gives $L_w = L_{xyv} = \varepsilon_{yv} = \emptyset$ because $x \in L$ implies $L_x = \varepsilon$. This is a contradiction, and so we must have $(L^*)_u^\varepsilon = \emptyset$. Hence, if L_w is non-empty, then $(L^*)_w = L_w L^*$, by the equation above. Now if L_w is accepting, then $L_w = \varepsilon$, and so $(L^*)_w = L^* = (L^*)_\varepsilon$. There are $n - 2$ choices for rejecting and non-empty quotients L_w . But, for a non-empty word w , we have $L_w \neq L$ since L is uniquely reachable by ε . This reduces the number of choices to $n - 3$ (since we have $n \geq 3$). If $L_w = \emptyset$, then by the observation above, $(L^*)_w = (L^*)_u^\varepsilon L_v L^*$, where $w = uv$ and v is the shortest word such that $L_v \neq \emptyset$. Such a quotient is either empty or has already been counted. In total, there are at most $n - 1$ quotients of L^* .

The subword-free language a^{n-2} over the alphabet $\{a, b\}$ meets the bound since the language $(a^{n-2})^*$ has $n - 2$ quotients of the form $a^{n-2-i}(a^{n-2})^*$ for $i = 1, 2, \dots, n - 2$, and it has the empty quotient, for a total of $n - 1$. \square

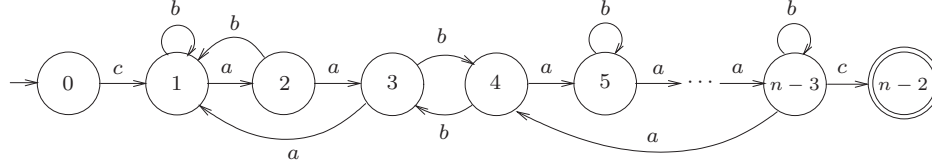


Figure 4: The ternary factor-free language meeting the $2^{n-3}+2$ bound on reversal.

6 Reversal

The last operation we consider is reversal. In [12, 13] it was shown that the complexity of reversal is $2^{n-2} + 1$ for suffix-free or prefix-free languages. We show that this bound can be reduced for bifix-free languages. We use the standard method of reversing the quotient dfa \mathcal{D} of L to obtain an nfa \mathcal{N} for L^R , and then we use subset construction to find the dfa \mathcal{D}^R for L^R .

Theorem 5 (Reversal: Bifix- and Factor-Free Languages). *If L is a bifix-free language with $\kappa(L) = n$, where $n \geq 3$, then $\kappa(L^R) \leq 2^{n-3} + 2$. Moreover, there exist ternary factor-free languages that meet this bound.*

Proof. If L is bifix-free, then so is L^R . Since L is prefix-free, it has exactly one accepting quotient, ε , and also has the empty quotient.

Consider the quotient automaton \mathcal{D} for L , and remove the empty quotient and all the transitions to the empty quotient. Reverse this incomplete dfa to get an $(n-1)$ -state nfa \mathcal{N} for L^R . Apply the subset construction to \mathcal{N} to get a dfa \mathcal{D}^R for L^R . The initial state of dfa \mathcal{D}^R is the singleton set $\{f\}$, where f is the ε quotient in \mathcal{D} . No other subset containing state f is reachable in \mathcal{D}^R since no transition goes to state f in nfa \mathcal{N} . This gives at most $2^{n-2} + 1$ reachable states. However, language L^R is prefix-free, and so all the accepting states of \mathcal{D}^R accept only the empty word, and can be merged into one state. Hence $\kappa(L^R) \leq 2^{n-3} + 2$.

If $n = 3$ or $n = 4$, then factor-free languages a and aa , respectively, meet the bounds. If $n \geq 5$, then consider the language $L = cKc$, where K is a regular language over the alphabet $\{a, b\}$ with $\kappa(K) = n - 3$ meeting the upper bound 2^{n-3} for reversal [25]. The quotient automaton of L without the empty state is shown in Figure 4. By Proposition 4, language L is factor-free, and $\kappa(L) = n$. Since $\kappa(K^R) = 2^{n-3}$, there exists a set S of 2^{n-3} words over $\{a, b\}$ that define distinct quotients of K^R . Then the quotients of cK^Rc by $2^{n-3} + 2$ words ε , cw with $w \in S$, and cuc for some word u in K^R , are distinct as well. This gives $\kappa(L^R) = 2^{n-3} + 2$. \square

Theorem 6 (Reversal: Subword-Free Languages). *If L is a subword-free language over an alphabet Σ with $\kappa(L) = n$, where $n \geq 4$, then $\kappa(L^R) \leq 2^{n-3} + 2$. The bound is tight if $|\Sigma| \geq 2^{n-3} - 1$, but cannot be met for smaller alphabets. The bound cannot be met if L contains a word of length at least 3.*

Proof. Suppose L is a subword-free language such that $\kappa(L^R) = 2^{n-3} + 2$. Let $\mathcal{D} = (Q, \Sigma, \delta, s, f)$ be the quotient dfa of L with $Q = \{s, q_1, \dots, q_{n-3}, f, e\}$ as the state set, where e and f correspond to the quotients \emptyset and ε . Construct a dfa \mathcal{D}^R for L^R as in the proof of Theorem 5. If $\kappa(L^R) = 2^{n-3} + 2$, then the state $\{q_1, q_2, \dots, q_{n-3}\}$ must be reachable. Therefore there must exist a non-empty word v such that, for all q_i , we have $\delta(q_i, v) = f$. Now suppose there exists a word w in L such that $|w| > 2$. Let $w = abx$ where $a, b \in \Sigma$ and $x \in \Sigma^+$. Also suppose $\delta(s, a) = q_i$ and $\delta(q_i, b) = q_j$. Then we have $av, abv \in L$, showing that L is not subword-free, which is a contradiction. Hence, if any word in L has length at least 3, then $\kappa(L^R) < 2^{n-3} + 2$. Now note that, if all the words in L have length at most 2, the only possible quotients of L^R are $L^R, (L^R)_a$ for all $a \in \Sigma, \varepsilon$, and \emptyset . Therefore $\kappa(L^R) \leq |\Sigma| + 3$, and the second claim follows.

Now consider tightness. If $n = 3$, then the bound is met by the unary subword-free language a . Let $n \geq 4$ and $\ell = 2^{n-3} - 1$. Also let $\Sigma = \{a_1, a_2, \dots, a_\ell\}$, and let S_1, S_2, \dots, S_ℓ be all the non-empty subsets of $\{1, 2, \dots, n-3\}$. Now let

$$L^R = a_1 \left(\bigcup_{j \in S_1} a_j \right) \cup a_2 \left(\bigcup_{j \in S_2} a_j \right) \cup \dots \cup a_\ell \left(\bigcup_{j \in S_\ell} a_j \right).$$

Since L^R only contains two-letter words, languages L^R and L are subword-free. The quotients of L^R are $L^R, (L^R)_{a_i} = \bigcup_{j \in S_i} a_j$ for $i = 1, 2, \dots, \ell, \varepsilon$, and \emptyset . Therefore $\kappa(L^R) = \ell + 3 = 2^{n-3} + 2$. But for L , the only possible and distinct quotients are L, L_{a_i} for $i = 1, 2, \dots, n-3, \varepsilon$, and \emptyset . Thus $\kappa(L) = n$. \square

7 Conclusions

Our results are summarized in Tables 1 and 2, where “B-, F-free” stands for bifix-free and factor-free, and “S-free” for subword-free. The bounds for operations on prefix-free languages are from [12, 14], for operations on suffix-free languages from [10, 13, 15], and those for regular languages, from [17, 18, 27]. For languages over a unary alphabet $\Sigma = \{a\}$, the concepts prefix-, suffix-, factor-, and subword-free coincide, and L is free with $\kappa(L) = n$ if and only if $L = \{a^{n-2}\}$.

In the case of subword-free languages the size of the alphabet cannot be decreased. In the other cases, whenever the size of the alphabet is greater than 2, we do not know whether or not the bounds are tight for smaller alphabets.

The fact that our bounds usually apply only when $m, n \geq 3$ is not a limitation, since bifix-free languages with smaller quotient complexities are simple. For $n = 1$, we have only \emptyset , for $n = 2$, only ε , and for $n = 3$, a subset of Σ . The complexities of operations on such languages can be computed directly.

	$K \cup L, K \oplus L$	$ \Sigma $	$K \cap L$	$ \Sigma $	$K \setminus L$	$ \Sigma $
free unary	$\max(m, n)$		$m = n$		m	
prefix	$mn - 2$	2	$mn - 2(m + n - 3)$	2	$mn - (m + 2n - 4)$	2
suffix	$mn - (m + n - 2)$	2	$mn - 2(m + n - 3)$	2	$mn - (m + 2n - 4)$	2
B-, F-free	$mn - (m + n)$	3	$mn - 3(m + n - 4)$	2	$mn - (2m + 3n - 9)$	2
S-free	$mn - (m + n)$	s_1	$mn - 3(m + n - 4)$	s_2	$mn - (2m + 3n - 9)$	s_3
regular	mn	2	mn	2	mn	2

Table 1: Complexities of boolean operations on free languages; $s_1 = m + n - 3$, $s_2 = m + n - 7$, $s_3 = m + n - 6$.

	KL	$ \Sigma $	L^*	$ \Sigma $	L^R	$ \Sigma $
free unary	$m + n - 2$		n		$2^{n-2} + 1$	
prefix-free	$m + n - 2$	1	n	2	$2^{n-2} + 1$	3
suffix-free	$(m - 1)2^{n-1} + 1$	3	$2^{n-2} + 1$	2	$2^{n-2} + 1$	3
B-, F-free	$m + n - 2$	1	$n - 1$	2	$2^{n-3} + 2$	3
S-free	$m + n - 2$	1	$n - 1$	2	$2^{n-3} + 2$	$2^{n-3} - 1$
regular	$(2m - 1)2^{n-1}$	2	$2^{n-1} + 2^{n-2}$	2	2^n	2

Table 2: Complexities of product, star, and reversal of free languages.

References

- [1] Ang, T., Brzozowski, J.: Languages convex with respect to binary relations, and their closure properties. *Acta Cybernet.* **19**(2) (2009) 445–464
- [2] Bassino, F., Giambruno, L., Nicaud, C.: Complexity of operations on cofinite languages. In López-Ortiz, A., ed.: *Proceedings of the 9th Latin American Theoretical Informatics Symposium, (LATIN)*. Volume 6034 of LNCS, Springer (2010) 222–233
- [3] Berstel, J., Perrin, D., Reutenauer, C.: *Codes and Automata (Encyclopedia of Mathematics and its Applications)*. Cambridge University Press (2010)

- [4] Brzozowski, J.: Derivatives of regular expressions. *J. ACM* **11**(4) (1964) 481–494
- [5] Brzozowski, J.: Quotient complexity of regular languages. In Dassow, J., Pighizzini, G., Truthe, B., eds.: *Proceedings of the 11th International Workshop on Descriptive Complexity of Formal Systems*, Magdeburg, Germany, Otto-von-Guericke-Universität (2009) 25–42. (Extended abstract at <http://arxiv.org/abs/0907.4547>).
- [6] Brzozowski, J., Jirásková, G. and Li, B.: Quotient complexity of ideal languages. In López-Ortiz, A., ed.: *Proceedings of the 9th Latin American Theoretical Informatics Symposium, (LATIN)*. Volume 6034 of LNCS, Springer (2010) 208–211
- [7] Brzozowski, J., Jirásková, G., Li, B., Smith, J.: Quotient complexity of bifix-, factor-, and subword-free languages. <http://arxiv.org/abs/1006.4843v3> (May 2011)
- [8] Brzozowski, J., Jirásková, G., Zou, C.: Quotient complexity of closed languages. In Ablayev, F., Mayr, E.W., eds.: *Proceedings of the 5th International Computer Science Symposium in Russia, (CSR)*. Volume 6072 of LNCS, Springer (2010) 84–95
- [9] Câmpeanu, C., Culik II, K., Salomaa, K., Yu, S.: State complexity of basic operations on finite languages. In Boldt, O., Jürgensen, H., eds.: *Revised Papers from the 4th International Workshop on Automata Implementation, (WIA)*. Volume 2214 of LNCS, Springer (2001) 60–70
- [10] Cmorik, R.: State complexity of basic operations on binary and ternary suffix-free languages. Manuscript (2011)
- [11] Han, Y.S., Salomaa, K.: State complexity of union and intersection of finite languages. *Internat. J. Found. Comput. Sci.* **19**(3) (2008) 581–595
- [12] Han, Y.S., Salomaa, K.: State complexity of basic operations on suffix-free regular languages. *Theoret. Comput. Sci.* **410**(27-29) (2009) 2537–2548
- [13] Han, Y.S., Salomaa, K., Wood, D.: Operational state complexity of prefix-free regular languages. In Ésik, Z., Fülöp, Z., eds.: *Automata, Formal Languages, and Related Topics*, University of Szeged, Hungary (2009) 99–115
- [14] Jirásková, G., Krausová, M.: Complexity in prefix-free regular languages. In McQuillan, I., Pighizzini, G., Trost, B., eds.: *Proceedings of the 12th*

International Workshop on Descriptive Complexity of Formal Systems (DCFS), University of Saskatchewan (2010) 236–244

- [15] Jirásková, G., Olejár, P.: State complexity of union and intersection of binary suffix-free languages. In Bordihn, H., Freund, R., Holzer, M., Kutrib, M., Otto, F., eds.: Proc. of the Workshop on Non-Classical Models for Automata and Applications (NCMA), Austrian Computer Society (2009) 151–166
- [16] Jürgensen, H., Konstantinidis, S.: Codes. In Rozenberg, G., Salomaa, A., eds.: Handbook of Formal Languages, Volume 1: Word, Language, Grammar. Springer (1997) 511–607
- [17] Leiss, E.: Succinct representation of regular languages by boolean automata. Theoret. Comput. Sci. **13** (2009) 323–330
- [18] Maslov, A.N.: Estimates of the number of states of finite automata. Dokl. Akad. Nauk SSSR **194** (1970) 1266–1268 (Russian). English translation: Soviet Math. Dokl. **11** (1970) 1373–1375.
- [19] Mirkin, B.G.: On dual automata. Kibernetika (Kiev) **2** (1966) 7–10 (Russian). English translation: Cybernetics **2** (1966) 6–9.
- [20] Perrin, D.: Finite automata. In van Leewen, J., ed.: Handbook of Theoretical Computer Science. Volume B. Elsevier (1990) 1–57
- [21] Pighizzini, G., Shallit, J.: Unary language operations, state complexity and Jacobsthal’s function. Internat. J. Found. Comput. Sci. **13** (2002) 145–159
- [22] Shyr, H.J.: Free Monoids and Languages. Hon Min Book Co, Taiwan (2001)
- [23] Shyr, H.J., Thierrin, G.: Hypercodes. Inform. and Control **24** (1974) 45–54
- [24] Thierrin, G.: Convex languages. In Nivat, M., ed.: Automata, Languages and Programming. North-Holland (1973) 481–492
- [25] Šebej, J.: Reversal of regular languages and state complexity. In Pardubská, D., ed.: Proc. 10th ITAT, Šafárik University, Košice (2010) 47–54
- [26] Yu, S.: Regular languages. In Rozenberg, G., Salomaa, A., eds.: Handbook of Formal Languages. Volume 1. Springer (1997) 41–110
- [27] Yu, S., Zhuang, Q., Salomaa, K.: The state complexities of some basic operations on regular languages. Theoret. Comput. Sci. **125** (1994) 315–328
- [28] Yu, S.: State complexity of regular languages. J. Autom. Lang. Comb. **6** (2001) 221–234