In Search of Most Complex Regular Languages*

Janusz Brzozowski

David R. Cheriton School of Computer Science, University of Waterloo,
Waterloo, ON, Canada N2L 3G1
brzozo@uwaterloo.ca

Abstract. Regular languages that are most complex under common complexity measures are studied. In particular, certain ternary languages $U_n(a,b,c)$, $n \geqslant 3$, over the alphabet $\{a,b,c\}$ are examined. It is proved that the state complexity bounds that hold for arbitrary regular languages are also met by the languages $U_n(a,b,c)$ for union, intersection, difference, symmetric difference, product (concatenation) and star. Maximal bounds are also met by $U_n(a,b,c)$ for the number of atoms, the quotient complexity of atoms, the size of the syntactic semigroup, reversal, and 22 combined operations, 5 of which require slightly modified versions. The language $U_n(a,b,c,d)$ is an extension of $U_n(a,b,c)$, obtained by adding an identity input to the minimal DFA of $U_n(a,b,c)$. The witness $U_n(a,b,c,d)$ and its modified versions work for 14 more combined operations. Thus $U_n(a,b,c)$ and $U_n(a,b,c,d)$ appear to be universal witnesses for alphabets of size 3 and 4, respectively.

Keywords: combined operation, finite automaton, operation, regular language, state complexity, syntactic semigroup, witness.

I dedicate this work to the memory of Sheng Yu whose extensive research on state complexity led to many questions studied in this paper.

1 Introduction

State complexity is currently an active area of research in the theory of formal languages; for references, see the surveys in [1,30] and the bibliography at the end of this paper. The *state complexity of a regular language* [30] L over a finite alphabet Σ is the number of states in the minimal (complete) deterministic finite automaton (DFA) recognizing the language. An equivalent notion is that of *quotient complexity* [1] of L, which is the number of distinct left quotients of L, where the quotient of $L \subseteq \Sigma^*$ by a word $w \in \Sigma^*$ is the language $w^{-1}L = \{x \in \Sigma^* \mid wx \in L\}$. This paper uses *complexity* for both of these equivalent notions, and this term will not be used for any other property here.

The *(state/quotient) complexity of an operation* on regular languages is the maximal complexity of the language resulting from the operation as a function of

* This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant No. OGP0000871.

N. Moreira and R. Reis (Eds.): CIAA 2012, LNCS 7381, pp. 5–24, 2012.

the complexities of the arguments. For example, for $K, L \subseteq \Sigma^*$, the complexity of the union $K \cup L$ is mn, if the complexities of K and L are m and n, respectively.

There are two parts to the process of establishing the complexity of an operation. First, one must find an *upper bound* on the complexity of the result of the operation by using quotient computations or automaton constructions. Second, one must find *witnesses* that meet this upper bound. One usually defines a sequence $(L_n \mid n \geqslant k)$ of languages, where k is some small positive integer. This sequence will be called a *stream*. The languages in a stream differ only in the parameter n. For example, one might study unary languages $(\{a^n\}^* \mid n \geqslant 1)$ that have zero a's modulo n. A unary operation then takes its argument from a stream $(L_n \mid n \geqslant k)$. For a binary operation, one adds as the second argument a stream $(K_n \mid n \geqslant k)$, usually different from the first. Also, the witness streams are normally different for different operations.

In this paper I pose the question: Is it possible to use the *same* stream for all the operations? In other words, is there a *universal witness?* The answer is "yes" for all of the common operations.

Section 2 describes common conditions that make a language difficult to handle, introduces the main witness stream $(U_n(a,b,c) \mid n \geqslant 3)$ (U for "universal"), and states the main theorem. Properties of a single language, unary operations, and binary operations are discussed in Sections 3–5, respectively. It is shown in Sections 6 and 7 that the bounds for several combined operations are also met by $U_n(a,b,c)$, or by other streams closely related to $U_n(a,b,c)$. Section 8 deals with combined operations that (seem to) require witnesses over four-letter alphabets. The witness $U_n(a,b,c)$ is then extended to $U_n(a,b,c,d)$, where d is an added identity input in the minimal DFA of $U_n(a,b,c)$.

If K and L are regular languages, let $K \cup L$, $K \cap L$, $K \setminus L$ and $K \oplus L$ be their union, intersection, difference, and symmetric difference, let L^R be the reverse of L, and let M be another regular language. Witnesses derived from $U_n(a,b,c)$ and $U_n(a,b,c,d)$ are presented for the following 36 combined operations:

$$
\begin{array}{l}
K \cup L^R,\ K \cap L^R,\ K \setminus L^R,\ K \oplus L^R,\ L^R \setminus K,\\
K^R \cup L^R,\ K^R \cap L^R,\quad KL^R,\ K^RL,\quad (KL)^R,\ (L^R)^*,\\
K \cup L^*,\ K \cap L^*,\ K \setminus L^*,\ K \oplus L^*,\ L^* \setminus K,\\
K^* \cup L^*,\ K^* \cap L^*,\quad KL^*,\ K^*L,\quad (KL)^*,\ (K \cup L)^*,\\
K^2 \cup L^2,\ K^2 \cap L^2,\ K^2 \setminus L^2,\ K^2 \oplus L^2,\\
(KL) \cup M,\ (KL) \cap M,\ (KL) \setminus M,\ (KL) \oplus M,\ M \setminus (KL),\\
(K \cup L)M,\ (K \cap L)M,\ K(L \cup M),\ K(L \cap M) \text{ and } K(L \setminus M).
\end{array}
$$

Section 9 concludes the paper.

2 Conditions for the Complexity of Languages

If a language L_n is most difficult, what properties should it have? Below are some suggestions to help answer this question.

2.1 Properties of a Single Language

Properties that make a single language L_n difficult to handle are discussed first.

A0: The (state/quotient) complexity of $L_n \subseteq \Sigma^*$ should be n. It is assumed that the complexity of the language is fixed at some integer $n \geqslant 1$, and all the other properties are expressed in terms of n.

A1: The complexity of each quotient of L_n should be n. The complexity of each quotient is bounded from above by n, because the DFA $\mathcal{D} = (Q, \Sigma, \delta, q_0, F)$ that defines L_n also defines $w^{-1}L_n$ for any word $w \in \Sigma^*$, if its initial state is changed to $\delta(q_0, w)$. This requirement is easy to meet, since every strongly connected DFA defines a language satisfying this condition.

A2: The number of atoms of L_n should be 2^n. Atoms of regular languages were introduced in 2011 by Brzozowski and Tamm [4], and the theory was slightly modified in [5]. The newer model, which admits up to 2^n atoms, is used here.

An *atom* of a regular language with quotients $K_0, \ldots, K_{n-1}$ is a non-empty intersection of the form $\widetilde{K_0} \cap \cdots \cap \widetilde{K_{n-1}}$, where $\widetilde{K_i}$ is either K_i or $\overline{K_i}$, $\overline{K_i}$ being the complement of K_i with respect to Σ^*. Thus the number of atoms is bounded from above by 2^n, and it was proved in [5] that this bound is tight. Since every quotient of L_n (including L_n itself) is a union of atoms, the atoms of L_n are its basic building blocks. So it is reasonable that L_n should have the maximal number of atoms.

A3: The complexity of each atom of L_n should be maximal. It was shown in [5] that the complexity of the atoms with 0 or n complemented quotients is bounded from above by $2^n - 1$, and the complexity of any atom with r complemented quotients, where $1 \leqslant r \leqslant n-1$, by

$$f(n,r) = 1 + \sum_{k=1}^{r} \sum_{h=k+1}^{n-r+k} C_h^n \cdot C_k^h,$$

where C_j^i is the binomial coefficient i *choose* j. It was also shown in [5] that these bounds are tight. It is reasonable to expect that the building blocks of a language should be as complex as possible.

A4: The syntactic semigroup of L_n should have cardinality n^n. The *Myhill congruence* [22] $\approx_L$ of $L \subseteq \Sigma^*$ is defined as follows: For $x, y \in \Sigma^*$,

$$x \approx_L y \text{ if and only if } uxv \in L \Leftrightarrow uyv \in L \text{ for all } u, v \in \Sigma^*.$$

The *syntactic semigroup* [20,24] of L is the quotient semigroup $\Sigma^+/\approx_L$. It is isomorphic to the *semigroup of transformations* by non-empty words in the minimal DFA of L [20]. The semigroup of transformations is normally used to represent the syntactic semigroup.

Since there are n^n possible transformations of a set of n elements, n^n is an upper bound on the size of the syntactic semigroup of L_n. That the bound is tight follows from the 1935 theorem of Piccard [23] who proved that three transformations of a set of n elements are sufficient to generate all n^n transformations. Also

in 1935, Eilenberg showed that fewer than three generators are not possible [28]. In the context of automata, it was first noted without proof by Maslov [19] in 1970 that n^n is a tight bound.

2.2 Unary Operations

B1: The complexity of the reverse of L_n should be 2^n. It follows from the 1959 subset construction of Rabin and Scott [25] that the upper bound is 2^n. It was first shown by Mirkin [21] in 1966 that this bound can be met.

B2: The complexity of the star of L_n should be $2^{n-1}+2^{n-2}$. It was first noted without proof by Maslov [19] in 1970 that this is a tight upper bound. A proof was provided by Yu, Zhuang and Salomaa [29] in 1994.

2.3 Binary Operations

Two types of binary operations are examined next: boolean operations and product (concatenation or catenation). Four boolean operations union ($\cup$), symmetric difference ($\oplus$), intersection ($\cap$) and difference ($\setminus$) are considered; they are chosen because the complexity of every other binary boolean operation can be obtained from the complexities of these four. Denote by $K_m \circ L_n$ any one of these four operations.

C1: The complexity of $K_m \circ L_n$ should be mn. The upper bound for the boolean operations is mn, since $w^{-1}(K_m \circ L_n) = (w^{-1}K_m) \circ (w^{-1}L_n)$. That the bound is tight for union was noted without proof by Maslov [19] in 1970, and proved for both union and intersection by Yu, Zhuang and Salomaa [29] in 1994. Symmetric difference and difference were treated by Brzozowski [1] in 2010.

C2: The complexity of the product K_mL_n should be $(m-1)2^n+2^{n-1}$. Maslov [19] stated without proof in 1970 that this bound is tight, and Yu, Zhuang and Salomaa [29] provided a proof in 1994.

2.4 The Witness

The language stream that turns out to be the universal witness for all the operations listed above is defined as follows:

Definition 1. *For $n \geqslant 3$, let $\mathcal{U}_n = \mathcal{U}_n(a,b,c) = (Q, \Sigma, \delta, q_0, F)$, where $Q = \{0,\ldots,n-1\}$ is the set of states*[1]*, $\Sigma = \{a,b,c\}$ is the alphabet, $q_0 = 0$ is the initial state, $F = \{n-1\}$ is the set of final states, $\delta(q,a) = q+1 \bmod n$, $\delta(0,b) = 1$, $\delta(1,b) = 0$, $\delta(q,b) = q$ for $q \notin \{0,1\}$, $\delta(n-1,c) = 0$, and $\delta(q,c) = q$ for $q \neq n-1$. Let $U_n = U_n(a,b,c)$ be the language accepted by $\mathcal{U}_n$.*

The structure of the DFA $\mathcal{U}_n(a,b,c)$ is shown in Fig. 1.

[1] Although Q, δ, and F depend on n, this dependence is not shown to keep the notation as simple as possible.

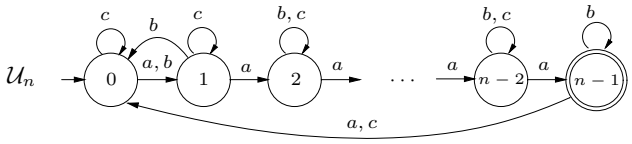

Fig. 1. DFA $\mathcal{U}_n$ of witness language U_n

A language $K \subseteq \Sigma^*$ is *permutationally equivalent* to a language $L \subseteq \Sigma^*$ if K can be obtained from L by permuting the letters of Σ. For example, let π be the permutation $\pi(a) = b$, $\pi(b) = c$ and $\pi(c) = a$; then $\pi(a(b^* \cup cc)) = b(c^* \cup aa)$. Similarly, let $\mathcal{K} = \mathcal{L}(\pi(a), \pi(b), \pi(c))$ be the DFA obtained from $\mathcal{L}(a, b, c)$ by changing the roles of the inputs according to permutation π. Then $\mathcal{K}$ is *permutationally equivalent* to $\mathcal{L}$. In such cases, K ($\mathcal{K}$) is essentially the same language (DFA) as L ($\mathcal{L}$), except that its inputs have been renamed. Obviously, if two languages are permutationally equivalent, then they have the same one-language complexity properties, and the same complexities of unary operations.

Specifically, for this paper let $\mathcal{U}_n(b, a, c)$ be the DFA obtained from $\mathcal{U}_n(a, b, c)$ by interchanging the roles of the inputs a and b. For some operations input c is not needed; then let $\mathcal{U}_n(a, b, \emptyset)$ be the DFA of Definition 1 restricted to inputs a and b, and let $U_n(a, b, \emptyset)$ be the language recognized by this binary DFA. Also, $\mathcal{U}_n(a, \emptyset, \emptyset)$ and $U_n(a, \emptyset, \emptyset)$ are $\mathcal{U}_n(a, b, c)$ and $U_n(a, b, c)$ restricted to a.

Theorem 1 (Main Theorem). *The stream* $(U_n(a, b, c) \mid n \geqslant 3)$ *meets conditions* **A0**–**A4**, **B1,B2** *and* **C2**, *whereas* **C1** *is met by two closely related streams* $(U_m(a, b, c) \mid m \geqslant 3)$ *and* $(U_n(b, a, c) \mid n \geqslant 3)$. *Moreover,*

- **A0** *and* **A1** *are met by* $(U_n(a, \emptyset, \emptyset) \mid n \geqslant 3)$.
- **B2** *is met by* $(U_n(a, b, \emptyset) \mid n \geqslant 3)$.
- **C1** *is met by* $(U_m(a, b, \emptyset) \mid m \geqslant 3)$ *and* $U_n(b, a, \emptyset) \mid n \geqslant 3)$.

3 Properties of a Single Language

Conditions **A0**–**A4** are now briefly discussed for the language U_n.

A0 Complexity of the Language: $U_n(a, \emptyset, \emptyset)$ has n quotients because DFA $\mathcal{U}_n(a, \emptyset, \emptyset)$ is minimal. This holds since state i accepts a^{n-1-i} and no other state accepts this word, for $0 \leqslant i \leqslant n-1$; hence no two states are equivalent.

A1 Complexity of Quotients: Each quotient of $U_n(a, \emptyset, \emptyset)$ has complexity n, since DFA $\mathcal{U}_n(a, \emptyset, \emptyset)$ is strongly connected.

A2 Number of Atoms: It was proved in [5] that U_n has 2^n atoms. This is discussed further below, in connection with **B1 Reversal**.

A3 Complexity of Atoms: The bounds given in the previous section were derived in [5].

Some background is needed before the next property can be discussed. A *transformation* of a set $Q = \{0, \ldots, n-1\}$ is a mapping of Q into itself [11]. If t is a

transformation of Q and $i \in Q$, then it is the *image* of i under t. An arbitrary transformation of Q can be represented by

$$t = \begin{pmatrix} 0 & 1 & \cdots & n-2 & n-1 \\ i_0 & i_1 & \cdots & i_{n-2} & i_{n-1} \end{pmatrix},$$

where $i_k = kt$, $0 \leqslant k \leqslant n-1$, and $i_k \in Q$. The notation $t = [i_0, i_1, \ldots, i_{n-1}]$ is also used for the transformation t above.

A *permutation* of Q is a mapping of Q *onto* itself. For $2 \leqslant k \leqslant n$, a permutation t is a *cycle* of length k, if there exist pairwise different elements $i_1, \ldots, i_k$ such that $i_1 t = i_2, i_2 t = i_3, \ldots, i_{k-1} t = i_k$, and $i_k t = i_1$. A cycle is denoted by $(i_1, i_2, \ldots, i_k)$. A *transposition* is the cycle (i, j) of length 2 that interchanges i and j and does not affect any other elements. A *singular* transformation, denoted by $\binom{i}{j}$, has $it = j$ and $ht = h$ for all $h \neq i$. The *identity* transformation of Q is denoted by $\mathbf{1}_Q$.

The set of all permutations of n elements is isomorphic to the symmetric group of degree n and has $n!$ elements. The following result is due to Piccard [23]:

Theorem 2 (Permutations). *For $n \geqslant 3$, the set of all $n!$ permutations of the set $\{0, \ldots, n-1\}$ is generated by a cycle of length n and a transposition (i, j).*

The set of all transformations of a finite set Q is a semigroup under composition, in fact, a monoid $\mathcal{T}_Q$ of n^n elements. In 1935 Piccard [23] proved that three transformations of Q are sufficient to generate $\mathcal{T}_Q$. Dénes [10] studied more general generators; his formulation is used here:

Theorem 3 (Transformations). *For $n \geqslant 3$, the set of all n^n transformations of the set $\{0, \ldots, n-1\}$ is generated by a cycle of length n, a transposition (i, j), and a singular transformation $\binom{k}{\ell}$.*

Every word w in Σ^+ performs a transformation of the set of states of a DFA defined by $q \to \delta(q, w)$. The set of all such transformations is the semigroup of transformations also called the *transition semigroup* of the DFA [24].

A4 Cardinality of Syntactic Semigroup: By Theorem 3, the syntactic semigroup of $U_n(a, b, c)$ has cardinality n^n, since the transformations performed by inputs a, b and c generate all possible transformations of Q.

4 Unary Operations

B1 Reversal: In 1966 Mirkin [21] used a DFA very similar to $U_n(a, b, c)$ to meet the 2^n bound for reversal. It is defined by inputs $a : (0, 1, \ldots, n-1)$, $b : (0, n-2)$ and $c : \binom{0}{n-1}$, with initial state 0 and final state 0. The syntactic semigroup of the language of this DFA has size n^n. Another similar DFA with inputs $a : (0, 1, \ldots, n-1)$, $b : (0, 1)$ and $c : \binom{0}{n-1}$ and initial state 0 and final state 0 was used by Leiss [18] in 1981; the semigroup is also of size n^n.

Salomaa, Wood, and Yu [27] showed the following result:

Theorem 4 (Transformations and Reversal). *Let $\mathcal{D}$ be a minimal DFA with n states accepting a language L. If the transformation semigroup of $\mathcal{D}$ has n^n elements, then the quotient complexity of L^R is 2^n.*

From this and **A4** it follows that U_n^R has 2^n quotients. In view of the following result proved by Brzozowski and Tamm [5], U_n has 2^n atoms.

Theorem 5 (Atoms). *The number of atoms of a regular language L is equal to the complexity of L^R.*

There are also binary witnesses that reach the bound 2^n for L^R. For a detailed discussion see the recent paper by Jirásková and Šebej [17]. Their witness has input a that performs the cycles $(0,1,2)$ and $(3,4,\ldots,n-1)$, and input b that has the cycles $(0,1)$ and $(2,3)$ and is an identity on the remaining states.

Is there a close relation between the size of the syntactic semigroup and the quotient complexity of reversal? Besides the result for regular languages in Theorem 4, there are other examples where the languages that have maximal syntactic semigroups also meet the maximal bound for reversal. This is the case for right ideals [6] and prefix-free languages [3]. For left and two-sided ideals [6] and for suffix-, bifix-, and factor-free languages [3] there are only conjectured upper bounds on the size of the syntactic semigroup, but the languages that meet these bounds also meet the maximal bounds for reversal.

The witness of Jirásková and Šebej [17] shows that it is possible for a language to reach the bound 2^n for reversal without having syntactic complexity of n^n. It is also possible for a language to have the maximal syntactic complexity for its class and not reach the bound for reversal. For example, it was shown by Brzozowski and Li [2] that the star-free language defined by the 3-state DFA with inputs $a : [0,0,1]$, $b : [1,1,2]$, $c : [0,2,2]$ and $d : [0,1,2]$ and final state 0 meets the maximal bound 10 for the size of the syntactic semigroup. But it does not meet the conjectured upper bound 7 for reversal. However, that bound is met if the final states are 0 and 2.

Does there always exists a language that meets both bounds?

B2 Star: The tightness of the bound for star is now proved. The language $(U_n(a,b,\emptyset))^*$ is accepted by the ε-NFA $\mathcal{S}_n = (Q_{\mathcal{S}}, \{a,b\}, \delta_{\mathcal{S}}, \{s\}, \{s, n-1\})$, where $Q_{\mathcal{S}} = Q \cup \{s\}$, $s \notin Q$, $\delta_{\mathcal{S}}(s,a) = \delta_{\mathcal{S}}(s,b) = \{1\}$, $\delta_{\mathcal{S}}(q,x) = \{\delta(q,x)\}$ for all $q \in Q$, $x \in \Sigma$, and $\delta_{\mathcal{S}}(n-1,\varepsilon) = \{0\}$. The ε-NFA S_4 is shown in Figure 2.

Theorem 6 (Star). *For $n \geqslant 3$, the complexity of $(U_n(a,b,\emptyset))^*$ is $2^{n-1}+2^{n-2}$.*

Proof. It will be proved that $\{s\}$, all 2^{n-1} subsets of Q containing 0, and all $2^{n-2}-1$ non-empty subsets of $\{1,\ldots,n-2\}$ are reachable and pairwise distinguishable, giving the DFA of $(U_n(a,b,\emptyset))^*$ a total of $2^{n-1}+2^{n-2}$ states.

Since s is the initial state, $\{s\}$ is reachable by ε, and $\{0\}$ by ab. It will be shown how to reach the remaining sets from $\{0\}$. Note that any subset containing $n-1$ must also contain 0.

First it is proved that all 2^{n-1} subsets of Q containing 0 are reachable. Since

$$\{0\} \xrightarrow{a^{n-1}} \{0,n-1\} \xrightarrow{a} \{0,1\} \xrightarrow{(ab)^{i-1}} \{0,i\},$$

for $2 \leqslant i \leqslant n-2$, all two-element subsets of Q containing 0 are reachable.

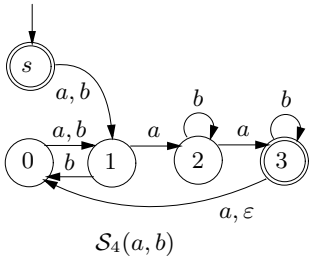

Fig. 2. NFA for $(U_4(a,b,\emptyset))^*$

For $k \geqslant 2$, if any k-element set containing 0 can be reached, then so can be any $(k+1)$-element set containing 0 and $n-1$, for if $i_1 < i_2 < \cdots < i_k$, then

$$\{0, i_2 - i_1, \ldots, i_{k-1} - i_1, n-1-i_1\} \xrightarrow{a^{i_1}} \{0, i_1, i_2, \ldots, i_{k-1}, n-1\}.$$

For $k \geqslant 3$, if any k-element set containing 0 and $n-1$ can be reached, then so can be any k-element set containing 0. This holds because

$$\{0, i_2 - i_1, \ldots, i_{k-1} - i_1, n-1\} \xrightarrow{a(ab)^{i_1-1}} \{0, i_1, \ldots, i_{k-1}\}.$$

It follows now that all 2^{n-1} subsets of Q containing 0 are reachable. Since also

$$\{0, i_2 - i_1, \ldots, i_k - i_1\} \xrightarrow{a^{i_1-1}} \{i_1, i_2, \ldots, i_k\},$$

all the $2^{n-2} - 1$ non-empty subsets of $\{1, \ldots, n-2\}$ are reachable.

It remains to prove that all subsets are pairwise distinguishable. Set $\{s\}$ and any subset of Q containing $n-1$ differ from any subset of Q not containing $n-1$, because they accept the empty word. Also, $\{s\}$ differs from any subset of Q containing $n-1$, because the latter accepts b. Finally, if set P contains $0 \leqslant i < n-1$ but set R does not, then P accepts a^{n-1-i}, and R does not. □

Since the required number of subsets can be reached by words in $\{a,b\}^*$, and they are pairwise distinguishable by words in $\{a,b\}^*$, it follows that the complexity of $(U_n(a,b,c))^*$ with the added input c is also $2^{n-1} + 2^{n-2}$.

Discussion: For $n = 1$, there are only two languages, $\emptyset$ and Σ^*. The complexity of $\emptyset^* = \varepsilon$ is 2, and that of $(\Sigma^*)^* = \Sigma^*$ is 1; the bound does not apply here.
For $n = 2$, the language of Definition 1 is well defined, but inputs a and b coincide. The star of U_2 has complexity 2 only; hence $U_2(a,\emptyset,c)$ is not most complex here. However, the bound $2^1 + 2^0 = 3$ is met by the language over $\{a,b\}$ of all the words with an odd number of a's [29].

5 Binary Operations

5.1 Boolean Operations

Since $K_n \cup K_n = K_n \cap K_n = K_n$, and $K_n \setminus K_n = K_n \oplus K_n = \emptyset$, two different languages have to be used to reach the bound mn if $m = n$. Figure 3 shows the DFA $\mathcal{U}_4(a, b, \emptyset)$ and the DFA $\mathcal{U}_5(b, a, \emptyset)$ permutationally equivalent to $\mathcal{U}_5(a, b, \emptyset)$. The direct product of $\mathcal{U}_4(a, b, \emptyset)$ and $\mathcal{U}_5(b, a, \emptyset)$ is in Figure 4.

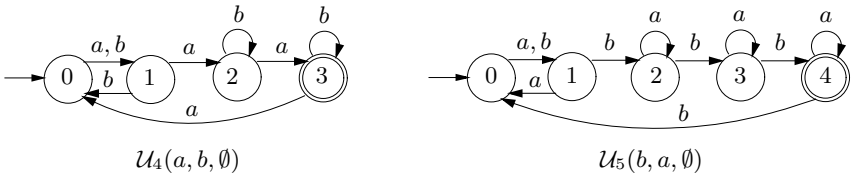

Fig. 3. DFA's of $U_4(a, b, \emptyset)$ and $U_5(b, a, \emptyset)$

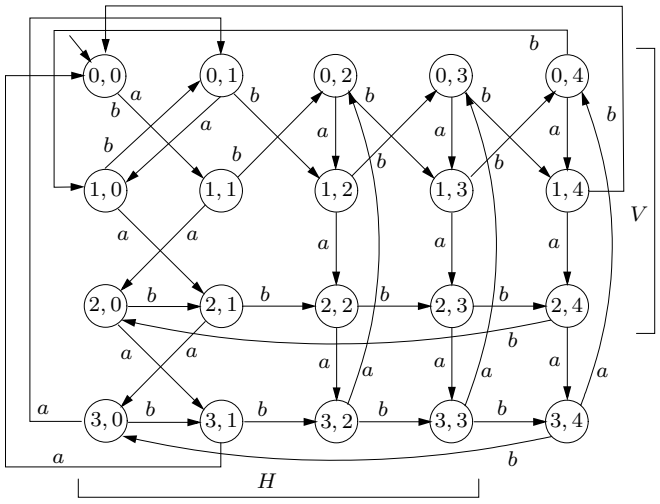

Fig. 4. Direct product of $\mathcal{U}_4(a, b, \emptyset)$ with $\mathcal{U}_5(b, a, \emptyset)$

Theorem 7 (Boolean Operations). *The complexity of* $U_m(a, b, \emptyset) \circ U_n(b, a, \emptyset)$ *is* mn *for* $m, n \geqslant 3$.

Proof. In the direct product, state $(0, 0)$ is the initial state, and state $(1, 1)$ is reached by a. From $(1, 1)$, state $(2, 0)$ is reached by a, and state $(0, 2)$, by b. From $(0, 2)$, state $(i, 2)$ is reached by a^i; hence all the states in column 2 are reachable. From $(1, 2)$, state $(0, 3)$ is reached by b and the states in column 3 are reached by words in a^*. This repeats until state $(0, n-1)$ is reached by b from $(1, n-2)$,

and other states in column $n-1$ are then reached by words in a^*. Thus all the states in columns $2, \ldots, n-1$ are reachable.

Next, $(1,0)$ is reached from $(0,n-1)$ by b, and $(0,1)$ from $(1,0)$ by b. The remaining states $(i,0)$, $i \geqslant 2$, are reached from $(1,1)$ by $a(ba)^{i-2}$ and states $(i,1)$, by $(ab)^{i-1}$. Thus all the states in columns 0 and 1 are also reachable.

It remains to prove that all the states are pairwise distinguishable. Let H (for *horizontal*) be the set $H = \{(m-1,0),(m-1,1),\ldots,(m-1,n-2)\}$, and let V (for *vertical*) be $V = \{(0,n-1),(1,n-1),\ldots,(m-2,n-1)\}$. The boolean operations are now considered one by one.

Union: The final states are $H \cup V \cup \{(m-1,n-1)\}$.

Consider the final states in $V' = V \cup \{(m-1,n-1)\}$. First, $(0,n-1)$ goes to $(m-1, m \bmod 2)$ by ba^{m-2}, and all other states in V' go to non-final states. Second, $(1,n-1)$ goes to $(m-1,(m-1) \bmod 2)$ by ba^{m-1}, and all other states in V' reject this word. For $i > 1$, $(i,n-1)$ goes to $(m-1,(m-1-i) \bmod 2)$ by ba^{m-1-i}, and all other states in V' reject this word. Thus all the states in column $n-1$ are distinguishable.

Next, take the final states in $H' = H \cup \{(m-1,n-1)\}$. By an argument symmetric to the one above, interchanging m and n and a and b, one concludes that all the states in row $m-1$ are distinguishable.

Each state in V accepts a but not b, each state in H accepts b but not a, and $(m-1,n-1)$ accepts both. Hence every state in V is distinguishable from every state in H, and all these states are distinguishable from $(m-1,n-1)$. Therefore all final states are pairwise distinguishable.

Any non-final state (i,j) accepts a^{m-1-i} and b^{n-1-j}, but no other non-final state accepts both of these words. So all non-final states are also distinguishable.

Symmetric Difference: The final states are those in $H \cup V$.

The final states are all distinguishable by the argument used for union. The non-final states other than $(m-1,n-1)$ are distinguishable by the same words as for union. State $(m-1,n-1)$ accepts both ab^n and ba^m, and no state other than $(m-2,n-2)$ accepts both of these words. But $(m-1,n-1)$ rejects aba, while $(m-2,n-2)$ accepts it. So all non-final states are also distinguishable.

Intersection: For intersection, there is only one final state $(m-1,n-1)$. The non-final states q and words w_q accepted only by those states are listed below:

1. $q = (0,j)$ with $n-1-j$ even, $w_q = b^{n-1-j}a^{m-1}$,
2. $q = (0,j)$ with $n-1-j$ odd, $w_q = b^{n-1-j}a^{m-2}$,
3. $q = (1,j)$ with $n-1-j$ even, $w_q = b^{n-1-j}a^{m-2}$,
4. $q = (1,j)$ with $n-1-j$ odd, $w_q = b^{n-1-j}a^{m-1}$,
5. for $i \geqslant 2$, $q = (i,j)$, $w_q = b^{n-1-j}a^{m-1-i}$.

Difference: For difference, the final states are H. State $(m-1,j)$ rejects b^{n-1-j}, but other final states accept it. So all final states are distinguishable.

For non-final states $q = (i, j)$ and (h, l), other than $(m-1, n-1)$:

1. If $i = h$ and $j \neq l$, then q rejects w_q, while (h, l) accepts it, where w_q is defined as for intersection. Thus all the non-final states in the same row are distinguishable.
2. Two non-final states $q = (i, j)$ and (h, j) in the same column, with $j < n-1$ and $i \neq h$, are distinguishable by a^{m-1-i}.
3. If $j = l = n-1$ and $i \neq h$, then q accepts $a^{m-1-i}b$, while (h, l) rejects it.

Any non-final state (i, j) with $j < n-1$ is distinguished from $(m-1, n-1)$ by a^{m-1-i}. State $(0, n-1)$ accepts ba^{m-2}, while $(m-1, n-1)$ rejects it. Similarly, $(1, n-1)$ accepts ba^{m-1}, while $(m-1, n-1)$ rejects it. For $2 \leqslant i \leqslant n-2$, $(i, n-1)$ accepts ba^{m-1-i}, but $(m-1, n-1)$ rejects it. □

Although it is impossible for the stream $(U_n(a, b, \emptyset), n \geqslant 3)$ to meet the bound for boolean operations when $m = n$, this stream is as complex as it could possibly be in view of the following:

Conjecture 1 ($K_m \circ L_n, m \neq n$)
If $m \neq n$, the complexity of $U_m(a, b, \emptyset) \circ U_n(a, b, \emptyset)$ is mn.
(Verified for $3 \leqslant m, n \leqslant 10$ and some higher values.)

Note about Conjectures: The 19 conjectures in this paper have 35 different claims. The proofs are not trivial; sometimes two such proofs constitute an entire paper, for example, in [7,9,16,26]. Because of the limitations of time, space, and the author's energy, the proofs are omitted, although some of the claims have been verified. The conjectures are supported by Grail and GAP computations.

5.2 Product

It is shown next that the complexity of the product of $U_m(a, b, c)$ with $U_n(a, b, c)$ reaches the maximal possible bound.

To avoid confusion of states, let $\mathcal{U}_m = \mathcal{U}_m(a, b, c) = (Q_m, \Sigma, \delta_m, q_0, \{q_{m-1}\})$, where $Q_m = \{q_0, \ldots, q_{m-1}\}$, and let $\mathcal{U}_n = \mathcal{U}_n(a, b, c)$, as in Definition 1. Define the ε-NFA $\mathcal{P} = (Q_m \cup Q_n, \Sigma, \delta_{\mathcal{P}}, \{q_0\}, \{n-1\})$, where $\delta_{\mathcal{P}}(q, a) = \{\delta_m(q, a)\}$ if $q \in Q_m$, $a \in \Sigma$, $\delta_{\mathcal{P}}(q, a) = \{\delta_n(q, a)\}$ if $q \in Q_n$, $a \in \Sigma$, and $\delta_{\mathcal{P}}(q_{m-1}, \varepsilon) = \{0\}$. This ε-NFA accepts $U_m U_n$, and is illustrated in Figure 5 for $m = 4$ and $n = 5$.

Theorem 8 (Product). *For $m, n \geqslant 2$, the complexity of $U_m(a, b, c)U_n(a, b, c)$ is $(m-1)2^n + 2^{n-1}$.*

Proof. It will be shown that all $(m-1)2^n$ subsets of states of $\mathcal{P}$ of the form $\{q_i\} \cup S$, where $i < m-1$ and S is any subset of Q_n, are reachable, as well as all 2^{n-1} subsets of the form $\{q_{m-1}, 0\} \cup S$, where S is any subset of $\{1, \ldots, n-1\}$. *All the arithmetic below is modulo n.*

First, study how states of the form $\{q_0\} \cup S$ can be reached. Since $\{q_0\}$ is the initial set of states, it is reached by ε. Sets $\{q_i\}$ are reached from $\{q_0\}$ by a^i, for $i = 1, \ldots, m-2$, and $\{q_{m-1}, 0\}$, by a^{m-1}.

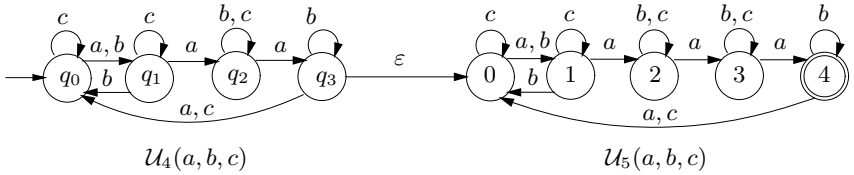

Fig. 5. ε-NFA $\mathcal{P}$ of $U_4(a,b,c)U_5(a,b,c)$

From $\{q_{m-1},0\}$, $\{q_0,0\}$ is reached by c, and $\{q_0,1\}$ by a. From $\{q_0,1\}$, $\{q_0,i\}$ is reached by $(ab)^{i-1}$, for $i=2,\ldots,n-1$. Hence all the sets of the form $\{q_0\}\cup S$, where $|S|\leqslant 1$ are reachable.

Second, it will be shown that, if $\{q_{m-1},0\}\cup S$ can be reached for all sets $S\subseteq\{1,\ldots,n-1\}$ with $|S|=k\geqslant 0$, then $\{q_0\}\cup T$ can be reached for all $T=\{t_0,t_1,\ldots,t_k\}\subseteq\{0,\ldots,n-1\}$ with $0\leqslant t_0<t_1<\cdots<t_k\leqslant n-1$. There are three cases to consider:

1. $t_0=0$: Use $\{q_{m-1},0,t_2-t_1,\ldots,t_k-t_1,n-1\}\stackrel{a(ab)^{t_1-1}}{\longrightarrow}\{q_0,t_1,t_2,\ldots,t_k,0\}$.
2. $t_0=1$: Use $\{q_{m-1},0,t_1-1,\ldots,t_k-1\}\stackrel{a}{\longrightarrow}\{q_0,1,t_1,\ldots,t_k\}$.
3. $t_0>1$: Use $\{q_{m-1},0,t_1-(t_0-1),\ldots,t_k-(t_0-1)\}\stackrel{bc(ab)^{t_0-1}}{\longrightarrow}\{q_0,t_0,t_1,\ldots,t_k\}$.

Third, consider sets $\{q_{m-1},0\}\cup S$, $S\subseteq\{1,\ldots,n-1\}$. It has already been shown that $\{q_{m-1},0\}$ is reachable. Suppose that all the sets of the form $\{q_0\}\cup S$ with $|S|=k\geqslant 1$, $0\notin S$ can be reached. Then to reach $\{q_{m-1},0,t_1,\ldots,t_k\}$ with $1\leqslant t_1<\cdots<t_k\leqslant n-1$, use $\{q_0,t_1-(m-1),\ldots,t_k-(m-1)\}\stackrel{a^{m-1}}{\longrightarrow}\{q_{m-1},0,t_1,\ldots,t_k\}$.

Finally, for $0<i<m-1$, $\{q_i,t_1,\ldots,t_k\}$ is reached by a^i from $\{q_0,t_1-i,\ldots,t_k-i\}$, where $0\leqslant t_1<\cdots<t_k\leqslant n-1$. Hence all the required states can be reached.

It will now be proved that all these subsets are pairwise distinguishable.

Consider $s=\{q_i\}\cup S$ and $t=\{q_j\}\cup T$, where $0\leqslant i,j\leqslant m-1$ and $S\neq T$, $S,T\subseteq Q_n$. If k is in $S\oplus T$, then a^{n-1-k} distinguishes s and t.

Next suppose $s=\{q_i\}\cup S$ and $t=\{q_j\}\cup S$ with $i<j<m-1$. Applying $(ca)^{m-1-j}$ sends $t=\{q_j\}\cup S$ to $t'=\{q_{m-1},0\}\cup S'$ for some $S'\subseteq\{1,\ldots,n-1\}$, but sends $s=\{q_i\}\cup S$ to $s'=\{q_{i+m-1-j}\}\cup S'$, and this pair can be distinguished since the subsets of Q_n are different. If $i>0$ and $j=m-1$, apply $(ca)^{m-1-i}$. Then $s=\{q_i\}\cup S$ is sent to $s'=\{q_{m-1},0\}\cup S'$, and $t=\{q_{m-1}\}\cup S$ is sent to $t'=\{q_k\}\cup S'$ for some $S'\subseteq\{1,\ldots,n-1\}$ and $k<m-1$.

This leaves the case where $i=0$ and $j=m-1$. Then use ba to send $t=\{q_j\}\cup S$ to $t'=\{q_0\}\cup S'$ and $s=\{q_i\}\cup S$ to $s'=\{q_2\}\cup S'$. Now $(ca)^{m-3}$ can be applied to make the subsets of Q_n different.

Since all reachable sets are pairwise distinguishable, the bound is met. □

Discussion. The restrictions of U_n to two letters do not meet the bound for product. This is a defect of U_n, since there exist binary witnesses for product. Maslov [19] used the DFA $\mathcal{K}_m$ with input a performing the cycle $(0, \ldots, m-1)$, with b being the identity, and a DFA $\mathcal{L}_n$ with a performing the transposition $(n-2, n-1)$, and b mapping i to $i+1$ for $i < n-1$, and $n-1$ to $n-1$. Yu, Zhuang and Salomaa [29] used ternary languages.

6 Combined Operations with $U_m(a, b, c)$ and $U_n(b, a, c)$

To simplify the notation, denote $U_n(b, a, c)$ by $\tilde{U}_n$.

Gao and Yu [15] studied the complexities of $K_m \cup L_n^R$ and $K_m \cap L_n^R$, and showed that they are both $m2^n - (m-1)$, and are met using a quaternary alphabet. Their results can be improved and extended as follows: (1) *ternary alphabets* suffice, (2) the *same language stream* can be used for K_m and L_n for both union and intersection, (2) the same language stream is also a witness for two *difference* operations and *symmetric difference*, and (4) the bound for symmetric difference is $m2^n$.

Conjecture 2 ($K_m \circ L_n^R$ **and** $L_n^R \setminus K_m$)
For $m, n \geqslant 3$, the complexities of $U_m \cup U_n^R$, $U_m \cap U_n^R$, $U_m \setminus U_n^R$, $U_n^R \setminus U_m$ are all $m2^n - (m-1)$, whereas that of $U_m \oplus U_n^R$ is $m2^n$. (Verified for $3 \leqslant m, n \leqslant 10$.)

It was shown in [12] by Gao, Kari, and Yu that the complexities of $K_m^R \cup L_n^R$ and $K_m^R \cap L_n^R$ are $(2^m-1)(2^n-1)+1$ with witnesses over a six-letter alphabet. The bound can also be met by ternary languages:

Conjecture 3 ($K_m^R \cup L_n^R$ **and** $K_m^R \cap L_n^R$)
For $m, n \geqslant 3$, the complexities of $U_m^R \cup \tilde{U}_n^R$ and $U_m^R \cap \tilde{U}_n^R$ are $(2^m-1)(2^n-1)+1$. (Verified for $3 \leqslant m, n \leqslant 7$.)

Incidentally, the same bound can be reached by these witnesses for difference.

It was shown in [9] by Cui, Gao, Kari, and Yu that the complexity of $K_m L_n^R$ is $(m-1)2^n + 2^{n-1} - (m-1)$ with ternary witnesses. Two permutationally equivalent witnesses also work:

Conjecture 4 ($K_m L_n^R$)
For $m, n \geqslant 3$, the complexity of $U_m U_n^R$ is $(m-1)2^n + 2^{n-1} - (m-1)$.
(Verified for $3 \leqslant m \leqslant 7$ and $3 \leqslant n \leqslant 6$.)

If the complexities of K_m, L_n and M_p are m, n and p, Cui, Gau, Kari, and Yu [8] showed that the complexity of $(K_m L_n) \cap M_p$ is $((m-1)2^n + 2^{n-1})p$; here the complexity of the result is the composition of the complexities of product and intersection. They also showed that the same bound holds for $(K_m L_n) \cup M_p$. This can be generalized to all binary boolean operations:

Conjecture 5 ($(K_m L_n) \circ M_p$ **and** $M_p \setminus (K_m L_n)$)
The complexities of $(U_m U_n) \circ \tilde{U}_p$ and $\tilde{U}_p \setminus (U_m U_n)$ are all $((m-1)2^n + 2^{n-1})p$ for $m, n, p \geqslant 3$. (Verified for various values of m, n, and p.)

Cui, Gao, Kari, and Yu [8] also proved that the complexity of $(K_m \cap L_n)M_p$ is the composition of complexities of intersection and product, $(mn-1)2^p + 2^{p-1}$. They used quaternary witnesses, but there are ternary witnesses:

Conjecture 6 ($(K_m \cap L_n)M_p$)
The complexity of $(U_m \cap \tilde{U}_n)U_p$ is $(mn-1)2^p + 2^{p-1}$ for $m, n, p \geqslant 3$.
(Verified for various values of m, n, and p.)

In the case of $(K_m \cap L_n)\Sigma^*$, the languages $K_m = U_m(a, b, \emptyset)$ and $L_n = U_n(b, a, \emptyset)$ also reach the bound mn.

Gao and Yu [15] showed that the complexity of $K_m \cup L_n^*$ is $3m2^{n-2} - (m-1)$. These results are extended here to symmetric difference and to one difference operation. For the remaining boolean operations see Conjecture 9.

Conjecture 7 ($K_m \cup L_n^*$, $K_m \oplus L_n^*$, and $L_n^* \setminus K_m$)
The complexities of the operations $U_m \cup \tilde{U}_n^*$, $U_m \oplus \tilde{U}_n^*$ and $\tilde{U}_n^* \setminus U_m$ are all $3m2^{n-2} - (m-1)$ for $m, n \geqslant 3$. (Verified for $3 \leqslant m, n \leqslant 10$.)

7 Combined Operations with "Dialects" of $U_n(a, b, c)$

For the combined operations in this section, the witness $U_n(a, b, c)$ no longer works. However, the class of witnesses can be extended beyond those permutationally equivalent to $U_n(a, b, c)$.

Definition 2. *A* dialect *of the language $U_n(a, b, c)$ is any ternary language $V_n(a, b, c)$ of complexity n, in which one input $a : (0, \ldots, n-1)$ performs a cyclic permutation of the n states in the minimal DFA of V_n, a second input $b : (i, j)$ performs a transposition of two states, and the third input is a singular transformation $c : \binom{k}{\ell}$.*

Clearly, a dialect $V_n(a, b, c)$ of $U_n(a, b, c)$ satisfies **A0** and **A1**. By Theorem 3, its syntactic semigroup is of size n^n, and so it satisfies **A4**. By Theorem 5, it satisfies **A2**. I conjecture that it satisfies **A3**. By Theorem 4, it satisfies **B1**. So it seems that dialects have many desirable single-language properties.

A Combined Operation with Binary Witnesses: In 2007 A. Salomaa, K. Salomaa, and S. Yu [26] showed that the complexity of $(K_m \cup L_n)^*$ is $2^{m+n-1} - (2^{m-1} + 2^{n-1} - 1)$ with ternary witnesses. Jirásková and Okhotin [16] used binary witnesses. It is shown below that permutationally equivalent binary dialects of $U_n(a, b, c)$ can also be used.

Let $\mathcal{S}_n = \mathcal{S}_n(a, b) = (Q, \Sigma, \delta_S, 0, \{0\})$, where $a : (0, 1, \ldots, n-1)$, and $b : \binom{0}{1}$. In this dialect, both the final state and the singular transformation have been changed. Let S_n be the language of $\mathcal{S}_n$, and let $\tilde{S}_n = S_n(b, a)$.

Theorem 9 ($(K_m \cup L_n)^*$)
For $m, n \geqslant 3$, the complexity of $(S_m \cup \tilde{S}_n)^*$ is $2^{m+n-1} - (2^{m-1} + 2^{n-1} - 1)$.

Proof The proof follows closely that of [16]. The DFA's of languages $S_4(a, b)$ and $S_5(b, a)$ are shown in Fig. 6. If the dotted transitions are added, the resulting

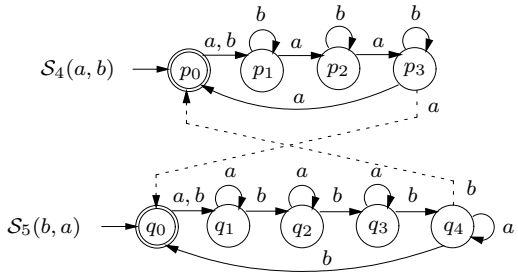

Fig. 6. DFA's of $S_4(a, b, \emptyset)$ and $S_5(b, a, \emptyset)$

NFA $\mathcal{N} = (Q_N, \Sigma, \delta_N, \{p_0, q_0\}, \{p_0, q_0\})$ in the general case accepts $(S_m \cup \tilde{S}_n)^*$. Note that all the reachable subsets of Q_N always contain at least one state from $\mathcal{S}_m(a, b)$ and at least one state from $\mathcal{S}_n(b, a)$. The reachable sets are (a) those consisting of a non-empty subset of $P \setminus \{p_0\}$ together with a non-empty subset of $Q \setminus \{q_0\}$ and (b), those consisting of $\{p_0, q_0\}$ together with any subset of $P \cup Q$. The number of reachable states is precisely the bound.

Any subset of cardinality 2 is reached as follows: $\{p_0, q_0\}$ is the initial state, and $\{p_i, q_j\}$, for $1 \leqslant i, j$, is reached by $ba^{i-1}b^{j-1}$. Next, use induction on the size of the set in several cases. To reach $\{p_{i_1}, \ldots, p_{i_k}, q_{j_1}, \ldots, q_{j_l}\}$, proceed as follows:

1. If $i_1 = j_1 = 0$ and ($i_2 > 1$ or $k = 1$), start with the set (of size $k + l - 1$) $\{p_{m-1}, p_{i_2-1} \ldots, p_{i_k-1}, q_{j_2}, \ldots, q_{j_l}\}$ and apply a.
2. If $i_1 = j_1 = 0$ and ($j_2 > 1$ or $l = 1$), this is symmetric to Case 1.
3. If $i_1 = j_1 = 0$ and $i_2 = j_2 = 1$, start with the set (of size $k + l - 1$) $\{p_{m-1}, p_0, p_{i_3-1}, \ldots, p_{i_k-1}, q_0, q_{j_3}, \ldots, q_{j_l}\}$ and apply a.
4. If $i_1 \geqslant 1$ and $j_1 \geqslant 1$, start with set S of size $k + l$, reachable by 1–3, where $S = \{p_0, p_{i_2-i_1}, \ldots, p_{i_k-i_1}, q_0, q_{j_2-(j_1-1)}, \ldots, q_{j_l-(j_1-1)}\}$, and apply $a^{i_1}b^{j_1-1}$.

Hence all the required states are reachable. Since only state p_i accepts a^{m-i}, and only p_j accepts b^{n-j}, all subsets are pairwise distinguishable. □

Incidentally, $(S_m \oplus \tilde{S}_n)^*$ also reaches the bound $2^{m+n-1} - (2^{m-1} + 2^{n-1} - 1)$.

The star of the reverse was studied by Gao, Salomaa, and Yu [14], who showed that the complexity of this operation is 2^n with a ternary witness. Here, a dialect $\mathcal{U}_{\{0\},n}(a, b, c)$ of $\mathcal{U}_n(a, b, c)$ with final state changed to 0 can be used:

Conjecture 8 ($(L_n^R)^*$)
For $n \geqslant 3$, the complexity of $(U_{\{0\},n}^R)^*$ is 2^n. (Verified for $3 \leqslant n \leqslant 7$.)

Gao and Yu [15] studied the intersection $K_m \cap L_n^*$; this result is extended here to a difference operation. The dialect which is the complement of $U_m(a, b, c)$ applies here.

Conjecture 9 ($K_m \cap L_n^*$ **and** $K_m \setminus L_n^*$)
For $m, n \geqslant 3$, the complexities of $\overline{U_m} \cap \tilde{U}_n^*$ and $\overline{U_m} \setminus \tilde{U}_n^*$ are both $3m2^{n-2} - (m-1)$. (Verified for $3 \leqslant m, n \leqslant 10$.)

The language $K_m L_n^*$ was studied by Cui, Gao, Kari, and Yu [9]. If the only final quotient of L_n is L_n itself, then $L_n = L_n^*$. The complexity of $K_m L_n^*$ is then that of $K_m L_n$; by Theorem 8, $U_m(a,b,c)$ and $U_n(a,b,c)$ meet this bound. Hence assume that there is at least one final quotient of L_n other than L_n. In that case, it was proved in [9] that the quotient complexity of $K_m L_n^*$ is at most $(3m-1)2^{n-2}$, and that this bound is tight with ternary witnesses.

Here one can use a dialect of $\mathcal{U}_n(a,b,c)$ with a different singular transformation. Let $\mathcal{T}_n = \mathcal{T}_n(a,b,c) = (Q, \Sigma, \delta_T, 0, \{n-1\})$, where $Q = \{0,\ldots,n-1\}$, $\Sigma = \{a,b,c\}$, $a : (0,1,\ldots,n-1)$, $b : (0,1)$, and $c : \binom{1}{0}$. Let T_n be the language of $\mathcal{T}_n$ and let $\tilde{T}_n(a,b,c) = T_n(b,a,c)$.

Conjecture 10 $(K_m L_n^*)$
For $m,n \geqslant 3$, the complexity of $T_m \tilde{T}_n^*$ is $(3m-1)2^{n-2}$.
(Verified for $3 \leqslant m,n \leqslant 6$.)

8 Witnesses Over Quaternary Alphabets

Operations that (appear to) require an alphabet of four letters are treated next.

8.1 Witnesses $U_n(a,b,c,d)$ and $\widehat{U}_n(a,b,c,d)$

Let $\mathcal{U}_n(a,b,c,d) = (Q, \Sigma, \delta_U, 0, \{n-1\})$, where $a : (0,1,\ldots,n-1)$, $b : (0,1)$, $c : \binom{n-1}{0}$, and $d : \mathbf{1}_Q$. Thus $\mathcal{U}_n(a,b,c) = \mathcal{U}_n(a,b,c,\emptyset)$. Let $U_n(a,b,c,d)$ be the language of $\mathcal{U}_n(a,b,c,d)$. Also let $\widehat{\mathcal{U}}_n(a,b,c,d) = \mathcal{U}_n(d,c,b,a)$; then $\widehat{\mathcal{U}}_n(a,b,c,d)$ and $\mathcal{U}_n(a,b,c,d)$ are permutationally equivalent.

From now on, the symbols $\mathcal{U}_n$ and U_n stand for $\mathcal{U}_n(a,b,c,d)$ and $U_n(a,b,c,d)$, and the same applies to the versions with the "hat".

It was shown by Cui, Gao, Kari, and Yu [8] that quaternary witnesses meet the bound $3 \cdot 2^{m+n-2} - 2^n + 1$ for $(K_m L_n)^R$. Here U_n and $\widehat{U}_n$ also work:

Conjecture 11 $((K_m L_n)^R)$
For $m,n \geqslant 3$, the complexity of $(U_m \widehat{U}_n)^R$ is $3 \cdot 2^{m+n-2} - 2^n + 1$.
(Verified for $3 \leqslant m,n \leqslant 7$.)

It was shown in [8] by Cui, Gao, Kari, and Yu that quaternary witnesses meet the bound $5 \cdot 2^{m+n-3} - (2^{m-1} + 2^n - 1)$ for $K_m^* L_n$. Here, one can also use U_n and $\widehat{U}_n$:

Conjecture 12 $(K_m^* L_n)$
For $m,n \geqslant 3$, the complexity of $U_m^* \widehat{U}_n$ is $5 \cdot 2^{m+n-3} - (2^{m-1} + 2^n - 1)$.
(Verified for $3 \leqslant m,n \leqslant 7$.)

It was shown Cui, Gao, Kari, and Yu in [8] that quaternary witnesses meet the bound $mn2^p - (m+n-1)2^{p-1}$ for $(K_m \cup L_n)M_p$. Here U_n and $\widehat{U}_n$ also work:

Conjecture 13 $((K_m \cup L_n)M_p)$
For $m,n,p \geqslant 3$, the complexity of $(U_m \cup \widehat{U}_n)U_p$ is $mn2^p - (m+n-1)2^{p-1}$.
(Verified for some values of m, n and p.)

The definition of "dialect" is extended to four inputs: $V_n(a,b,c,d)$ is a *dialect* of $U_n(a,b,c,d)$ if $V_n(a,b,c,\emptyset)$ is a dialect of $U_n(a,b,c,\emptyset)$ and $d : \mathbf{1}_Q$.

8.2 Witnesses $V_n(a,b,c,d)$ and $\widehat{V}_n(a,b,c,d)$

Let $\mathcal{V}_n = \mathcal{V}_n(a,b,c,d) = (Q, \Sigma, \delta_{\mathcal{V}}, 0, \{n-1\})$, where $a : (0,1,\ldots,n-1)$, $b : (n-2,n-1)$, $c : \binom{n-1}{n-2}$, and $d : \mathbf{1}_Q$. Let $V_n = V_n(a,b,c,d)$ be the language of $\mathcal{V}_n$; so $V_n(a,b,c,d)$ is a dialect of $U_n(a,b,c,d)$. Also let $\widehat{\mathcal{V}}_n(a,b,c,d) = \mathcal{V}_n(d,c,b,a)$; then $\widehat{\mathcal{V}}_n$ and $\mathcal{V}_n$ are permutationally equivalent.

It was shown in [8] by Cui, Gao, Kari and Yu that quaternary witnesses meet the bound $3 \cdot 2^{m+n-2}$ for $K_m^R L_n$. Here V_n and $\widehat{V}_n$ can be used:

Conjecture 14 ($K_m^R L_n$)
The complexity of $V_m^R \widehat{V}_n$ is $3 \cdot 2^{m+n-2}$ for $m, n \geqslant 3$. (Verified for $3 \leqslant m, n \leqslant 7$.)

8.3 Witnesses $W_n(a,b,c,d)$ and $\widehat{W}_n(a,b,c,d)$

Let $\mathcal{W}_n = \mathcal{W}_n(a,b,c,d) = (Q, \Sigma, \delta_{\mathcal{W}}, 0, \{n-1\})$, where $a : (0,1,\ldots,n-1)$, $b : (n-2,n-1)$, $c : \binom{1}{0}$, and $d : \mathbf{1}_Q$. Let $W_n = W_n(a,b,c,d)$ be the language of $\mathcal{W}_n$; so $W_n(a,b,c,d)$ is a dialect of $U_n(a,b,c,d)$. Also let $\widehat{\mathcal{W}}_n(a,b,c,d) = \mathcal{W}_n(d,c,b,a)$; then $\widehat{\mathcal{W}}_n$ and $\mathcal{W}_n$ are permutationally equivalent.

It was shown in [13] by Gao, Kari and Yu that quaternary witnesses meet the bound $9 \cdot 2^{m+n-4} - (3 \cdot 2^{m-2} + 3 \cdot 2^{n-2} - 2)$ for $K_m^* \cup L_n^*$ and $K_m^* \cap L_n^*$. Here W_n and $\widehat{W}_n$ apply:

Conjecture 15 ($K_m^* \cup L_n^*$ **and** $K_m^* \cap L_n^*$)
The complexities of $W_m^* \cup \widehat{W}_n^*$ and $W_m^* \cap \widehat{W}_n^*$ are $9 \cdot 2^{m+n-4} - (3 \cdot 2^{m-2} + 3 \cdot 2^{n-2} - 2)$ for $m, n \geqslant 3$. (Verified for $3 \leqslant m, n \leqslant 7$.)

It was shown in [12] by Gao, Kari and Yu that quaternary witnesses meet the bound $[(m-1)2^m + 2^{m-1}] \cdot [(n-1)2^n + 2^{n-1}]$ for $K_m^2 \cup L_n^2$ and $K_m^2 \cap L_n^2$. This is extended here to all four boolean operations:

Conjecture 16 ($K_m^2 \circ L_n^2$)
The complexity of $W_m^2 \circ \widehat{W}_n^2$ is $[(m-1)2^m + 2^{m-1}] \cdot [(n-1)2^n + 2^{n-1}]$ for $m, n \geqslant 3$. (Verified for $3 \leqslant m, n \leqslant 5$.)

It was shown in [14] by Gao, Salomaa and Yu that quaternary witnesses meet the bound $2^{m+n-1} + 2^{m+n-4} - (2^{m-1} + 2^{n-1} - m - 1)$ for $(K_m L_n)^*$. Here W_n and $\widehat{W}_n$ can also be used:

Conjecture 17 ($(K_m L_n)^*$)
The complexity of $(W_m \widehat{W}_n)^*$ is $2^{m+n-1} + 2^{m+n-4} - (2^{m-1} + 2^{n-1} - m - 1)$ for $m, n \geqslant 3$. (Verified for $3 \leqslant m, n \leqslant 5$.)

It was shown in [7] by Cui, Gao, Salomaa and Yu that quinary witnesses meet the bound $(m-1)(2^{n+p} - 2^n - 2^p + 2) + 2^{n+p-2}$ for $K_m(L_n \cup M_p)$. Here the alphabet size is reduced to 4:

Conjecture 18 ($K_m(L_n \cup M_p)$)
The complexity of $W_m(W_n \cup \widehat{W}_p)$ is $(m-1)(2^{n+p} - 2^n - 2^p + 2) + 2^{n+p-2}$, for $m, n, p \geqslant 3$. (Verified for $3 \leqslant m, n, p \leqslant 4$ and several larger values.)

It was shown in [7] by Cui, Gao, Salomaa and Yu that quaternary witnesses meet the bound $(m-1)2^{np} + 2^{np-1}$ for $K_m(L_n \cap M_p)$. This result is extended here also to $K_m(L_n \setminus M_p)$:

Conjecture 19 ($K_m(L_n \cap M_p)$ **and** $K_m(L_n \setminus M_p)$)
The complexities of $W_m(W_n \cap \widehat{W}_p)$ and $W_m(W_n \setminus \widehat{W}_p)$ are $(m-1)2^{np} + 2^{np-1}$ for $m, n, p \geqslant 3$. (Verified for $3 \leqslant m, n, p \leqslant 4$ and several larger values.)

9 Conclusions

It is clear that a witness over an alphabet of three or four letters cannot be a witness when a larger alphabet is required. Also, $U_n(a, b, c)$ and $U_n(a, b, c, d)$ cannot be witnesses in a proper subclass of regular languages, since they do not possess the special properties of that class. However, in several cases it was possible to use $U_n(a, b, c)$ by "embedding" it in larger witnesses. For example, the DFA of a right ideal—a language L_n satisfying $L_n\Sigma^* = L_n$—can be constructed as shown in Fig. 7 to meet the upper bound on the size of the syntactic semigroup [6]. Similar constructions have been used for left ideals, and two-sided ideals [6], and for prefix-free, suffix-free, bifix-free and factor-free languages [3].

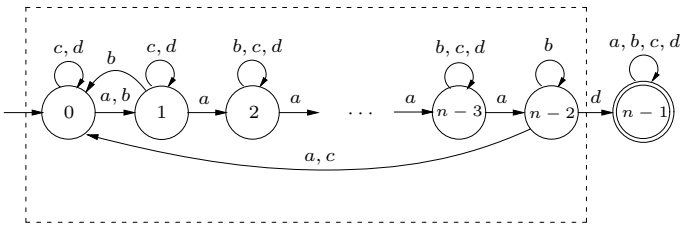

Fig. 7. Right ideal with n^{n-1} transformations

Although $U_n(a, b, c)$ and $U_n(a, b, c, d)$ succeed in all the cases I tried, they do have some shortcomings:

1. No binary language related to $U_n(a, b, c)$ seems to satisfy the reversal bound.
2. No binary languages related to $U_n(a, b, c)$ seem to satisfy the product bound.
3. Dialects of $U_n(a, b, c)$ and $U_n(a, b, c, d)$ had to be used for some operations.

In spite of these shortcomings, the results presented here strongly suggest that these witnesses ought to be considered when one is looking at new operations. The main remaining open questions are:

1. Does there exist a better ternary (quaternary) witness that would overcome the shortcomings listed above?
2. In general, does there exist a universal n-ary witness for operations that require witnesses over alphabets of n letters?

Finally, I remark that this paper must surely have a record number of conjectures! This would not have been possible without computer programs.

Acknowledgment. I am very grateful to Baiyu Li for helping me debug several proofs, for carrying out some computations, and for multiple proofreadings. I thank David Liu and Hellis Tamm for proofreading, and Lila Kari for providing references to work on combined operations.

References

1. Brzozowski, J.: Quotient complexity of regular languages. J. Autom. Lang. Comb. 15(1/2), 71–89 (2010)
2. Brzozowski, J., Li, B.: Syntactic complexities of some classes of star-free languages. In: Proceedings of the 14th International Workshop on Descriptional Complexity of Formal Systems (DCFS). LNCS. Springer, Heidelberg (to appear, 2012)
3. Brzozowski, J., Li, B., Ye, Y.: Syntactic complexity of prefix-, suffix-, bifix-, and factor-free regular languages. Theoret. Comput. Sci. (in press, 2012)
4. Brzozowski, J., Tamm, H.: Theory of Átomata. In: Mauri, G., Leporati, A. (eds.) DLT 2011. LNCS, vol. 6795, pp. 105–116. Springer, Heidelberg (2011)
5. Brzozowski, J., Tamm, H.: Quotient complexity of atoms of regular languages. In: Proceedings of the 16th International Conference on Developments in Language Theory (DLT). LNCS. Springer, Heidelberg (to appear, 2012)
6. Brzozowski, J., Ye, Y.: Syntactic Complexity of Ideal and Closed Languages. In: Mauri, G., Leporati, A. (eds.) DLT 2011. LNCS, vol. 6795, pp. 117–128. Springer, Heidelberg (2011)
7. Cui, B., Gao, Y., Kari, L., Yu, S.: State complexity of two combined operations: catenation-union and catenation-intersection. Int. J. Found. Comput. Sc. 22(8), 1797–1812 (2011)
8. Cui, B., Gao, Y., Kari, L., Yu, S.: State complexity of combined operations with two basic operations. Theoret. Comput. Sci. 437, 82–102 (2012)
9. Cui, B., Gao, Y., Kari, L., Yu, S.: State complexity of two combined operations: catenation-star and catenation-reversal. Int. J. Found. Comput. Sc. 23(1), 51–66 (2012)
10. Dénes, J.: On transformations, transformation semigroups and graphs. In: Erdös, P., Katona, G. (eds.) Theory of Graphs. Proceedings of the Colloquium on Graph Theory held at Tihany 1966, pp. 65–75. Akadémiai Kiado (1968)
11. Ganyushkin, O., Mazorchuk, V.: Classical Finite Transformation Semigroups: An Introduction. Springer (2009)
12. Gao, Y., Kari, L., Yu, S.: State complexity of union and intersection of square and reversal on k regular languages. Theoret. Comput. Sci. (in press, 2012)
13. Gao, Y., Kari, L., Yu, S.: State complexity of union and intersection of star on k regular languages. Theoret. Comput. Sci. 429, 98–107 (2012)

14. Gao, Y., Salomaa, K., Yu, S.: The state complexity of two combined operations: star of catenation and star of reversal. Fund. Inform. 83(1-2), 75–89 (2008)
15. Gao, Y., Yu, S.: State complexity of combined operations with union, intersection, star, and reversal. Fund. Inform. 116, 1–12 (2012)
16. Jirásková, G., Okhotin, A.: On the state complexity of star of union and star of intersection. Fund. Inform. 109, 1–18 (2011)
17. Jirásková, G., Šebej, J.: Note on Reversal of Binary Regular Languages. In: Holzer, M., Kutrib, M., Pighizzini, G. (eds.) DCFS 2011. LNCS, vol. 6808, pp. 212–221. Springer, Heidelberg (2011)
18. Leiss, E.: Succinct representation of regular languages by boolean automata. Theoret. Comput. Sci. 13, 323–330 (1981)
19. Maslov, A.N.: Estimates of the number of states of finite automata. Dokl. Akad. Nauk SSSR 194, 1266–1268 (1970) (Russian); English translation: Soviet Math. Dokl. 11, 1373–1375 (1970)
20. McNaughton, R., Papert, S.A.: Counter-Free Automata. M.I.T. Research Monographs, vol. 65. The MIT Press (1971)
21. Mirkin, B.G.: On dual automata. Kibernetika (Kiev) 2, 7–10 (1966) (Russian); English translation: Cybernetics 2, 6–9 (1966)
22. Myhill, J.: Finite automata and representation of events. Wright Air Development Center Technical Report 57–624 (1957)
23. Piccard, S.: Sur les fonctions définies dans les ensembles finis quelconques. Fund. Math. 24, 298–301 (1935)
24. Pin, J.E.: Syntactic semigroups. In: Handbook of Formal Languages. Word, Language, Grammar, vol. 1, pp. 679–746. Springer, New York (1997)
25. Rabin, M., Scott, D.: Finite automata and their decision problems. IBM J. Res. and Dev. 3, 114–129 (1959)
26. Salomaa, A., Salomaa, K., Yu, S.: State complexity of combined operations. Theoret. Comput. Sci. 383, 140–152 (2007)
27. Salomaa, A., Wood, D., Yu, S.: On the state complexity of reversals of regular languages. Theoret. Comput. Sci. 320, 315–329 (2004)
28. Sierpiński, W.: Sur les suites infinies de fonctions définies dans les ensembles quelconques. Fund. Math. 24, 209–212 (1935)
29. Yu, S., Zhuang, Q., Salomaa, K.: The state complexities of some basic operations on regular languages. Theoret. Comput. Sci. 125, 315–328 (1994)
30. Yu, S.: State complexity of regular languages. J. Autom. Lang. Comb. 6, 221–234 (2001)