Descriptional Complexity of Deterministic Regular Expressions

Katja Losemann[1,*], Wim Martens[1], and Matthias Niewerth[2,**]

[1] Universität Bayreuth
[2] Technische Universität Dortmund

Abstract. We study the descriptional complexity of regular languages that are definable by deterministic regular expressions. First, we examine possible blow-ups when translating between regular expressions, deterministic regular expressions, and deterministic automata. Then we give an overview of the closure properties of these languages under various language-theoretic operations and we study the descriptional complexity of applying these operations. Our main technical result is a general property that implies that the blow-up when translating a DFA to an equivalent deterministic expression can be exponential.

1 Introduction

Deterministic or *one-unambiguous* regular expressions have been a topic of research since they were formally defined by Brüggemann-Klein and Wood in order to investigate a requirement in the ISO standard for the Standard Generalized Markup Language (SGML), where they were introduced to ensure efficient parsing. Today, the prevalent schema languages for XML data, such as Document Type Definition (DTD) and XML Schema, require that the regular expressions in their specification be deterministic. From a more foundational point of view, one-unambiguity is a natural manner in which to define determinism in regular expressions. As such, several decision problems behave better for deterministic regular expressions than for general ones. For example, language inclusion for regular expressions is PSPACE-complete but is tractable when the expressions are deterministic.

Although deterministic regular expressions are rather widespread and have been around for quite some time, they are not yet well-understood. This motivates us to study various foundational properties. In particular, we investigate the differences in the descriptional complexity between regular expressions (REs), deterministic regular expressions (DREs), and deterministic finite automata (DFA). Our initial motivation for this work was an unproved claim in

* Supported by grant number MA 4938/21 of the Deutsche Forschungsgemeinschaft (Emmy Noether Nachwuchsgruppe).

** Supported by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under the FET-Open grant agreement FOX, number FP7-ICT-233599

[2] which states that, for expressions of the form $\Sigma^* w$, where w is a Σ-string, every equivalent DRE is at least exponential in w. However, to the best of our knowledge, no proof for this result exists in the literature and proving it turned out to be rather non-trivial. Since this language has a polynomial-size RE and DFA, we needed to develop new techniques for proving lower bounds on the size of DREs.

A second set of contributions in this paper is a study of the effect of language-theoretic operations on languages that are definable by a DRE. In particular, we consider union, intersection, difference, concatenation, star, and reversal, for unary and arbitrary alphabets. We provide a complete overview of the closure properties of DRE-definable languages under these operations and we study the descriptional complexity of applying such operations on DREs and their DFAs. Several of these operations are relevant in XML schema management [7, 17].

Until now, research on descriptional complexity of regular languages focused mainly on REs and DFAs. It is well-known that an exponential blow-up cannot be avoided when translating an RE into a DFA [12]. Ehrenfeucht and Zeiger [5] proved that there also exist DFAs which are exponentially more succinct than each equivalent RE. Gruber and Holzer [9, 11] showed that there exist certain characteristics of automata which make equivalent regular expressions large. However, these characteristics cannot naïvely be transferred to DREs. For example, the languages used in the literature for proving lower bounds on the size of REs (e.g. [5, 9, 11]) are not definable by DREs.

The state complexity of boolean operations on DFAs is studied in [15, 18, 20], where in [18] the focus is on unary languages. In Section 4.2 we see that many results in [20] directly apply for DRE-definable languages, since they are on finite languages and every finite language is DRE-definable [1]. Gelade and Neven [8] and Gruber and Holzer [10] independently examined the descriptional complexity of complementation and intersection for REs. They showed that the size of the smallest RE for the intersection of a fixed number of REs can be exponential; and that the size of the smallest RE for the complement of an RE can be double-exponential. Furthermore, these bounds are tight. Gelade and Neven also investigate these operations on DREs and proved that the exponential bound on intersection is also tight when the input is given as DREs instead of REs [8]. Furthermore, they proved that the complement of a DRE can always be described by a polynomial-size RE. However, in their proofs, the languages of the resulting REs are not DRE-definable. Concatenation and reversal operations on regular languages are studied in [3, 13, 14, 19, 21], where in [21] also unary languages are examined.

2 Definitions

By Σ we always denote a finite alphabet of symbols. A (Σ-)*word* w over alphabet Σ is a finite sequence of symbols $a_1 \cdots a_n$, where $a_i \in \Sigma$ for each $i = 1, \ldots, n$. The set of all Σ-words is denoted by Σ^*. The *length* of a word $w = a_1 \cdots a_n$ is n and is denoted by $|w|$. The empty word is denoted by ε.

A *(deterministic, finite) automaton* (or *DFA*) A is a tuple $(Q, \Sigma, \delta, q_0, F)$, where Q is a finite set of states, the transition function $\delta \subseteq Q \times \Sigma \to Q$ is a partial function, q_0 is the initial state and $F \subseteq Q$ is the set of accepting states. We sometimes abuse notation and denote a transition $\delta(q_1, a) = q_2$ by a tuple (q_1, a, q_2). We say that the aforementioned transition is q_1-*outgoing*, q_2-*incoming*, or a-*labeled*. The *run of* A *on word* $w = a_1 \cdots a_n$ is a sequence $q_0 \cdots q_n$ where, for each $i = 1, \ldots, n$, $\delta(q_{i-1}, a_i) = q_i$. Word w is *accepted* by A if the run is *accepting*, i.e., if $q_n \in F$. By $L(A)$ we denote the *language of* A, i.e., the set of words accepted by A. By δ^* we denote the extension of δ to words, i.e., $\delta^*(q, w)$ is the state which is reached from q by reading w. In this paper we assume that all states of automata are *useful*, that is, every state can appear in some accepting run. This implies that, from each state in an automaton, an accepting state can be reached. The *size* $|A|$ of a DFA is the cardinality of $\{(q, a) \mid \delta(q, a)$ is defined$\}$.

The *regular expressions (RE)* over Σ are defined as follows: ε and every Σ-symbol is a regular expression; and whenever r and s are regular expressions then so are $(r \cdot s)$, $(r + s)$, and $(s)^*$. In addition, we allow $\emptyset$ as a regular expression, but we do not allow $\emptyset$ to occur in any other regular expression. We refer to Σ-symbols, ε, and $\emptyset$ as *atomic* expressions. For readability, we usually omit concatenation operators and parentheses in examples. The *language* defined by an RE r, denoted by $L(r)$, is defined as usual. Whenever we say that expressions or automata are *equivalent*, we mean that they define the same language. The *size* $|r|$ of r is defined to be the total number of occurrences of alphabet symbols, epsilons, and operators, i.e., the number of nodes in its parse tree. A regular expression r is *minimal* if there does not exist a regular expression r' with $L(r') = L(r)$ and $|r'| < |r|$. By *first(L)* we denote the set of all symbols $a \in \Sigma$, such that there is a word $aw \in L$. For a regular expression r, we define first(r) as first$(L(r))$.

Deterministic regular expressions are defined as follows. Let $\bar{r}$ stand for the RE obtained from r by replacing, for every i and a, the i-th occurrence of alphabet symbol a in r (counting from left to right) by a_i. For example, for $r = b^*a(b^*a)^*$ we have $\bar{r} = b_1^*a_1(b_2^*a_2)^*$. A regular expression r is *deterministic* (or *one-unambiguous* [2] or a *DRE*) if there are no words wa_iv and wa_jv' in $L(\bar{r})$ such that $i \neq j$. The expression $(a + b)^*a$ is not deterministic since both strings a_2 and a_1a_2 are in $L((a_1 + b_1)^*a_2)$. The equivalent expression $b^*a(b^*a)^*$ is deterministic. Brüggemann-Klein and Wood showed that not every regular expression is equivalent to a deterministic one [2]. We call a regular language *DRE-definable* if there exists a DRE that defines it. The canonical example for a language that is not DRE-definable is $(a + b)^*a(a + b)$ [2].

3 Descriptional Complexity of DFAs, REs, and DREs

We consider the relative descriptional complexity of REs, DREs and DFAs. An overview of our results is shown in Figure 1. Since every DRE is an RE, we know that every minimal RE for a language L is smaller or equal to a minimal DRE

Finite Languages					Infinite Languages				
RE	DRE	DFA	Case exists?	Ref	RE	DRE	DFA	Case exists?	Ref
$\Theta(n)$	$\Theta(n)$	$\Theta(n)$	yes	Obs.1	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$	yes	Obs.1
$\Theta(n)$	$2^{\Omega(n)}$	$2^{\Omega(n)}$	yes	[15, 2]	$\Theta(n)$	$2^{\Omega(n)}$	$2^{\Omega(n)}$	yes	Th.6
$2^{\Omega(n)}$	$2^{\Omega(n)}$	$\Theta(n)$	no	[6]	$2^{\Omega(n)}$	$2^{\Omega(n)}$	$\Theta(n)$	?	
$\Theta(n)$	$2^{\Omega(n)}$	$\Theta(n)$	?		$\Theta(n)$	$2^{\Omega(n)}$	$\Theta(n)$	yes	Th.15
$\Omega(n^{\log n})$	$\Omega(n^{\log n})$	$\Theta(n)$	yes	[11]					

Fig. 1. Overview descriptional complexity.

for L. Furthermore, Brüggemann-Klein and Wood showed that, given a DRE r, one can construct a DFA A for $L(r)$ with size $O(|\Sigma||r|)$. Thus the table contains all substantial cases that ought to be considered.

We start with a trivial observation that shows that there are languages that do not cause any significant blow-up between the different representations. For example, consider the singleton $\{a^n\}$ and the infinite language $\{a^k \mid k \equiv 0 \mod n\} = L((aa\cdots a)^*)$ in which the latter expression has n occurrences of a.

Observation 1. There exists a class of finite languages $(L_n)_{n\in\mathbb{N}}$ and a class of infinite languages $(L'_n)_{n\in\mathbb{N}}$ such that, for each $n \in \mathbb{N}$, the minimal DFAs, minimal REs, and minimal DREs for L_n and L'_n have size $\Theta(n)$.

3.1 Finite Languages

We present an overview of what is known in the case of finite languages. For the language $(0+1)^{\leq n}1(0+1)^n$, Kintala and Wotschke, and Brüggemann-Klein and Wood showed that every DFA and every DRE has size exponential in n.

Theorem 2 ([15, 2]). *For each $n \in \mathbb{N}$, the minimal DFA (and therefore every minimal DRE) for the language $(a+b)^{\leq n}a(a+b)^n$ have size $2^{\Omega(n)}$.*

Ellul et al. [6] showed that, for each DFA (or even non-deterministic automaton) A of size n that defines a finite language $L(A)$, there exists an RE for $L(A)$ of size $O(n^{\log n})$. Gruber and Johannsen showed that this upper bound is also tight. However, this problem was open for quite some time [11].

Theorem 3 ([6]). *Let A be a DFA of size n and let $L(A)$ be finite. Then there exists an RE r for $L(A)$ such that $|r| \leq O(n^{\log n})$.*

Theorem 4 ([11]). *There exists a family of finite languages $(L_n)_{n\in\mathbb{N}}$, such that the minimal DFA for L_n has $\Theta(n)$ states but every minimal RE for L_n has size $\Theta(n^{\log n})$.*

It remains open whether there exists a class of finite languages $(L_n)_{n\in\mathbb{N}}$, such that the minimal REs and the minimal DFA for L_n are exponentially more succinct than a minimal DRE for L_n.

3.2 Infinite Languages

In the case of infinite languages, it is well known that an exponential blow-up can occur when translating between REs and DFAs:

Theorem 5 ([12, 5]).
- *The minimal DFA for* $(a+b)^*a(a+b)^n$ *has size* $2^{\Theta(n)}$.
- *There exists a family of infinite regular languages* $(L_n)_{n\in\mathbb{N}}$*, s.t. the minimal DFA for* L_n *has size* $\Theta(n^2)$ *and every minimal RE for* L_n *has size* $2^{\Omega(n)}$.

However, to the best of our knowledge, all languages that are used in the literature to prove those blow-ups are not DRE-definable. Here, we prove that those blow-ups cannot be avoided for DRE-definable languages, too. For an exponential blow-up when translating an RE for a DRE-definable language to a DFA, we can extend the language of Theorem 2 to an infinite language.

Theorem 6. *For each* $n \in \mathbb{N}$*, the minimal DFA and every minimal DRE for the DRE-definable language* $(a+b)^{\leq n}a(a+b)^n\#^*$ *have size* $2^{\Omega(n)}$.

Next, we prove that there can be an exponential blow-up when translating a DFA to a DRE. The main idea of the proof is to identify concatenations of a minimal DRE in a DFA. Therefore, we search for *bottleneck states*, which are states through which every accepting run needs to go.

Definition 7. Let $A = (Q, \Sigma, \delta, q_0, Q_f)$ be a DFA. A state $q \in Q \setminus \{q_0\}$ is a *bottleneck state* of A if
- for every $w \in L(A)$ there are $v, z \in \Sigma^*$, s.t. $w = v \cdot z$ and $\delta^*(q_0, v) = q$, and
- if $q \in Q_f$, then $Q_f = \{q\}$ and there are $a \in \Sigma$ and $p \in Q$ s.t. $\delta(q, a) = p$.

Notice that we explicitly define initial states not to be bottleneck states.

Lemma 8. *Let* $A = (Q, \Sigma, \delta, q_0, Q_f)$ *be a DFA with a bottleneck state* q*. Then* A *has no equivalent DRE that is atomic or of the form* s^*.

In the following we show that accepting bottleneck states in a DFA identify concatenations in an equivalent minimal DRE. Therefore, let $A = (Q, \Sigma, \delta, q_0, Q_f)$ be a DFA. Then an equivalent DRE r is a *q-concatenation* if and only if $r = r_1 \cdot r_2$ and for every $v \in L(r_1)$ it holds that $\delta^*(q_0, v) = q$ in A. If r is a DRE with these conditions such that $L(r) \subsetneq L(A)$, then r is a *partial q-concatenation for A*.

Lemma 9. *Let* $A = (Q, \Sigma, \delta, q_0, \{q_f\})$ *be a DFA for a DRE-definable language* L*, such that* q_f *is a bottleneck state of* A*. Then every minimal DRE* r *for* L *is a* q_f*-concatenation* $r_1 \cdot r_2$ *with* $\mathrm{first}(r_2) = \{a \in \Sigma \mid \delta(q_f, a)$ *is defined*$\}$.

Proof. By Lemma 8 it holds that r is neither atomic nor an expression s^*. It remains to show that r is neither a disjunction nor a concatenation which is not a q_f-concatenation. We can prove the following claim:

Claim 10. Let $A = (Q, \Sigma, \delta, q_0, \{q_f\})$ *be a DFA for a DRE-definable language* L*, such that* q_f *is a bottleneck state of* A*. Let* $\emptyset \neq S \subseteq \mathrm{first}(L)$ *and* $r = r_1 r_2 \cdots r_n$ *(with* $n > 1$*) be a minimal DRE for* $L \cap S\Sigma^*$*, such that no* r_i *is a concatenation. Then there exists an* $i \in \{1, \ldots, n-1\}$ *such that,*

– *for every word* $w \in L(r_1 \cdots r_i)$*, it holds that* $\delta^*(q_0, w) = q_f$*, and*
– $\text{first}(r_{i+1} \cdots r_n) = \{a \in \Sigma \mid \delta(q_f, a) \text{ is defined}\}$.

In particular, this means that r *is a partial* q_f*-concatenation for* A.

We show why Claim 10 implies Lemma 9. From the discussion above, we know that r is either a concatenation or a disjunction. In the case that r is a concatenation, Claim 10 clearly implies the lemma (if $S = \text{first}(r)$, then $L \cap S\Sigma^* = L$).

We now show that, if r is a disjunction $(s_1 + \cdots + s_k)$, then r is not minimal, which contradicts the assumption we made about r. As an intermediate step we want to apply Claim 10 to every s_i. We therefore have to show that, for every i, (a) $L(s_i) = L \cap S_i\Sigma^*$ with $\emptyset \subsetneq S_i \subseteq \text{first}(L)$ and (b) s_i is a concatenation.

Since r is a DRE, it holds that $\text{first}(s_i) \cap \text{first}(s_j) = \emptyset$ for all $i \neq j$. Furthermore, we know that $\varepsilon \notin L$ and therefore $\varepsilon \notin L(s_i)$ for every i. Thus we can conclude that $L(s_i) = L \cap S_i\Sigma^*$ with $S_i = \text{first}(s_i) \subseteq \text{first}(r)$ for every i. This proves (a). Notice that $S_i \neq \emptyset$ because r is minimal. Next we prove that every s_i is a concatenation. W.l.o.g., s_i is not a disjunction. Since $\varepsilon \notin L(s_i)$, s_i is not of the form t^*. Now take an arbitrary $a \in S_i$. Then there exists a word $aw \in L(r)$ with $w \neq \varepsilon$, because q_f has at least one outgoing transition. Since r is a DRE, $L(s_i)$ contains all words $b \cdot v \in L$ where $b \in S_i$ and $v \in \Sigma^*$, and therefore $aw \in L(s_i)$. As $|aw| > 1$, s_i cannot be atomic. The only remaining possibility is that s_i is a concatenation, which proves (b). Also, s_i is a minimal DRE.

We can now apply Claim 10 to every s_i and conclude that we can write every s_i as $s_i^a s_i^b$ such that (i) $\delta(q_0, w) = q_f$ for every $w \in L(s_i^a)$ and (ii) $\text{first}(s_i^b) = \{a \in \Sigma \mid \delta(q_f, a) \text{ is defined}\}$. Notice that s_i^a and s_i^b can be concatenations again.

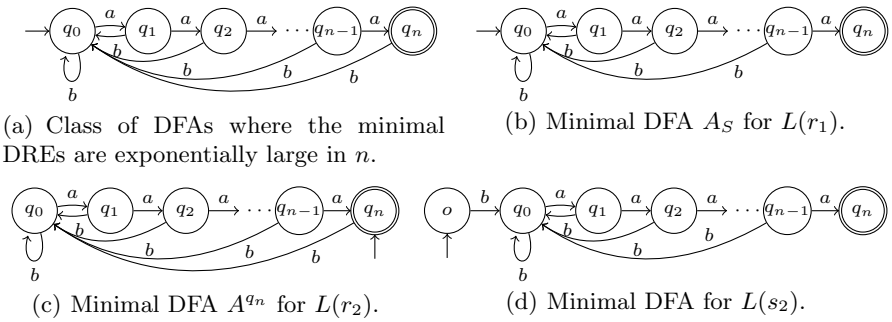
Let $A^{q_f} = (\Sigma, Q, \delta, q_f, \{q_f\})$ be the automaton A where the initial state is q_f. From (i) and (ii), we can conclude that $L(s_i^b) = L(A^{q_f})$ for every i. Therefore all expressions s_i^b are equivalent. Thus, r can equivalently be written as $(s_1^a + \cdots + s_k^a)s_1^b$, which is strictly smaller than r. This contradicts the minimality of r and therefore contradicts that r is a disjunction, which concludes the proof. □

Notice that a DRE can have multiple q_f-concatenations. For example, the expression $a \cdot b^* \cdot (c \cdot b^*)^*$ has a DFA with a unique accepting state q_f and has two q_f-concatenations. However, a DRE can only have one q_f-concatenation of the form $r_1 \cdot r_2$ where $\text{first}(r_2) = \{a \mid \delta(q_f, a) \text{ is defined}\}$. Furthermore, if $A = (Q, \Sigma, \delta, q_0, \{q_f\})$ is a DFA with a bottleneck state q_f, it holds that $L(A)$ is infinite. Lemma 9 gives us a rather precise structure of each minimal DRE $r_1 \cdot r_2$. The following lemma also clarifies $L(r_1)$ and $L(r_2)$.

Lemma 11. *For a DFA* $A = (Q, \Sigma, \delta, q_0, \{q_f\})$ *with a bottleneck state* q_f *let the* q_f*-concatenation* $r_1 \cdot r_2$ *be an equivalent minimal DRE with* $\text{first}(r_2) = \{a \in \Sigma \mid \delta(q_f, a) \text{ is defined}\}$*. Then*

(1) $L(r_1) = L(A_S)$ *where* $A_S = (Q, \Sigma, \delta - S, q_0, \{q_f\})$, $S = \{(q_f, a, q) \in \delta \mid a \in \Sigma, q \in Q\}$*; and*
(2) $L(r_2)$ *is infinite where* $L(r_2) = L(A^{q_f})$ *with* $A^{q_f} = (Q, \Sigma, \delta, q_f, \{q_f\})$.

Note that (2) follows from the proof of Lemma 9.



(a) Class of DFAs where the minimal DREs are exponentially large in n.

(b) Minimal DFA A_S for $L(r_1)$.

(c) Minimal DFA A^{q_n} for $L(r_2)$.

(d) Minimal DFA for $L(s_2)$.

Fig. 2. Minimal DFAs for subexpressions from the proof of Lemma 14.

Before we can finally prove the blow-up from DFA to DRE, we need two general results on minimal DREs. The first is a very straightforward relation between a state and a concatenation in the DRE. Therefore, we say that a regular language L is *prefix-free* if and only if, for every word $v \in L$, there exists no $z \in \Sigma^*$ such that $v \cdot z \in L$.

Lemma 12. *Let $L_a = L \cdot \{a\}$ be a prefix-free regular language. Then there exists a minimal DRE for L_a which is either a or of the form $r \cdot a$.*

Lemma 13. *If r^* is a minimal DRE, then $\varepsilon \notin L(r)$.*

However, the DRE r in a minimal DRE r^* in Lemma 13 can still contain ε-symbols. This holds, for example, for the DRE $r = (a(b + \varepsilon))^*$.

Now we are ready to prove the exponential blow-up when translating DFAs to DREs. In particular, we prove that every minimal DRE for the DFA in Figure 2(a) is exponential in n. We denote the language of this DFA with $L^{[n]}$.

Lemma 14. *There is a minimal DRE r for $L^{[n]}$ containing at least 2^n concatenations.*

Proof. Let A be the minimal DFA for $L^{[n]}$ (see Figure 2(a)). The proof is by induction on n. For the induction basis, let $n = 1$. Since A has an accepting bottleneck state, we know by Lemma 9 that r is a concatenation $r_1 \cdot r_2$ with $\text{first}(r_2) = \{b\}$. By Lemma 11, it follows that $L(r_1) = L(b^*a)$. This implies that there is a minimal DRE for $L^{[1]} = L(b^* \cdot a \cdot r_2)$, with at least 2 concatenations.

For the induction step, assume that there exists a minimal DRE for $L^{[n-1]}$ containing at least 2^{n-1} concatenations. By Lemma 9, r is a q_n-concatenation $r_1 \cdot r_2$ with $\text{first}(r_2) = \{b\}$. Lemma 11 implies that the automaton in Figure 2(b) is a DFA for $L(r_1)$ and the automaton in Figure 2(c) is a DFA for $L(r_2)$.

Next we show that r_1 and r_2 each contain a subexpression for the language $L^{[n-1]}$. For r_1 we observe that $L(r_1)$ (see Figure 2(b)) is prefix-free and a language of the form $L' \cdot \{a\}$. Thus there exists a minimal DRE r_1 of the form $s_1 \cdot a$ by Lemma 12, where $L(s_1)$ is defined by the DFA of Figure 2(b) without

the transition $\delta(q_{n-1}, a) = q_n$ and with q_{n-1} as accepting state. As we can see, this is a DFA for $L^{[n-1]}$; hence $L(s_1) = L^{[n-1]}$. Then, by induction hypothesis there exists a DRE s_1, such that s_1 and therefore r_1 contain at least 2^{n-1} concatenations.

For r_2, we observe that $L(r_2)$ is infinite (see A^{q_n} in Figure 2(c)), which implies that r_2 is not an atomic expression. Furthermore, it holds that $|\operatorname{first}(r_2)| = 1$ and $\varepsilon \in L(r_2)$. It follows that r_2 cannot be a concatenation $r_2 = r_3 \cdot r_4$, as the first sets of r_3 and r_4 would have to be disjunct because of $\varepsilon \in r_3$. Next we show that r_2 cannot be a disjunction. Since $\operatorname{first}(r_2) = \{b\}$, the only possible disjunction is $r_2 = b \cdot r_3 + \varepsilon$ for some DRE r_3. As $\delta(q_n, b) = q_0$ in A^{q_n}, we observe that $L(r_3) = L^{[n]}$, which directly contradicts that r is a minimal DRE for $L^{[n]}$.

Thus r_2 is an expression of the form s_2^*. Next we investigate the structure of a DFA for $L(s_2)$. For every word $v \in L(s_2)$, it holds that $\delta^*(q_n, v) = q_n$ in A^{q_n}. Since s_2^* is a DRE and $\operatorname{first}(r_2) = \{b\}$, $L(s_2)$ cannot contain a word v such that $v = w \cdot z$ with $w, z \neq \varepsilon$ and $\delta^*(q_n, w) = q_n$. These properties characterize $L(s_2)$, for which the minimal DFA is shown in Figure 2(d). Because the DFA has a bottleneck state q_1, s_2 cannot be atomic or an expression t^* by Lemma 8. Furthermore, s_2 is not a disjunction, because $|\operatorname{first}(s_2)| = 1$, $\varepsilon \notin L(s_2)$, and s_2 is a DRE. Thus s_2 is a concatenation $b \cdot t$, where $L(t)$ is definded by the DFA from Figure 2(d) without the transition (o, b, q_0) and with q_0 as initial state. By Lemma 12, it follows that $s_2 = b \cdot t \cdot a$, where $L(t) = L^{[n-1]}$. Thus, by induction hypothesis, t and therefore r_2 contain at least 2^{n-1} concatenations.

Finally, it holds that r_1 and r_2 contain at least 2^{n-1} concatenations each, i.e., $r = r_1 \cdot r_2$ contains at least 2^n concatenations. This concludes the proof. □

Since we can write $L^{[n]} = L((b + ab + \cdots + a^n b)^* a^n) = L((b(a + b(\cdots(ab + b)\cdots)))^* a^n)$, we obtain the following theorem:

Theorem 15. *For each $n \in \mathbb{N}$, the minimal DFA for $L^{[n]}$ has size $\Theta(n)$, every minimal RE for $L^{[n]}$ has size $\Theta(n)$, and every minimal DRE has size $2^{\Omega(n)}$.*

3.3 Application on an Example from the Literature

Brüggemann-Klein and Wood claimed that every minimal DRE for languages of the form $\Sigma^* a_1 \cdots a_n$, where $a_1 \cdots a_n$ is a fixed Σ-word, is exponential [2]. However, to the best of our knowledge, no proof for this result exists in the literature. We prove this claim by using bottleneck states. Therefore we will generalize the special structure of the automata of languages $L^{[n]}$ (see Figure 2(a)) and $L(\Sigma^* a_1 \cdots a_n)$ to provide a formal proof.

Definition 16. Let $A = (Q, \Sigma, \delta, o, \{q_n\})$ be a DFA with $\Sigma \supseteq \{a_1, \ldots, a_n\}$ and $Q \supseteq \{q_0, \ldots, q_n\}$. Then A contains a *bottleneck tail* of length n, if all of the following hold:

1. q_i is a bottleneck state for every $i \in \{0, \ldots, n\}$;
2. $(q_{i-1}, a_i, q_i) \in \delta$ for all $i \in \{1, \ldots, n\}$;
3. for every $i \in \{0, \ldots, n\}$ there is an $a \in \Sigma$ and a transition (q_i, a, o) in A; and
4. for every $i \in \{1, \ldots, n\}$, if $(q, a, q_i) \in \delta$ then $q = q_{i-1}$ and $a = a_i$.

Op.	$\lvert\Sigma\rvert = 1$	$\lvert\Sigma\rvert \geq 1$	Op.	$\lvert\Sigma\rvert = 1$	$\lvert\Sigma\rvert \geq 1$	Op.	$\lvert\Sigma\rvert = 1$	$\lvert\Sigma\rvert \geq 1$
$\setminus$	no	no	$\cup$	no	no	$\cdot$	no	no
Rev	yes	no	$\cap$	yes	no	*	yes	no

Fig. 3. Closure Properties of DRE-definable languages.

For example, the automaton in Figure 2(a) and the minimal DFA for $L(\Sigma^* a_1 \cdots a_n)$ each contain a bottleneck tail of length $n-1$. We prove that a bottleneck tail causes a blow-up in a DRE, exponential in the length of the tail.

Theorem 17. *Let $A = (Q, \Sigma, \delta, o, \{q_n\})$ be a DFA for a DRE-definable regular language L with a bottleneck tail of length n. Then there exists a minimal DRE r for L which contains at least 2^n concatenations.*

Theorem 18. *Every minimal DRE for $L(\Sigma^* a_1 \cdots a_n)$ has size $2^{\Omega(n)}$.*

4 Operations on DRE-Definable Languages

We investigate the descriptional complexity of several language-theoretic operations on DREs and their DFAs. Most results concern DFAs for DRE-definable languages, which allows us to infer lower bounds for DREs as well. First, we present an overview of the closure properties of DRE-definable languages.

4.1 Closure Properties of DRE-Definable Languages

It has been observed that DRE-definable languages are not closed under union [2], intersection [16, 4] or complement [8]. DRE-definable languages are also not closed under concatenation [2], reversal[3] (take $L((a+b)^* a(a+b))$) or Kleene star [2]. These results hold for alphabets with at least two symbols. For unary alphabets, the same results hold, except for reversal, intersection and star. In these three cases, we prove that DRE-definable languages are closed. It is easy to see that DRE-definable languages over unary alphabets are closed under reversal, since for unary alphabets the language and its reversal are equal. The other two cases are non-trivial. The results are summarized in Figure 3.

Theorem 19. *DRE-definable regular languages over a unary alphabet are closed under reversal, intersection, and Kleene star.*

4.2 Descriptional Complexity of Operations on DRE-Definable Languages

We are now ready to apply previously obtained results to prove lower bounds on the descriptional complexity of operations on DREs. From Section 4.1 we know that we need to be careful that the language after performing the operations is indeed DRE-definable. We first prove some lower bounds directly on DREs. For DRE-definable languages we get the following by Theorem 2 and 15.

[3] The reversal of a language L is the set of strings $\{a_n \cdots a_1 \mid a_1 \cdots a_n \in L\}$.

Theorem 20. *There exist regular languages $(L_n)_{n\in\mathbb{N}}$ such that, for each $n \in \mathbb{N}$, the minimal DREs for L_n have size $\Theta(n)$, whereas the minimal DREs for the reversal of L_n have size $2^{\Theta(n)}$. This holds in the case where all L_n are finite languages and in the case where all L_n are infinite languages.*

Indeed, in the finite case one could take L_n to be $L((a+b)^{\leq n}a(a+b)^n)$ and in the infinite case take the language $L^{[n]}$ from Theorem 15.

Theorem 21. *There exist regular languages $(L^1_n)_{n\in\mathbb{N}}$ and $(L^2_n)_{n\in\mathbb{N}}$ such that, for each $n \in \mathbb{N}$, the minimal DREs for L^1_n and L^2_n have size $\Theta(n)$ and the minimal DREs for $L^1_n \cdot L^2_n$ have size $2^{\Theta(n)}$. This holds in the case where all L^1_n and L^2_n are finite languages and in the case where all L^1_n and L^2_n are infinite languages.*

Indeed, in the finite case we can take $L^1_n = (a+b)^{\leq n}$ and $L^2_n = a(a+b)^n$ and in the infinite case we can take $L^1_n = (a+b)^*$ and $L^2_n = a^n cc^*$.

The following results do not immediately concern the minimal size of DREs after performing an operation, but focus on the minimal size of the DFAs for the DREs. For DRE-definable languages, lower bounds can always be transfered. In some cases, we can even infer upper bounds on the DRE size. Consider, for example, the case of languages over a unary alphabet: For those languages all minimal DREs have size linear in the minimal DFA.

Theorem 22. *Let A be a minimal DFA with m states for a DRE-definable language L over a unary alphabet. Then there exists a minimal DRE r for L, such that r is of size $O(m)$.*

To this end, for a DRE-definable language L, we write DDFA(L) for the minimal DFA defining L. We summarize our results in Figure 4 and 5, where in each case we consider a single use of a boolean operation and a k-times application. In Figure 5 the resulting DFA has to define an infinite language.

It is well-known that for the complement on DFAs there is no blow-up [12]. Since all finite languages are DRE-definable, we provide the known results of Yu [20] separated in Figure 4. For all remaining operations the upper bounds are obtained by the standard product construction [12]. For the union and intersection of two finite languages an exact result is as far as we know still open.

Theorem 23. *For every $k \in \mathbb{N}$ there exists finite languages $L_1, \ldots, L_k$, such that the minimal DFA for every L_i has $\Theta(k)$ states and the minimal DFA for $L_1 \cap \cdots \cap L_k$ or $L_1 \cup \cdots \cup L_k$ has at least $2^{\Theta(k)}$ states.*

The theorem is obtained by taking $L_i = \{x_1 \ldots x_k y_k \ldots y_1 \in \{a,b\}^* \mid x_i = y_i\}$.

Now we examine DDFAs which are the result of a boolean operation on $k \geq 2$ infinite DRE-definable languages. For general DFAs the descriptional complexity is studied in [18, 20]. In the following we show that for infinite DRE-definable languages the complexity remains the same in almost all cases. Only for the union of two DDFAs the descriptional complexity is strictly lower than for DFAs.

	$\lvert\Sigma\rvert = 1$		$\lvert\Sigma\rvert \geq 1$	
	1	k	1	k
$\setminus$	$\Theta(m)$ [12]	—	$\Theta(m)$ [12]	—
$\cap$	$\Theta(\min\{m_1, m_2\})$ [20]	$\Theta(\min\{m_1, ..., m_k\})$ [20]	$O(m_1 m_2)$ [20]	$2^{\Omega(k)}$ (Th. 23)
$\cup$	$\Theta(\max\{m_1, m_2\})$ [20]	$\Theta(\max\{m_1, ..., m_k\})$ [20]	$O(m_1 m_2)$ [20]	$2^{\Omega(k)}$ (Th. 23)

Fig. 4. Descriptional complexity of minimal DFAs for finite languages.

	$\lvert\Sigma\rvert = 1$		$\lvert\Sigma\rvert \geq 1$	
	1	k	1	k
$\setminus$	$\Theta(m)$ [12]	—	$\Theta(m)$ [12]	—
$\cap$	$\Theta(m_1 m_2)$ (Th. 24)	$2^{\Omega(k)}$ (Th. 24)	$\Theta(m_1 m_2)$ (Th. 24)	$2^{\Omega(k)}$ (Th. 24)
$\cup$	$\Theta(\max\{m_1 m_2\})$ (Th. 25)	$\Theta(\max\{m_1, ..., m_k\})$(Th. 25)	$O(m_1 m_2)$	$2^{\Omega(k)}$ (Th. 26)

Fig. 5. Descriptional complexity of minimal DFAs for inifinite DRE-definable languages

Theorem 24. *For each $k \in \mathbb{N}$, there exist infinite DRE-definable languages $L_1, \ldots, L_k$, such that, for every $i \in \{1, \ldots, k\}$ the minimal DFA for L_i has $O(k \log k)$ states and DDFA$(L_1 \cap \cdots \cap L_k)$ has $k^{\Omega(k)}$ states. This holds even when the alphabet is unary.*

The theorem is obtained by k languages $L_i = (a^{m_i})^*$ with $1 \leq i \leq k$ and k different m_i, such that $\gcd(m_i, m_j) = 1$ for each pair (m_i, m_j).

At last we examine the union of DFAs for DRE-definable languages where the result still describes a DRE-definable language. We get that for DFAs over unary alphabets the complexity is only linear; hence is strictly lower than for intersection. For arbitrary alphabets the complexity is again exponential.

Theorem 25. *Let $L_1, \ldots, L_k$ be infinite languages over a unary alphabet, such that the minimal DFAs for every L_i with $i \in \{1, \ldots, k\}$ has m_i states. Then the DDFA A for $L_1 \cup \cdots \cup L_k$ has $\Theta(\max\{m_1, \ldots, m_k\})$ states.*

Theorem 26. *For each $k \in \mathbb{N}$, there are infinite DRE-definable languages L_1, $\ldots, L_k$ such that, for each $i \in \{1, \ldots, k\}$ there is a DFA of size $\Theta(k)$ for L_i, but DDFA$(L_1 \cup \cdots \cup L_k)$ has size $2^{\Theta(k)}$.*

The theorem follows from taking $L_i = \{x_1 \ldots x_k y_k \ldots y_1 w \in \{a, b\}^* \mid x_i = y_i\}$.

5 Conclusions and further work

In this paper we were motivated by the aim to come to a better understanding of DRE-definable languages. For example, we developed a new technique to prove lower bounds on the size of DREs by using bottleneck states and tails in a DFA. As a consequence of this technique, we now know that, when translating an RE into a DFA and when translating a DFA into a DRE, an exponential blow-up cannot be avoided. However, we do not know yet whether there are DRE-definable languages for which a translation from an RE to a DRE causes a double exponential blow-up.

Finally we examine several operations on DRE-definable languages. We obtain an overview of the closure properties and the descriptional complexity of these operations on DRE-definable languages. A tight lower bound for the union of two DFAs for DRE-definable languages remains open.

References

1. G. J. Bex, W. Gelade, W. Martens, and F. Neven. Simplifying XML Schema: effortless handling of nondeterministic regular expressions. In *SIGMOD*, 2009.
2. A. Brüggemann-Klein and D. Wood. One-unambiguous regular languages. *Information and Computation*, 142(2):182–206, 1998.
3. C. Câmpeanu, K. Culik, K. Salomaa, and Sheng Y. State complexity of basic operations on finite languages. In *Automata Implementation*. 2001.
4. P. Caron, Y. Han, and L. Mignot. Generalized one-unambiguity. In *DLT*, pages 129–140. 2011.
5. A. Ehrenfeucht and H. Zeiger. Complexity measures for regular expressions. *JCSS*, 12(2):134–146, 1976.
6. K. Ellul, B. Krawetz, J. Shallit, and M. Wang. Regular expressions: new results and open problems. *JALC*, pages 233–256, 2004.
7. W. Gelade, T. Idziaszek, W. Martens, and F. Neven. Simplifying XML Schema: Single-type approximations of regular tree languages. In *PODS*, 2010.
8. W. Gelade and F. Neven. Succinctness of the complement and intersection of regular expressions. In *TOCL*, pages 4:1–4:19, 2012.
9. H. Gruber and M. Holzer. Finite automata, digraph connectivity, and regular expression size. In *ICALP*, pages 39–50, 2008.
10. H. Gruber and M. Holzer. Tight bounds on the descriptional complexity of regular expressions. In *DLT*, pages 276–287, 2009.
11. H. Gruber and J. Johannsen. Optimal lower bounds on regular expression size using communication complexity. In *FoSSaCS*, pages 273–286, 2008.
12. J.E. Hopcroft, R. Motwani, and J.D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 2007.
13. J. Jirásek, G. Jirásková, and A. Szabari. State complexity of concatenation and complementation of regular languages. In *CIAA*. 2005.
14. G. Jirásková. On the state complexity of complements, stars, and reversals of regular languages. In *DLT*, pages 431–442. 2008.
15. C. Kintala and D. Wotschke. Amounts of nondeterminism in finite automata. *Acta Informatica*, 13:199–204, 1980.
16. K. Losemann. Boolesche Operationen auf deterministischen regulären Ausdrücken. Master's thesis, TU Dortmund, October 2010.
17. W. Martens, M. Niewerth, and T. Schwentick. Schema design for XML repositories: Complexity and tractability. In *PODS*, 2010.
18. G. Pighizzini and J. Shallit. Unary language operations, state complexity and Jacobsthal's function. *IJFCS*, pages 145–159, 2002.
19. A. Salomaa, D. Wood, and S. Yu. On the state complexity of reversals of regular languages. *TCS*, pages 315 – 329, 2004.
20. S. Yu. State complexity of regular languages. *JALC*, pages 221–, 2001.
21. S. Yu, Q. Zhuang, and K. Salomaa. The state complexities of some basic operations on regular languages. *TCS*, pages 315 – 328, 1994.