Prime Decompositions of Regular Languages

Yo-Sub Han[1], Kai Salomaa[2,*], and Derick Wood[3,**]

[1] System Technology Division, Korea Institute of Science and Technology,
P.O. Box 131, Cheongryang, Seoul, Korea
emmous@kist.re.kr

[2] School of Computing, Queen's University, Kingston, Ontario K7L 3N6, Canada
ksalomaa@cs.queensu.ca

[3] Department of Computer Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, SAR
dwood@cs.ust.hk

Abstract. We investigate factorizations of regular languages in terms of prime languages. A language is said to be strongly prime decomposable if any way of factorizing the language yields a prime decomposition in a finite number of steps. We give a characterization of the strongly prime decomposable regular languages and using the characterization we show that every regular language over a unary alphabet has a prime decomposition. We show that there exist co-context-free languages that do not have prime decompositions.

1 Introduction

A language is said to be prime [12, 14] if it cannot be written as a catenation of two languages neither one of which is the singleton language consisting of the empty word. A prime decomposition of a language is a factorization where all the components are prime languages. The original work on prime decompositions concentrated mainly on finite languages [12]. Factorizations of prefix-free or infix-free regular languages into prime components that in turn are required to be prefix-free or infix-free, respectively, are considered in [3, 6]. Decompositions of factorial languages are investigated in [1].

Any finite language always has a prime decomposition, although it need not be unique [12, 14]. Work on factorizations of finite languages leads to nontrivial questions concerning commutativity. Recent work in this direction and more references can be found e.g. in [10].

Generally the decomposition of a language can be chosen in very different ways and it turns out to be somewhat difficult to find languages without any prime decompositions. We give a construction of a nonregular language that provably does not have any prime decomposition.

[*] The author was supported by the Natural Sciences and Engineering Research Council of Canada Grant OGP0147224.

[**] The author was supported under the Research Grants Council of Hong Kong Competitive Earmarked Research Grant HKUST6197/01E.

O.H. Ibarra and Z. Dang (Eds.): DLT 2006, LNCS 4036, pp. 145–155, 2006.

We consider also a stronger factorization property that requires that any refinement of a decomposition of the language leads to a prime decomposition in a finite number of steps. We call such languages strongly prime decomposable. We give necessary and sufficient conditions for a regular language to be strongly prime decomposable. The characterization establishes that the property is decidable for regular languages.

Using the characterization of the strongly prime decomposable languages we show that every regular language over a unary alphabet has a prime decomposition. As a by-product of the proof we establish the existence of prime decompositions for context-free languages over arbitrary alphabets where, roughly speaking, the set of "short words" of the corresponding length set is not closed under any multiple of the cycle of the length set.

The main open question remaining is whether all regular languages have prime decompositions.

2 Language Decompositions

Let Σ be a finite alphabet. A language is any subset of Σ^*. The length of a word $w \in \Sigma^*$ is $|w|$. The catenation of languages L_1 and L_2 over Σ is $L_1 \cdot L_2 = \{w \in \Sigma^* \mid (\exists u_i \in L_i, i = 1, 2)\ w = u_1 u_2\}$. For all unexplained notions in language theory we refer the reader e.g. to [9, 16, 17].

We say that a language L has a non-trivial decomposition if we can write $L = A \cdot B$ where $A, B \neq \{\varepsilon\}$. In the following, unless otherwise mentioned, by a decomposition or a factorization of a language we always mean a non-trivial decomposition.

A nonempty language $L \neq \{\varepsilon\}$ is said to be *prime* if L has no decompositions. For a given regular language L it is decidable whether or not L has a decomposition [11, 12], i.e., whether or not L is prime. More generally, the regular language decomposition problem is decidable for all operations defined by letter-bounded regular sets of trajectories [5].

Definition 2.1. [12] *A* prime decomposition *of a language L is a factorization*

$$L = L_1 \cdot \ldots \cdot L_m, \tag{1}$$

where each of the languages L_i, $i = 1, \ldots, m$, is prime.

A language, unlike an integer, can have also infinite factorizations, that is, decompositions into an infinite product of nontrivial factors. Here we restrict consideration to decompositions having finitely many components. Infinite factorizations obviously would involve interesting and different types of questions.

A finite language (distinct from $\emptyset$, $\{\varepsilon\}$) clearly always has a prime decomposition. On the other hand, a prime decomposition need not be unique even for finite languages [12]. Any prefix-free regular language has a unique decomposition in terms of prime languages if it is additionally required that the components are regular and prefix-free [3, 7]. Interestingly, the analogous property does not hold for decompositions of infix-free regular languages [6]. A factorial language is a language that is closed under the subword operation. In [1] it is shown that a factorial language has a unique canonical decomposition, where the components satisfy certain minimality conditions, into indecomposable factorial components.

Example 2.1. Let $H \subseteq \Sigma^n$, $n \geq 1$, be a set of words of length n. We show that H^* has the following prime decomposition

$$H^* = (\{\varepsilon\} \cup H) \cdot (\bigcup_{i=1}^{\infty} H^{2i-1} \cup \{\varepsilon\}) \tag{2}$$

Since the equality obviously holds, it is sufficient to verify that the two factors on the right side are prime.

In any decomposition $\{\varepsilon\} \cup H = AB$ both of the sets A and B must contain ε. Then the equality can hold only if one of A and B contains all words of H and the other set is $\{\varepsilon\}$, that is, $\{\varepsilon\} \cup H$ has only trivial decompositions.

In order to see that the second language on the right side of (2) is prime, assume that we can write

$$\bigcup_{i=1}^{\infty} H^{2i-1} \cup \{\varepsilon\} = AB \tag{3}$$

for some $A, B \subseteq \Sigma^*$. Again ε has to be in both A and B. Thus A or B cannot contain any nonempty words shorter than n and all words of H must be in A or B. If both A and B contain words of H then AB would have some word of length $2n$. We assume that $H \subseteq A$, the other possibility being symmetric. Again all words of H^3 must be in A or B, and similarly as above we see that the only possibility is that $H^3 \subseteq A$ since otherwise the catenation of A and B would have some word of length $4n$. By induction it follows that $A = \bigcup_{i=1}^{\infty} H^{2i-1} \cup \{\varepsilon\}$ and $B = \{\varepsilon\}$.

It seems that earlier work [12] did not expect that the Kleene-star of languages as in Example 2.1 could have prime decompositions. In fact, we do not know any regular language L such that L provably has no prime decompositions. In Section 4 we show that every regular language over a unary alphabet has a prime decomposition.

Next we show that there exist nonregular languages without any prime decompositions. Let $\Sigma = \{a, b\}$. We define $H_0 \subseteq \Sigma^*$ as follows:

$$H_0 = \{a^{i_1}b^{i_1}a^{i_2}b^{i_2} \cdots a^{i_k}b^{i_k} \mid k \geq 0, 1 \leq i_1 < i_2 < \ldots < i_k\}.$$

Lemma 2.1. *The language H_0 does not have any prime decomposition.*

Proof. Consider an arbitrary decomposition of H_0,

$$H_0 = L_1 \cdot \ldots \cdot L_m, \tag{4}$$

$m \geq 1$. For the sake of contradiction assume that (4) is a prime decomposition.

By the maximal *ab*-prefix, *mab-prefix,* (respectively, *mab*-suffix) of a word w we mean the longest prefix (respectively, longest suffix) of w that is in a^*b^*.

Consider a fixed $i \in \{1, \ldots, m\}$. We claim that if the *mab*-prefix of some word in L_i is of a form

$$a^j b^k, \ j \neq k, \ j, k \geq 0, \tag{5}$$

then all words in L_i must have the same *mab*-prefix $a^j b^k$. This follows from the observation that if L_i has two words u_1, u_2 where the *mab*-prefix u_{mp} of u_1 is as in (5) and the *mab*-prefix of u_2 is distinct from u_{mp}, then for any fixed $v \in L_1 \cdots L_{i-1}$ and $w \in L_{i+1} \cdots L_m$ only one of the words vu_1w and vu_2w can be in H_0.

Now if all words in L_i have the same *mab*-prefix as in (5) (which is not the empty word since $j \neq k$) we get a decomposition for L_i by factoring out the common prefix.

Since L_i is prime, the above means that we need to consider only the case where the *mab*-prefix of all words in L_i, $i = 1, \ldots, m$, is of a form $a^j b^j$, $j \geq 1$. (Note that in this case the *mab*-prefixes need not be identical, e.g., it is possible that $L_i = \{a^j b^j a^{j+1} b^{j+1}, a^j b^j, a^{j+1} b^{j+1}, \varepsilon\}$.) With a completely symmetric argument we see that the same property holds for *mab*-suffixes.

By a balanced word we mean a word of the form $a^j b^j$, $j \geq 0$. From the above we can conclude that for all $i \in \{1, \ldots, m\}$,

$$\text{the } mab\text{-prefix and the } mab\text{-suffix of any word in } L_i \text{ is balanced.} \tag{6}$$

Thus all words occurring in L_i, $1 \leq i \leq m$, are of the form

$$w_i = a^{k_{1,i}} b^{k_{1,i}} \cdots a^{k_{r,i}} b^{k_{r,i}}, \ \ 0 < k_{1,i} < \ldots < k_{r,i}, \ \ r \geq 0. \tag{7}$$

Now if we consider an arbitrary word $w_{i+1} = a^{k_{1,i+1}} b^{k_{1,i+1}} \cdots a^{k_{s,i+1}} b^{k_{s,i+1}} \in L_{i+1}$, the equation $k_{r,i} < k_{1,i+1}$ has to hold since otherwise $w_i w_{i+1}$ cannot occur as a subword of a word in H_0.

Now the equation (4) implies that, for all $i = 1, \ldots m-1$, there exist integers M_i and N_i ($M_1 = 1$, $N_i = M_{i+1} - 1$) such that L_i consists of exactly all the words as in (7) where $M_i \leq k_{1,i}$ and $k_{r,i} \leq N_i$, and L_m consists of all words as in (7) where $k_{1,i} > N_{m-1}$.

It follows that (4) is not a prime decomposition since, for example,

$$L_m = \{\varepsilon, a^{N_{m-1}+1} b^{N_{m-1}+1}\} \cdot A,$$

where A consists of all words as in (7) where $k_{1,i} > N_{m-1} + 1$. This concludes the proof. □

The language H_0 used in Lemma 2.1 is not context-free but its complement is context-free. It should be noted that Lemma 2.1 does not require any assumptions concerning the component languages, that is, H_0 doesn't have a prime decomposition even if the components could be non-recursively enumerable languages.

We conclude with the following question.

Open Problem. *Does there exist a context-free (or even a regular) language L such that L has no prime decomposition.*

3 Strong Prime Decomposition Property

In the previous section we saw (in Example 2.1) that regular languages can have artificial prime decompositions even if the natural way of decomposing the language does not result in a prime decomposition, i.e., the components could always be factorized further.

Example 3.1. Let $L = \varepsilon + a^2a^*$. We note that $L = L \cdot L$ or $L = (\varepsilon + a^2) \cdot L$ so obviously L has many different factorizations with arbitrarily many components. However, L has also the following prime decomposition

$$(\varepsilon + a^2)(\varepsilon + a^3)(\varepsilon + \bigcup_{i=1}^{\infty} (a^2)^{2i-1}).$$

Note that the last component is an instance of the left side of (3) that was shown to be prime in Example 2.1.

Here we consider a stronger version of the prime decomposition property that prevents situations as in Example 3.1.

Definition 3.1. *Let $L \subseteq \Sigma^*$. The* index *of a non-trivial decomposition of L,*

$$L = L_1 \cdot \ldots \cdot L_m \tag{8}$$

is m. The decomposition index *of L is the maximum index of any non-trivial decomposition of L if the maximum exists. Otherwise, we say that the decomposition index of L is infinite.*

If a language L has a finite decomposition index, we say that L is *strongly prime decomposable*. When L is strongly prime decomposable, any way of iteratively decomposing L has to stop after a finite number of steps, i.e., the refinement of any decomposition results in a prime decomposition in a finite number of steps.

Clearly all finite languages are strongly prime decomposable since the decomposition index of a finite language L is at most the length of the longest word in L. The language L considered in Example 3.1 has a prime decomposition but it is not strongly prime decomposable. An example of a strongly prime decomposable infinite language is $a^* + b^*$. This follows from Theorem 3.1 below.

For presenting a characterization of the strongly prime decomposable regular languages we recall some notation and a result from [12, 14]. Let $A = (Q, \Sigma, \delta, q_0, Q_F)$ be a deterministic finite automaton (DFA). For a subset $P \subseteq Q$ we define the languages

$$R_1^P = \{w \in \Sigma^* \mid \delta(q_0, w) \in P\},$$

$$R_2^P = \bigcap_{p \in P} \{w \in \Sigma^* \mid \delta(p, w) \in Q_F\}.$$

Proposition 3.1. [12] *Let $A = (Q, \Sigma, \delta, q_0, Q_F)$ be the minimal DFA for a language L and assume that we can write $L = L_1L_2$. Then*

$$L = R_1^P R_2^P,$$

where $P \subseteq Q$ is defined by

$$P = \{p \in Q \mid (\exists w \in L_1)\ \delta(q_0, w) = p\}.$$

Furthermore, we know that $L_i \subseteq R_i^P$, $i = 1, 2$.

Theorem 3.1. *A regular language L is not strongly prime decomposable if and only if there exist regular languages H_1, H_2, H_3, where H_2 contains some nonempty word such that*

$$L = H_1(H_2)^*H_3. \tag{9}$$

Proof. The "if"-direction follows from the observation that, for any $k \geq 1$, the equation (9) gives for L a decomposition of index at least k:

$$L = H_1(H_2 \cup \{\varepsilon\})^{k-1}(H_2)^*H_3. \tag{10}$$

(The index of the decomposition (10) is between k and $k+2$ depending on whether or not H_1 or H_3 is the trivial language $\{\varepsilon\}$.)

Next we prove the "only-if"-direction. Let $A = (Q, \Sigma, \delta, q_0, Q_F)$ be the minimal DFA for L. Since L is not strongly prime decomposable, we can write

$$L = L_1L_2 \cdot \ldots \cdot L_m,$$

where $m = 2^{|Q|} + 1$ and $L_i \neq \{\varepsilon\}$, $i = 1, \ldots, m$. Furthermore, by [12] (Proposition 3.1 above) we know that the languages L_i can be chosen to be regular.

Define $P_i = \{p \in Q \mid (\exists w \in L_1 \cdot \ldots \cdot L_i)\ \delta(q_0, w) = p\}$, $i = 1, \ldots m-1$. By Proposition 3.1,

$$L = R_1^{P_i} R_2^{P_i}, \quad i = 1, \ldots, m-1. \tag{11}$$

Here $R_j^{P_i}$, $j = 1, 2$, is as defined in Proposition 3.1.

Since $m-1 \geq 2^{|Q|}$ and $P_i \neq \emptyset$, $i = 1, \ldots, m-1$, there exist $j, k \in \{1, \ldots, m-1\}$, $j < k$, such that $P_j = P_k$. This means that for all $p \in P_j$ and $w \in L_{j+1} \cdot \ldots \cdot L_k$ we have

$$\delta(p, w) \in P_j\ (= P_k).$$

Thus (11) implies that for all $r \geq 1$,

$$R_1^{P_j}(L_{j+1} \cdot \ldots \cdot L_k)^r R_2^{P_j} \subseteq L.$$

Consequently, $L = R_1^{P_j}(L_{j+1} \cdot \ldots \cdot L_k)^* R_2^{P_j}$ and $L_{j+1} \cdot \ldots \cdot L_k$ is not empty or $\{\varepsilon\}$ since $j < k$. □

It is known that primality is decidable for regular languages [12]. As a corollary of the proof of Theorem 3.1 we see that also the strong prime decomposition property is decidable for regular languages.

Corollary 3.1. *Given a regular language L it is decidable whether or not L is strongly prime decomposable.*

Proof. Let $A = (Q, \Sigma, \delta, q_0, Q_F)$ be the minimal DFA for L. In the "only if" part of the proof of Theorem 3.1 it is established that if L is not strongly prime decomposable, there exist $P \subseteq Q$ and a nonempty language $K_P \neq \{\varepsilon\}$ such that $L = R_1^P R_2^P$ (where R_j^P, $j = 1, 2$, is defined as in Proposition 3.1) and $\delta(p, w) \in P$ for all $p \in P$ and $w \in K_P$. Conversely, the existence of P and K_P as above implies that $L = R_1^P (K_P)^* R_2^P$ and hence, by the first part of Theorem 3.1, L is not strongly prime decomposable.

Given $P \subseteq Q$, a language K_P as above exists if and only if some nonempty word of length at most $s = |Q|^{|P|}$ takes each state of P to a state in P. Note that if this property holds for some word of length greater than s, using a pumping argument it follows that the property has to hold for a word of length at most s. Hence we can determine whether P and K_P as above exist by testing the required property for all subsets of Q. □

The algorithm given by Corollary 3.1 is extremely inefficient since it relies on an exhaustive search of subsets of the state set of the minimal DFA for L. It is probable that an efficient (e.g. a polynomial time) algorithm cannot be found since there is no known polynomial time algorithm even to test primality of a regular language [12].

4 Unary Regular Languages

We want to show that every regular language over a unary alphabet has a prime decomposition. First we recall some terminology concerning regular languages over a unary alphabet. A standard reference is [2], and references to more recent work on unary regular languages can be found e.g. in [4, 8].

A DFA A with a unary input alphabet can be divided into a *tail* which has the states that are not reachable from themselves with any non-empty word, and the *cycle* consisting of the remaining states of A. Naturally, A has no accepting states in the cycle if the language recognized by it is finite. If A is minimal, it is additionally required that all states are pairwise inequivalent. If the tail of A accepts words $a^{j_1}, \ldots a^{j_{r-1}}$ and the length of the cycle of A is m, the language accepted by A is denoted by a regular expression

$$a^{j_1} + \ldots + a^{j_{r-1}} + a^{j_r}(a^{i_1} + \ldots a^{i_{s-1}})(a^m)^*, \tag{12}$$

$0 \leq j_1 < \ldots j_{r-1} < j_r$, $0 \leq i_1 < \ldots < i_{s-1} < m$, $r, s \geq 0$. We use the names "tail" and "cycle" also when referring to the corresponding parts of a regular expression as in (12).

Lemma 4.1. *Let $L \subset \{a\}^*$ be any unary language. Then L^* is the union of a finite language and a linear language, that is, $L^* = F \cup \{a^{i \cdot p} \mid i \geq 0\}$ where $p \geq 0$ and $F \subseteq \{a\}^*$ is finite. Furthermore, p divides the length of any word in F.*

Proof. If L is empty or $L = \{\varepsilon\}$, the property holds by choosing $F = \emptyset$ and $p = 0$. Otherwise, if p is the greatest common divisor of the lengths of all words in L, there exists $M_p \geq 1$ such that for all $n > M_p$, $a^n \in L$ if and only if n is a multiple of p. We can choose F as the set of all words in L of length at most M_p. The length of any word in F is divided by p. □

Lemma 4.2. *Let $L \subseteq \{a\}^*$ be a regular language such that*

$$L = LR^* \tag{13}$$

where R contains a nonempty word. Then L has a prime decomposition.

Proof. Let L be denoted by a regular expression as in (12). By factoring out the shortest word we can assume without loss of generality that $\varepsilon \in L$, that is, $j_1 = 0$. We assume that m (using the notations of (12)) is the cycle length of the minimal DFA for L and all words ε, a^{j_2}, $\ldots a^{j_{r-1}}$, $a^{j_r+i_1}$, $\ldots$, $a^{j_r+i_{s-1}}$ are pairwise inequivalent. These properties hold if the tail and cycle of (12) are as in the minimal DFA for L. Note that (13) implies that L is infinite and hence the minimal DFA has a cycle containing an accepting state, that is, $m \geq 1$.

By Lemma 4.1 we can write

$$R^* = \varepsilon + a^{k_1} + \ldots + a^{k_{t-1}} + a^{k_t}(a^n)^*, \tag{14}$$

where $0 < k_1 < \ldots < k_t$, $t \geq 1$, are all multiples of n. Here we require that $k_t \geq 1$ and as the word a^{k_t} we can choose the first nonempty word that is in the cycle of R^*. (The expression (14) does not need to correspond to the minimal DFA for R^*. This would be the case, for example, if the minimal DFA is cyclic, i.e., it has no tail.) Since R contains a nonempty word, it follows that $n \geq 1$.

By (13), $uv \in L$ for all $u \in L$ and $v \in R^*$. Since m is the cycle length of the minimal DFA for L, this implies that m divides n, and consequently the length of any word in R^* is a multiple of m. Write

$$a^{k_t} = c \cdot m, \quad c \geq 1.$$

Then

$$\begin{aligned} L = (\varepsilon + a^{j_2} + \ldots + a^{j_{r-1}} + a^{j_r}(a^{i_1} + \ldots + a^{i_{s-1}} + a^{i_1+m} + \ldots \\ + a^{i_{s-1}+m} + \ldots + a^{i_1+(c-1)m} + \ldots + a^{i_{s-1}+(c-1)m}))(a^{k_t})^*. \end{aligned} \tag{15}$$

In (15) the inclusion from right to left follows by (13) since all words in the first factor are in L and $(a^{k_t})^* \subseteq R^*$ because k_t is a multiple of n. The inclusion from left to right follows using the simple observation that the right side of (15) is obtained from the regular expression (12) for L with cycle length m by repeating the original cycle c times and taking $c \cdot m$ to be the new cycle length.

In the right side of (15) the first component has a prime decomposition since it is a finite language. The second component has a prime decomposition by Example 2.1. □

The construction of Lemma 4.2 is illustrated in the next example. In particular, the example shows that in the factorization (15) we could not use $(a^n)^*$ as a factor for L where n is the cycle length of the minimal DFA for R^*.

Example 4.1. Let

$$L = \varepsilon + a^5 + a^{12} + a^{17}(a^3)^* + a^{18}(a^3)^*,$$

and let $R = (a^{12} + a^{18})^*$. Now $L = LR^*$ and the the construction from the proof of Lemma 4.2 gives for L the factorization

$$L = (\varepsilon + a^5 + a^{12} + a^{17} + a^{18} + a^{20} + a^{21} + a^{23} + a^{24} + a^{26} + a^{27})(a^{12})^*.$$

It can be noted that the cycle length of R^* is 6. However, $(a^6)^*$ is not a factor of L since $\varepsilon, a^5 \in L$ and $a^6, a^{11} \notin L$.

Theorem 4.1. *Every regular language over a unary alphabet has a prime decomposition.*

Proof. Let $L \subseteq \{a\}^*$ be regular. If we can write $L = L_1(L_2)^*$ for regular languages L_1 and L_2, where L_2 contains a nonempty word, then also $L = L(L_2)^*$ holds and, by Lemma 4.2, L has a prime decomposition.

If there exist no regular languages L_i, $i = 1, 2$, $L_2 \neq \{\varepsilon\}$, $L_2 \neq \emptyset$, such that $L = L_1(L_2)^*$, then using the commutativity of catenation of unary languages and Theorem 3.1 we get that L is strongly prime decomposable. □

Let Σ be an arbitrary finite alphabet and $L \subseteq \Sigma^*$. The *length set* of L is the language over the unary alphabet $\{a\}$ defined by

$$\text{length}(L) = \{a^k \mid (\exists w \in L)\, |w| = k\}.$$

A language L over a non-unary alphabet may have more structure than the corresponding length set and decompositions of the length set of L do not necessarily yield a factorization of L. For example, the language $\{bc, cb\}$ is prime but its length set has the factorization $\{aa\} = \{a\} \cdot \{a\}$. Conversely, however, corresponding to any decomposition of L there exists a decomposition of the length set of L. This gives the following lemma.

Lemma 4.3. *Let Σ be a finite alphabet and $L \subseteq \Sigma^*$. If* length(L) *is strongly prime decomposable, then the same holds for L.*

Proof. If L has a non-trivial decomposition $L = L_1 \cdot L_2$, then length$(L_1) \cdot$ length(L_2) is a non-trivial decomposition of length(L). Hence, if L has an infinite decomposition index, the same holds for length(L). In other words, if length(L) is strongly prime decomposable, so is L. □

The result of Lemma 4.3 can be used to show the existence of prime decompositions for context-free languages where the tail of the length set is "not closed" under any multiple of the cycle length of the minimal DFA for the length set. Note that the length set of a context-free language is always regular [9, 16].

Theorem 4.2. *Let L be a context-free language and let m be the cycle length of the minimal DFA for* $\text{length}(L)$*. If for some $d \geq 0$ and $M_d \geq 1$, $a^d \in \text{length}(L)$ and, for all $i \geq M_d$, $a^{d+i \cdot m} \notin \text{length}(L)$, then L has a prime decomposition.*

Proof. Assume that $\text{length}(L)$ has a decomposition $\text{length}(L) = MR^*$ in terms of regular languages M and R, where R contains a nonempty word. Then $\text{length}(L) = \text{length}(L)R^*$ and, by the proof of Lemma 4.2, we know that there is a constant c such that $a^d \in \text{length}(L)$ implies that, for all $i \geq 1$, $a^{d+i \cdot c \cdot m}$ is in $\text{length}(L)$. This contradicts the assumptions for $\text{length}(L)$.

Hence there do not exist regular languages M and R, $R \neq \emptyset$, $R \neq \{\varepsilon\}$, such that $\text{length}(L) = MR^*$. By Theorem 3.1, $\text{length}(L)$ is strongly prime decomposable and Lemma 4.3 implies that also L is strongly prime decomposable. □

The conditions of Theorem 4.2 apply, for example, to any context-free language L such that L has a word of odd length and there exists a constant $M_L \geq 1$ such that all words of L of length greater than M_L have even length. The assumption that L is context-free is needed to guarantee that the length set of the language is regular.

Finally, we note that Theorem 4.1 cannot be extended for arbitrary unary languages. Recently, Rampersad and Shallit [13], and A. Salomaa and Yu [15], have independently given examples of non-regular unary languages that provably do not have a prime decomposition. The first mentioned language consists of all words over a unary alphabet whose length when represented in ternary notation does not contain a 2. The second mentioned language consists of all words whose length when represented in binary has no 1's in odd positions from the right. These languages are higher in the complexity hierarchy than the language of Lemma 2.1 in the sense that they are not co-context-free.

5 Conclusions

We have established an effective characterization of the strongly prime decomposable regular languages. Using the characterization it is easy to construct regular languages (over a unary or a non-unary alphabet) that are not strongly prime decomposable, i.e., that have an infinite decomposition index. We have shown that every regular language over a unary alphabet has a prime decomposition. The main open problem remaining is whether all regular languages over arbitrary alphabets have at least one prime decomposition. We conjecture a positive answer to this question.

References

1. Avgustinovich, S.V., Frid, A.E.: A unique decomposition theorem for factorial languages. Intern. J. of Algebra and Computation **15** (2005) 149–160
2. Chrobak, M.: Finite automata and unary languages. Theoret. Comput. Sci. **47** (1986) 149–158

3. Czyzowicz, J., Fraczak, W., Pelc, A., Rytter, W.: Linear-time prime decomposition of regular prefix codes. Internat. J. Foundations of Computer Science **14** (2003) 1019–1031
4. Domaratzki, M., Ellul, K., Shallit, J., Wang, M.-W.: Non-uniqueness and radius of cyclic unary NFAs. Internat. J. Foundations of Computer Science **16** (2005) 883–896
5. Domaratzki, M., Salomaa, K.: Decidability of trajectory based equations. Theoret. Comput. Sci. **345** (2005) 304–330
6. Han, Y.-S., Wang, Y., Wood, D.: Infix-free regular expressions and languages, Internat. J. Foundations of Computer Science, to appear.
7. Han, Y.-S., Wood, D.: The generalization of generalized automata: Expression automata. In: Domaratzki, M., Okhotin, A., Salomaa, K., Yu, S. (eds.): Implementation and Application of Automata, CIAA'04. Lecture Notes in Computer Science, Vol. 3317. Springer-Verlag (2005) 156–166
8. Holzer, M., Kutrib, M.: Unary language operations and their nondeterministic state complexity. In: Ito, M., Toyama, M. (eds.): Developments in Language Theory, DLT'02. Lecture Notes in Computer Science, Vol. 2450. Springer-Verlag (2003) 162–172
9. Hopcroft, J.E., Ullman, J.D.: Introduction to Automata Theory, Languages, and Computation. Addison-Wesley Publishing Company (1979)
10. Karhumäki, J.: Finite sets of words and computing. In: Margenstern, M. (ed.): Machines, Computations and Universality, MCU04. Lecture Notes in Computer Science, Vol. 3354. Springer-Verlag (2005) 36–49
11. Kari, L., Thierrin, G.: Maximal and minimal solutions to language equations, J. Comput. System Sci. **53** (1996) 487–496
12. Mateescu, A., Salomaa, A., Yu, S.: Factorizations of languages and commutativity conditions. Acta Cybernetica **15** (2002) 339–351
13. Rampersad, N., Shallit, J.: private communication.
14. Salomaa, A., Yu, S.: On the decomposition of finite languages. In: Rozenberg, G., Thomas, W. (eds.): Developments in Language Theory, DLT'99. World Scientific (2000) 22–31
15. Salomaa, A., Yu, S.: private communication.
16. Wood, D.: Theory of Computation. John Wiley & Sons, New York, NY, (1987)
17. Yu, S.: Regular languages. In: Rozenberg, G., Salomaa, A. (eds.): Handbook of Formal Languages, Vol. I. Springer-Verlag (1997) 41–110