Available online at www.sciencedirect.com
ScienceDirect
Theoretical Computer Science 376 (2007) 60–69
Theoretical Computer Science
www.elsevier.com/locate/tcs

On the existence of prime decompositions

Yo-Sub Han[a], Arto Salomaa[b], Kai Salomaa[c,*], Derick Wood[d,e], Sheng Yu[e]

[a] System Technology Division, Korea Institute of Science and Technology, P.O. Box 131, Cheongryang, Seoul, Republic of Korea
[b] Turku Centre for Computer Science, Joukahaisenkatu 3–5 B, 20520 Turku, Finland
[c] School of Computing, Queen's University, Kingston, Ontario K7L 3N6, Canada
[d] Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[e] Department of Computer Science, University of Western Ontario, London, Ontario N6A 5B7, Canada

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 it 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 non-regular unary languages that do not have prime decompositions. We also consider infinite factorizations of unary languages.

Keywords: Language decompositions; Primality; Unary languages

1. Introduction

A non-empty language is said to be prime [15,20] 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 [15]. 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 [5,9]. Decompositions of factorial languages, that is, languages closed under the subword operation, are investigated in [1]. Unambiguous square roots of regular and more general languages have been studied in [2].

Any finite language always has a prime decomposition, although it need not be unique [15,20]. Work on factorizations of finite languages leads to non-trivial questions concerning commutativity. Recent work in this direction and more references can be found, e.g., in [14]. 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.

[*] Corresponding author.
E-mail addresses: emmous@kist.re.kr (Y.-S. Han), asalomaa@utu.fi (A. Salomaa), ksalomaa@cs.queensu.ca (K. Salomaa), dwood@cs.ust.hk (D. Wood), syu@csd.uwo.ca (S. Yu).

0304-3975/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2007.01.013

We consider 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 can show that languages (over arbitrary alphabets) where the words satisfy certain length conditions always have a prime decomposition. On the other hand, we show that there exist even unary languages that provably have no prime decompositions. A different construction of a unary language without any prime decomposition has been obtained independently in [17].

We consider properties of infinitary decompositions of unary languages and the uniqueness of such factorizations. We construct a unary language where the infinitary prime factorization is not unique even if we disregard the order of the components.

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 denoted $|w|$. The catenation of languages L_1 and L_2 over Σ is $L_1 \cdot L_2 = \{w \in \Sigma^* \mid (\exists u_i \in L_i, i = 1, 2)\ w = u_1u_2\}$. For all unexplained notions in language theory we refer the reader, for example, to [12,18,22,23].

We say that a non-empty 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 non-empty 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 [4,15], 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 [7].

Definition 2.1 (*[15]*). 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 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 [15]. The situation is essentially different if we restrict consideration to prefix-free languages since it is known that the monoid of prefix codes is a free monoid [16] (see also [13,21]). 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 [5,10]. Interestingly, the analogous property does not hold for decompositions of infix-free regular languages [9].

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

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

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

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

for some $A, B \subseteq \Sigma^*$. Again ε has to be in both A and B. Thus A or B cannot contain any non-empty 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 $B = \{\varepsilon\}$.

Generally it is not easy to find languages that do not 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. On the other hand, in Section 6 we show that there exist non-regular unary languages with no finitary prime decomposition.

A language, unlike an integer, can also have infinitary factorizations, that is, decompositions into an infinite product of non-trivial factors. Unless otherwise mentioned, by a prime factorization we mean a decomposition as given in Definition 2.1. Below we introduce some definitions concerning infinitary (prime) decompositions. Infinitary decompositions will be considered in Section 6.

It is shown in [8] that the language

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

does not have a prime factorization. (The language H_0 is not context-free but its complement is context-free.) This reflects the fact that Definition 2.1 requires the prime factorization to be finitary. If this requirement is relaxed, we can write

$$H_0 = \prod_{i=1}^{\infty} (\varepsilon + a^i b^i),$$

where each factor is prime.

When we consider infinite products $\prod_{i=1}^{\infty} L_i$, where each L_i is a language, we assume that each L_i contains the empty word. Indeed, an infinite product of languages defines finite words only if all of these languages, with at most finitely many exceptions, contain the empty word. In this case there is a language K and an integer $m \geq 1$ such that the original product can be written as

$$\prod_{i=1}^{\infty} L_i = K \prod_{i=m}^{\infty} L_i,$$

where each language in the product on the right side contains the empty word.

In the following definition we assume that each of the languages L_i and K_i *properly* contains the empty word.

Definition 2.2. A language L has a *unique infinitary prime factorization* if $L = \prod_{i=1}^{\infty} L_i$, where each L_i is prime and, whenever $L = \prod_{i=1}^{\infty} K_i$, where each K_i is prime, then $L_i = K_i$, for all i. If L is over a one-letter alphabet, it is only required that the languages K_i are the languages L_i in some order.

Since languages over one letter are commutative, the relaxation of uniqueness given in the definition is very natural. A language can have both a (finite) prime factorization and an infinitary prime factorization. For instance, as seen above, Σ^* has a prime factorization. It has also an infinitary prime factorization

$$\Sigma^* = \prod_{w} (\varepsilon + w),$$

where w runs through all non-empty words over Σ.

Question. Can a language have both a prime factorization and a unique infinitary prime factorization?

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

$$(\varepsilon + a^2)(\varepsilon + a^3)\left(\varepsilon + \bigcup_{i=1}^{\infty} a^{4i-2}\right).$$

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,

$$L = L_1 \cdot \ldots \cdot L_m \tag{4}$$

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 [15,20]. 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

$$R_1^P = \{w \in \Sigma^* \mid \delta(q_0, w) \in P\}, \quad R_2^P = \bigcap_{p \in P} \{w \in \Sigma^* \mid \delta(p, w) \in Q_F\}.$$

Proposition 3.1 *([15]). 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, the inclusion $L_i \subseteq R_i^P$ holds, $i = 1, 2$.

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 non-empty word such that*

$$L = H_1 H_2^* H_3. \tag{5}$$

Proof. The "if"-direction follows from the observation that, for any $k \geq 1$, the Eq. (5) gives for L a decomposition of index at least k:

$$L = H_1(H_2 \cup \{\varepsilon\})^{k-1}(H_2)^* H_3. \tag{6}$$

(The index of the decomposition (6) is between k and $k + 2$ depending on whether 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

$$L = L_1 L_2 \cdot \ldots \cdot L_m,$$

where $m = 2^{|Q|} + 1$ and $L_i \neq \{\varepsilon\}$, $i = 1, \ldots, m$. Furthermore, by [15] (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,

$$L = R_1^{P_i} R_2^{P_i}, \quad i = 1, \ldots, m - 1. \tag{7}$$

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 (7) 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 [15]. As a corollary of the proof of Theorem 3.1 we see that the strong prime decomposition property is also 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 non-empty 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 non-empty 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 the primality of a regular language [15].

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. Some older references are [3,19], and references to more recent work on unary regular languages can be found, e.g., in [6,11].

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

$$a^{j_1} + \cdots + a^{j_{r-1}} + a^{j_r}(a^{i_1} + \cdots + a^{i_{s-1}})(a^m)^*, \tag{8}$$

$0 \leq j_1 < \cdots < j_{r-1} < j_r$, $0 \leq i_1 < \cdots < i_{s-1} < m$, $r, s \geq 0$. We use the names "tail" and "cycle" also when referring to the corresponding parts of a regular expression as in (8).

The lemma below is well-known but for the sake of completeness we include the short proof.

Lemma 4.1. *Let $L \subset \{a\}^*$ be any unary language. Then L^* is the union of a finite language and a language consisting of powers of some word, 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 divisible by p. ■

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

$$L = LR^* \tag{9}$$

where R contains a non-empty word. Then L has a prime decomposition.

Proof. Let L be denoted by a regular expression as in (8). 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 notation of (8)) is the cycle length of the minimal DFA for L and all words $\varepsilon, a^{j_2}, \ldots, a^{j_{r-1}}, a^{j_r+i_1}, \ldots, a^{j_r+i_{s-1}}$ are pairwise inequivalent. These properties hold if the tail and cycle of (8) are as in the minimal DFA for L. Note that (9) 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

$$R^* = \varepsilon + a^{k_1} + \cdots + a^{k_{t-1}} + a^{k_t}(a^n)^*, \tag{10}$$

where $0 < k_1 < \cdots < k_t$, $t \geq 1$, 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 non-empty word that is in the cycle of R^*. (The expression (10) 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 non-empty word, it follows that $n \geq 1$.

By (9), $uv \in L$ 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

$$\begin{aligned} L = {} & (\varepsilon + a^{j_2} + \cdots + a^{j_{r-1}} + a^{j_r}(a^{i_1} + \cdots + a^{i_{s-1}} + a^{i_1+m} \\ & + \cdots + a^{i_{s-1}+m} + \cdots + a^{i_1+(c-1)m} + \cdots + a^{i_{s-1}+(c-1)m}))(a^{k_t})^*. \end{aligned} \tag{11}$$

In (11) the inclusion from right to left follows from (9) 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 (11) is obtained from the regular expression (8) 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 (11) 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 (11) 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 non-empty word, then $L = L(L_2)^*$ also 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. ■

5. Length sets and prime decompositions

Here, using the results of Section 4, we give criteria that can be used to show that certain context-free languages are guaranteed to have prime decompositions. 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 5.1. *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 5.1 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 [18,22].

Theorem 5.1. *Let L be a context-free language and let m be the cycle length of the minimal DFA for* length(L). *If for some $d \geq 0$ and $M_d \geq 1$, $a^d \in$* length(L) *and, for all $i \geq M_d$, $a^{d+i\cdot m} \notin$* length(L), *then L is strongly prime decomposable.*

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

Hence there do not exist regular languages M and R, $R \neq \emptyset$, $R \neq \{\varepsilon\}$, such that length$(L) = MR^*$. By Theorem 3.1, length(L) is strongly prime decomposable and Lemma 5.1 implies that also L is strongly prime decomposable. ■

The conditions of Theorem 5.1 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.

6. Non-regular unary languages

We show that the result of Theorem 4.1 cannot be extended for arbitrary unary languages. We consider also infinitary factorizations of unary languages and give methods for constructing, for instance, languages possessing no unique infinitary prime factorization.

The following notion is a useful tool for our constructions.

Definition 6.1. The *binary indicator* $\beta(i)$ of a non-negative integer i is the (finite) set consisting of positive integers j such that the jth bit from the right in the binary representation of i equals 1.

Thus, $\beta(4) = \{3\}$, $\beta(19) = \{1, 2, 5\}$, $\beta(0) = \emptyset$. Clearly, β constitutes a bijection between non-negative integers and finite sets of positive integers. We may identify words a^i over $\{a\}$ with the exponent i or $\beta(i)$. Then the catenation $a^i a^j$ is associated with $\beta(i + j)$.

We are now ready for the result settling the existence of one-letter languages without a prime factorization.

Theorem 6.1. *There is a language over the alphabet $\{a\}$ with no prime factorization but with a unique infinitary prime factorization.*

Proof. We show that the language

$$L_1 = \{a^i \mid \text{ no odd number is in } \beta(i)\}$$

has the required properties. Thus,

$$\varepsilon,\ a^2,\ a^8,\ a^{10},\ a^{32},\ a^{34},\ a^{40}$$

are the seven shortest words in L_1. We shall establish the following assertion.

Claim. The sets $f(\nu) = \{\varepsilon, a^{2^{2\nu+1}}\}$, $\nu \geq 0$, constitute the collection of prime languages appearing in any decomposition of L_1.

To prove the Claim, we first observe that, for any ν, the set $f(\nu)$ is prime, and

$$L_1 = f(\nu) \prod_{\mu \neq \nu} f(\mu).$$

Indeed, $f(\nu)$ contributes the factor $a^{2^{2\nu+1}}$ to the words in L_1. If the factor is not needed, ε is taken from $f(\nu)$.

Let now $L_1 = AB$ be any non-trivial decomposition. Then the empty word is contained in both A and B and, hence, both A and B are subsets of L_1. Choose arbitrary words $a^x \in A$, $a^y \in B$, where $x, y \neq 0$. Both $\beta(x)$ and $\beta(y)$ consist of even numbers. Assume that some number occurs in both of them, and let $2i$ be the smallest of such numbers. But this means that the odd number $2i + 1$ is in $\beta(x + y)$. This is impossible, since $a^{x+y} \in L_1$. This contradiction shows that the sets $\beta(x)$ and $\beta(y)$ are disjoint. Consequently, also the sets

$$\beta(A) = \bigcup_{a^x \in A} \beta(x) \qquad \text{and} \qquad \beta(B) = \bigcup_{a^y \in B} \beta(y)$$

are disjoint. Hence, if $a^z \in A$, then also $a^{z_1} \in A$, whenever $\beta(z_1) \subseteq \beta(z)$. (This follows because $a^{z_1} \in L_1$ and $\beta(z_1) \cap \beta(B) = \emptyset$.) This means that A is of the form

$$A = A_1 \prod (\varepsilon + a^{2^{x-1}}),$$

where x runs through all elements of $\beta(z)$. (For instance, if $z = 162$, then $\beta(z) = \{2, 6, 8\}$ and we have $A = A_1(\varepsilon + a^2)(\varepsilon + a^{32})(\varepsilon + a^{128})$.) The same analysis applies to the language A_1, as well as B. The Claim now follows, since L_1 contains all words $a^{2^{2\nu+1}}$, $\nu \geq 0$.

But the Claim clearly implies Theorem 6.1. Every decomposition of L_1 can be continued up to prime factors, and their order is immaterial in the case of the alphabet $\{a\}$. ■

Instead of the language L_1, numerous other languages can be used for the proof of Theorem 6.1. The argument with the binary indicator β remains exactly the same if we consider the language of all words a^i where no even number is in $\beta(i)$, or the language of all words a^i where every number in $\beta(i)$ is an odd prime.

A different construction of a unary language admitting no prime decomposition has been given independently by Rampersad and Shallit [17]. The language used there consists of all words over a unary alphabet whose length when represented in ternary notation does not contain a 2.

We present, finally, a couple of somewhat more involved constructions. In the proof of Theorem 6.1, the cardinality of each of the prime factors equals 2. We now show that the minimal cardinality of the prime factors can be arbitrarily large.

Theorem 6.2. *Consider an arbitrary integer $k \geq 3$. There is a language L_k over the alphabet $\{a\}$ with no prime factorization but with a unique infinitary prime factorization, where each factor is of cardinality k.*

Proof. We again define a language L_k by imposing conditions on the binary indicator. We define L_k to consist of all words a^i such that: (i) no number jk, $j \geq 1$, is in $\beta(i)$, and (ii) for each $j \geq 1$, at most one of the numbers $jk - 1$, $jk - 2, \ldots,$ $jk - (k - 1)$ is in $\beta(i)$. Thus, we divide the sequence of positive integers into blocks of length k. At most one number from each block is in $\beta(i)$, and this number is never divisible by k. To improve readability, we assume in the sequel that $k = 5$. This is no loss of generality, since everything works in the same way in the general case.

Thus, for an arbitrary word a^i in our language, at most one of the numbers $1, 2, 3, 4$ is in $\beta(i)$, similarly at most one of the numbers $6, 7, 8, 9$ and at most one of the numbers $11, 12, 13, 14$. None of the numbers $5, 10, 15$ is in $\beta(i)$. Thus, the empty word is in our language, and the other words a^i in the language are obtained by catenating at most one word from each of the following sets:

$$\{a, a^2, a^4, a^8\},\ \{a^{32}, a^{64}, a^{128}, a^{256}\},\ \{a^{1024}, a^{2048}, a^{4096}, a^{8192}\}, \ldots.$$

This means that we have

$$L_5 = \prod_{\nu=0}^{\infty} (\varepsilon + a^{2^{5\nu}} + a^{2^{5\nu+1}} + a^{2^{5\nu+2}} + a^{2^{5\nu+3}}).$$

Clearly, each of the factors in the product is prime. For instance, if $C_1 = (\varepsilon + a + a^2 + a^4 + a^8) = AB$, then both A and B must contain ε and be subsets of C_1. This is possible only if one of them is trivial. The same argument applies to all factors. In an arbitrary decomposition $L_5 = AB$, we obtain $\beta(A) \cap \beta(B) = \emptyset$. (Indeed, we first consider words $a^x \in A$ and $a^y \in B$ as in the the proof of Theorem 6.1. The set $\beta(x) \cap \beta(y)$ cannot contain an element of the form $5j - 1$ because then $\beta(x + y)$ would contain an element of the form $5j$. Also elements of other forms in the set $\beta(x) \cap \beta(y)$ lead to a contradiction with the definition of the language L_5.) Consequently, any decomposition finally leads to the product above, where the prime factors are of cardinality 5 (or k). ■

Note that the language L_5 consists of all words a^i such that the binary representation of i is in the regular language

$$(00000 + 00001 + 00010 + 00100 + 01000)^*.$$

The results above reflect the rich possibilities in the construction of languages using the binary indicator. We construct, finally, a language having (no prime factorization and) no unique infinitary prime factorization. In fact, the language constructed has non-denumerably many infinitary prime factorizations.

Theorem 6.3. *There is a language K over the alphabet $\{a\}$ with no prime factorization and no unique infinitary prime factorization.*

Proof. We define K to consist of all words a^i such that the binary representation of i is in the regular language

$$(000 + 010 + 011 + 100 + 101 + 110 + 111)^*.$$

Equivalently, for all $j \geq 0$, whenever $3j + 1$ is in $\beta(i)$, then at least one of the numbers $3j + 2$ and $3j + 3$ is in $\beta(i)$. Thus, K misses words such as a and a^8. It is not difficult to see that K can be represented as the product

$$(\varepsilon + a^2 + a^3 + a^4 + a^5 + a^6 + a^7)(\varepsilon + a^{16} + a^{24} + \cdots + a^{56})(\varepsilon + a^{128} + \cdots + a^{448}) \ldots$$

or, formally,

$$K = \prod_{j=0}^{\infty} \left(\varepsilon + \sum_{k=2}^{7} a^{k2^{3j}} \right),$$

and that any decomposition of K leads to this product. But now the individual factors of the product are not prime. Each of them can be factorized in two different ways. For instance, the first factor equals

$$(\varepsilon + a^2 + a^3)(\varepsilon + a^3 + a^4) = (\varepsilon + a^2)(\varepsilon + a^3 + a^4 + a^5),$$

and the general factor equals

$$(\varepsilon + a^{2 \cdot 2^{3j}} + a^{3 \cdot 2^{3j}})(\varepsilon + a^{3 \cdot 2^{3j}} + a^{4 \cdot 2^{3j}}),$$

as well as

$$(\varepsilon + a^{2 \cdot 2^{3j}})(\varepsilon + a^{3 \cdot 2^{3j}} + a^{4 \cdot 2^{3j}} + a^{5 \cdot 2^{3j}}).$$

Hence, our theorem follows. ■

7. 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. On the other hand, we have constructed non-regular unary languages having no prime decompositions. We have considered infinitary prime factorizations in the context of unary languages. A topic for further research is to study infinitary decompositions over general alphabets.

The main open problem remaining is whether all regular, or even context-free, languages over arbitrary alphabets have at least one prime decomposition. We conjecture a positive answer for the case of regular languages.

References

[1] S.V. Avgustinovich, A.E. Frid, A unique decomposition theorem for factorial languages, Internat J. Algebr. Comput. 15 (2005) 149–160.

[2] A. Bertoni, P. Massazza, On the square root of languages, in: D. Krob, A.A. Mikhalev, A.V. Mikhalev (Eds.), Proc. of 12th International Conference FPSAC'00, Springer, 2000, pp. 125–134.

[3] M. Chrobak, Finite automata and unary languages, Theoret. Comput. Sci. 47 (1986) 149–158.

[4] J.H. Conway, Regular Algebra and Finite Machines, Chapman and Hall, 1971.

[5] J. Czyzowicz, W. Fraczak, A. Pelc, W. Rytter, Linear-time prime decomposition of regular prefix codes, Internat. J. Found. Comput. Sci. 14 (2003) 1019–1031.

[6] M. Domaratzki, K. Ellul, J. Shallit, M.-W. Wang, Non-uniqueness and radius of cyclic unary NFAs, Internat. J. Found. Comput. Sci. 16 (2005) 883–896.

[7] M. Domaratzki, K. Salomaa, Decidability of trajectory based equations, Theoret. Comput. Sci. 345 (2005) 304–330.

[8] Y.-S. Han, K. Salomaa, D. Wood, Prime decompositions of regular languages, in: O.H. Ibarra, Z. Dang (Eds.), Developments in Language Theory 2006, in: LNCS, vol. 4036, Springer, 2006, pp. 145–155.

[9] Y.-S. Han, Y. Wang, D. Wood, Infix-free regular expressions and languages, Internat. J. Found. Comput. Sci. 17 (2006) 379–393.

[10] Y.-S. Han, D. Wood, The generalization of generalized automata: Expression automata, in: M. Domaratzki, A. Okhotin, K. Salomaa, S. Yu (Eds.), Implementation and Application of Automata, CIAA'04, in: LNCS, vol. 3317, Springer, 2005, pp. 156–166.

[11] M. Holzer, M. Kutrib, Unary language operations and their nondeterministic state complexity, in: M. Ito, M. Toyama (Eds.), Developments in Language Theory, DLT'02, in: LNCS, vol. 2450, Springer, 2003, pp. 162–172.

[12] J.E. Hopcroft, J.D. Ullman, Introduction to Automata Theory, Languages, and Computation, Addison-Wesley Publishing Company, 1979.

[13] J. Karhumäki, Equations over finite sets of words and equivalence problems in automata theory, Theoret. Comput. Sci. 108 (1993) 103–118.

[14] J. Karhumäki, Finite sets of words and computing, in: M. Margenstern (Ed.), Machines, Computations and Universality, MCU04, in: LNCS, vol. 3354, Springer, 2005, pp. 36–49.

[15] A. Mateescu, A. Salomaa, S. Yu, Factorizations of languages and commutativity conditions, Acta Cybern. 15 (2002) 339–351.

[16] D. Perrin, Codes conjugués, Inform. and Control 20 (1972) 221–231.

[17] N. Rampersad, J. Shallit, Private communication, February 2006.

[18] G. Rozenberg, A. Salomaa (Eds.), Handbook of Formal Languages, vol. I, Springer-Verlag, 1997.

[19] A. Salomaa, Theorems on the representation of events in Moore-automata, Ann. Univ. Turku, Ser. AI 69 (1964).

[20] A. Salomaa, S. Yu, On the decomposition of finite languages, in: G. Rozenberg, W. Thomas (Eds.), Developments in Language Theory, DLT'99, World Scientific, 2000, pp. 22–31.

[21] H.J. Shyr, Free Monoids and Languages, 3rd ed., Hon Min Book Company, Taichung, Taiwan ROC, 2001.

[22] D. Wood, Theory of Computation, John Wiley & Sons, New York, NY, 1987.

[23] S. Yu, Regular languages, in: [18], pp. 41–110.