ELSEVIER

Discrete Applied Mathematics 117 (2002) 239–252

DISCRETE APPLIED MATHEMATICS

On the robustness of primitive words ☆

Gheorghe Păun[a], Nicolae Santean[b], Gabriel Thierrin[b,*], Sheng Yu[b]

[a] *Institute of Mathematics of the Romanian Academy, P.O. Box 1-764, 70700 Bucureşti, Romania*
[b] *Department of Computer Science, University of Western Ontario, London, Ontario, Canada N6A 5B7*

Received 23 June 1997; revised 18 September 2000; accepted 2 October 2000

Abstract

We investigate the effect on primitive words of *point mutations* (inserting or deleting symbols, substituting a symbol for another one), of *morphisms*, and of the operation of *taking prefixes*. Several ways of producing primitive words are obtained in this framework.

Keywords: Word; Language; Primitive; Density; Morphism; Insertion; Deletion

1. Primitive words; prerequisites

The primitive words, those which cannot be written as the power of another word, play an important role in formal language theory, coding theory, combinatorics on words, etc. In spite of the fact that many researches were devoted to them, still the subject is far from being exhausted. It is enough to mention the open problem whether or not the language of all primitive words over a given alphabet is context-free; see [2,5], and their references.

For an alphabet V, we denote by V^* the free monoid generated by V under the operation of concatenation; the empty word is denoted by λ and $V^* - \{\lambda\}$ is denoted by V^+. For $x \in V^*$, $|x|$ is the length of x, $|x|_a$ is the number of occurrences in x of the symbol $a \in V$, and $alph(x)$ is the set of symbols appearing in x. We also denote by $pref(x)$ the set of all non-empty prefixes of x. For a language $L \subseteq V^*$, we put $length(L) = \{|x| \mid x \in L\}$. The cardinality of a set X is denoted by $card(X)$.

For elementary notions and results of formal language theory, we refer to [7].

In what follows, V is always an alphabet containing *at least two symbols*.

☆ Research supported by Grants OGP0007877 and OGP0041630 of the Natural Sciences and Engineering Research Council of Canada.

* Correspondence address: Department of Mathematics, University of Western Ontario, London, Ontario, Canada N6A 5B7.

E-mail address: gab@csd.uwo.ca (G. Thierrin).

0166-218X/02/$ - see front matter
PII: S0166-218X(00)00338-3

A word $x \in V^+$ is said to be *primitive* if $x = u^n, u \in V^+$, implies $n = 1$ (hence $x = u$). Note the fact that the empty word is not primitive. For every word $x \in V^+$ there is a unique primitive word u and a unique integer $m \geqslant 1$ such that $x = u^m$; the word u is called the *root* of x.

The language of all primitive words over an alphabet V is denoted by Q_V; when V is understood, we write simply Q. The language of non-primitive words over V is denoted by Z_V (by Z when V is understood).

A language L is called *reflective* if $uv \in L$ implies $vu \in L$, for any $u, v \in V^*$.

Several facts about primitive words are known; we recall some of them which will be useful below; proofs can be found in [1,8].

Lemma 1. (i) *If* $uv = vu, u, v \in V^+$, *then* $u = p^m, v = p^n$, *for some* $p \in Q$ *and* $n, m \geqslant 1$.

(ii) (*The periodicity theorem of Fine and Wilf*). *If* u^m, v^n *have a common prefix of length at least* $|u| + |v| - 1$, *then* u *and* v *are powers of the same word.*

(iii) *The languages* Q *and* Z *are reflective*

Lemma 2. (i) *If* $u \in V^+$, $u \neq a^n$ *for some* $a \in V$, $n \geqslant 1$, *then at least one of the words* u, ua *is primitive.*

(ii) *If* $u, v \in Q, u \neq v$, *then* $u^m v^n \in Q$, *for all* $m \geqslant 2, n \geqslant 2$.

Point (i) in Lemma 2 has the following immediate consequence.

Corollary 1. *If* $u_1, u_2 \in V^+, u_1 u_2 \neq a^n, a \in V, n \geqslant 1$, *then at least one of* $u_1 u_2, u_1 a u_2$ *is primitive*

This corollary suggests the following problem: start with a primitive word w and try to obtain new primitive words by inserting symbols into w. More generally, consider *point mutations* (as, for instance, in the genetic area), that is insertions, deletions, or substitutions of a single symbol each. The next sections are devoted to investigating such operations. Still more generally, any operation with words can be considered from the point of view of primitivity preserving. Two natural cases investigated here are the morphisms and the operation *pref*, of taking prefixes.

2. Insertion, deletion and primitivity

We start by a result related to Lemma 2(i) and Corollary 1 above.

Proposition 1. *For every word* $u \in V^+$ *and every symbols* $a, b \in V, a \neq b$, *at least one of the words* ua, ub *is primitive.*

Proof. If $u = a^i, i \geqslant 0$, then ub is primitive, and if $u = b^i, i \geqslant 0$, then ua is primitive. Thus, we may assume that $alph(u)$ is different from both $\{a\}$ and $\{b\}$.

Suppose that none of ua, ub is primitive, hence there are $f, g \in Q$ such that $ua = f^n, ub = g^m$ for some $n \geqslant 2, m \geqslant 2$.

We cannot have $n = m$: if $n = m$, then $f = g$, which is impossible because $a \neq b$. Therefore, at least one of n, m must be strictly greater than 2. Let us assume that $m \geqslant 3$; the case $n \geqslant 3$ is similar.

From the equations $ua = f^n, ub = g^m$, we obtain $|u| = n|f| - 1$, and $|u| = m|g| - 1$, hence $2|u| = n|f| + m|g| - 2$, that is $|u| = n/2|f| + m/2|g| - 1$. Therefore, making use of the relations $n \geqslant 2, m \geqslant 3$, we obtain $|u| \geqslant |f| + |g| - 1$.

Consequently, the words f^n, g^m have a common prefix u of length greater than $|f| + |g| - 1$. In view of Lemma 1(ii), it follows that $f = p^r, g = p^s$, for some $p \in V^+, r, s \geqslant 1$. Therefore, $ua = p^{rn}$, $ub = p^{sm}$, which is contradictory, because $a \neq b$. In conclusion, at least one of ua, ub is primitive. □

This result has several rather interesting consequences, proving in some sense that "there are very many primitive words".

Corollary 2. (i) *For every word $u \in V^*$, at most one of the words ua, with $a \in V$, is not primitive.* (ii) *For every words $u_1, u_2 \in V^*$, at most one of the words $u_1 a u_2$, with $a \in V$, is not primitive.*

Proof. (i) For a given $u \in V^*$, apply Proposition 1 for two symbols $a, b \in V$. If ua is primitive, mark the symbol a, if ub is primitive, mark the symbol b. At least a symbol is marked in this way. Continue by considering any two non-marked symbols. Eventually, at most one symbol remains unmarked, and this completes the proof. Assertion (ii) follows now from the reflectivity of Q. □

Corollary 3. *If the language $L \subseteq V^*$ is infinite, then there is at least one letter $a \in V$ such that $L\{a\} \cap Q$ is infinite.*

Proof. Let $a, b \in V$. If both $L\{a\}$ and $L\{b\}$ contain only a finite number of primitive words, then for some integer n all the words of the form ua, ub with $|u| \geqslant n$ will be non-primitive. However, by Proposition 1, $\{u\}V$ contains at most one non-primitive word, a contradiction. □

Corollary 2(ii) also has interesting consequences concerning the possibility of obtaining primitive words by deletion operations (of single symbols).

Proposition 2. *Every word $w \in Q, |w| \geqslant 2$, can be written in the form $w = u_1 a u_2$, for some $u_1, u_2 \in V^*, a \in V$, such that $u_1 u_2 \in Q$.*

Proof. Take $w \in Q, |w| \geqslant 2$. Because of the primitivity, we can write $w = v_1 a b v_2$, for some $v_1, v_2 \in V^*, a, b \in V$, with $a \neq b$. According to Corollary 2(ii), at least one of $v_1 a v_2$ and $v_1 b v_2$ is primitive. The word $v_1 a v_2$ is obtained by erasing the symbol b from $w = v_1 a b v_2$ (hence, $u_1 = v_1 a, u_2 = v_2$), whereas the word $v_1 b v_2$ is obtained by erasing

the symbol a from $w = v_1abv_2$ (hence $u_1 = v_1, u_2 = bv_2$). In each case, a symbol can be deleted from w and the obtained word is primitive. □

This proposition suggests to consider a robustness notion of the following type: a primitive word $w \in V^*$ is said to be *del-robust* if each symbol of w can be erased such that the obtained words are all primitive. Symmetrically, we can consider *ins-robust* words, those which lead to primitive words irrespective where we insert irrespective which symbol of V.

It is worth noticing that there are primitive words of arbitrary length which are at the same time del-robust and ins-robust. An example is provided by the sequence of words

$$w_n = aba^2b^2 \dots a^nb^n, \quad n \geqslant 2$$

over $V = \{a, b\}$. Removing any symbol from w_n or inserting a symbol a or b in any position, we obtain primitive words: either a^n or b^n remains unchanged, hence no possibility appears of writing the obtained words as powers of other words.

3. Primitivity and density

Proposition 1 also suggests a series of developments related to the density of words in a language.

Let k be a positive integer. A language $L \subseteq V^*$ is called *right k-dense* (*strictly right k-dense*) (see [3]) if for every $u \in V^*$ there exists a word $x \in V^*$, $|x| \leqslant k$ ($1 \leqslant |x| \leqslant k$), such that $ux \in L$. A language L is *right dense* if it is k-dense for all $k \geqslant 1$.

Thus, Proposition 1 says that the language Q is right 1-dense and therefore right k-dense for every k.

A *minimal* (*strictly*) *right k-dense* language L is a (strictly) right k-dense language that contains no proper (strictly) right k-dense language. Remark that $\lambda \notin L$.

It is proved in [3] that every right k-dense language contains a minimal right k-dense language. Consequently, also the language Q contains a minimal right k-dense language, for each $k \geqslant 1$.

Proposition 3. *Let M be a minimal right 1-dense language of primitive words over some alphabet V. Then*:

(i) $M \cap V \neq \emptyset$.

(ii) *For every integer $n \geqslant 1$ and every $a \in V$ there exists $b \in V$, $b \neq a$, such that $a^nb \in M$.*

(iii) *If M is minimal strictly right 1-dense and if $ua, ub \in M$ with $u \in V^*, a, b \in V$, then $a = b$.*

Proof. (i) Since M is right 1-dense, there exists $x \in V^*, |x| \leqslant 1$, such that $\lambda x \in M$. Since λ is not primitive, it follows that $|x| = 1$ and $x \in M \cap V$.

(ii) If $a \in V$ and $n \geqslant 2$, then a^n is not primitive. Because of the right 1-density of M, there is x with $|x| \leqslant 1$ such that $a^n x \in M$. This implies $x = b$ for some $b \in V, b \neq a$.

(iii) Suppose that $a \neq b$. The set $M - \{ub\}$ is not strictly right 1-dense and hence there is $v \in V^*$ such that $\{v\}V \cap (M - \{ub\}) = \emptyset$. Since $\{v\}V \cap M \neq \emptyset$, it follows that $vx \in M$ for some $x \in V$. From $vx \notin M - \{ub\}$ it follows that $vx = ub$. Since $x \in V$, then $x = b$ and $v = u$. Hence $\{u\}V \cap (M - \{ub\}) = \emptyset$ and, since $ua \in M - \{ub\}$, we have a contradiction. □

Proposition 4. *Let M be a strictly right 1-dense language of primitive words over some alphabet V. Then M is minimal if and only if $ua, ub \in M$ for $u \in V^*$, $a, b \in V$, implies $a = b$.*

Proof. ($\Rightarrow$) By Proposition 3(iii).

($\Leftarrow$) Suppose that M is not minimal. Then there exists a language $L \subset M$ which is strictly right 1-dense. Let $v \in M - L, v \neq \lambda$. Then $v = ua$ for some $a \in V$. Since L is right 1-dense, there is $b \in V$ such that $ub \in L$. Hence $ua, ub \in M$ and therefore $a = b$. This implies that $ua \in L$ and $ua \in M - L$, a contradiction. □

Using the previous result, it is possible to construct a minimal strictly 1-dense language of primitive words.

Let $V = \{a_1, a_2, \ldots, a_n\}, n \geqslant 2$. Consider the sequence of languages $U_i, i \geqslant 1$, defined as follows:

$$
\begin{aligned}
U_1 &= \{u \in V^* \mid ua_1 \in Q\}, \quad T_1 = V^* - U_1, \\
U_2 &= \{u \in T_1 \mid ua_2 \in Q\}, \quad T_2 = T_1 - U_2, \\
&\vdots \\
U_i &= \{u \in T_{i-1} \mid ua_i \in Q\}, \; T_i = T_{i-1} - U_i, \\
&\vdots
\end{aligned}
$$

The sequence $U_i, i \geqslant 1$, is finite because the alphabet V is finite (the sequence must reach a set U_{i+1} which is empty or $i = n$).

Let $U = U_1\{a_1\} \cup U_2\{a_2\} \cup \cdots \cup U_i\{a_i\}$. The language U is a subset of Q. It is strictly right 1-dense, because for every $u \in V^*$ there is at least one a_j such that ua_j is primitive and it is minimal by construction and by Proposition 4.

4. Symbol substitutions and primitivity

For $x \in V^+$, consider the set

$$one(x) = \{x_1 b x_2 \mid x = x_1 a x_2, \; x_1, x_2 \in V^*, a, b \in V, a \neq b\}.$$

Consider a language $L \subseteq V^*$ and a word $x \in L$. We say that x is *subst-robust* (with respect to L) if $one(x) \subseteq L$.

It is easy to see that Q contains both subst-robust and non-subst-robust words of arbitrary length. For instance,

$$w_n = aba^2b^2 \dots a^nb^n, \quad n \geqslant 3$$

are subst-robust, whereas

$$z_n = ba^n, \quad n \geqslant 1$$

are not. (Note that $w_2 = abaabb$ is not subst-robust: replacing the second occurrence of a by b, we get the non-primitive word $abbabb$.)

Lemma 3. *If L consists of only subst-robust words, then $L = \{w \in V^* \mid |w| \in length(L)\}$.*

Proof. Take $x \in L, x = a_1a_2 \dots a_n$. Because x is subst-robust, each word $a_1 \dots a_{i-1} b_i a_{i+1} \dots a_n$ is in L. In turn, these words are subst-robust, hence one more symbol of them can be changed. Continuing in this way, each word y such that $|x| = |y|$ is proved to be in L, which implies the inclusion $\{w \in V^* \mid |w| \in length(L)\} \subseteq L$. The converse inclusion is obvious. □

Corollary 4. *If $L \subseteq V^*$ is a context-free language such that all its words are subst-robust, then L is regular.*

Although not all primitive words are subst-robust in the sense defined above, a weaker notion of robustness is verified at least for alphabets with more than two symbols.

Proposition 5. *If V contains at least three symbols, then for each word $x \in V^*$ and for each decomposition $x = x_1ax_2$, $x_1, x_2 \in V^*, a \in V$, there is $b \in V$, $b \neq a$, such that x_1bx_2 is primitive.*

Proof. Consider a word $x \in V^*$ and a decomposition $x = x_1ax_2$ as above. Consider the words x_1cx_2 with $c \in V$. According to Corollary 2(ii), at most one is not primitive. Because V contains at least three distinct symbols, there is $b \in V - \{a\}$ for which x_1bx_2 is primitive. □

If we start with $x \in V^*, x \in Z$, then all substitutions lead to primitive words: we know that $x = x_1ax_2$ is not primitive, hence from Corollary 2(ii) all words $x_1bx_2, b \neq a$, are primitive. The argument holds even for $V = \{a, b\}$.

The assertion in Proposition 5 does not hold true for $V = \{a, b\}$: in $abaabb$ we cannot replace the second occurrence of a by b, or the last occurrence of b by a without losing the primitivity. However, we have the following result.

Proposition 6. *If $V = \{a, b\}$, then for each word $x \in V^*, |x| \geqslant 3$, and for each decomposition $x = x_1\alpha\beta x_2$, $x_1, x_2 \in V^*, \alpha, \beta \in V$, at least one of the words $x_1\alpha'\beta x_2, x_1\alpha\beta' x_2$ is primitive, where $\alpha, \beta \in V$, $\alpha' \neq \alpha$ and $\beta' \neq \beta$.*

Proof. Take $x = x_1\alpha\beta x_2$ as above $\alpha, \beta \in V, |x_1x_2| \geqslant 1$. In view of the reflectivity of Q, it is enough to prove that at least one of $x_2x_1\alpha'\beta, x_2x_1\alpha\beta'$ is primitive. Assume the contrary, that is $x_2x_1\alpha'\beta = f^n$, and $x_2x_1\alpha\beta' = g^m$, for some $n \geqslant 2, m \geqslant 2$ and $f, g \in Q$. As in the proof of Proposition 1, we may assume that at least one of n, m is greater than or equal to 3; assume that $m \geqslant 3$.

If $|g| = 1$, then $x_2x_1\alpha\beta' = g^m$, for some $g \in \{a, b\}$. Because $\alpha' \neq \alpha, \beta' \neq \beta$, it follows that $x_2x_1\alpha'\beta$ is primitive, a contradiction to the previous assumption. Thus, we may assume that $|g| \geqslant 2$.

We have $|x_2x_1| = n|f| - 2$, and $|x_2x_1| = m|g| - 2$, which implies $|x_2x_1| = n/2|f| + m/2|g| - 2$. Because $n \geqslant 2$ and $m \geqslant 3$, we get $|x_2x_1| \geqslant |f| + |g| + 1/2|g| - 2$. Because $|g| \geqslant 2$, we conclude that $|x_2x_1| \geqslant |f| + |g| - 1$. From Lemma 1(ii), we obtain that f and g are powers of the same primitive word, which is impossible because $\alpha' \neq \alpha$ and $\beta' \neq \beta$. In conclusion, at least one of $x_1\alpha'\beta x_2$, $x_1\alpha\beta' x_2$ is primitive. □

The condition $|x| \geqslant 3$ in the theorem is necessary: for $x = ab$, none of the words aa and bb is primitive. Note also that ab is primitive, hence the condition of x being primitive does not help.

5. Taking prefixes and primitivity

Consider now the operation *pref*. Let $L \subseteq V^*$ be a given language.

We say that $x \in L$ is *pref-robust* (with respect to L) if $pref(x) \subseteq L$. We denote by L_{pr} the set of all pref-robust words in L.

It is easy to see that there are primitive words of arbitrary length which are pref-robust: $ba^n, n \geqslant 1$, and $aba^2b^2 \ldots a^nb^n, n \geqslant 1$, are such examples.

Proposition 7. *The language Q_{pr} is not context-free.*

Proof. Consider the regular language $R = ba^+ba^+ba^+ba^+$. Intersecting Q_{pr} with R we obtain

$$Q_{\mathrm{pr}} \cap R = \{ba^{n_1}ba^{n_2}ba^{n_3}ba^{n_4} \mid n_1, n_2, n_3, n_4 \geqslant 1,$$
$$n_1 > n_2 \quad \text{and} \quad (n_1 \neq n_3 \text{ or } n_2 > n_4)\}.$$

The inclusion $\supseteq$ is obvious. Conversely, take a word $x = ba^{n_1}ba^{n_2}ba^{n_3}ba^{n_4} \in Q_{\mathrm{pr}} \cap R$. Because $pref(x) \subseteq Q$, we must have

1. $n_1 > n_2$ (otherwise $ba^{n_1}ba^{n_1} \in pref(x) - Q$) and
2. $n_1 \neq n_3$ or $n_2 > n_4$ (otherwise $ba^{n_1}ba^{n_2}ba^{n_1}ba^{n_2} \in pref(x) - Q$).

Because $x \in pref(x)$, no further restriction is entailed by the fact that $x \in Q$. Consequently, also the inclusion $\subseteq$ holds.

Using a sequential transducer (a *gsm*), we can now translate $Q_{\rm pr} \cap R$ into the language

$$L = \{a^{n_1} b^{n_2} c^{n_3} d^{n_4} \mid n_1, n_2, n_3, n_4 \geqslant 1, \\ n_1 > n_2 \quad \text{and} \quad (n_1 \neq n_3 \text{ or } n_2 > n_4)\}.$$

This is not a context-free language.

Indeed, assume that L is context-free. Then L has to satisfy Ogden's lemma, that is there is a constant N such that every $z \in L$ in which at least N symbols are considered marked can be written in the form $z = uvwxy$ such that the following conditions hold: (i) vwx contains at most N marked symbols, (ii) v and x together have at least one marked symbol, (iii) for all $i \geqslant 0$, $uv^i wx^i y \in L$.

Consider a constant N associated to L as in Ogden's lemma and let $z = a^{n_1} b^{n_2} c^{n_3} d^{n_4}$ with $n_1 = N + 1$, $n_2 = N$, $n_3 = (N + 1)! + N + 1$, and $n_4 = (N + 1)! + N + 1$. It is clear that $z \in L$ (we have $n_1 > n_2$ and $n_1 \neq n_3$). We mark all occurrences of the symbol b in z (there are at least N such occurrences).

Let $z = uvwxy$ be a decomposition of z which satisfies the conditions (i)–(iii) above. It is clear that neither v nor x may contain two different symbols. Condition (ii) implies that $|vx|_b \geqslant 1$. It suffices to consider the following two cases.

(1) vx does not contain any occurrence of a. Then $z' = uv^2 wx^2 y$ should be in L by condition (iii) in Ogden's lemma. However, $|z'|_a \leqslant |z'|_b$, a contradiction.

(2) vx contains occurrences of a. Then, we obviously have $v = a^i$ and $x = b^j$ for some $i, j \geqslant 1$. If $i < j$, then for large enough k we get words $z'' = uv^k wx^k y$ with $|z''|_a < |z''|_b$, again contradicting point (iii) above. Therefore, we must have $i \geqslant j$. Let

$$k = \frac{(N + 1)!}{i} + 1$$

and consider the word $z''' = uv^k wx^k y = a^{m_1} b^{m_2} c^{m_3} d^{m_4}$. We have

$$m_1 = ik + N + 1 - i = (N + 1)! + i + N + 1 - i = (N + 1)! + N + 1,$$

$$m_2 = jk + N - j \leqslant (N + 1)! + j + N - j = (N + 1)! + N,$$

$$m_3 = n_3 = (N + 1)! + N + 1, \text{ and}$$

$$m_4 = n_4 = (N + 1)! + N + 1.$$

Consequently, $m_1 = m_3$ and $m_2 < m_4$, that is $z''' \notin L$. Again, this contradicts Ogden's lemma.

We conclude that L is not context-free. Because the family of context-free languages is closed under intersection with regular languages and under sequential transducers, the language Q_{pr} is not context-free. □

Unfortunately, proving that Q_{pr} is not context-free does not imply that Q is not context-free: there are context-free languages L such that L_{pr} is not context-free. An example is

$$L = \{a^n b^m a^m b \mid n, m \geqslant 1\} \cup pref(\{a^n b^n a^m \mid n, m \geqslant 1\}).$$

This is a linear language (over $V=\{a,b\}$). However,

$$L_{pr}\cap a^+b^+a^+b=\{a^nb^na^nb\mid n\geqslant 1\},$$

which is non-context-free. (The equality above is obvious: the prefix $a^nb^ma^m$ of $a^nb^ma^mb$ is in L only if $a^nb^ma^m\in pref(\{a^pb^pa^q\mid p,q\geqslant 1\})$, which implies $n=m$.)

6. Morphisms and primitivity

The behavior of morphisms with respect to primitivity has been considered previously, in particular in [4,10]. For a morphism f that is injective, it has been shown in [4] that $f(Q)\cap Q$ is infinite when $|V|=2$. As shown below, this is in fact a characteristic property of injective morphisms for alphabets containing two letters.

Recall (see [9]) that a morphism f is injective if and only if $f(V)$ is a code and $|f(V)|=|V|$. Also a language $\{u,v\}\subseteq V^+$ is a code if and only if $uv\neq vu$.

A morphism f is said to be *λ-free* if $\lambda\notin f(V)$.

Lemma 4. *Let $V=\{a,b\}$ and let f be a λ-free morphism such that $|f(V)|=|V|$. If f is not injective, then $f(a)=p^m$, $f(b)=p^n$ for some $p\in Q$.*

Proof. Since f is not injective, then $f(V)=\{f(a),f(b)\}$ is not a code. Therefore, $f(a)f(b)=f(b)f(a)$ and, by Lemma 1(i), $f(a)=p^m$, $f(b)=p^n$ for some $p\in Q$. □

Proposition 8. *Let f be a λ-free morphism of V^* with $V=\{a,b\}$. Then the following properties are equivalent*:

(i) *f is injective*,

(ii) *$f(Q)\cap Q$ is infinite*.

Proof. (i) → (ii). The words a^nb^n are primitive for all $n\geqslant 1$. Let $u=a^mb^m$ with $m\geqslant 2$ and suppose that $f(u)$ is not primitive. Let $f(a)=p^r$, $f(b)=q^s$ where $p,q\in Q$. Then $f(u)=p^{rm}q^{sm}$ with $rm,sm\geqslant 2$. Since $f(u)$ is not primitive, this is only possible if $p=q$ (Lemma 2(ii)). Hence $f(ab)=f(ba)$, a contradiction.

(ii) → (i) Suppose that f is not injective. If $|f(V)|=|V|$, then by Lemma 4 $f(a)=p^m$, $f(b)=p^n$. Hence $f(V^+)\subseteq p^+$ and $f(Q)\cap Q$ is finite, a contradiction.

If $|f(V)|=1$, then $f(a)=f(b)$. Let $u=f(a)$. Then $f(V^+)=u^+$ and $f(Q)\cap Q$ is finite, a contradiction. □

If the alphabet V has more than two letters, the above proposition is no longer true. For example, let $V=\{a,b,c\}$ and let f be defined by $f(a)=a$, $f(b)=ab$, $f(c)=aba$. This morphism is not injective because $f(ba)=f(c)$. However, $f(Q)\cap Q$ is infinite because $a^nb^n\in Q$ and $f(a^nb^n)\in Q$ for $n\geqslant 1$.

Lemma 5. *Let f be a λ-free morphism of V^* with $|V| \geqslant 2$. Then the following properties are equivalent*:
(i) *f is injective*;
(ii) *$f(uv) = f(vu)$ implies $uv = vu$ for all $u, v \in V^*$.*

Proof. (i) $\rightarrow$ (ii). Immediate.

(ii) $\rightarrow$ (i) Suppose that $f(u) = f(v)$. Then $f(uv) = f(vu)$, $uv = vu$ and $u = p^m$, $v = p^n$ with $p \in Q$. Hence $f(p)^m = f(u) = f(v) = f(p)^n$. Since $f(p) \neq \lambda$, we have $m = n$ and $u = v$. □

Corollary 5. *If f is injective, then $f(uv) = f(vu)$ implies $f(u) = p^m$, $f(v) = p^n$ with $p \in Q$, $m, n \geqslant 1$.*

Proof. Since $uv = vu$, then $u = q^r$, $v = q^s$ and $f(q) = p^k$ for some $q, p \in Q$, $r, s, k \geqslant 1$. Therefore, $f(u) = p^m, f(v) = p^n$ with $m = kr$, $n = ks$. □

Let f be a morphism of V^*. Denote by f_Q the set of all the primitive words $u \in Q$ such that $f(u) \in Q$ and by f_Z the set of all the primitive words $u \in Q$ such that $f(u) \in Z$, i.e. $f(u) = p^n$, $p \in Q, n \geqslant 2$.

Proposition 9. *Let f be a morphism of V^*. Then*:
(i) *The languages f_Q and f_Z are reflective.*
(ii) *If f is injective, then $u, v \in f_Z$, $u \neq v$ imply $uv \notin f_Z$ and $uv \in Q$.*

Proof. (i) If $f_Q = \emptyset$, this is immediate. Suppose $f_Q \neq \emptyset$ and let $uv \in f_Q$. Then $uv \in Q$ and $f(u)f(v) = f(uv) \in Q$. Since Q is reflective, then $vu \in Q$ and $f(vu) = f(v)f(u) \in Q$. Therefore, $vu \in f_Q$.

If $f_Z = \emptyset$, this is immediate. Suppose $f_Z \neq \emptyset$ and let $uv \in f_Z$. Then $uv \in Q$, $vu \in Q$ and $f(u)f(v) = f(uv) \in Z$. Since Z is reflective, then $f(vu) = f(v)f(u) \in Z$. Therefore, $vu \in f_Z$.

(ii) We have $f(u) = p^m$, $f(v) = q^n$ with $p, q \in Q$ and $m, n \geqslant 2$. Since $u \neq v$ and $u, v \in Q$, then $p \neq q$. Indeed if $p = q$, then $f(uv) = f(vu)$ and $uv = vu$ because of the injectivity of f. This implies that u, v are powers of the same primitive word, a contradiction. Hence $f(uv) = p^m q^n$ with $p \neq q$. By Lemma 2(ii), this implies $f(uv)$ primitive. Therefore, $uv \notin f_Z$ and $uv \in Q$. □

If f is a morphism of V^*, the word u is said to be *f-reductible* if $f(u) = p^m$, $p \in Q$, $m \geqslant 2$. Since Q and Z are reflective, then uv is f-reductible if and only if vu is f-reductible. Remark that if $\lambda \notin f(V)$, then every non-primitive word is f-reductible.

Proposition 10. *Let f be an injective morphism of V^* and let u be an f-reductible word. If $u = u_1 w u_2$ with $w u_2 u_1 \neq u_2 u_1 w$ and if w is f-reductible, then $v = u_1 u_2$ is primitive and v is not f-reductible.*

Proof. Suppose that $v = u_1u_2$ is f-reductible. Then u_2u_1 is also f-reductible and $f(u_2u_1)=r^{k_1}$, $r \in Q, k_1 \geqslant 2$. We have $f(u)=f(u_1)f(w)f(u_2)=p^m$, $p \in Q$, $m \geqslant 2$. Since $f(u) \in Z$ which is reflective, then $f(w)f(u_2)f(u_1) \in Z$ i.e. $f(w)f(u_2)f(u_1)=q^n$, $q \in Q$, $n \geqslant 2$. Since w is f-reductible, then $f(w) = s^{k_2}$, $s \in Q$, $k_2 \geqslant 2$. Hence,

$$f(w)f(u_2u_1) = s^{k_2}f(u_2u_1) = s^{k_2}r^{k_1} = q^n.$$

We have $s \neq r$. Indeed if $s = r$, then $f(wu_2u_1) = f(u_2u_1w)$ and, since f is injective, $wu_2u_1 = u_2u_1w$, against our hypothesis. By Lemma 2(ii), it follows then that q^n, $n \geqslant 2$, is primitive, a contradiction.

Therefore, v is not f-reductible and hence v is primitive. □

Corollary 6. *Let f be an injective morphism and let u be an f-reductible word such that $u = u_1x^mu_2$, $m \geqslant 2$, $x \neq \lambda$ and $x^mu_2u_1 \neq u_2u_1x^m$. Then u_1u_2 is primitive and not f-reductible.*

Proposition 11. *Let f be an injective morphism of V^*.*

(i) *If $u, v \in V^+$ with $uv \neq vu$ are f-reductible, then uv is not f-reductible.*

(ii) *If uv is f-reductible and $uv \neq vu$, then either u or v is not f-reductible.*

Proof. (i) Since u and v are f-reductible, then $f(u) = p^m$, $f(v) = q^n$ where $p, q \in Q$, $m, n \geqslant 2$. If $p = q$, then $f(uv) = p^{m+n} = f(vu)$. Since f is injective, then $uv = vu$, a contradiction. Hence $f(uv) = p^mq^n$ with $p \neq q$, $m, n \geqslant 2$. From Lemma 2(ii), it follows then that $f(uv) \in Q$ and therefore uv is not f-reductible.

(ii) Suppose that both u and v are f-reductible. Then $f(u) = p^m$, $f(v) = q^n$, $p, q \in Q$, $m, n \geqslant 2$. From $uv \neq vu$ and the injectivity of f follows $p \neq q$. By Lemma 2(ii), $f(uv) = p^mq^n$ is primitive, a contradiction. □

A word $u \in V^+$ is said to be *universally-primitive* or simply *u-primitive* if for every injective morphism f of V^*, the word $f(u)$ is primitive. Hence a u-primitive word is a word that is not f-reductible for every injective morphism of V^*. Let Q_U denote be the set of all the u-primitive words of V^*. Clearly $Q_U \subseteq Q$. Since $V \cap Q_U = \emptyset$, the inclusion is strict.

Proposition 12. *Let $w = u^mv^n$ with $u, v \in V^+$ and $m, n \geqslant 2$. Then the following properties are equivalent*:

(i) *w is u-primitive*;

(ii) *u and v have different roots*;

(iii) *$uv \neq vu$.*

Proof. (i) → (ii) If u and v have the same root p, then $w = p^k$ for some $k \geqslant 2$ and hence w is not primitive, a contradiction.

(ii) → (iii) If $uv = vu$, then $u = p^r$, $v = p^s$ for some primitive word p. This again implies that w is not primitive.

(iii) → (i) If $uv \neq vu$, then $u = p^r$, $v = q^s$ for some $p, q \in Q$ and $p \neq q$. Therefore, $w = p^{rm}q^{sn}$ with $rm, sn \geqslant 2$. This implies that w is primitive. Let f be an injective morphism and let $f(p) = p_1^{m_1}$, $f(q) = q_1^{n_1}$ with $p_1, q_1 \in Q$, $m_1, n_1 \geqslant 1$. We have $p_1 \neq q_1$. If not, then $f(uv) = f(vu)$ and since f is injective, $uv = vu$, a contradiction. Hence, $f(w) = p_1^{m_1 rm} q_1^{n_1 sn}$ with $p_1, q_1 \in Q$, $p_1 \neq q_1$, $m_1 rm, n_1 sn \geqslant 2$. By Lemma 2(ii), this implies $f(w)$ primitive and therefore w u-primitive. □

Corollary 7. *The set Q_U is infinite.*

Proof. Let $a, b \in V$. Then the set $\{a^n b^n \mid n \geqslant 2\}$ contains only u-primitive words and hence Q_U is infinite. □

Proposition 13. *If f is an injective morphism, then $T = f(Q - Q_U) \cap Q \neq \emptyset$.*

Proof. Without loss of generality, we can assume that $V = \{a, b\}$. The words a, b, ab are not u-primitive. Indeed, the morphisms g defined by $g(a) = a^2$, $g(b) = b^2$ and h defined by $h(a) = aba$, $h(b) = b$ are injective. This implies that a and b are not u-primitive. Since $h(ab) = (ab)^2$, then ab is also not u-primitive.

Suppose now that $T = \emptyset$. Then, since a, b, ab are not u-primitive, $f(a), f(b)$ and $f(ab)$ are not primitive. This implies in particular that $f(a) = p^m$ and $f(b) = q^n$ for some $p, q \in Q$, $m, n \geqslant 2$. Since f is injective, we have $p \neq q$. From $f(ab) = f(a)f(b) = p^m q^n$, it follows by Lemma 2(ii) that $f(ab) \in Q$, a contradiction. □

Remark 1. Words of the form $p^m q^n$, $p, q \in Q$, $p \neq q$, $m, n \geqslant 2$ are u-primitive (Proposition 12). However, not all u-primitive words are of this form.

For instance, the word $u = a^r b^n a^s$ with $a, b \in V$, $r + s, n \geqslant 2$ is u-primitive. Suppose not. Then for some injective morphism f, the word $f(u) = f(a)^r f(b)^n f(a)^s$ is not primitive. Hence $f(u) \in Z$ which is reflective. This implies that

$$w = f(b)^n f(a)^s f(a)^r = f(b)^n f(a)^{s+r} \in Z$$

and therefore w is not primitive. By Lemma 2(ii), this is only possible if $f(a)$ and $f(b)$ have the same roots. This implies $f(ab) = f(ba)$, a contradiction since f is injective.

In particular, the word ab^2a is u-primitive. However, this word cannot be written in the form $p^m q^n$, $m, n \geqslant 2$.

Remark 2. If a u-primitive word u has a decomposition of the type $p^m q^n$, $p, q \in Q$, $p \neq q$, $m, n \geqslant 2$, this decomposition is not necessarily unique.

For example, let $V = \{a, b\}$ and let $p_1 = ababb$, $p_2 = bababb$, $q_1 = ab$, $q_2 = bababb$. Then $p_1, p_2, q_1, q_2 \in Q$, $p_1 \neq p_2$, $q_1 \neq q_2$. Then the following word u is u-primitive and has at least two such different decompositions:

$$u = p_1^2 p_2^2 = ababbababbbababbbababb = q_1^2 q_2^3.$$

Remark 3. Words of the type $p_1^m p_2^n$ with $p_1, p_2 \in Q$, $p_1 \neq p_2$, $m,n \geqslant 2$ are u-primitive and hence primitive. This result cannot be extended further. For example, if $a,b \in V$, the words a, b and a^2b^2 are primitive, but the word $u = a^2b^2(a^2b^2)^2$ is not primitive.

The word u in the above example is of the form $p^iq^j(p^rq^j)^m$ and it is not primitive. However, as shown below, words of this form are in general u-primitive.

Proposition 14 (Shyr and Yu [9]). *Let V with $|V| \geqslant 2$ and let $p,q \in Q$, $p \neq q$. Then*:

$$u = p^iq^j(p^rq^j)^m \in Q \quad for\ j,m \geqslant 1,\ i \geqslant 0,\ r \geqslant 2,\ i \neq r. \tag{1}$$

Proposition 15. *Every word u of the form* (1) *is u-primitive.*

Proof. Let f be an injective morphism and let $f(p) = p_1^s$, $f(q) = q_1^t$ with $p_1, q_1 \in Q$, $s,t \geqslant 1$. Since f is injective, then $f(p) \neq f(q)$ and $f(pq) \neq f(qp)$. This implies that $p_1 \neq q_1$. We have $f(u) = p_1^{si}q_1^{tj}(p_1^{sr}q_1^{tj})^m$, with $tj, m \geqslant 1$, $si \geqslant 0$, $sr \geqslant 2$ and $si \neq sr$. By Proposition 14, it follows then that $f(u)$ is primitive. Since this is true for every injective morphism, the word u is u-primitive. □

7. Final remarks

One can consider other operations with words from the point of view of primitivity preserving.

One candidate is the *interchanging of adjacent symbols*. We believe that a result of the following type is true: for every $w \in Q_V, w \notin \{abba \mid a,b \in V, a \neq b\}$, there is a decomposition $w = x_1cdx_2, x_1,x_2 \in V^*, c,d \in V, c \neq d$, such that $x_1dcx_2 \in Q_V$.

Acknowledgements

Useful remarks by two referees are gratefully acknowledged. In particular, a variant of Proposition 6 and Ref. [6] (where a result similar to our Proposition 1 can be found) were suggested by them.

References

[1] C. Choffrut, J. Karhumaki, Combinatorics on words, Handbook of Formal Languages, in Vol. 1, Springer, Berlin, 1997, pp. 329–438 (Chapter 6).
[2] P. Dömösi, S. Horvath, M. Ito, Formal languages and primitive words, Publ. Math. Debrecen 42 (3–4) (1993) 315–321.
[3] M. Ito, G. Thierrin, Right k-dense languages, Semigroup Forum 48 (1994) 313–325.
[4] Gh. Păun, G. Thierrin, Morphisms and primitivity, Bulletin of the EATCS 61 (1997) 85–88.
[5] H. Petersen, On the language of primitive words, Theor. Comput. Sci. 161 (1996) 141–156.
[6] J.M. Robson, Separating strings with small automata, Inform. Process. Lett. 30 (1989) 209–214.
[7] G. Rozenberg, A. Salomaa (Eds.), Handbook of Formal Languages, 3 Vols., Springer, Berlin, 1997.

[8] H. J. Shyr, Free Monoids and Languages, Hon Min Book Company, Taichung, Taiwan, 1991.
[9] H.J. Shyr, S.S. Yu, Non-primitive words in the language p^+q^+, Soochow J. Math. 20 (1994) 535–546.
[10] H.J. Shyr, G. Thierrin, Codes, languages and MOL schemes, RAIRO. Theor. Computer Sci. 1 (4) (1977) 293–301.

For Further Reading

M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading, MA, 1983.
Y.S. Tai, H.L. Wu, Primitive properties of words in $\{uv, u\}$, Soochow J. Math. 20 (1994) 83–100.