Word insertions and primitivity *

Lila Kari and Gabriel Thierrin
Department of Mathematics, University of Western Ontario
London, Ontario, Canada N6A 5B7
email: lkari@julian.uwo.ca

June 14, 2010

Abstract

In this paper we consider relations between the operation of word insertion and primitivity. A necessary and sufficient condition under which the insertion $u \leftarrow u$ of the word u into itself has maximum cardinality is obtained. The notion of insertion sequence is introduced and sufficient conditions under which an insertion sequence is a special type of language (regular, context-free, biprefix code) are obtained. Based on the operations of insertion, shuffle and commutative shuffle (which generalize catenation), the notions of ins-primitive words, shuffle-primitive words, and com-shuffle-primitive words are defined and investigated. These notions turn out to be generalizations of the classical notion of primitive words.

1 Introduction

Insertion is a word operation that has been introduced in [1] as a generalization of catenation. Given two words, u and v, instead of catenating v to the right extremity of u, the *insertion* $u \leftarrow v$ consists of the words obtained by inserting v into an arbitrary place in u:

$$u \leftarrow v = \{u_1 v u_2 |\ \ u = u_1 u_2\}.$$

The result of the insertion $u \leftarrow v$ is thus a set with cardinality greater than one, and less than or equal to $|u| + 1$, where $|u|$ denotes the length of u. A natural question is under what conditions does the set $u \leftarrow v$ have maximum cardinality. The cardinality of $u \leftarrow v$ is in direct connection with the structure of u and v : if they contain repetitive patterns, then chances are that there are

*This research was supported by Grant OGP0007877 of the Natural Sciences and Engineering Research Council of Canada

two different places of insertion of v into u that yield the same word, reducing thus the cardinality of $u \leftarrow v$. To formalize this idea, we introduce the notion of insertion of rank i, $u \leftarrow_i v$: the word v is inserted exactly after the ith letter of u. (The notion makes sense only if i is less than or equal to $|u|$.)

Note that now we have $\text{card}(u \leftarrow v) < |u| + 1$ iff there exist $0 \leq i, j \leq |u|$ such that $u \leftarrow_i v = u \leftarrow_j v$. Based on the properties of the insertion of rank i obtained in Section 2, we show that a necessary and sufficient condition for $u \leftarrow u$ to have maximum cardinality, $|u|$, is that the word u is primitive. (A word $u \in X^+$ is said to be *primitive* if $u = v^n, v \in X^+, n \geq 1$, implies $n = 1$.)

The idea of controlling the position where the insertion takes place carries over to the iterated insertion. The *iterated insertion* is obtained from insertion in a similar way the Kleene closure is obtained from catenation:

$$u \leftarrow^0 v = u,$$

$$u \leftarrow^{i+1} v = (u \leftarrow^i v) \leftarrow v, i \geq 0,$$

$$u \leftarrow^* v = \bigcup_{i=0}^{\infty} (u \leftarrow^i v).$$

If we add the restriction of controlling the position of insertion, we obtain the notion of inserting sequences. If u is a given word, the *inserting sequence* $\sigma(u)$ determined by the word u and the *pilot sequence* of integers $\sigma = \{i_1, \ldots, i_k, \ldots\}$, $0 \leq i_j \leq |u|$, is obtained by starting with the word u and inserting at each step the word u into the preceding inserted u. The position of insertion is determined by the pilot sequence. More precisely, if the word obtained at step j is xuy, where we emphasized the last inserted u, then the next word is obtained by inserting u after the i_{j+1}th letter of the marked occurence of u.

Section 2 investigates properties of inserting sequences. Namely, sufficient conditions under which an inserting sequence is a special kind of language (regular language, context-free language, biprefix code) are obtained.

Section 3 considers properties of a generalization of the notion of primitive word, induced by the insertion operation. A word u is called *ins-primitive* if $u \notin v \leftarrow^n v$ for any $n \geq 1$ and any $v \in X^+$. As insertion is a generalization of catenation, it follows that ins-primitive words are primitive, but the converse is not true. We show that the set of ins-primitive words over an alphabet X with at least two letters is right and left dense. (A language $L \subseteq X^*$ is called right/left dense if for every $u \in L$ there exists $w \in X^*$ such that $uw \in L$, respectively $wu \in L$.) If the word u is not ins-primitive, an ins-primitive $v \in X^+$ with $u \in v \leftarrow^n v$, $n \geq 1$, is called an *ins-root* of u. It turns out that every word that is not ins-primitive has an ins-root, but, in contrast with the primitive case, a word can have several ins-root.

Finally, based on the fact that the shuffle operation and the commutative shuffle operation also generalize catenation, the notions of *shuffle-primitive* and *com-shuffle-primitive* words are introduced and investigated in Section 4. For

example, a necessary and sufficient condition under which a word is com-shuffle-primitive is obtained.

In the following, an alphabet X is a finite nonempty set and X^* is the free monoid generated by X under the catenation operation. 1 denotes the empty word and $X^+ = X^* \setminus \{1\}$. For a word w and a letter $a \in X$, $N_a(w)$ is the number of occurrences of the letter a in w. For futher notions of formal language theory and theory of codes the reader is referred to [5], respectively [6].

2 Word insertions

The insertion of a word v into a word u consists of all the words obtained by inserting v in arbitrary places in u. For example, let $X = \{a, b\}$ and let $u = bab^2$. We have $u \leftarrow u = \{bab^3ab^2, b^2ab^2ab^2, babab^4, bab^2ab^3\}$. Here, card$(u \leftarrow u)$ is maximum, i.e., equals $|u| = 4$. This is not always necessarily the case. For example, in the extreme case when $u = a^n$ and $v = a^m$, $u \leftarrow v = \{a^{n+m}\}$ is a singleton set. We are looking in the following for necessary and sufficient conditions under which the cardinality of $u \leftarrow v$ equals $|u| + 1$. Note that two words in $u \leftarrow v$ coincide iff there exist two positions i and j in u such that the insertion of v after the ith letter of u yields the same word as the insertion of v after the jth letter of u. This notion of controlling the position of insertion can be formalized as follows.

Definition 2.1 *Let $u = a_1a_2 \dots a_k$ with $a_i \in X$, $0 \leq i \leq k$. For $1 \leq i \leq k$, the insertion of rank i of v into u is defined as:*

$$u \leftarrow_i v = a_1a_2 \dots a_iva_{i+1} \dots a_k.$$

For $i = 0$, $u \leftarrow_0 v = vu$.

Remark that if $|u| = k$, then $u \leftarrow_0 u = u \leftarrow_k u = u^2$. Note also that the catenation uv equals the insertion of rank $|u|$, and the catenation vu equals the insertion of rank 0 of v into u. The following proposition summarizes some necessary conditions for the insertions of certain ranks to yield the same word.

Proposition 2.1 *Let $u, v \in X^+$.*

(i) If $u \leftarrow_i v = v \leftarrow_j u$ and $0 \leq i < j \leq min\{|u|, |v|\}$, then either u is an infix of v or $vy = xu$ for some $x, y \in X^*$.

(ii) If $u \leftarrow_i v = v \leftarrow_i u$, $0 \leq i \leq min\{|u|, |v|\}$, then either u is a prefix of v or v is a prefix of u.

(iii) If $u \leftarrow_i u = u \leftarrow_j u$, $0 < i < j < |u|$, then $u = g^k$ for some $g \in X^+$ and $k \geq 2$, i.e., the word u is not primitive.

Proof. (i) We have $u \leftarrow_i v = u_1vu_2$ with $u = u_1u_2$ and $v \leftarrow_j u = v_1uv_2$ with $v = v_1v_2$. Therefore $u_1vu_2 = v_1uv_2$ and $|u_1| < |v_1|$ because $i < j$. Hence $v_1 = u_1x$, $u_1vu_2 = u_1xuv_2$ and $vu_2 = xuv_2$.

If $|u_2| \leq |v_2|$, then $v_2 = yu_2$ which implies $vu_2 = xuyu_2$, that means u is an infix of v. On the other hand, if $|u_2| > |v_2|$, then $u_2 = yv_2$, which implies $vyv_2 = xuv_2$, that is $vy = xu$.

(ii) Assume that $|u| < |v|$. Then $u = u_1u_2$, $v = v_1v_2$ with $u_1vu_2 = v_1uv_2$, where $|u_1| = |v_1| = i$. We deduce that $u_1 = v_1$ and $vu_2 = uv_2$. Therefore $v = ux$ and u is a prefix of v.

(iii) We have $u_1uu_2 = u \leftarrow_i u = u \leftarrow_j u = v_1uv_2$ with $|u_1| < |v_1|$, $|u_2| > |v_2|$ where $u = u_1u_2 = v_1v_2$. Hence $v_1 = u_1x$ and $u_2 = yv_2$ with $x \neq 1$, $y \neq 1$ and $|x| < |u|$, $|y| < |u|$. As $u = u_1u_2 = v_1v_2$, we deduce $u = u_1yv_2 = u_1xv_2$, which further implies $x = y$. From $u_1uxv_2 = u_1xuv_2$ follows $ux = xu$. Since $|x| < |u|$, $u = xu' = u'x$, which by a known result (see [6]) implies that x and u' are powers of a common word $g \in X^+$, i.e., $x = g^i$, $u' = g^j$, $i, j > 0$. Consequently, $u = g^{i+j}$, $i + j > 1$, that is, the word u is not primitive. □

Part *(iii)* of Proposition 2.1 indicates that a sufficient condition for $u \leftarrow_i u$ and $u \leftarrow_j u$ to yield different words for $i \neq j$, $0 < i, j < |u|$ is that u is a primitive word. This condition is also necessary, as seen in the following proposition.

Proposition 2.2 *Let $u \in X^+$ with $|u| = k$. Then card$(u \leftarrow u) = k$ iff u is a primitive word.*

Proof. " $\Rightarrow$ " Let u be a primitive word and assume, for the sake of contradiction, that card$(u \leftarrow u) < k$, i.e., there exist $0 < i < j < |u|$ such that $u \leftarrow_i u = u \leftarrow_j u$. According to Proposition 2.1 *(iii)*, this implies that the word u is not primitive – a contradiction.

For the reverse implication, let $u \in X^+$ with card$(u \leftarrow u) = k$ and assume that u is not primitive. Then $u = g^i$, $i > 1$, with g a primitive word. This implies $gug^{i-1} = g^{i-1}ug \in u \leftarrow u$, i.e., $u \leftarrow_{|g|} u = u \leftarrow_{(i-1)|g|} u$, which further means that card$(u \leftarrow u) < k$ – a contradiction. □

In the remainder of this section we consider carrying over the idea of controlling the position of insertion to iterated insertion. Let $u \in X^+$ and consider the following sequence:

$$u_0 = u, u_1 = v_0uw_0, u_2 = v_0v_1uw_1w_0, \ldots, u_{i+1} = v_0v_1 \ldots v_iuw_i \ldots w_1w_0, \ldots$$

where $v_iw_i = u$ for $i \geq 0$. Such a sequence is called an *inserting sequence of the word u*. Intuitively, an inserting sequence is obtained by starting with a word u and inserting at each step the word u into the preceding inserted u.

The inserting sequence is completely determined by the word u and the sequence of integers:

$$\sigma = \{|v_0|, |v_1|, |v_2|, \ldots, |v_i|, \ldots\}, 0 \leq |v_i| \leq |u|.$$

The sequence of non-negative integers associated with an inserting sequence is called the *pilot sequence* of the inserting sequence and the word u is called the

germ of the sequence. Hence, given a word u, every pilot sequence generates an inserting sequence and vice versa.

For example, the inserting sequence $\{ab, a^2b^2, \ldots, a^nb^n, \ldots\}$ is generated by the germ $u = ab$ and the pilot sequence $\{1, 1, \ldots, 1, \ldots\}$.

The inserting sequence $\{ab, (ab)^2, \ldots, (ab)^n, \ldots\}$ is generated by the germ $u = ab$ and the pilot sequence $\{0, 0, \ldots, 0, \ldots\}$ or $\{2, 2, \ldots, 2, \ldots\}$. More generally, for $u \in X^+$, $\{u, u^2, \ldots, u^n, \ldots\}$ is the inserting sequence generated by the germ u and either the pilot sequence $\{0, 0, \ldots\}$ or the pilot sequence $\{|u|, |u|, \ldots\}$. If the pilot sequence is $\{0, 1, 2, 0, 1, 2, \ldots\}$ then the first words of the inserting sequence are $abab$, $aabbab$, $aababbab$, $a(ab)^3bab$, $(aab)^2(bab)^2$, $(aab)^2ab(bab)^2$, etc.

The following proposition gives a sufficient condition under which an insertion sequence is a context-free language. (See [5] for the definition of regular and context-free languages.)

Proposition 2.3 *Let $w \in X^+$. If there exists an integer N such that for $n \geq N$ the pilot sequence is periodical, then the corresponding inserting sequence is a context-free language.*

Proof. Consider the word $w \in X^*$ and the pilot sequence

$$\{j_1, j_2, \ldots, j_N, i_1, i_2, \ldots i_k, i_1, i_2, i_k, \ldots\}$$

where $0 \leq j_p, i_q \leq |w|$, $1 \leq p \leq N$, $1 \leq q \leq k$.

The corresponding inserting sequence will be

$$\begin{array}{l}\{u_{j_1}v_{j_1},\ \ u_{j_1}u_{j_2}v_{j_2}v_{j_1},\ \ \ldots u_{j_1}\ldots u_{j_N}v_{j_N}\ldots v_{j_1}\}\cup \\ \{u_{j_1}\ldots u_{j_N}(u_{i_1}\ldots u_{i_k})^n u_{i_1}\ldots u_{i_q}v_{i_q}\ldots v_{i_1}(v_{i_k}\ldots v_{i_1})^n v_{j_N}\ldots v_{j_1}| \\ n \geq 0, 0 \leq q \leq k\}\end{array}$$

where $|u_{j_p}| = j_p, |u_{i_q}| = i_q$, and $w = u_{j_p}v_{j_p}$, $w = u_{i_q}v_{i_q}$, $1 \leq q \leq k$, $1 \leq p \leq N$.

The inserting sequence is context-free as it can be generated by the context-free grammar $G = (\{S, S'\}, X, S, P)$ where

$$\begin{array}{l}\{S \longrightarrow u_{j_1}\ldots u_{j_p}v_{j_p}\ldots v_{j_1} | 1 \leq p \leq N\}\cup \\ \{S \longrightarrow u_{j_1}\ldots u_{j_N}S'v_{j_N}\ldots v_{j_1}\}\cup \\ \{S' \longrightarrow (u_{i_1}\ldots u_{i_k})S'(v_{i_k}\ldots v_{i_1})\}\cup \\ \{S' \longrightarrow u_{i_1}\ldots u_{i_q}v_{i_q}\ldots v_{i_1} |\ \ 0 \leq q \leq k-1\}\end{array}$$

□

Note that if $u \in X^*$ and the pilot sequence is either $\sigma_1 = \{|u|, |u|, \ldots, |u| \ldots\}$ or $\sigma_2 = \{0, 0, \ldots, 0, \ldots\}$, then the inserting sequences are $\sigma_1(u) = \sigma_2(u) = \{u^n |\ \ n \geq 1\}$, which is a regular language.

Proposition 2.4 *If $w \in X^+$ is a primitive word and the pilot sequence is constant, not equal to 0 or $|w|$, then the corresponding inserting sequence is a non-regular context-free language.*

Proof. Let $w \in X^*$ and consider the pilot sequence $\sigma = \{i, i, \ldots, i \ldots\}$, where $w = uv$, $|u| = i$, $0 < i < |w|$. The corresponding sequence is $\sigma(w) = \{u^n v^n | \ n > 1\}$.

As $w = uv$ is primitive, we have $uv \neq vu$. This is further equivalent to the fact that $\{u, v\}$ is a code. If we consider two letters $a, b \notin X$ and the morphism $h : \{a, b\} \longrightarrow X^*$ defined by $h(a) = u$ and $h(b) = v$, then the fact that $\{u, v\}$ is a code is equivalent to h being injective. This implies $h^{-1}(\{u^n v^n | \ n > 0\}) = \{a^n b^n | n > 0\}$, which is a non-regular context-free language.

If $\sigma(w)$ were regular, as the family of regular languages is closed under inverse morphisms, the language $h^{-1}(\sigma(w)) = \{a^n b^n | \ n > 0\}$ would be regular – a contradiction. According to the preceding proposition, $\sigma(w)$ is context-free, therefore we conclude that $\sigma(w)$ is a non-regular context-free language.

□

A pilot sequence $\sigma = \{s_0, s_1, s_2, \ldots, s_n \ldots\}$ for a word u is said to be *internal* if $0 < s_n < |u|$ for $n \geq 1$. The following proposition connects the notions of internal pilot sequence, biprefix code and dipolarity.

Recall (see for example [7]) that a word u is called *unipolar* (or bordered) if $u = vx = yv$ for some $v, x, y \in X^+$. A word that is not unipolar is called *dipolar* (or unbordered). A language $L \subseteq X^*$ is a *prefix (suffix) code* if $u, ux \in L$ $(u, xu \in L)$ imply $x = 1$. A language is *biprefix code* if it is both a prefix and a suffix code.

Proposition 2.5 *Let $u \in X^+$. The word u is a dipolar word $\Leftrightarrow$ for every internal pilot sequence σ, the corresponding inserting sequence $\sigma(u) = \{u = u_0, u_1, u_2, \ldots, u_k, \ldots\}$ is a language that is a biprefix code.*

Proof. ($\Rightarrow$) We have $\sigma(u) \subseteq (u \leftarrow^* u)$. Let $w = vx$ with $w, v \in \sigma(u)$. Assume that $x \neq 1$. Since u is dipolar, by a result of ([7], Proposition 3.2), $1.v.x \in (u \leftarrow^* u)$ and $v \in (u \leftarrow^* u)$ with $1.x \neq 1$ imply $1.x = x \in (u \leftarrow^* u)$. Since u is a dipolar word, by a result of ([7], Proposition 3.5), $v, x \in (u \leftarrow^* u)$ imply $vx \notin (u \leftarrow^* u)$, a contradiction since $vx \in \sigma(u) \subseteq (u \leftarrow^* u)$. It follows then that $x = 1$ and hence $\sigma(u)$ is a prefix code. A similar argument will show that $\sigma(u)$ is also a suffix code and hence a biprefix code.

($\Leftarrow$) Suppose that u is not a dipolar word. We have to consider two cases depending if u is primitive or not.

Case 1. The word u is primitive. It is known (see [7], Lemma 3.1) that if a unipolar word u is primitive, then u can be expressed as $u = vwv$ for some $v, w \in X^+$. Let $n = |v|$ and let σ be the pilot sequence $\sigma = \{n, n, \ldots n, \ldots\}$. This sequence σ is internal and the corresponding inserting sequence is :

$$\sigma(u) = \{vwv, v.vwv.wv = vvwvwv, \ldots\}$$

Since vwv, $vvw.vwv \in \sigma(u)$, $\sigma(u)$ is not biprefix code.

Case 2. The word u is not primitive. Then $u = p^m$ where p is a primitive word and $m \geq 2$. Let $n = |p|$ and let σ be the internal pilot sequence $\sigma = \{n, n, \ldots, n, \ldots\}$. Then the corresponding inserting sequence is:

$$\sigma(u) = \{p^m, p.p^m.p^{m-1} = p^{2m}, \ldots\}$$

Clearly $\sigma(u)$ is not a biprefix code – a contradiction. □

Examples. Let $X = \{a, b\}$.

(1) The word $u = ab$ is dipolar. If the pilot sequence is $\sigma = \{1, \ldots, 1, \ldots\}$, then the language $\sigma(ab) = \{a^n b^n | n \geq 1\}$ is a biprefix code.

(2) The word $u = aba$ is not dipolar. Taking the same pilot sequence σ as above, the language $\sigma(aba) = \{aba, aababa, \ldots\}$ is not a biprefix code because $aababa = aab.aba$ with $aba \in \sigma(aba)$. This example can be extended in the following way. If u is a primitive word that is not dipolar, then u can be written as $u = vwv$ with $v, w \in X^+$. If $|v| = n$ and the pilot sequence is $\sigma = \{n, \ldots, n, \ldots\}$ then $\sigma(u)$ is not a biprefix code since its second word is $vvwvwv$ which has u as its suffix.

(3) The word $u = abba$ is primitive and unipolar. Using the internal pilot sequence $\sigma_1 = \{1, 1, \ldots, 1, \ldots\}$, we obtain the language

$$\sigma_1(u) = \{abba, aabbabba, \ldots\}$$

that is not a biprefix code. However if we use instead the pilot sequence $\sigma_2 = \{2, \ldots, 2, \ldots\}$, we obtain the language:

$$\sigma_2(u) = \{abba, ababbaba, \ldots, (ab)^n abba(ba)^n, \ldots\}$$

that is a biprefix code.

3 Ins-primitive words

Closely connected to the operation of catenation is the notion of primitivity. A word $u \in X^+$ is termed primitive if $u = g^n$ for some $g \in X^+$ implies $n = 1$, i.e., $u = g$. If instead of catenation we consider its generalization, the insertion operation, we obtain the notion of ins-primitivity. A word $u \in X^+$ is said to be *ins-primitive* if, for every $v \in X^+$, $n \geq 1$, $u \notin v \leftarrow^n v$. Clearly, ins-primitive words are primitive, but the converse is not true. For example $a^n b$ is ins-primitive and, for $n \geq 1$, $a^n b^n$ is primitive, but not ins-primitive as $a^n b^n \in ab \leftarrow^{n-1} ab$.

One of the main results about primitivity is that for every $u \in X^+$ there exists a unique primitive word g and a unique integer $n \geq 1$ such that $u = g^n$. The following lemma aids in showing that a similar result holds also in the case of ins-primitivity, with the exception that the corresponding ins-primitive word is not anymore unique.

Recall that a language L is *ins-closed* ([3]) if $u_1 u_2, v \in L$ imply $u_1 v u_2 \in L$.

Lemma 3.1 *If u is a word in X^* then $(u \leftarrow^n u) \leftarrow^m (u \leftarrow^p u) \subseteq (u \leftarrow^* u)$, for all $m, n, p \geq 0$.*

Proof. Let $\alpha \in (u \leftarrow^n u) \leftarrow^m (u \leftarrow^p u)$. There exists $v \in (u \leftarrow^n u)$ and $w \in (u \leftarrow^p u)$ such that $\alpha \in (v \leftarrow^m w)$. As $u \leftarrow^* u = \cup_{k \geq 0}(u \leftarrow^k u)$, we have that $v, w \in (u \leftarrow^* u)$.

The Lemma will be proved if we show that $(v \leftarrow^m w) \subseteq (u \leftarrow^* u)$.

The latter statement will be proved by induction on m. If $m = 0$ then $v \in (u \leftarrow^* u)$. The insertion closure $I(L)$ of a language L is the smallest ins-closed language containing L and it has been proven in [3] that $I(L) = L \leftarrow^* L$. By taking $L = \{u\}$ we conclude that $u \leftarrow^* u$ is ins-closed, therefore $v, w \in (u \leftarrow^* u)$ imply $v \leftarrow w \in (u \leftarrow^* u)$.

Assume the statement true for m and take a word $\beta \in (v \leftarrow^{m+1} w) = (v \leftarrow^m w) \leftarrow w$. There exists a word $\gamma \in (v \leftarrow^m w)$ such that $\beta \in \gamma \leftarrow w$. As, according to the induction hypothesis, $\gamma \in (v \leftarrow^m w) \subseteq (u \leftarrow^* u)$ and $w \in (u \leftarrow^* u)$, the fact that $(u \leftarrow^* u)$ is ins-closed implies that $\beta \in (u \leftarrow^* u)$. This implies $(v \leftarrow^{m+1} w) \subseteq (u \leftarrow^* u)$. The proof of the induction step and therefore of the lemma is thus complete. □

Proposition 3.1 *For every word $u \in X^+$ there exists an ins-primitive word v and a positive integer n such that $u \in v \leftarrow^n v$.*

Proof. Suppose u is not ins-primitive. Then there exists $v_1 \in X^+$ such that $u \in v_1 \leftarrow^{n_1} v_1$. If v_1 is not ins-primitive, then $v_1 \in v_2 \leftarrow^{n_2} v_2$ for some $v_2 \in X^*$. This implies $u \in (v_2 \leftarrow^{n_2} v_2) \leftarrow^{n_1} (v_2 \leftarrow^{n_2} v_2)$ which, according to Lemma 3.1, implies $u \in (v_2 \leftarrow^* v_2)$, where $|v_2| < |v_1|$. By repeatedly applying the procedure and Lemma 3.1, after a finite number of steps, we will get an ins-primitive word v such that $u \in v \leftarrow^* v$. According to the definition of the iterated insertion, this means that there exists a number n such that $u \in (v \leftarrow^n v)$. □

An ins-primitive word v such that $u \in v \leftarrow^n v$ is called an *ins-root* of u. In contrast with the case of primitivity, where the root is unique, a word can have *several ins-roots.* For example, let $X = \{a, b\}$ and let $w = bbabbabb$. Then:

$$w = b.babb.abb = babb \leftarrow babb \in babb \leftarrow^* babb$$

$$w = bba.bbab.b = bbab \leftarrow bbab \in bbab \leftarrow^* bbab$$

The words *babb* and *bbab* are ins-primitive words and they are both ins-roots of w.

Let $u = babb$ and $v = bbab$. Then:

$$b^2ab^2ab^2 \in (u \leftarrow^1 u) \cap (v \leftarrow^1 v)$$

$$b^2ab^2ab^2ab^3 \in (u \leftarrow^2 u) \cap (v \leftarrow^2 v)$$

$$b^2ab^2ab^2ab^2ab^4 \in (u \leftarrow^3 u) \cap (v \leftarrow^3 v)$$

$$\dots$$

$$(b^2a)^{i+1}b^{i+1} \in (u \leftarrow^i u) \cap (v \leftarrow^i v)$$

$$\dots$$

We show, by induction on i, that

$$(b^2a)^{i+1}b^{i+1} \in (u \leftarrow^i u) \cap (v \leftarrow^i v)$$

for all $i \geq 1$. Indeed, the relation holds for $i = 1$. Suppose it holds for $i-1$, i.e.

$$(b^2a)^ib^i \in (u \leftarrow^{i-1} u) \cap (v \leftarrow^{i-1} v).$$

As $(b^2a)^ib^i \in (u \leftarrow^{i-1} u)$, by inserting $u = babb$ we obtain that

$$(b^2a)^ib.babb.b^{i-1} = (b^2a)^i.b^2a.b^2.b^{i-1} = (b^2a)^{i+1}b^{i+1},$$

is a word in $(u \leftarrow^i u)$. As $(b^2a)^ib^i \in (v \leftarrow^{i-1} v)$, by inserting $v = bbab$ we obtain that

$$(b^2a)^i.bbab.b^i = (b^2a)^i.b^2a.b.b^i = (b^2a)^{i+1}b^{i+1}$$

is a word in $(v \leftarrow^i v)$. Hence the relation is true for all $i \geq 1$. This shows that u and v are both ins-roots of $(b^2a)^{n+1}b^{n+1}$ for every $n \geq 1$.

The result can be generalized to ins-primitive words of the type $u = wxww$, $v = wwxw$, $x, w \in X^+$. Namely, if u and v are ins-primitive, they are both ins-roots of the words $(w^2x)^{n+1}w^{n+1}$ for all $n \geq 1$.

The following results show that there are quite many ins-primitive words. In fact, as Corollary 3.1 states, the set of all ins-primitive words over an alphabet is right and left dense.

Lemma 3.2 *If $w \in u \leftarrow^n u$ then $N_a(w) = (n+1)N_a(u)$ for all $a \in X$.*

Proposition 3.2 *If $u \in X^+$, $a \in X$, $u \neq a^n$ then either u is ins-primitive or all the words $w \in (u \leftarrow a)$ are ins-primitive.*

Proof. If u is not ins-primitive, then $u \in v \leftarrow^n v$, where v is an ins-primitive word. According to Lemma 3.2, $N_a(u) = (n+1)N_a(v)$ for all $a \in X$. On the other hand, if $b \neq a$ is a letter occurring in u, $N_a(w) = (n+1)N_a(v) + 1$ whereas $N_b(w) = (n+1)N_b(v)$. If $w \in w' \leftarrow^m w'$, $m > 0$ with w' ins-primitive, then the numbers of $a's$ and $b's$ in w would have as common factor $m+1$. We deduce that w cannot be ins-primitive, as the number of letters a and b in w are relatively prime. □

Corollary 3.1 *Let $IQ(X)$ be the set of ins-primitive words over an alphabet X with card$(X) \geq 2$. Then $IQ(X)$ is right and left dense.*

Proof. Let $u \in X^+$. If $u = a^n$ for some $a \in X$ and if $b \in X$, $b \neq a$ then $a^n b \in IQ(X)$. If $u \neq a^n$ then, according to Proposition 3.2 all the words in $(u \leftarrow a)$ are ins-primitive. In particular, $ua \in IQ(X)$. This proves that $IQ(X)$ is right dense. By symmetry, one can show that $IQ(X)$ is also left dense. □

We conclude this section with a result connecting the notions of ins- primitivity, ins-closed language and infix order. Recall that the infix order is defined as follows: $u \leq_i v$ if u is an infix of v, that is $v = v_1 u v_2$ for some $v_1, v_2 \in X^*$.

Proposition 3.3 *Let $L \subseteq X^+$ be an ins-closed language such that L^c is ins-closed. Let $IP(L)$ be the set of ins-primitive words and let $IF(L)$ be the set of minimal words of L relatively to the infix order $\leq_i$. Then:*

(i) If $u \in L$ and if v is an ins-root of u, then $v \in L$.

(ii) If L' is an ins-closed language containing $IP(L)$ the $L \subseteq L'$.

(iii) Every word $u \in IF(L)$ is ins-primitive.

Proof. (i) Since v is an ins-root of u, $u \in v \leftarrow^n v$. If $v \in L^c$, then, since L^c is ins-closed, $v \leftarrow^n v \subseteq L^c$ and $u \in L$, a contradiction. Hence $v \in L$.

(ii) This follows from *(i)*.

(iii) Suppose u is not ins-primitive. Then $u \in v \leftarrow^n v$, $n \geq 1$ for some ins-primitive $v \in L$ (see *(i)*). Therefore $u = xvy$. Since u is minimal relatively the infix order $\leq_i$, $x = y = 1$, a contradiction. □

4 Shuffle- and com-shuffle-primitivity

Recall that the *shuffle product* $\amalg$ of two words u and v is the set:

$$u \amalg v = \{w \in X^* | w = u_1 v_1 u_2 v_2 \dots u_n v_n, u = u_1 u_2 \dots u_n, v = v_1 v_2 \dots v_n,$$
$$u_i, v_j \in X^*, 1 \leq j \leq n\}$$

Since the shuffle product is associative and commutative, the n^{th} shuffle power $u \amalg^n u$ of a word u can be defined in the usual way.

Note that the shuffle product operation generalizes the notion of catenation in the sense that the catenation uv is an element of the set $u \amalg v$. Consequently, we can use the notion of shuffle to obtain yet another generalization of the notion of primitivity, namely the shuffle-primitivity. A word $u \in X^+$ is *shuffle-primitive* (or shortly, *shf-primitive*) if $u \notin (v \amalg^n v)$ for any $v \in X^+$ and $n \geq 1$. Clearly every shf-primitive word is primitive and ins-primitive.

The converse is not true. For example, the word *aabbaa* is not shf-primitive because $aabbaa \in aba \amalg aba$. On the other hand, *aabbaa* is primitive and ins-primitive. Indeed, from the form of the word we see that the only possibility for it to be obtained as an insertion, is that $aabbaa \in u \leftarrow u$ where u contains two a's and one b. (Two is the only common divisor of the number of a's and b's, therefore only one insertion is performed). By exploring all the possibilities, we deduce that no u of this form produces *aabbaa*, therefore *aabbaa* is ins-primitive.

Proposition 4.1 *For every word $u \in X^+$ there exists a shf-primitive word v and a positive integer such that $u \in v \amalg^n v$.*

Proof. Suppose that u is not shf-primitive. Then there exists $v_1 \in X^+$ such that $u \in v_1 \amalg^{n_1} v_1$. If v_1 is not shf-primitive, then $v_1 \in v_2 \amalg^{n_2} v_2$. This implies $u \in (v_2 \amalg^{n_2} v_2) \amalg^{n_1} (v_2 \amalg^{n_2} v_2)$ which, because the the associativity of the shuffle operation, is included in $(v_2 \amalg^* v_2)$. By repeatedly applying the procedure, after a finite number of steps, we will get a shf-primitive word v such that $u \in v \amalg^* v$. According to the definition of the iterated shuffle, this means that there exists a number n such that $u \in (v \amalg^n v)$. □

A shf-primitive word v such that $u \in v \amalg^n v$ is called an *shf-root* of u.

A word u can have *several shf-root.* Let $X = \{a, b\}$ and let $u = bbabbabb$. Then:

$$u \in (babb \amalg babb) \cap (bbab \amalg bbab)$$

The words *babb* and *bbab* are shf-primitive words and they both are shf-roots of u.

Recall that a language L is shuffle-closed if $u, v \in L$ imply $u \amalg v \subseteq L$.

Proposition 4.2 *Let $L \subseteq X^+$ be a shuffle-closed language such that L^c is also shuffle-closed. Let $SH(L)$ be the set of shf-primitive words in L and let $EB(L)$ be the set of the minimal words of L relatively to the embedding order $\leq_e$. Then:*

(i) If $u \in L$ and if v is a shf-root of u then $v \in L$.

(ii) If L' is a shuffle-closed language containing $SH(L)$, then $L \subseteq L'$.

(iii) Every word $u \in EB(L)$ is shf-primitive.

Proof. The proof is similar to the proof of Proposition 3.3. □

The shuffle operation can be further generalized if we consider relaxing the condition that the order of the letters in the words is preserved. The *com-shuffle* of u and v, $u@v$, is defined by:

$$u@v = com(u) \amalg com(v),$$

where $com(v)$ is the commutative closure of v, i.e., the set of all words obtained by arbitrarily permuting the letters of v. A language L is *com-shuffle-closed* if $u, v \in L$ implies $u@v \in L$. Since the com-shuffle product is associative, the n^{th} com-shuffle power $u@^n u$ of a word u can be defined in the usual way. A word $u \in X^+$ is *com-shf-primitive if $u \notin v@^n v$* for every $v \in X^+$ and $n \geq 1$.

Proposition 4.3 *For every word $u \in X^+$ there exists a com-shf-primitive word v and a positive integer such that $u \in v@^n v$.*

Proof. Proof similar to the proof of Proposition 4.1, by using the fact that the com-shuffle is an associative operation. □

A com-shf-primitive word v such that $u \in v@^n v$ is called a *com-shf-root* of u.

Proposition 4.4 *Let $u \in X^*$ be a word and let alph$(u) = \{a_1, \ldots a_n\}$, $n \geq 2$. Then u is com-shf-primitive iff the numbers $N_{a_1}(u)$,$N_{a_2}(u)$,$\ldots$,$N_{a_n}(u)$ are relatively prime (i.e. not having any common divisors other than the unity).*

Proof. Assume u is com-shf-primitive. Suppose that the numbers $N_{a_i}(u)$, $1 \leq i \leq n$ have m as the greatest common divisor. If we take now the word w consisting of $N_{a_i}(u)/m$ letters a_i for each $1 \leq i \leq n$ (their order does not matter) then we have that $u \in w@^{m-1}w$ which contradicts the fact that u is com-shf-primitive.

For the converse implication take a word u with the property that $N_{a_i}(u)$ are relatively prime, $1 \leq i \leq n$. The word is com-shf-primitive because, if $u \in w@^m w$ for some com-shf-primitive word w and number m, then $m+1$ would be a common divisor of $N_{a_i}(u)$, $1 \leq i \leq n$.

References

[1] L. Kari. *On Insertion and Deletion in Formal Languages.* Ph.D. Thesis, University of Turku, 1991.

[2] L.Kari. On language equations with invertible operations. *Theoretical Computer Science*, 132(1994), 129-150.

[3] M. Ito, L. Kari and G. Thierrin. Insertion and Deletion Closure of Languages. To appear in *Theoretical Computer Science*.

[4] M. Ito, G. Thierrin and S.S. Yu. Shuffle-closed Languages, To appear in *Publicatione Mathematicae*.

[5] A.Salomaa. *Formal Languages*. Academic Press, New York, 1973.

[6] H.J.Shyr. *Free monoids and languages*. Institute of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan, 1991.

[7] H. J. Shyr, G. Thierrin and S.S. Yu. Monogenic e-closed languages and dipolar words, *Discrete Mathematics* 126(1994), 339-348.

[8] G. Thierrin and S.S. Yu. Shuffle relations and codes. *J. Information & Optimization Sciences* 12(1991), 441-449.