



HAL
open science

Invariance: a Theoretical Approach for Coding Sets of Words Modulo Literal (Anti)Morphisms

Jean Néraud, Carla Selmi

► **To cite this version:**

Jean Néraud, Carla Selmi. Invariance: a Theoretical Approach for Coding Sets of Words Modulo Literal (Anti)Morphisms. 2017. hal-01519557v1

HAL Id: hal-01519557

<https://hal.science/hal-01519557v1>

Preprint submitted on 15 May 2017 (v1), last revised 27 Jul 2017 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Invariance: a Theoretical Approach for Coding Sets of Words Modulo Literal (Anti)Morphisms

Jean Néraud and Carla Selmi

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes, University of Rouen,
France

`neraud.jean@free.fr, carla.selmi@univ-rouen.fr`

Abstract. Let A be a finite or countable alphabet and let θ be literal (anti)morphism onto A^* (by definition, such a correspondence is determined by a permutation of the alphabet). This paper deals with sets which are invariant under θ (θ -invariant for short). We establish an extension of the famous defect theorem. Moreover, we prove that for the so-called thin θ -invariant codes, maximality and completeness are two equivalent notions. We prove that a similar property holds for some special families of θ -invariant codes such as prefix (bifix) codes, codes with a finite deciphering delay, uniformly synchronous codes and circular codes. For a special class of involutive antimorphisms, we prove that any regular θ -invariant code may be embedded into a complete one.

Keywords: word, code, variable length code, morphism, antimorphism, literal, prefix, bifix, deciphering delay, synchronizing delay, circular, defect, equation, maximal, complete.

1 Introduction

During the last decade, in the free monoid theory, due to their powerful applications, in particular in DNA-computing, one-to-one *morphic* or *antimorphic* correspondences play a particularly important part. Given a finite or countable *alphabet*, say A , any such mapping is a substitution which is completely determined by extending a unique permutation of A onto A^* (the *free monoid* that it generates). The resulting mapping is commonly referred to as *literal* (or *letter-to-letter*) moreover, in the case of a finite alphabet, it is well known that such a correspondence is idempotent with respect to the composition.

In the special case of involutive morphisms or antimorphisms -we write (anti)morphisms for short, lots of successful investigations have been done for extending the now classical combinatorial properties on words: we mention the study of the so-called pseudo-palindromes [2, 6], or that of pseudo-repetitions [1, 8, 10]. The framework of some peculiar families of codes [9] and equations in words [5, 13] were also concerned. Moreover, in the more general family of idempotent (anti)morphisms, a nice generalization of the famous theorem of Fine and Wilf [11, Proposition 1.3.5] has been recently established in [12].

Equations in words are also the starting point of the study in the present paper, where we adopt the point of view from [11, Chap. 9]. Let A be a finite or countable alphabet; a literal (anti)morphism, namely θ , being fixed, consider a finite collection of unknown words, say Z . In view of making the present foreword more readable, in the first instance we take θ as an involutive literal substitution (that is $\theta^2 = id_{A^*}$). We assign that the words in Z and their images by θ to satisfy a given equation, and we ask for the computation of a finite set of words, say Y , such that all the words of Z can be expressed as a concatenation of words in Y . Actually, such a question might be more complex than in the classical configuration,

where θ does not interfere: in that classical case, according to the famous defect theorem [11, Theorem 1.2.5], it is well known that at most $|Z| - 1$ words allow to compute the words in Z . At the contrary, in [10], examples where $|Y| = |Z|$ are provided by the authors.

Along the way, for solving our problem, applying the defect theorem to the set $X = Z \cup \theta(Z)$ might appear as natural. Such a methodology guarantees the existence of a set Y , with $|Y| \leq |X| - 1$ and whose elements allow by concatenation to rebuild all the words in X . It is also well known that Y can be chosen in such a way that only trivial equations hold among its elements: with the terminology of [3], Y is a *code*, or equivalently Y^* , the submonoid that it generates, is *free*. Unfortunately, since both the words in Z and $\theta(Z)$ are expressed as concatenations of words in Y , among the words of $Y \cup \theta(Y)$ non-trivial equations can hold; in other words, by applying that methodology, the initial problem would be transferred among the words in $Y \cup \theta(Y)$.

An alternative methodology will consist in asking for codes Y which are invariant under θ (θ -invariant for short), that is $\theta(Y) = Y$. Returning to the general case where θ is a literal idempotent (anti)morphism, this is equivalent to say that the union of the sets $\theta^i(Y)$, for all $i \in \mathbb{Z}$, is itself θ -invariant. By the way, it is straightforward to prove that the intersection of an arbitrary family of free θ -invariant submonoids is itself a free θ -invariant submonoid. In the present paper we prove the following result:

Theorem 1. *Let A be a finite or countable alphabet, let θ be a literal (anti)morphism onto A^* , and let X be a finite θ -invariant set. If X is not a code, then the smallest θ -invariant free submonoid of A^* containing X is generated by a θ -invariant code Y which satisfies $|Y| \leq |X| - 1$.*

For illustrating this result in term of equations, we refer to [5, 13], where the authors considered generalizations of the famous three unknown variables equation of Lyndon-Shützenberger [11, Sect. 9.2]. They proved that, an involutive (anti)morphism θ being fixed, given such an equation with sufficiently long members, a word t exists such that any 3-uple of “solutions” can be expressed as a concatenation of words in $\{t\} \cup \{\theta(t)\}$. With the notation of Theorem 1, the elements of the θ -invariant set X are $x, y, z, \theta(x), \theta(y), \theta(z)$ and those of Y are t and $\theta(t)$: we verify that Y is a θ -invariant code with $|Y| \leq |X| - 1$.

In the sequel, we will continue our investigation by studying the properties of complete θ -invariant codes: a code X is *complete* if any word of A^* is a factor of some words in X^* . From this point of view, a famous result from Schützenberger states that, for the wide family of the so-called *thin* codes (which contains regular codes) [3, Sect. 2.5], maximality and completeness are two equivalent notions. In the framework of invariant codes, we prove the following result:

Theorem 2. *Let A be a finite or countable alphabet. Given a thin θ -invariant code $X \subseteq A^*$, the three following conditions are equivalent:*

- (i) X is complete
- (ii) X is a maximal code
- (iii) X is maximal in the family of the θ -invariant codes.

In the proof, the main feature consists in establishing that a non-complete θ -invariant code X cannot be maximal in the family of θ -invariant codes: actually, the most delicate step consists in constructing a convenient θ -invariant set $Z \subseteq A^*$, with $X \cap Z = \emptyset$ and such that $X \cup Z$ remains itself a θ -invariant code.

It is well known that the preceding result from Schützenberger has been successfully extended to some famous families of thin codes, such as *prefix (bifix, uniformly synchronous, circular)* codes (cf [3, Proposition 3.3.8], [3, Proposition 6.2.1], [3, Theorem 10.2.11], [4,

Proposition 3.6] and [14, Theorem 3.5]) and codes with a *finite deciphering delay* (f.d.d. codes, for short) [3, Theorem 5.2.2]. From this point of view, we will examine the behavior of corresponding families of θ -invariant codes.

Actually we establish a result similar to Theorem 2 in the framework of the family of prefix (bifix, f.d.d., two-way f.d.d, uniformly synchronized, circular codes). In the proof, a construction very similar to the previous one may be used in the case of prefix, bifix, f.d.d., two-way f.d.d codes. At the contrary, investigating the behavior of circular codes with regards to the question necessitates the computation of a more sophisticated set; moreover the family of uniformly synchronized codes itself impose to make use of a significantly different methodology.

In the last part of our study, we address to the problem of embedding a non-complete invariant code into a complete one. For the first time, this question was stated in [15], where the author asked whether any finite code can be imbedded into a regular one. A positive answer was provided in [7], where was established a formula for embedding any regular code into a complete one. From the point of view of θ -invariant codes, we obtain a positive answer only in the case where θ is an involutive antimorphism which is different of the so-called mirror image; actually the general question remains open.

We now describe the contents of the paper. Section 2 contains the preliminaries: the terminology of the free monoid is settled, and the definitions of some classical families of codes are recalled. Theorem 1 is established in Section 3, where an original example of equation is studied. The proof of Theorem 2 is done in Section 3, and extensions for special families of θ -invariant codes are studied in Section 4. The question of embedding a regular θ -invariant code into a complete one is examined in Section 5.

2 Preliminaries

We adopt the notation of the free monoid theory: given an alphabet A , we denote by A^* the free monoid that it generates. Given a word w , we denote by $|w|$ its length, the empty word, that we denote by ε , being the word with length 0. We denote by w_i the letter of position i in w : with this notation we have $w = w_1 \cdots w_{|w|}$. We set $A^+ = A^* \setminus \{\varepsilon\}$. Given $x \in A^*$ and $w \in A^+$, we say that x is a *prefix* (*suffix*) of w if a word u exists such that $w = xu$ ($w = ux$). Similarly, x is a *factor* of w if a pair of words u, v exists such that $w = uxv$. Given a non-empty set $X \subseteq A^*$, we denote by $P(X)$ ($S(X), F(X)$) the set of the words that are prefix (suffix, factor) of some word in X . Clearly, we have $X \subseteq P(X)$ ($S(X), F(X)$). A set $X \subseteq A^*$ is *complete* iff $F(X^*) = A^*$. Given a pair of words w, w' , we say that it *overlaps* if words u, v exist such that $uw' = wv$ or $w'u = vw$, with $1 \leq |u| < |w|$ and $1 \leq |v| < |w'|$; otherwise, the pair is *overlapping-free* (in such a case, if $w = w'$, we simply say that w is overlapping-free).

It is assumed that the reader has a fundamental understanding with the main concepts of the theory of variable length codes: we only recall some of the main definitions and we suggest, if necessary, that he (she) report to [3]. A set X is a *variable length code* (a *code* for short) iff any equation among the words of X is trivial, that is, for any pair of sequences of words in X , namely $(x_i)_{1 \leq i \leq m}, (y_j)_{1 \leq j \leq n}$, the equation $x_1 \cdots x_m = y_1 \cdots y_n$ implies $m = n$ and $x_i = y_i$ for each integer $i \in [1, m]$. By definition X^* , the submonoid of A^* which is generated by X , is *free*. Equivalently, X^* satisfies the property of *equidivisibility*, that is $(X^*)^{-1}X^* \cap X^*(X^*)^{-1} = X^*$.

Some famous families of codes that have been studied in the literature: X is a *prefix* (*suffix, bifix*) *code* iff $X \neq \{\varepsilon\}$ and $X \cap XA^+ = \emptyset$ ($X \cap A^+X = \emptyset, X \cap XA^+ = X \cap A^+X = \emptyset$). X is a code with a *finite deciphering delay* (*f.d.d. code* for short) if it is a code and if a non-

negative integer d exists such that $X^{-1}X^* \cap X^d A^+ \subseteq X^+$. With this condition, if another integer d' exists such that we have $X^* X^{-1} \cap A^+ X^{d'} \subseteq X^+$, we say that X is a *two-way f.d.d. code*. X is a *uniformly synchronized code* if it is a code and if a positive integer k exists such that, for all $x, y \in X^k, u, v \in A^+$: $uxyv \in X^* \implies ux, xv \in X^*$. X is a *circular code* if for any pair of sequences of words in X , namely $(x_i)_{1 \leq i \leq m}, (y_j)_{1 \leq j \leq n}$, and any pair of words s, p , with $s \neq \varepsilon$, the equation $x_1 \cdots x_m = sy_2 \cdots y_n p$, with $y_1 = ps$, implies $m = n, p = \varepsilon$ and $x_i = y_i$ for each $i \in [1, m]$.

In the whole paper, we consider a *finite* or *countable* alphabet A and a mapping θ which satisfies each of the three following conditions:

- (a) θ is a one-to-one correspondence onto A^*
- (b) θ is *literal*, that is $\theta(A) \subseteq A$
- (c) either θ is a *morphism* or it is an *antimorphism* (it is an antimorphism if $\theta(\varepsilon) = \varepsilon$ and $\theta(xy) = \theta(y)\theta(x)$, for any pair of words x, y); for short in any case we write that θ is an *(anti)morphism*.

In the case where A is a finite set, it is well known that the literal (anti)morphism θ is idempotent (that is, an integer n exists such that $\theta^n = id_{A^*}$). In the whole paper, we are interested in the family of sets $X \subseteq A^*$ that are invariant under the mapping θ (θ -invariant for short), that is $\theta(X) = X$.

3 A defect effect for invariant sets

Informally, the famous defect theorem says that if some words of a set X satisfy a non-trivial equation, then these words may be written upon an alphabet of smaller size. In this section, we examine whether a corresponding result may be stated in the framework of θ -invariant sets. The following property comes from the definition:

Proposition 1. *Let M be a submonoid of A^* and let $S \subseteq A^*$ be such that $M = S^*$. Then M is θ -invariant if and only if S is θ -invariant.*

Clearly the intersection of a non-empty family of θ -invariant free submonoids of A^* is itself a θ -invariant free submonoid. Given a submonoid M of A^* , recall that its *minimal generating set* is $(M \setminus \{\varepsilon\}) \setminus (M \setminus \{\varepsilon\})^2$.

Theorem 1. *Let A be a finite or countable alphabet, let $X \subseteq A^*$ be a θ -invariant set and let Y be the minimal generating set of the smallest θ -invariant free submonoid of A^* which contains X . If X is not a code, then we have $|Y| \leq |X| - 1$.*

Proof. With the notation of Theorem 1, since Y is a code, each word $x \in X$ has a unique factorization upon the words of Y , namely $x = y_1 \cdots y_n$, with $y_i \in Y$ ($1 \leq i \leq n$). In a classical way, we say that y_1 (y_n) is the *initial* (*terminal*) factor of x (with respect to such a factorization). Before to prove our result, we shall establish the following lemma:

Lemma 1. *With the preceding notation, each word in Y is the initial (terminal) factor of a word in X .*

Proof. By contradiction, assume that a word $y \in Y$ that is never initial of any word in X exists. Set $Z_0 = (Y \setminus \{y\})\{y\}^*$ and $Z_i = \theta^i(Z_0)$, for each integer $i \in \mathbb{Z}$. In a classical way (cf e.g. [11, p. 7]), since Y is a code, Z_0 itself is a code. Since θ^i is a one-to-one correspondence, for each integer $i \in \mathbb{Z}$, Z_i is a code, that is Z_i^* is a free submonoid of A^* . Consequently, the intersection, namely M , of the family $(Z_i^*)_{i \in \mathbb{Z}}$ is itself a free submonoid of A^* . Moreover, since Y is θ -invariant, we have $\theta(M) \subseteq M$ therefore, since θ is onto, we obtain $\theta(M) = M$.

Let x be an arbitrary word in X . Since $X \subseteq Y^*$, and according to the definition of y , we have $x = (z_1 y^{k_1})(z_2 y^{k_2}) \cdots (z_n y^{k_n})$, with $z_1, \dots, z_n \in Y \setminus \{y\}$ and $k_1, \dots, k_n \geq 0$. Consequently x belongs to Z_0^* , therefore we have $X \subseteq Z_0^*$. Since X is θ -invariant, this implies $X = \theta(X) \subseteq Z_i^*$ for each $i \in \mathbb{Z}$, thus $X \subseteq M$.

But the word y belongs to Y^* and doesn't belong to Z_0^* thus it doesn't belong to M . This implies $X \subseteq M \subsetneq Y^*$: a contradiction with the minimality of Y^* . ■

Proof of Theorem 1. Let α be the mapping from X onto Y which, with every word $x \in X$, associates the initial factor of x in its (unique) factorization over Y^* . According to Lemma 1, α is onto. We will prove that it is not one-to-one. Classically, since X is not a code, a non-trivial equation may be written among its words, say:

$x_1 \cdots x_n = x'_1 \cdots x'_m$, with $x_i, x'_j \in X$ $x_1 \neq x'_1$ ($1 \leq i \leq n, 1 \leq j \leq m$). Since Y is a code, a unique sequence of words in Y , namely y_1, \dots, y_p exists such that:
 $x_1 \cdots x_n = x'_1 \cdots x'_m = y_1 \cdots y_p$. This implies $y_1 = \alpha(x_1) = \alpha(x'_1)$ and completes the proof. ■

In what follows we discuss some interpretation of Theorem 1 with regards to equations in words. For this purpose, we assume that A is finite, θ being idempotent of order n , and we consider a finite set of words, say Z . Let X be the union of the sets $\theta^i(Z)$, for $i \in [1, n]$, and assume that a non-trivial equation holds among the words of X , namely $x_1 \cdots x_m = y_1 \cdots y_p$. By construction X is θ -invariant therefore, according to Theorem 1, a θ -invariant code Y exists such that $X \subseteq Y^*$, with $|Y| \leq |X| - 1$. This means that each of the words in X can be expressed by making use of at most $|X| - 1$ words of type $\theta^i(u)$, with $u \in Y$ and $1 \leq i \leq n$. It will be easily verified that the examples from [5, 10, 13] corroborate this fact, moreover below we mention an original one:

Example 1. Let θ be a literal antimorphism of order 3. Consider two different words x, y , with $|x| > |y|$, satisfying the equation: $x\theta(y) = \theta^2(y)\theta(x)$. With this condition, a pair of words u, v exists such that $x = uv$, $\theta^2(y) = u$, thus $y = \theta(u)$, moreover we have $v = \theta(v)$ and $u = \theta(u) = \theta^2(u)$. With the preceding notation, we have $Z = \{x, y\}$, $X = Z \cup \theta(Z) \cup \theta^2(Z)$, $Y = \{u\} \cup \{v\} \cup \{\theta(u)\} \cup \{\theta(v)\} \cup \{\theta^2(u)\} \cup \{\theta^2(v)\}$. It follows from $y = \theta(y) = \theta^2(y)$ that $X = \{x\} \cup \{\theta(x)\} \cup \{\theta^2(x)\} \cup \{y\}$.

- At first, assume that no word t may exists such that $u, v \in t^+$. In a classical way, we have $uv \neq vu$, thus $X = \{x, \theta(x), \theta^2(x), y\}$ and $Y = \{u, v\}$. We verify that $|Y| \leq |X| - 1$.

- Now, assume that we have $u, v \in t^+$. We obtain $X = Z = \{x, y\}$ and $Y = \{t\}$. Once more we have $|Y| \leq |X| - 1$.

4 Maximal θ -invariant codes

Given set $X \subseteq A^*$, it is *thin* iff $A^* \neq F(X)$. Regular codes are well known examples of thin codes. From the point of view of maximal codes, let's recall one of the famous result stated by Schützenberger:

Theorem 2. [3, Theorem 2.5.16] *Let X be an thin code. Then the following conditions are equivalent:*

- (i) X is complete
- (ii) X is a maximal code.

The aim of this section is to examine whether a similar result may be stated in the family of θ -invariant codes. In the case where $|A| = 1$, we have $\theta = id_{A^*}$, moreover the codes are all the singletons in A^+ . Therefore any code is θ -invariant, maximal and complete. In the rest of the paper, we assume that $|A| \geq 2$.

Some notations. Let X be a non-complete θ -invariant code, and let $y \notin F(X^*)$. Without loss of generality, we may assume that the initial and the terminal letters of y are different (otherwise, substitute to y the word $ay\bar{a}$, with $a, \bar{a} \in A$ and $a \neq \bar{a}$), we may also assume that $|y| \geq 2$. Set:

$$y = ax\bar{a}, \quad z = \bar{a}^{|y|}y\bar{a}^{|y|} = \bar{a}^{|y|}ax\bar{a}^{|y|}. \quad (1)$$

Since θ is a literal (anti)morphism, for each integer $i \in \mathbb{Z}$, a pair of different letters b, \bar{b} and a word x' exist such that $|x'| = |x| = |y| - 2$, and:

$$\theta^i(z) = \bar{b}^{|y|}\theta^i(y)b^{|y|} = \bar{b}^{|y|}bx'\bar{b}b^{|y|}. \quad (2)$$

Given two (not necessarily different) integers $i, j \in \mathbb{Z}$, we will accurately study how the two words $\theta^i(z), \theta^j(z)$ may overlap.

Lemma 2. *With the notation in (2), let $u, v \in A^+$ and $i, j \in \mathbb{Z}$ such that $|u| \leq |z| - 1$ and $\theta^i(z)v = u\theta^j(z)$. Then we have $|u| = |v| \geq 2|y|$, moreover a letter b and a unique positive integer k (depending of $|u|$) exist such that we have $\theta^i(z) = ub^k$, $\theta^j(z) = b^k v$, with $k \leq |y|$.*

Proof. According to (2), we set $\theta^i(z) = \bar{b}^{|y|}bx'\bar{b}b^{|y|}$ and $\theta^j(z) = \bar{c}^{|y|}cx''\bar{c}c^{|y|}$, with $b, \bar{b}, c, \bar{c} \in A$ and $b \neq \bar{b}, c \neq \bar{c}$. Since θ is a literal (anti)morphism, we have $|\theta^i(z)| = |\theta^j(z)|$ thus $|u| = |v|$; since we have $1 \leq |u| \leq 3|y| - 1$, exactly one of the following cases occurs:

Case 1: $1 \leq |u| \leq |y| - 1$. With this condition, we have $(\theta^i(z))_{|u|+1} = \bar{b} = \bar{c} = (u\theta^j(z))_{|u|+1}$ and $(\theta^i(z))_{|y|+1} = b = \bar{c} = (u\theta^j(z))_{|y|+1}$, which contradicts $b \neq \bar{b}$.

Case 2: $|u| = |y|$. This condition implies $(\theta^i(z))_{|u|+1} = b = \bar{c} = (u\theta^j(z))_{|u|+1}$ and $(\theta^i(z))_{2|y|} = \bar{b} = \bar{c} = (u\theta^j(z))_{2|y|}$, which contradicts $b \neq \bar{b}$.

Case 3: $|y| + 1 \leq |u| \leq 2|y| - 1$. We obtain $(\theta^i(z))_{2|y|} = \bar{b} = \bar{c} = (u\theta^j(z))_{2|y|}$ and $(\theta^i(z))_{2|y|+1} = b = \bar{c} = (u\theta^j(z))_{2|y|+1}$ which contradicts $b \neq \bar{b}$.

Case 4: $2|y| \leq |u| \leq 3|y| - 1$. With this condition, necessarily we have $b = \bar{c}$, therefore an integer $k \in [1, |y|]$ exists such that $\theta^i(z) = ub^k$ and $\theta^j(z) = b^k v$. ■

Set $Z = \{\theta^i(z) | i \in \mathbb{Z}\}$. Since $y \notin F(X^*)$ and since X is θ -invariant, for any integer $i \in \mathbb{Z}$ we have $\theta^i(z) \notin F(X^*)$, hence we obtain $Z \cap F(X^*) = \emptyset$. By construction, all the words in Z have length $|z|$ moreover, as a consequence of Lemma 2:

Lemma 3. *With the preceding notation, we have $A^+ZA^+ \cap ZX^*Z = \emptyset$.*

Proof. By contradiction, assume that $z_1, z_2, z_3 \in Z$, $x \in X^*$ and $u, v \in A^+$ exist such that $uz_1v = z_2xz_3$. By comparing the lengths of u, v with $|z|$, exactly one of the three following cases occurs:

Case 1: $|z| \leq |u|$ and $|z| \leq |v|$. With this condition, we have $z_2 \in P(u)$ and $z_3 \in S(v)$, therefore the word z_1 is a factor of x : this contradicts $Z \cap F(X^*) = \emptyset$.

Case 2: $|u| < |z| \leq |v|$. We have in fact $u \in P(z_2)$ and $z_3 \in S(v)$. We are in the condition of Lemma 2: the words z_2, z_1 overlap. Consequently, $u \in A^+$ and $b \in A$ exist such that $z_2 = ub^k$ and $z_1 = b^k z'_1$, with $1 \leq k \leq |y|$. But by construction we have $|uz_1| = |z_2xz_3| - |v|$.

Since we assume $|v| \geq |z|$, this implies $|uz_1| \leq |z_2xz_3| - |z| = |z_2x|$, hence we obtain $uz_1 = ub^kz'_1 \in P(z_2x)$. It follows from $z_2 = ub^k$ that $z'_1 \in P(x)$. Since $z_1 \in Z$ and according to (2), $i \in \mathbb{Z}$ and $\bar{b} \in A$ exist such that we have $z_1 = b^kz'_1 = b^{|y|}\theta^i(y)\bar{b}^{|y|}$. Since by Lemma 2 we have $|z'_1| = |u| \geq 2|y|$, we obtain $\theta^i(y) \in F(z'_1)$, which contradicts $y \notin F(X^*)$.

Case 3: $|v| < |z| \leq |u|$. Same arguments on the reversed words lead to a conclusion similar to that of Case 2.

Case 4: $|z| > |u|$ and $|z| > |v|$. With this condition, both the pairs of words z_2, z_1 and z_1, z_3 overlap. Once more we are in the condition of Lemma 2: letters c, d , words u, v, s, t , and integers h, k exist such that the two following properties hold:

$$z_2 = uc^h, \quad z_1 = c^hs, \quad |u| = |s| \geq 2|y|, \quad h \leq |y|, \quad (3)$$

$$z_1 = td^k, \quad z_3 = d^kv, \quad |v| = |t| \geq 2|y|, \quad k \leq |y|. \quad (4)$$

It follows from $uz_1v = z_2xz_3$ that $uz_1v = (uc^h)x(d^kv)$, thus $z_1 = c^hd^k$. Once more according to (2), $i \in \mathbb{Z}$ and $\bar{c} \in A$ exist such that we have $z_1 = c^{|y|}\theta^i(y)\bar{c}^{|y|}$. Since we have $h, k \leq |y|$, this implies $d = \bar{c}$ moreover $\theta^i(y)$ is a factor of x . Once more, this contradicts $y \notin F(X^*)$. ■

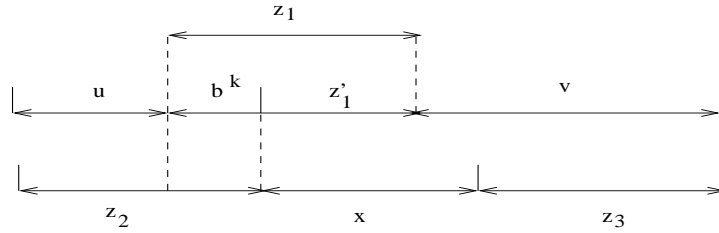


Fig. 1. Proof of Lemma 3: Case 2.

Thanks to Lemma 3 we will prove some meaningful results in Section 5. Presently, we will apply it in a special context:

Corollary 1. *With the preceding notation, X^*Z is a prefix code.*

Proof. Let $z_1, z_2 \in Z, x_1, x_2 \in X^*, u \in A^+$, such that $x_1z_1u = x_2z_2$. For any word $z_3 \in Z$, we have $(z_3x_1)z_1(u) = z_3x_2z_1$, a contradiction with Lemma 3. ■

We are now ready to prove the main result of the section:

Theorem 3. *Let A be a finite or countable alphabet and $X \subseteq A^*$ be a thin θ -invariant code. Then the following conditions are equivalent:*

- (i) X is complete
- (ii) X is a maximal code
- (iii) X is maximal in the family θ -invariant codes.

Proof. Let X be a θ -invariant code. According to Theorem 2, if X is thin and complete, then it is a maximal code, therefore X is maximal in the family of θ -invariant codes. For proving

the converse, we consider a set X which is maximal in the family of θ -invariant codes. Assume that X is not complete and let $y \notin F(X^*)$. Define the word z as in (1) and consider the set $Z = \{\theta^i(z) | i \in \mathbb{Z}\}$. At first, we will prove that $X \cup Z$ remains a code. In view of that, we consider an arbitrary equation between the words in $X \cup Z$. Since X is a code, without loss of generality, we may assume that at least one element of Z has at least one occurrence in one of the two sides of this equation. As a matter of fact, with such a condition and since $Z \cap F(X^*) = \emptyset$, two sequences of words in X^* , namely $(x_i)_{1 \leq i \leq n}, (x'_j)_{1 \leq j \leq p}$ and two sequences of words in Z , namely $(z_i)_{1 \leq i \leq n-1}, (z'_j)_{1 \leq j \leq p-1}$ exist such that the equation takes the following form:

$$x_1 z_1 x_2 z_2 \cdots x_{n-1} z_{n-1} x_n = x'_1 z'_1 x'_2 z'_1 \cdots x'_{p-1} z'_{p-1} x'_p. \quad (5)$$

Without loss of generality, we assume $n \geq p$. At first, according to Corollary 1, necessarily, we have $x_1 = x'_1$, therefore Equation (5) is equivalent to: $z_1 x_2 z_2 \cdots x_{n-1} z_{n-1} x_n = z'_1 x'_2 z'_2 \cdots x'_{p-1} z'_{p-1} x'_p$, however, since all the words in Z have a common length, we have $z_1 = z'_1$ hence our equation is equivalent to $x_2 z_2 \cdots x_{n-1} z_{n-1} x_n = x'_2 z'_2 \cdots x'_{p-1} z'_{p-1} x'_p$. Consequently, by applying iteratively the result of Corollary 1, we obtain: $x_2 = x'_2, \dots, x_p = x'_p$, which implies $x_{p+1} z_{p+1} \cdots z_{n-1} x_n = \varepsilon$, thus $n = p$. In other words Equation (5) is trivial, thus $X \cup Z$ is a code.

Next, since we have $\theta(X \cup Z) \subseteq \theta(X) \cup \theta(Z) = X \cup Z$, the code $X \cup Z$ is θ -invariant. It follows from $z \in Z \setminus X$ that X is strictly included in $X \cup Z$: this contradicts the maximality of X in the whole family of θ -invariant codes, and completes the proof of Theorem 3. ■

Example 2. Let $A = \{a, b, c\}$. Consider the antimorphism θ which is generated by the permutation $\sigma(a) = b, \sigma(b) = c, \sigma(c) = a$. Consider the set $X = \{ab, cb, ca, ba, bc, ac\}$ which is a code invariant under θ . We have $a^3 \notin F(X^*)$ by taking $y = a^3b, z = b^4 \cdot a^3b \cdot a^4$, we are in Condition (1). The corresponding set Z is $\{\theta^i(z) | i \in \mathbb{Z}\} = \{b^4cb^3a^4, b^4c^3ac^4, a^4ba^3c^4, a^4b^3cb^4, c^4ac^3b^4, c^4a^3ba^4\}$. Since $X \cup Z$ is a prefix set, this guarantees that $X \cup Z$ remains a θ -invariant code.

5 Maximality in some families of θ -invariant codes

In the literature, statements similar to Theorem 2 were established in the framework of some special families of thin codes. In this section we will draw similar investigations with regards to θ -invariant codes. We will establish the following result:

Theorem 4. *Let A be a finite or countable alphabet and let $X \subseteq A^*$ be a thin θ -invariant prefix (resp. bifix, f.d.d., two-way f.d.d., uniformly synchronized, circular) code. Then the following conditions are equivalent:*

- (i) X is complete
- (ii) X is a maximal code
- (iii) X is maximal in the family of prefix (bifix, f.d.d., two-way f.d.d., uniformly synchronized, circular) codes
- (iv) X is maximal in the family θ -invariant codes
- (v) X is maximal in the family of θ invariant prefix (bifix, f.d.d., two-way f.d.d., uniformly synchronized, circular) codes.

Sketch proof. According to Theorem 3, and thanks to [3, Proposition 3.3.8], [3, Proposition 6.2.1], [3, Theorem 5.2.2], [4, Proposition 3.6] and [14, Theorem 3.5], if X is complete then

it is maximal in the family of θ -invariant codes and maximal in the family of θ -invariant prefix (bifix, f.d.d., two-way f.d.d, uniformly synchronized, circular) codes. Consequently, the proof of Proposition 4 comes down to establish that if X is not complete, then it cannot be maximal in the family of θ -invariant prefix (bifix, f.d.d., wo-way f.d.d, uniformly synchronized, circular) codes.

1) Let's begin by θ -invariant prefix codes. At first, we assume that θ is an antimorphism. Since $X \cap XA^+ = \emptyset$, and since θ is injective, we have $\theta(X) \cap \theta(XA^+) = \emptyset$, thus $X \cap A^+X = \emptyset$, hence X is also a suffix code. Assume that X is not complete. According to [3, Proposition 3.3.8], it is non-maximal in both the families of prefix codes and suffix codes. Therefore a pair of words $y, y' \in A^+ \setminus X$ exists such $X \cup \{y\}$ ($X \cup \{y'\}$) remains a prefix (suffix) code. By construction, $X \cup \{yy'\}$ remains a code which is both prefix and suffix. Set $Y = \{\theta^i(yy') \mid i \in \mathbb{Z}\}$: since all the words in Y have same positive length, Y is a prefix code. From the fact that θ is one-to-one, for any integer $i \in \mathbb{Z}$ we obtain $\theta^i(\{yy'\}) \cap \theta^i(P(X)) = \theta^i(X) \cap P(\theta^i(yy')) = \emptyset$, consequently $X \cup Y$ remains a prefix code. By construction, Y is θ -invariant and it is not included in X , thus X is not a maximal prefix code.

In the case where θ is a morphism, the preceding arguments may be simplified. Actually, a word $y \in A^+ \setminus X$ exists such that $X \cup \{y\}$ remains a prefix code, thereforere by setting $Y = \{\theta^i(y) \mid i \in \mathbb{Z}\}$, $X \cup Y$ remains a prefix code.

2) (sketch) The preceding arguments may be applied for proving that in any case, if X is a non-complete bifix code, then it is maximal.

3,4) (sketch) In the case where X is a (two-way) f.d.d.-code, according to [3, Proposition 5.2.1], similar arguments leads to a similar conclusion.

5) In the case where X is a θ -invariant uniformly synchronized code with *verbal delay* k (cf [3, Section 10.2]), we must make use of different arguments. Actually, according to [4, Theorem 3.10], a complete uniformly synchronized code X' exists, with synchronizing delay k , and such that $X \subsetneq X'$. More precisely, X' is the minimal generating set of the following submonoid of A^* : $M = (X^{2k}A^* \cap A^*X^{2k}) \cup X^*$. According to Proposition 1, X' is θ -invariant. Since X is strictly included in X' , it cannot be maximal in the family of θ -invariant uniformly synchronized codes with delay k .

6) It remains to study the case where X is a non-complete θ -invariant circular code. Let $y \notin F(X^*)$ and let z and Z be computed as in Section 3: this guarantees that $X \cup Z$ is a θ -invariant set. For proving that $X \cup Z$ is a circular code, by contradiction we assume that words $y_1, \dots, y_n, y'_1, \dots, y'_m \in X \cup Z$, $p \in A^*$, $s \in A^+$, with $m + n$ minimal, exist such that the following equation holds:

$$y_1 y_2 \cdots y_n = s y'_2 y'_3 \cdots y'_m p \quad \text{and} \quad y'_1 = ps. \quad (6)$$

Once more since X is a code, and since $Z \cap F(X^*) = \emptyset$, without loss of generality we assume that at least one integer $i \in \mathbb{Z}$ exists such that $y_i \in Z$; similarly, at least one integer $j \in [1, m]$ exists such that $y'_j \in Z$. By construction, we have $y_i \in F(y'_j \cdots y'_m y'_1 \cdots y'_j \cdots y'_m y'_1 \cdots y'_j)$; consequently, since all the words in Z have the same length, a pair of integers $h, k \in [1, m]$ and a pair of words u, v exist such that $u y_i v \in y'_h X^* y'_k$. According to Lemma 3, necessarily we have either $u = \varepsilon$ or $v = \varepsilon$; this implies $y_i = y'_h$ or $y_i = y'_k$, which contradicts the minimality of $m + n$, therefore $X \cup Z$ is a circular code. ■

6 Embedding a regular invariant code into a complete one

In this section, we consider a non-complete regular θ -invariant code X and we are interested in the problem of computing a complete regular θ -invariant code Y such that $X \subseteq Y$. Historically, such a question appears for the first time in [15], where the author asked for the possibility of embedding a finite code into a regular complete one. With regards to θ -invariant codes, it seems natural to generalize the formula from [7] by making use of the code Z that was introduced in Section 4. More precisely we would consider the set $X' = (ZU)^*Z$, with $U = A^* \setminus (X^* \cup A^*ZA^*)$. Unfortunately, in such a construction, we observe that some pairs of words in Z may overlap, therefore a non-trivial equation could exist among the words of X' .

Nevertheless, in the special case where θ is an involutive antimorphism, convenient invariant overlapping-free words can be computed:

Proposition 2. *Let A be a finite alphabet and let θ be an antimorphism onto A^* whose restriction on A is different of the identity. If θ is involutive, then any non-complete regular θ -invariant code can be embedded into a complete one.*

Proof. Let X be such that $\theta(X) = X$. Assume that X is not complete. We will construct an overlapping-free word $t \notin F(X^*)$ such that $\theta(t) = t$. At first, we consider a word x such that $x \notin F(X^*)$ and $|x| \geq 2$. Without loss of generality, we assume that x is overlapping-free (otherwise, as in [3, Proposition 1.3.6], a word s exists such that xs is overlapping-free). If $\theta(x) = x$, then we set $t = x$, otherwise let $y = cx$, where c stands for the initial letter of x . Once more, without loss of generality we assume that y is overlapping-free. By construction we have $y \in ccA^+$, thus $|y| \geq 3$ and $y_1 = y_2 = c$. If $\theta(y) = y$, then we set $t = y$. Now assume $\theta(y) \neq y$; according to the condition of Proposition 2, we have $\theta|_A \neq id_A$, therefore a pair of letters a, b exists such that the following property holds:

$$a \neq b, \quad b \neq c, \quad \theta(a) = b, \quad \theta(b) = a. \quad (7)$$

Set $t = a^{|y|}b\theta(y)yab^{|y|}$. By construction, we have $\theta(t) = t$, moreover the following property holds:

Claim. t is an overlapping-free word.

Proof. Let $u, v \in A^*$ such that $ut = tv$, with $1 \leq |u| \leq |t| - 1$. According to the length of u , exactly one of the following cases occurs:

Case 1: $1 \leq |u| \leq |y|$. With this condition, we obtain $t_{|y|+1} = b = (ut)_{|y|+1} = a$: a contradiction with $a \neq b$.

Case 2: $|y| + 1 \leq |u| \leq 2|y|$. This condition implies $\theta(y_1) = t_{2|y|+1} = a$, therefore we obtain $c = y_1 = \theta(a) = b$: a contradiction with (7).

Case 3: $|u| = 2|y| + 1$. We have $y = a^{|y|}$: since we have $|y| \geq 3$, this contradicts the fact that y is overlapping-free.

Case 4: $|u| = 2|y| + 2$. We have $t_{2|y|+3} = y_2 = c = (ut)_{2|y|+3} = a$. It follows from $y_1 = y_2 = c$ that $y = a^{|y|}$: once more this contradicts the fact that y is overlapping-free.

Case 5: $2|y| + 3 \leq |u| \leq 3|y| + 2$. By construction, we have $t_{|u|+|y|} = b = (ut)_{|u|+|y|} = a$, a contradiction with (7).

Case 6: $3|y| + 3 \leq |u| \leq |t| - 1 = 4|y| + 1$. We obtain $t_{|u|+1} = b = (ut)_{|u|+1} = a$: once more this contradicts (7).

In any case we obtain a contradiction: this establishes the claim.

Since we have $t \notin F(X^*)$, and since t is overlapping-free, the classical method from [7] may be applied without any modification to ensure that X may be embedded into a complete code, say X' . Recall that it computes in fact a code X' as $X \cup V$, with $V = t(Ut)^*$ and $U = A^* \setminus (X^* \cup A^*tA^*)$. Moreover, since $\theta(t) = t$, it is straightforward to verify that $\theta(X') = X'$. ■

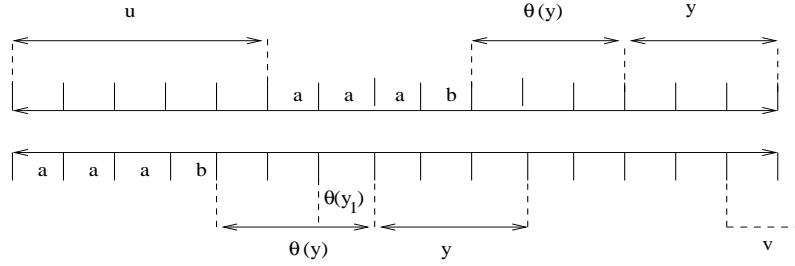


Fig. 2. Proof of Proposition 2: Case 2 with $|y| = 3$ and $|u| = 5$.

Note that if the restriction of θ to A is id_{A^*} , then for any non-empty word w , the word $\theta(w)$ is the returned word of w : necessarily $w, \theta(w)$ is an overlapping pair; actually, the preceding methodology appears inefficient in the most general case. We finish our paper by stating the following open problem:

Problem. Let A be a finite alphabet and let θ be an (anti)morphism onto A^* . Given a non-complete regular θ -invariant code $X \subset A^*$, can we compute a complete regular θ -invariant code Y such that $X \subseteq Y$?

References

1. Annal Deva Priya Darshini, C., Rajkumar Dare, V., Venkat I. and K.G. Subramanian: Factors of Words under an Involution, *J. of Math. and Inf.* **1** (2014) 52–59.
2. Büchi, M., de Luca, A., De Luca, A. and L. Q. Zamboni: On θ -episturmian words, *Eur. J. of Comb.* **30** (2009) 473–479.
3. Berstel, J., Perrin, D. and C. Reutenauer: Codes and Automata, Encyclopedia of Mathematics and Applications 129, Cambridge University Press (1985).
4. Bruyère, V.: On maximal codes with bounded synchronization delay, *Theoret. Comp. Sci.* **204** (1998) 11–28.
5. Czeizler, E., Czeizler, E., Kari, L. and S. Seki: An extension of the Lyndon- Schützenberger result to pseudoperiodic words, *Inf. Comput.* **209**(4) (2011) 717–730.
6. de Luca, A., De Luca, A., Pseudopalindrome closure operators in free monoids, *Theoret. Comp. Sci.* **362** (2006) 282–300.
7. Ehrenfeucht, A. Rozenberg, S.: Each regular code is included in a regular maximal one, *Theor. Inform. Appl.* **20** (1)(1985) 89–96.
8. Gawrychowski, P. , Manea, F., Mercas, R., Nowotka, D. and C. Tisceanu: Finding Pseudo-repetitions, in the proceedings of 30th Symposium on Theoretical Aspects of Computer Science (STACS13). N. Portier and T. Wilke ed., pp. 257–268 Leibniz International Proceedings in Informatics Schloss Dagstuhl Leibniz-Zentrum fr Informatik, Dagstuhl Publishing, Germany (2013).

9. Kari, L., Mahalingam, K.: DNA codes and their properties, 12th International Meeting on DNA Computing, DNA12, Seoul, Korea, June 5-9, 2006, Revised Selected Papers, C. Mao, T. Yokomori eds., *Lect. Notes in Comp. Sci.* **4287** (2006) 127–142.
10. Kari, L., Mahalingam, K.: Watson-Crick conjugate and commutative words, in acts of DNA 13, M.H. Garzon and H. Yan(eds.), *Lect. Notes in Comp. Sci.* **4848** (2008) 273–283.
11. Lothaire M.: Combinatorics on words, Encyclopedia of mathematics and its applications, Volume 17, Addison Westley Pub. Company (1983) (2nd edition Cambridge University Press 1997).
12. Manea, F., Mercas, R and D. Nowotka: Fine and Wilf's theorem and pseudo-repetitions, in acts of Mathematical Foundations of Computer Science 2012, B. Rován, V. Sassone, P. Widmayer eds., *Lect. Notes in Comp. Sci.* **7464** (2012) 668–680.
13. Manea, F., Muller, M. , Nowotka, D and S. Seki: Generalised Lyndon-Schützenberger Equations, in acts of Mathematical Foundations of Computer Science 2014 (MFCS 2014) *Lect. Notes in Comp. Sci.* **8634** (2014) 402–413.
14. Néraud J.: Completing circular codes in regular submonoids, *Theoret. Comp. Sci.* **391** (2008) 90–98.
15. Restivo, A.: On codes having no finite completion, *Discr. Math.* **17** (1977) 309-316.