Embedding a θ-invariant code into a complete one

Jean Néraud, Carla Selmi

Embedding a θ-invariant code into a complete one

Jean Néraud, Carla Selmi

Laboratoire d'Informatique, de Traitemement de l'Information et des Systèmes (LITIS), Université de Rouen Normandie, UFR Sciences et Techniques Avenue de l'université, 76830 Saint Etienne du Rouvray, France

Abstract

Let A be a finite or countable alphabet and let θ be a literal (anti)morphism onto A^* (by definition, such a correspondence is determinated by a permutation of the alphabet). This paper deals with sets which are invariant under θ (θ-invariant for short). We establish a formula which allows to embed any non-complete θ-invariant code into a complete one: this brings a positive answer to the open question that was stated in [11].

Keywords: word, equation, code, variable length code, prefix, suffix, regular, maximal, thin, complete, morphism, antimorphism, (anti)morphism, invariant, involutive, idempotent, literal, letter-to-letter

1. Introduction

In the free monoid theory, during the last decade, research involving one-to-one *morphic* or *antimorphic* correspondences have played a particularly important part: this is due to the powerful applications of these objects, in particular in the framework of DNA-computing. 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 is generated by A), the resulting mapping being commonly referred to as *literal* (or *letter-to-letter*).

In the special case of involutive morphisms or antimorphisms -we write (anti)morphisms for short, lots of successful investigations were done for extending the now classical combinatorial properties on words. The topics of the so-called pseudo-palindromes [5], that of θ-episturmian words [14], and the one of pseudo-repetitions [3, 16] were particularly involved. The framework of some peculiar families of codes [12] and that of equations in words [6, 4, 13, 8] were also concerned. In the family of *involutive* (anti)morphisms, generalizations of the famous theorem of Fine and Wilf [15, Proposition 1.3.5] were also established [7, 9].

Also starting with equations in words, in [11] we initiated a study of sets that are invariant under a given (anti)morphism θ (θ-invariant sets for short). In particular, we proved that those sets satisfy an extension of the famous Defect theorem (eg [15, Theorem 1.2.5]). We established also that, for any θ-invariant code in the large class of the so-called *thin* codes (which contains the *regular* ones), being *maximal* is equivalent to being *complete* [11, Theorem 11]. This last result consists in an extension of a famous result due to Schützenberger [10, Theorem 2.5.16]. As an aside, define a θ-code as a set X such that $\bigcup_{i\in\mathbb{Z}} \theta^i(X)$ is a code.

Email addresses: jean.neraud@univ-rouen.fr, neraud.jean@gmail.com (Jean Néraud), carla.selmi@univ-rouen.fr (Carla Selmi)

It is straightforward to prove that, in the family of those θ-codes, the maximal members are necessarily θ-invariant: this underscores the importance of the notion of θ-invariance.

In [11], we also interested in the problem of embedding a non-complete code into a complete one. For the first time, this question was stated in [2], where the author asked whether any finite code can be embedded into a regular one. A positive answer was provided in [1], where was established a formula for embedding any regular code into a complete one. From the point of view of θ-invariant codes, in [11, Proposition 12], we obtained a positive answer in the special case where θ is an involutive antimorphism different from the so-called mirror image, and we let open the general problem:

Given a regular non-complete θ-invariant code X, is there a regular complete one that contains X?

In the present paper, by establishing the following result, we bring a positive anwser:

Theorem *Given a finite alphabet A, any non-complete regular θ-invariant code $X \subseteq A^*$ may be embedded into a complete regular one.*

As a consequence, we obtain a very shorter alternative proof of [11, Theorem 11].

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 notions concerning codes are recalled. The main result is established in Section 3.

2. Preliminaries

2.1. Words and free monoid

We adopt the notation of the free monoid theory. In the whole paper, we consider a finite or countable alphabet A, and we denote by A^* the free monoid that it generates. Given a subset X of A^*, we denote by X^* the submonoid of A^* that is generated by X, moreover we set $X^+ = X^* \setminus \{\varepsilon\}$. Given a word w, we denote by $|w|$ its length, the empty word, which 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 $w = w_1 \cdots w_{|w|}$.

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 two words u, v exist such that $w = uxv$. Given a non-empty set $X \subseteq A^*$, we denote by $P(X)$ ($F(X)$) the set of the words that are prefix (factor) of some word in X. Clearly, we have $X \subseteq P(X) \subseteq F(X)$. 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 $1 \leq |u| \leq |w| - 1$ and $1 \leq |v| \leq |w'| - 1$; otherwise, the pair is *overlapping-free* (in such a case, if $w = w'$, we simply say that w is overlapping-free).

2.2. Variable length codes

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 [10]. A set X is a *variable length code* (a *code* for short) if any equation among the words of X is trivial, that is, for any pair of sequences of words in X, namely $(x_i)_{1 \leq i \leq n}$, $(y_j)_{1 \leq j \leq p}$, the equation $x_1 \cdots x_n = y_1 \cdots y_p$ implies $n = p$ and $x_i = y_i$, for each integer $i \in [1, n]$. By definition X^*, is a *free* submonoid. In the present paper the so-called *prefix* codes play a peculiar part in the proof of the main result: a code $X \subseteq A^*$ is prefix if $X \cap XA^+ = \emptyset$.

A code $X \subseteq A^*$ is *maximal* if it is not strictly included in another code of A^*. Given a code X, it is *complete* if $A^* = F(X^*)$; X is *thin* if $A^* \neq F(X)$. Regular codes are well known examples of thin codes [10, Proposition 2.5.20]. From this point of view, the following result was established by Schützenberger:

Theorem 2.1. [10, Theorem 2.5.16] *Let $X \subseteq A^*$ be a thin code. Then the two following conditions are equivalent:*

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

In the framework of θ-invariant codes, Theorem 2.1 admits the following extension:

Theorem 2.2. [11, Theorem 9] *Let $X \subseteq A^*$ 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 of θ-invariant codes.*

2.3. Literal (anti)morphisms

In the whole paper, we consider 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 antimophism 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 particular, if $|A| = 2$, then θ is *involutive.* 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 $\theta(X) = X$.

Example 1. Let $A = \{a, b, c, d\}$. Consider the (unique) antimorphism θ that is defined by $\theta(a) = a, \theta(b) = b, \theta(c) = d, \theta(d) = c$. The mapping θ is involutive, moreover the sets $\{aba\}$ and $\{abcd, cdba\}$ are θ-invariant.

Remark 1. In the spirit of the families of codes that were introduced in [12], given an (anti)morphism θ, define a θ-*code* as a set X such that $\bigcup_{i\in\mathbb{Z}} \theta^i(X)$ is a code. Clearly, with this definition any θ-code is a code; moreover any θ-code that is a maximal code, is necessarily θ-invariant. Indeed, if X is not θ-invariant, then we have $X \subsetneq X \cup \theta(X)$, thus X is strictly included in the code $\bigcup_{i\in\mathbb{Z}} \theta^i(X)$.

A similar argument proves that if X is maximal as a θ-code, then it is θ-invariant (indeed, $\bigcup_{i\in\mathbb{Z}} \theta^i(X)$ itself is a θ-code).

Taking account of the fundamental importance of the concept of maximality in the theory of codes, such properties reinforces the relevance of the notion of θ-invariant code.

3. Embedding a regular invariant code into a complete one

In this section, we address to the problem of embedding an invariant code into a complete one. Historically, such a question appears for the first time in [2], where the author asked for the possibility of embedding a finite code into a complete regular code. A positive answer was given in [1], where the authors provided a regularity preserving method.

Presently, we consider a non-complete regular θ-invariant code X and we ask for a complete regular θ-invariant code Y such that $X \subseteq Y$. In a special case of involutive antimorphisms, in [11, Proposition 12] we gave a method of embedding; however, in the general case of literal (anti)morphisms, the problem remained open.

In the present paper, we will bring a positive answer. Let's begin by describing our construction.

3.1. Some notation

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

$$z = \overline{a}^{|y|} y a^{|y|} \quad (y \in aA^*\overline{a}). \tag{1}$$

Since θ is a literal (anti)morphism, for each integer $i \in \mathbb{Z}$, two different letters $b, \overline{b}$ exist such that the following property holds:

$$\theta^i(z) = \overline{b}^{|y|} \theta^i(y) b^{|y|}. \tag{2}$$

Finally, we introduce the three following sets:

$$Z = \bigcup_{i \in \mathbb{Z}} \{\theta^i(z)\}, \tag{3}$$

$$W = ZA^* \cap A^*Z, \tag{4}$$

$$T = W \setminus (W \cup X)(W \cup X)^+. \tag{5}$$

By construction, the following inclusion holds:

$$W \subseteq (X \cup T)^+. \tag{6}$$

3.2. Basic properties of Z

By construction, each element of the preceding set Z has length $3|y|$. The following statement summarizes the results of [11, Lemma 6, Lemma 7, Corollary 8]:

Lemma 3.1. *With the preceding notation each of the three following conditions holds:*

(i) *Let $u, v \in A^+$ such that $uZ \cap Zv \neq \emptyset$, with $|u| \leq |z| - 1$. Then we have $|u| = |v| \geq 2|y|$.*
(ii) *We have $A^+ZA^+ \cap ZX^*Z = \emptyset$.*
(iii) *The set X^*Z is a prefix code.*

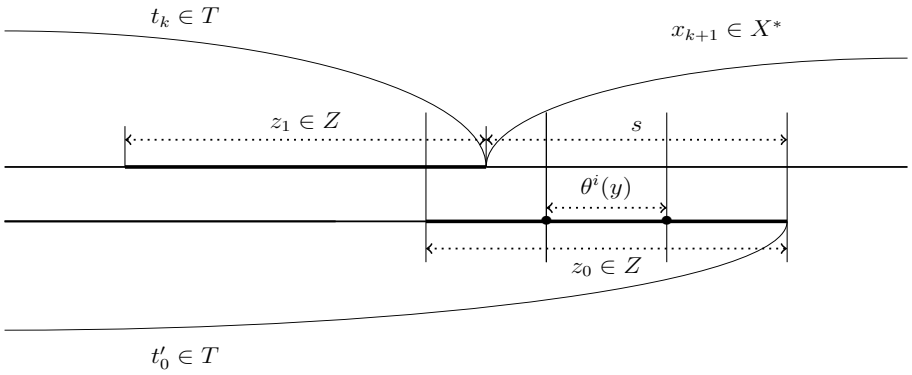

Figure 1: Proof of Lemma 3.2 - the case where $s \in P(x_{k+1})$.

3.3. The consequences for the set $X \cup T$

Lemma 3.2. *The set $X \cup T$ is a θ-invariant code.*

Proof of Lemma 3.2. The fact that $X \cup T$ is θ-invariant comes from its construction. For proving that it is a code, we consider an arbitrary equation among the words in $X \cup T$. Since X is a code, and since $z \notin F(X^*)$, we may assume that at least one occurrence of a word in T appears in each side of the equation, therefore this equation takes the following form:

$$x_0 t_0 x_1 t_1 \cdots t_{n-1} x_n = x'_0 t'_0 \cdots t'_{p-1} x_p, \tag{7}$$

with $x_i, x'_j \in X^*$ $(0 \leq i \leq n,\ 0 \leq j \leq p)$ and $t_i, t'_j \in T$ $(0 \leq i \leq n-1,\ 0 \leq j \leq p-1)$. Since by construction we have $T \subseteq W \subseteq ZA^*$, each side of the equation contains a word in X^*Z as a prefix: according to Property (iii) of Lemma 3.1, this implies $x_0 = x'_0$, therefore our equation is equivalent to:

$$t_0 x_1 t_1 \cdots t_{n-1} x_n = t'_0 \cdots t'_{p-1} x_p. \tag{8}$$

Without loss of generality, we assume that $|t_0| \leq |t'_0|$; let k be the greatest non-negative integer such that a word s exists with $t_0 x_1 t_1 \cdots x_k t_k s = t'_0$, with $s \in A^*$. By contradiction, we assume $s \neq \varepsilon$. Let $z_0 \in Z$ $(z_1 \in Z)$ be the unique word such that $t'_0 \in A^* z_0$ $(t_k \in A^* z_1)$. According to the preceding property (2), an integer $i \in \mathbb{Z}$ and two letters b, $\overline{b}$ exist such that $z_0 = \overline{b}^{|y|} \theta^i(y) b^{|y|}$. Moreover, since we have $y \notin F(X^*)$, and since X is θ-invariant, we have $\theta^i(y) \notin F(X^*)$. By construction, the set with elements $t_0 x_1 \cdots t_k x_{k+1}$ and t'_0 cannot be a prefix code; more precisely, exactly one of the two following main conditions holds:

1. At first, we assume that $t'_0 \in P(t_0 x_1 \cdots t_k x_{k+1})$, that is $s \in P(x_{k+1})$ (cf Figure 1). By construction, at least one of the two words s, z_0 is a suffix of the other one. Actually, since we have $z_0 \notin F(X^*)$, necessarily s is a proper suffix of z_0, therefore $(z_1,\ z_0)$ is an overlapping pair of words. According to Property (i) of Lemma 3.1, necessarily we have $|s| \geq 2|y|$, which implies $\theta^i(y) \in F(s)$: a contradiction with $\theta^i(y) \notin F(X^*)$.
2. Now, we assume that $t_0 x_1 \cdots t_k x_{k+1}$ is a proper prefix of t'_0, thus we have $s = x_{k+1} s_1$, with $s_1 \neq \varepsilon$. Let $z_2 \in Z$ be the unique word such that $t_{k+1} \in z_2 A^*$. By construction the set with elements $t_0 x_1 \cdots t_k x_{k+1} z_2$ and t'_0 cannot be a prefix code. More precisely, exactly one of the two following cases occurs:

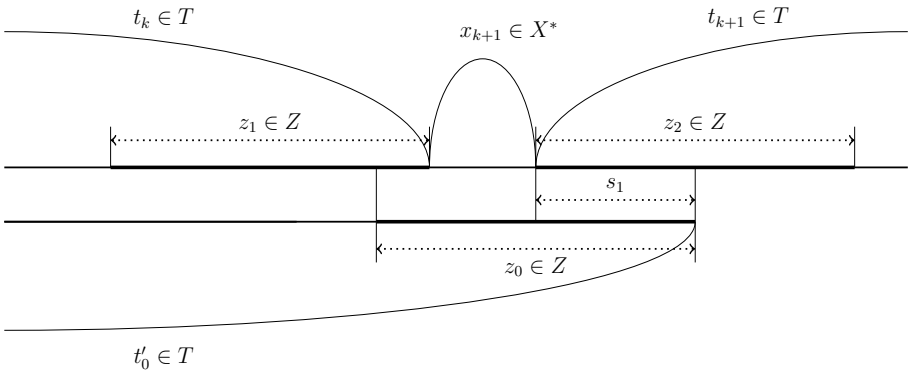

Figure 2: Proof of Lemma 3.2 -the case where $x_{k+1} \in P(s)$ and $s_1 \in P(z_2) \setminus \{z_2\}$.

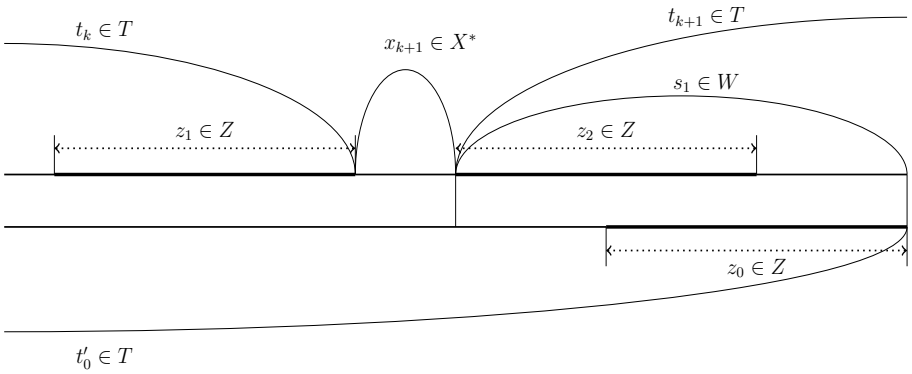

Figure 3: The case where $x_{k+1} \in P(S)$, with $z_2 \in P(s_1)$ in the proof of Lemma 3.2.

2.1 The first case corresponds to t'_0 being a proper prefix of $t_0x_1 \cdots t_kx_{k+1}z_2$, that is s_1 being a proper prefix of z_2 (cf Figure 2). With this condition, the word z_0 is necessarily a factor of $z_1x_{k+1}z_2$. According to Property (ii) of Lemma 3.1, since we have $s_1 \neq \varepsilon$, this implies $z_1 = z_0$: this contradicts $s \neq \varepsilon$.

2.2 It remains to consider the case where z_2 is a prefix of s_1 (cf Figure 3). Actually, since $t_0x_1 \cdots t_kx_{k+1}z_2$ is a prefix of $t_0x_1 \cdots t_kx_{k+1}s_1 = t'_0$, with $|z_2| = |z_0|$, necessarily z_0 is a suffix of s_1: we obtain $s_1 \in z_2A^* \cap A^*z_0$ thus, according to (4), $s_1 \in W$. We obtain $t'_0 = t_0x_1 \cdots t_kx_ks_1 \in (TX^*)^+W$, thus $t'_0 \in (T \cup X)^+W$: this contradicts (5).

In each case we obtain a contradiction: as a consequence we have $s = \varepsilon$, thus $t'_0 = t_0x_1 \cdots x_kt_k$. Once more according to (5), it follows from $t'_0 \in T$ that we have $k = 0$, thus $t'_0 = t_0$. As a consequence, Equation (8) is equivalent to the following one:

$$x_1t_1 \cdots t_{n-1}x_n = x'_1 \cdots t'_{p-1}x_p. \tag{9}$$

By iterating these arguments, we shall obtain: $n = p$ and $x_i = x'_i$, $t'_j = t_j$ $(0 \leq i \leq n$, $0 \leq j \leq n-1)$, therefore $X \cup T$ is a code: this completes the proof of Lemma 3.2.

■

Lemma 3.3. *The code $X \cup T$ is complete.*

Proof of Lemma 3.3. Let $w \in A^*$. According to the construction of W we have $ZwZ \subseteq W$. According to (6) this implies $ZwZ \subseteq (X \cup T)^*$, therefore we have $w \in F(X^*)$.

■

Notice that, in the case where A is a finite alphabet, by starting with $y \notin F(X^*)$, the construction in (3) leads to a finite set Z: this guarantees the regularity of the sets W and T. As a direct consequence, we obtain the following result:

Theorem 3.4. *Given a countable (finite) alphabet A, any non-complete (regular) θ-invariant code may embedded into a complete (regular) θ-invariant code.*

Remark 2. According to the result of Theorem 3.4, we can draw an alternative proof of Theorem 2.2. Indeed, given a non-complete θ-invariant code X, the existence of a complete θ-invariant code strictly containing X is guaranteed, thus X is not maximal in the family of θ-invariant codes. In other words, in the statement of Theorem 2.2, Property (iii) implies Property (i).

[1] Ehrenfeucht A. and Rozenberg S. Each regular code is included in a regular maximal one. *Theor. Inform. Appl.*, 20:89–96, 1985.

[2] Restivo A. On codes having no finite completion. *Discr. Math.*, 17:309–316, 1977.

[3] 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.*, 1:52–59, 2013 -2014.

[4] Day J. D., Fleishmann P., Manea F., and D. Nowotka. Equations enforcing repetitions under permutations. In S. Brlek, F. Dolce, C. Reutenauer, and E. Vandomme, editors, *Combinatorics on Words*, volume 10432, pages 72–84. Lect. Notes in Comp. Sci., sept 2017.

[5] de Luca A. and De Luca A. Pseudopalindrome closure operators in free monoids. *Theoret. Comp. Sci.*, 362:282–300, 2006.

[6] Czeizler E., Czeizler E., Kari L., and S. Seki. An extension of the Lyndon-Schützenberger result to pseudoperiodic words. *Inf. Comput.*, 209:717–730, 2011.

[7] Czeizler E., Kari L., and S. Seki. On a special class of primitive words. *Theoret. Comp. Sci.*, 411:617–630, 2010.

[8] Manea F., Müller M., Nowotka D., and S. Seki. Generalised Lyndon-Schützenberger equations. In E. Csuhaj-Varú, M. Dietzfelbinger, and Z. Ésik, editors, *Acts of Mathematical Foundations of Computer Science 2014 (MFCS 2014)*, volume 8634, pages 402–413. Lect. Notes in Comp. Sci., 2014.

[9] Manea F., Mercas R., and Nowotka D. Fine and Wilf' theorem and pseudo-repetitions. In B. Rovan, V. Sassone, and P. Widmayer, editors, *Acts of Mathematical Foundations of Computer Science 2012 (MFCS 2012)*, volume 7464, pages 668–680. Lect. Notes in Comp. Sci., 2012.

[10] Berstel J., Perrin D., and C. Reutenauer. *Codes and Automata.* Cambridge University Press, 2010.

[11] Néraud J. and Selmi C. Invariance: a theoretical approach for coding sets of words modulo literal (anti)morphisms. In S. Brlek, F. Dolce, C. Reutenauer, and E. Vandomme, editors, *Combinatorics on Words*, volume 10432, pages 214–227. Lect. Notes in Comp. Sci., sept 2017.

[12] Kari L. and Mahalingam K. Dna codes and their properties. In C. Mao and T. Yokomori, editors, *12th International Meeting on DNA Computing (DNA12), Revised Selected Papers*, volume 4287, pages 127–142. Lect. Notes in Comp. Sci., june 2006.

[13] Kari L and Mahalingam K. Watson-Crick conjugate and commutative words. In M.H. Garzon and H. Yan, editors, *13th International Meeting on DNA Computing (DNA13)*, volume 4848, pages 273–283. Lect. Notes in Comp. Sci., 2008.

[14] Bucci M., de Luca A., De Luca A., and L. Q. Zamboni. On θ-episturmian words. *Eur. J. of Comb.*, 30:473–479, 2009.

[15] Lothaire M. *Combinatorics on Words.* Cambridge University Press, (2nd edition Cambridge University Press 1997), 1983.

[16] Gawrychowski P., Manea F., Mercas R., Nowotka D., and C. Tiseanu. Finding pseudo-repetitions. In *30th International Symposium on Theoretical Aspects of Computer Science (STACS 2013)*, pages 257–26.