Words with the Smallest Number of Closed Factors (extended abstract)

Gabriele Fici[1] and Zsuzsanna Lipták[2]

[1] I3S, CNRS & Université Nice Sophia Antipolis, France, gabriele.fici@unice.fr
[2] Università di Verona, Italy, zsuzsanna.liptak@univr.it

Abstract. A word is closed if it contains a factor that occurs both as a prefix and as a suffix but does not have internal occurrences. We show that any word of length n contains at least $n+1$ closed factors (i.e., factors that are closed words). We investigate the language $\mathcal{L}$ of words over the alphabet $\{a,b\}$ containing exactly $n+1$ closed factors. We show that a word belongs to $\mathcal{L}$ if and only if its closed factors and its palindromic factors coincide (and therefore the words in $\mathcal{L}$ are rich words). We also show that $\mathcal{L}$ coincides with the language of conjugates of words in a^*b^*.

Keywords: Closed word, closed factor, rich word, bitonic word.

1 Introduction

A *word* is a finite sequence of elements from a finite set Σ. We refer to the elements of Σ as *letters* and to Σ as the *alphabet*. The i-th letter of a word w is denoted by w_i. Given a word $w = w_1 w_2 \cdots w_n$, with $w_i \in \Sigma$ for $1 \leq i \leq n$, the nonnegative integer n is the *length* of w, denoted by $|w|$. The empty word has length zero and is denoted by ε. The set of all words over Σ is denoted by Σ^*. Any subset of Σ^* is called a *language*.

A *prefix* (resp. a *suffix*) of a word w is any word u such that $w = uz$ (resp. $w = zu$) for some word z. A *factor* of w is a prefix of a suffix (or, equivalently, a suffix of a prefix) of w. The set of prefixes, suffixes and factors of the word w are denoted by $\mathit{Pref}(w)$, $\mathit{Suff}(w)$ and $\mathit{Fact}(w)$ respectively. A *border* of a word w is any word in $\mathit{Pref}(w) \cap \mathit{Suff}(w)$ different from w. From the definitions, we have that ε occurs as a prefix, suffix and factor in any word. An *occurrence* of a factor u in a word w is a pair of positions (i,j) such that $w_i \ldots w_j = u$. An occurrence is *internal* if $i > 1$ and $j < |w|$.

The word $\tilde{w} = w_n w_{n-1} \cdots w_1$ is called the *reversal* (or *mirror image*) of w. A *palindrome* is a word w such that $\tilde{w} = w$. In particular, the empty word is a palindrome. A *conjugate* of a word w is any word of the form vu such that $uv = w$, for some $u, v \in \Sigma^*$. A conjugate of a word w is also called a *rotation* of w.

Let w be a word. We denote by $PAL(w)$ the set of factors of w that are palindromes. Droubay, Justin and Pirillo showed [4] that for any word w of length n, one has $|PAL(w)| \leq n+1$. Consequently, w is called *rich* [6] (or *full* [1]) if $|PAL(w)| = n+1$, that is, if it contains the largest number of palindromes a word of length n can contain.

A language L is called *factorial* if $L = \mathit{Fact}(L)$, i.e., if L contains all the factors of its words. A language L is *extendible* if for every word $w \in L$, there exist letters $a, b \in \Sigma$ such that $awb \in L$. The language of rich words over a fixed alphabet Σ is an example of factorial and extendible language.

We now recall the definition of closed word [5]:

Definition 1. *A word w is* closed *if it is empty or has a factor occurring exactly twice in w, as a prefix and as a suffix of w.*

The word aba is closed, since its factor a appears only as a prefix and as a suffix. The word $abaa$, instead, is not closed. Note that for any letter $a \in \Sigma$ and for any $n > 0$, the word a^n is closed, a^{n-1} being a factor occurring only as a prefix and as a suffix in it. More generally, any word w that is a power of a shorter word, i.e., $w = v^n$ for a non-empty v and $n > 1$, is closed.

There exist closed words that are not palindromes, for example the word $abab$. Conversely, there exist palindromes that are not closed, but it is worth noticing that a shortest palindrome over a two-letter alphabet that is not closed has length 14. An example is $aabbabaababbaa$.

Remark 1. The notion of closed word is closely related to the concept of *complete return* to a factor, as considered in [6]. A complete return to the factor u in a word w is any factor of w having exactly two occurrences of u, one as a prefix and one as a suffix. Hence w is closed if and only if it is a complete return to one of its factors; such a factor is clearly both the longest repeated prefix and the longest repeated suffix of w (that is, the longest border of w). The notion of closed word is also equivalent to that of *periodic-like* word [3]. A word w is periodic-like if its longest repeated prefix does not have two occurrences in w followed by different letters.

Observation 1 *Let w be a non-empty word over Σ. The following characterizations of closed words follow easily from the definition:*

1. *w has a factor occurring exactly twice in w, as a prefix and as a suffix of w;*
2. *the longest repeated prefix of w does not have internal occurrences in w, that is, occurs in w only as a prefix and as a suffix;*
3. *the longest repeated suffix of w does not have internal occurrences in w, that is, occurs in w only as a suffix and as a prefix;*
4. *the longest repeated prefix of w does not have two occurrences in w followed by different letters;*
5. *the longest repeated suffix of w does not have two occurrences in w preceded by different letters;*
6. *w has a border that does not have internal occurrences in w;*
7. *the longest border of w does not have internal occurrences in w;*
8. *w is the complete return to its longest prefix;*
9. *w is the complete return to its longest border;*
10. *$w = uv = zu$, with v, z non-empty, and $Fact(w) \cap \Sigma u \Sigma = \emptyset$.*

For more details on closed words and related results cf. [3, 2, 5].

2 Closed factors

Let w be a word. A factor of w that is a closed word is called a *closed factor* of w. The set of closed factors of the word w is denoted by $C(w)$.

Lemma 1. *For any non-empty word w of length n, one has $|C(w)| \geq n + 1$.*

Lemma 2. *Let u, v be non-empty words. Then $|C(u)| + |C(v)| \leq |C(uv)| + 1$.*

Proposition 1. *Let w be a non-empty word of length n. If $C(w) \subseteq PAL(w)$, then $C(w) = PAL(w)$ and $|C(w)| = |PAL(w)| = n + 1$. In particular, w is a rich word.*

Bucci et al. showed [2, Proposition 4.3] that a word w is rich if and only if every closed factor v of w has the property that the longest palindromic prefix (or suffix) of v is unrepeated in v. Moreover, if w is a palindrome, then it is rich if and only if $PAL(w) \subseteq C(w)$ [2, Corollary 5.2].

In Section 4, we shall prove that the condition $PAL(w) = C(w)$ characterizes the words having the smallest number of closed factors over a binary alphabet.

3 Words with the smallest number of closed factors

By Lemma 1, we have that $n+1$ is a lower bound on the number of closed factors of a word of length $n>0$. We introduce the following definition:

Definition 2. *A word $w \in \Sigma^*$ is* C-poor *if $|C(w)| = |w|+1$. We also set*

$$\mathcal{L}_{\Sigma} = \{w \in \Sigma^* : |C(w)| = |w|+1\}$$

the language of C-poor words over the alphabet Σ.

Remark 2. If $|\Sigma| = 1$, then $\mathcal{L}_{\Sigma} = \Sigma^*$. So in what follows we will suppose $|\Sigma| \geq 2$.

Lemma 3. *The language $\mathcal{L}_{\Sigma}$ is closed under reversal.*

Lemma 4. *Let w be a C-poor word over the alphabet Σ and $x \in \Sigma$. The word wx (resp. xw) is C-poor if and only if it has a unique suffix (resp. prefix) that is closed and is not a factor of w.*

Proposition 2. *A word $w \in \Sigma^*$ belongs to $\mathcal{L}_{\Sigma}$ if and only if every factor of w belongs to $\mathcal{L}_{\Sigma}$. That is, $\mathcal{L}_{\Sigma}$ is a factorial language.*

4 Binary words

In this section, we fix the alphabet $\Sigma = \{a,b\}$. For simplicity of exposition, we will denote the language of C-poor words over $\{a,b\}$ by $\mathcal{L}$ rather than by $\mathcal{L}_{\{a,b\}}$. We first recall the definition of bitonic word.

Definition 3. *A word $w \in \{a,b\}^*$ is* bitonic *if it is a conjugate of a word in a^*b^*, i.e., if it is of the form $a^ib^ja^k$ or $b^ia^jb^k$ for integers $i,j,k \geq 0$.*

The following lemma, the proof of which is straightforward, relates bitonic words to closed factors.

Lemma 5. *If a word $w \in \{a,b\}^*$ does not contain any complete return to ab or ba as a factor, then it is bitonic.*

Lemma 6. *Let w be a bitonic word. Then $C(w) \subseteq PAL(w)$.*

Thus, by Proposition 1, any bitonic word w of length $n>0$ contains exactly $n+1$ closed factors and so is a C-poor word. In the rest of the section we shall prove the converse, that is, we shall prove that if w is a C-poor word over $\{a,b\}$, then w is bitonic.

Consider the word $w = abab$. The word w does not belong to $\mathcal{L}$, since it has 6 closed factors, namely ε, a, b, aba, bab and $abab$. In fact, it has two suffixes (bab and $abab$) that are closed and do not appear before in w, and hence by Proposition 2 it cannot belong to $\mathcal{L}$. More generally, any word u such that u is the complete return to ab or ba does not belong to $\mathcal{L}$ for the same reason. So, using Proposition 2, we get:

Lemma 7. *If $w \in \mathcal{L}$, then w does not contain any complete return to ab or ba as a factor.*

We summarize the characterizations of $\mathcal{L}$ in the following theorem:

Theorem 1. *Let $w \in \{a,b\}^*$. The following are equivalent:*

1. $w \in \mathcal{L}$;

2. $C(w) = PAL(w)$*;*
3. $C(w) \subseteq PAL(w)$*;*
4. w *is a bitonic word;*
5. w *does not contain any complete return to* ab *or* ba*.*

Proof. 1) $\Rightarrow$ 5) by Lemma 7; 5) $\Rightarrow$ 4) by Lemma 5; 4) $\Rightarrow$ 3) by Lemma 6; finally, 3) $\Rightarrow$ 2) and 2) $\Rightarrow$ 1) by Proposition 1. $\square$

So, by Theorem 1 and Proposition 1, every word in $\mathcal{L}$ is rich. Notice that there exist rich words that are not in $\mathcal{L}$, for example the word $w = abab$, which has 6 closed factors, namely ε, a, b, aba, bab and $abab$. Another consequence of Theorem 1 is that $\mathcal{L}$ is extendible, since the language of bitonic words is clearly extendible. Thus, the language $\mathcal{L}$ is a factorial and extendible subset of the language of (binary) rich words.

It further follows from Theorem 1 that $\mathcal{L}$ is a regular language, since the language of the conjugates of words of a regular language is regular [7]. In the following proposition we exhibit a closed enumerative formula for the language $\mathcal{L}$.

Proposition 3. *For every* $n > 0$*, there are exactly* $n^2 - n + 2$ *distinct words in* $\mathcal{L}$*.*

Proof. Each of the $n-1$ words of length $n > 0$ in a^+b^+ has n distinct rotations, while for the words a^n and b^n all the rotations coincide. Thus, there are $n(n-1)+2$ bitonic words of length n, and the statement follows from Theorem 1. $\square$

5 Conclusion and open problems

In this paper we studied words with the smallest number of closed factors, which we referred to as C-poor words. We gave some interesting characterizations in the case of a binary alphabet. In particular, we showed that the language of binary C-poor words coincides with the language of bitonic words. A natural direction of further investigation is finding a characterization for C-poor words over alphabets larger than 2.

An enumerative formula for rich words is not known, not even in the binary case. A possible approach to this problem is to separate rich words in subclasses to be enumerated separately. Our enumerative formula for C-poor words given in Proposition 3 constitutes a step in this direction.

References

1. S. Brlek, S. Hamel, M. Nivat, and C. Reutenauer. On the palindromic complexity of infinite words. *Internat. J. Found. Comput. Sci.*, 15:293–306, 2004.
2. M. Bucci, A. de Luca, and A. De Luca. Rich and Periodic-Like Words. In *DLT 2009, 13th International Conference on Developments in Language Theory*, volume 5583 of *Lecture Notes in Comput. Sci.*, pages 145–155. Springer, 2009.
3. A. Carpi and A. de Luca. Periodic-like words, periodicity and boxes. *Acta Inform.*, 37:597–618, 2001.
4. X. Droubay, J. Justin, and G. Pirillo. Episturmian words and some constructions of de Luca and Rauzy. *Theoret. Comput. Sci.*, 255(1-2):539–553, 2001.
5. G. Fici. A Classification of Trapezoidal Words. In *WORDS 2011, 8th International Conference on Words*, number 63 in Electronic Proceedings in Theoretical Computer Science, pages 129–137, 2011.
6. A. Glen, J. Justin, S. Widmer, and L. Q. Zamboni. Palindromic richness. *European J. Combin.*, 30:510–531, 2009.
7. J.E. Hopcroft, R. Motwani, and J.D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, 2001.