Journal of Computer and System Sciences 80 (2014) 711–719

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

On the structure of bispecial Sturmian words ☆

Gabriele Fici

Dipartimento di Matematica e Informatica, Università di Palermo, Via Archirafi 34, 90123 Palermo, Italy

ARTICLE INFO

Article history:
Received 1 January 2013
Received in revised form 21 September 2013
Accepted 1 November 2013
Available online 19 November 2013

Keywords:
Sturmian words
Christoffel words
Special factors
Minimal forbidden words
Enumerative formula

ABSTRACT

A balanced word is one in which any two factors of the same length contain the same number of each letter of the alphabet up to one. Finite binary balanced words are called Sturmian words. A Sturmian word is bispecial if it can be extended to the left and to the right with both letters remaining a Sturmian word. There is a deep relation between bispecial Sturmian words and Christoffel words, that are the digital approximations of Euclidean segments in the plane. In 1997, J. Berstel and A. de Luca proved that *palindromic* bispecial Sturmian words are precisely the maximal internal factors of *primitive* Christoffel words. We extend this result by showing that bispecial Sturmian words are precisely the maximal internal factors of *all* Christoffel words. Our characterization allows us to give an enumerative formula for bispecial Sturmian words. We also investigate the minimal forbidden words for the language of Sturmian words.

1. Introduction

A word w is balanced if and only if for any u, v factors of w of the same length, and for any letter a, one has $||u|_a - |v|_a| \leqslant 1$, where $|z|_a$ denotes the number of a's in the word z.

Balanced words appear in several problems in Computer Science. For example, Altman, Gaujal and Hordijk [1] proved that balanced words are optimal sequences for some classes of scheduling problems, such as routing among several systems. An interesting problem arising in this context is that of constructing infinite balanced words with assigned frequencies of letters. There is a conjecture of A.S. Fraenkel [18], originally stated in the context of Number Theory, that is equivalent to the following: for any fixed $k > 2$, there is only one infinite balanced word (up to letter permutation) over an alphabet of size k, in which all letters have different frequencies, and this word is periodic. The Fraenkel Conjecture has been proved true for small alphabet sizes (see [29,2] and references therein), but the general problem remains open.

For any alphabet Σ of size at least two, there exist infinite words over Σ that are balanced and aperiodic. When $|\Sigma| = 2$, these are called infinite Sturmian words. Sturmian words are very rich from the combinatorial point of view, and because of this fact they have a lot of equivalent definitions and characterizations (see, as a classical reference, [22, Chapter 2]). However, if the Fraenkel Conjecture is true for every $k > 2$, the only balanced infinite words that are aperiodic and have different letter frequencies are the infinite Sturmian words.

A finite Sturmian word (or, briefly, a Sturmian word) is a factor of some infinite Sturmian word. The set St of Sturmian words therefore coincides with the set of binary balanced finite words.

If one considers extendibility within the set St of Sturmian words, one can define left special Sturmian words (resp. right special Sturmian words) [15] as those words w over the alphabet $\Sigma = \{a, b\}$ such that aw and bw (resp. wa and wb) are both Sturmian words. For example, the word aab is left special since $aaab$ and $baab$ are both Sturmian words, but is not right special since $aabb$ is not a Sturmian word.

☆ A preliminary version of this paper was presented at the *37th International Symposium on Mathematical Foundations of Computer Science, MFCS 2012* [17].
E-mail address: gabriele.fici@unipa.it.

0022-0000/$ – see front matter
http://dx.doi.org/10.1016/j.jcss.2013.11.001

Left special Sturmian words are precisely the binary words having suffix automaton[1] with minimal state complexity (cf. [28,16]). From combinatorial considerations one has that right special Sturmian words are the reversals of left special Sturmian words.

The Sturmian words that are both left and right special are called bispecial Sturmian words. They are of two kinds: strictly bispecial Sturmian words, that are the words w such that awa, awb, bwa and bwb are all Sturmian words (e.g. aa), or non-strictly bispecial Sturmian words otherwise (e.g. ab). Strictly bispecial Sturmian words are also called central words, and have been deeply studied (see for example [15,10]) because they constitute the kernel of the theory of Sturmian words. Non-strictly bispecial Sturmian words, instead, received less attention.

One important field in which Sturmian words arise naturally is Discrete Geometry. Indeed, infinite Sturmian words can be viewed as the digital approximations of Euclidean straight lines in the plane. It is known that given a point (p,q) in the grid $\mathbb{Z}\times\mathbb{Z}$, with $p,q>0$, there exists a unique path that approximates from below (resp. from above) the Euclidean segment joining the origin $(0,0)$ to the point (p,q). If one encodes horizontal and vertical unitary segments with the letters a and b respectively, this path is called the lower (resp. upper) Christoffel word[2] associated to the pair (p,q), and is denoted by $w_{p,q}$ (resp. $w'_{p,q}$). By elementary geometrical considerations, one has that for any $p,q>0$, $w_{p,q}=aub$ for some word u, and $w'_{p,q}=b\tilde{u}a$, where $\tilde{u}$ is the reversal of u. If (and only if) p and q are coprime, the Christoffel words $w_{p,q}$ and $w'_{p,q}$ are primitive (that is, they are not a concatenation of copies of a shorter word).

A well known result of Jean Berstel and Aldo de Luca [5] is that a word u is a strictly bispecial Sturmian word if and only if aub is a primitive lower Christoffel word (or, equivalently, if and only if bua is a primitive upper Christoffel word). As a main result of this paper, we show that this correspondence holds in general between bispecial Sturmian words and Christoffel words. More precisely, we prove (in Theorem 3.11) that u is a bispecial Sturmian word if and only if there exist letters x,y in $\{a,b\}$ such that xuy is a Christoffel word.

This characterization allows us to prove an enumerative formula for bispecial Sturmian words (Corollary 4.2): there are exactly $2n+2-\phi(n+2)$ bispecial Sturmian words of length n, where ϕ is the Euler totient function, i.e., $\phi(n)$ is the number of positive integers smaller than or equal to n and coprime with n. Surprisingly, enumerative formulae for left special, right special and strictly bispecial Sturmian words were known [15], but to the best of our knowledge we exhibit the first proof of an enumerative formula for non-strictly bispecial (and therefore for bispecial) Sturmian words.

We then investigate the minimal forbidden words for the set of finite Sturmian words. Recall that the set of minimal forbidden words of a factorial language is the set of words of minimal length that do not belong to the language [25]. More precisely, given a factorial language L over an alphabet Σ, a word $v=v_1v_2\cdots v_n$, with $v_i\in\Sigma$, is a minimal forbidden word for L if $v_1\cdots v_{n-1}$ and $v_2\cdots v_n$ are in L, but v is not.

Minimal forbidden words represent a powerful tool to investigate the structure of a factorial language (see [3,4,13]), such as the language of factors of a (finite or infinite) word, or of a set of words. They also appear in different contexts in Computer Science, such as symbolic dynamics [4], data compression (where the set of minimal forbidden words is often called an antidictionary) [12], or bio-informatics (where they are also called minimal absent words) [11].

We give a characterization of minimal forbidden words for the language St of Sturmian words in Theorem 5.1. We show that they are precisely the words of the form ywx such that xwy is a non-primitive Christoffel word, where $\{x,y\}=\{a,b\}$. This characterization allows us to give an enumerative formula for the set of minimal forbidden words of St (Corollary 5.2): there are exactly $2(n-1-\phi(n))$ minimal forbidden words of length n for every $n>1$.

The paper is organized as follows. In Section 2 we recall standard definitions on words and factors. In Section 3 we deal with Sturmian words and Christoffel words, and present our main result, and in Section 4 we give an enumerative formula for bispecial Sturmian words. Finally, in Section 5, we investigate minimal forbidden words for the language of finite Sturmian words.

2. Words and special factors

We give here basic definitions on words and fix the notation.

An alphabet, denoted by Σ, is a finite set of symbols, called letters. A word over Σ is a finite sequence of letters from Σ. The length of a word w is denoted by $|w|$. The only word of length 0 is called the empty word and is denoted by ε. The set of all words over Σ is denoted by Σ^*. The set of all words over Σ having length n is denoted by Σ^n. Any subset X of Σ^* is called a language, and we note $X(n)=|X\cap\Sigma^n|$ the set of words of length n in X.

Given a non-empty word w, we denote its i-th letter by $w[i]$, $1\leqslant i\leqslant|w|$. The reversal of the word $w=w[1]w[2]\cdots w[n]$ is the word $\tilde{w}=w[n]w[n-1]\cdots w[1]$. We set $\tilde{\varepsilon}=\varepsilon$. A palindrome is a word w such that $\tilde{w}=w$. A word is called a power if it is the concatenation of copies of another word; otherwise it is called primitive. For a letter $a\in\Sigma$, $|w|_a$ is the number of a's occurring in w. A positive integer p is a period of a word w if $p>|w|$ or $w[i]=w[i+p]$ for every $i=1,\dots,|w|-p$. For a word au, $a\in\Sigma$, $u\in\Sigma^*$, we define $\rho(au)=ua$. The set of rotations of a word w of length n is the set $\{\rho^i(w)\mid 1\leqslant i\leqslant n\}$. Note that the rotations of a word w are all different if and only if w is primitive.

[1] The suffix automaton of a finite word w is the minimal deterministic finite state automaton accepting the set of suffixes of w.

[2] Some authors require that p and q be coprime in the definition of Christoffel word. Here we follow the definition given in [5] and do not require this condition.