THE NUMBER OF BINARY ROTATION WORDS *

A. Frid[1] and D. Jamet[2]

Abstract. We consider binary rotation words generated by partitions of the unit circle to two intervals and give a precise formula for the number of such words of length n. We also give the precise asymptotics for it, which happens to be $\Theta(n^4)$. The result continues the line initiated by the formula for the number of all Sturmian words obtained by Lipatov [*Problemy Kibernet.* **39** (1982) 67–84], then independently by Mignosi [*Theoret. Comput. Sci.* **82** (1991) 71–84], and others.

Mathematics Subject Classification. 68R15, 37B10.

1. Introduction

Infinite words arising from rotations of the circle belong to the same family of infinite words defined by the means of dynamical systems as Sturmian words and interval exchange words in general. They were considered by Rote in 1992 [11] and can be defined using three parameters $\alpha, \beta, \gamma \in [0, 1)$ as $r = r_0 r_1 \ldots$, where for all i we have

$$r_i = \begin{cases} 1, & \text{if } \{i\alpha\} \in [\beta, \gamma), \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

(Here the interval $[\beta, \gamma)$ is denoted as usual if $\beta < \gamma$ and as $[\beta; 1) \cup [0, \gamma)$ otherwise).

In the particular case when $\gamma - \beta = \alpha \pmod 1$, w is a *Sturmian* word. The family of Sturmian words is very well studied (see Chapt. 2 of [9]); in particular,

[1] Sobolev Institute of Mathematics SB RAS, Koptyug Av. 4, 630090, Novosibirsk, Russia, and Université de Lorraine, 34 cours Léopold, CS 25233, 54052 Nancy cedex, France. anna.e.frid@gmail.com

[2] LORIA, UMR 7503, Campus scientifique BP 239, 54506 Vandoeuvre-lès-Nancy cedex, France.

the total number of factors of all Sturmian words taken together is known to be

$$1 + \sum_{p=1}^{n} (n - p + 1)\varphi(p),$$

where φ is the Euler's totient function. This formula was rediscovered several times [2, 3, 8, 10]; the order of growth of this function is $\Theta(n^3/\pi^2)$.

In [6, 7] Cassaigne and the first author estimated and for some cases found the number of factors of length n of all rotation words with a given length $\gamma - \beta$ of the interval; it happens that it also grows as $\Theta(n^3)$. In [1], Ambrož, Frid, Masáková and Pelantová estimated the number of all words arising from three-interval exchange, which continues the same line since Sturmian words are exactly two-interval exchange words; it happens that the number of three-interval exchange words grows as $\Theta(n^4)$. In [5], Berstel and Vuillon coded rotation words by Sturmian words.

In this paper, we find a precise formula for the number of all rotation words (1.1), predictably involving sums of the Euler's function. To write down the formula, we had to understand very clearly the structure of the set of rotation words, which is of independent interest.

2. Main statement

The main result of the paper is the following

Theorem 2.1. *Starting from* $n = 3$, *the number of binary rotation words of length* $n + 1$ *is*

$$f(n+1) = n^2 + 3n + 4 + \frac{1}{2} \sum_{p=3}^{n} \varphi(p) \left(n^2 - p^2 + n + p\right) - f_1(n) - 2 \sum_{l=2}^{n-1} f_2(n, l), \quad (2.1)$$

where

$$f_1(n) = \begin{cases} 2 \sum_{i=k}^{2k} \sum_{p=1}^{i+1} \varphi(p), & \text{if } n = 2k + 1, \\ 2 \sum_{i=k}^{2k-1} \sum_{p=1}^{i+1} \varphi(p) + \sum_{p=1}^{k} \varphi(p), & \text{if } n = 2k, \end{cases} \quad (2.2)$$

$$g(n, l) = n - l + 1 + (n \bmod (l + 1)), \quad (2.3)$$

$$h(n, l) = \min(l + 1, n - l),$$

and

$$f_2(n, l) = \left(\frac{1}{2} \left\lfloor \frac{n}{l+1} \right\rfloor g(n, l) - h(n, l) \right) (\varphi(l + 1) - 1) + h(n, l) \left(\frac{\varphi(l+1)}{2} - 1 \right). \quad (2.4)$$

Note that the only addend of this formula growing faster than than $O(n^3)$ is the sum

$$\sum_{p=3}^{n} \varphi(p) \left(n^2 - p^2\right).$$

So, the asymptotics of the number of binary rotation words is equal to the asymptotics of this addend, which means that

$$f(n) = \frac{3n^4}{4\pi^2} + O(n^3 \log n).$$

The values of $f(n)$ for some values of n are shown in the table below.

n	6	7	10	15	20	30	50	75	100
$f(n)$	64	112	504	2804	9442	51306	423814	2222984	7155096
$\dfrac{4\pi^2 f(n)}{3n^4} \approx$	0.65	0.61	0.66	0.73	0.78	0.83	0.89	0.92	0.94

The rest of the paper is devoted to the proof of Theorem 2.1 and thus to a study of internal structure of the set of rotation words.

3. ROTATIONS AND STURMIAN WORDS

Denote the prefix $r_0 \ldots r_{n-1}$ of length n of the word r defined in (1.1) by $r(\alpha, \beta, \gamma, n)$. The parameter α is called the *slope* of the rotation word r. The set of all rotation words $r(\alpha, \beta, \gamma, n)$ of length n is denoted by $R(n)$, so, the searched function is $f(n) = \#R(n)$.

Lemma 3.1. *It is sufficient to consider rotation words of slopes not greater than 1/2:*

$$R(n) = \{r(\alpha, \beta, \gamma, n) | \alpha \in (0, 1/2), \beta, \gamma \in \mathbb{R}/\mathbb{Z}\}.$$

Proof. Due to the symmetry, we have $r(\alpha, \beta, \gamma, n) = r(1 - \alpha, 1 - \beta, 1 - \gamma, n)$ if $\{k\alpha\} \neq \beta$ or γ for all $k = 0, \ldots, n-1$, that is, if the point $k\alpha$ is never equal to the end of an interval. But if it is, we can just take $r(1 - \alpha, 1 - \beta, 1 - \gamma, n)$ and then slightly shift the interval to avoid its ends. So, slopes less than 1/2 and greater than 1/2 give exactly the same set of all rotation words. $\square$

The following lemma is a particular case of the result of Berstel and Vuillon [5]. We give its proof for the sake of clarity.

Lemma 3.2. *For any two-interval rotation word $r = r(\alpha, \beta, \gamma, n)$, where $\alpha \leq 1/2$, we have*

$$r_k = r_{k-1} + u_k - v_k \tag{3.1}$$

for the Sturmian words $u = r(\alpha, \beta, \beta+\alpha, n)$ and $v = r(\alpha, \gamma, \gamma+\alpha, n)$ of the slope α.

Proof. The fact that $u_k = 1$ is equivalent to the fact that $\beta \in [\{(k-1)\alpha\}, \{k\alpha\})$; the fact that $v_k = 1$ is equivalent to the fact that $\gamma \in [\{(k-1)\alpha\}, \{k\alpha\})$. So, if $u_k = v_k = 0$, the interval $[\{(k-1)\alpha\}, \{k\alpha\})$ contains neither β nor γ, and thus $r_k = r_{k-1}$; if $u_k = v_k = 1$, the interval $[\{(k-1)\alpha\}, \{k\alpha\})$ contains both β and γ, and thus $r_k = r_{k-1}$ again; if $u_k = 1$ and $v_k = 0$, then $[\{(k-1)\alpha\}, \{k\alpha\})$ contains β but not γ, and thus $r_{k-1} = 0$ and $r_k = 1$; at last, if $u_k = 0$ and $v_k = 1$, then $[\{(k-1)\alpha\}, \{k\alpha\})$ contains γ but not β, and thus $r_{k-1} = 1$ and $r_k = 0$. In all the four cases (3.1) holds. □

Note that the letters u_0 and v_0 are not used in the previous lemma, so, we see that a rotation word of length $n+1$ is uniquely defined by its first symbol and two Sturmian words of the same slope of length n. If these two Sturmian words are distinct, we can uniquely reconstruct from them the symbol r_0; if they are equal, both rotation words 0^{n+1} and 1^{n+1} can appear. It is clear also that each pair of Sturmian words of the same slope gives some rotation word (of that slope). This gives us the next lemma:

Lemma 3.3. *The number of binary rotation words is bounded as*

$$f(n+1) \leq \#\{(u,v)|u, v \in St(n, \alpha), \alpha \in (0, 1/2), u \neq v\} + 2. \qquad (3.2)$$

Here $St(n, \alpha)$ is the set of all Sturmian words of length n and of slope α. The addend 2 in the formula above corresponds to all possible pairs of equal Sturmian words which all correspond to two rotation words, 0^{n+1} and 1^{n+1}.

Denote by $f_{pairs}(n)$ the number of such pairs of distinct Sturmian words of length n of the same slope $\alpha \in (0, 1/2)$, so that (3.2) can be rewritten as

$$f(n+1) \leq f_{pairs}(n) + 2. \qquad (3.3)$$

Lemma 3.4. *For all $n \geq 1$ we have*

$$f_{pairs}(n) = n(n+1) + \frac{1}{2} \sum_{p=3}^{n} \varphi(p)(n^2 - p^2 + n + p).$$

Proof. Recall that the Farey series F_n of order n is the increasing sequence of all fractions between 0 and 1 whose denominators are at most n. The intervals between consecutive Farey fractions are called *Farey intervals*. The first Farey fraction is taken to be $0 = 0/1$; all the others are of the form q/p, where $1 \leq q < p \leq n$, $\gcd(q, p) = 1$.

It is well-known that the sets $St(n, \alpha)$ coincide for all α from the same Farey interval of order n (see, *e.g.* Lem. 11 of [6]); if the beginning of the interval is the fraction q/p, we can denote this set as $St(n, \alpha) = St(n, q, p)$.

Let us say that a Sturmian word of length n is *new* in the Farey interval starting from q/p if it belongs to $St(n, q, p)$ but does not belong to any $St(n, q', p')$ for $q'/p' < q/p$, where $q/p, q'/p' \in F_n$. Denote the set of all new Sturmian words from

$St(n, q, p)$ by $N(n, q, p)$; all other words from $St(n, q, p)$ are called *old*, and their set is denoted by $Old(n, q, p)$.

As it follows directly from the results by Berstel and Pocchiola [4], for all $q/p \in F_n$ with $p > 1$ we have $\#N(n, q, p) = n - p + 1$. As a corollary, we immediately see that $\#Old(n, q, p) = p$.

Now let us count $f_{pairs}(n)$ starting from the minimal slope and going on along the Farey series. In the interval starting from 0, all the words are new, and they give $n(n + 1)$ pairs. In any other interval, we are interested only in pairs where at least one of the words is new, since the pairs where both words are old have been counted before. So, after excluding pairs of old words, we see in the interval starting from q/p the following number of new Sturmian pairs:

$$n(n + 1) - \#Old(n, q, p)(\#Old(n, q, p) - 1) = n^2 - p^2 + n + p.$$

Now note that the number of Farey fractions whose denominator is p and which are less than $1/2$ is $\varphi(p)/2$ for all $p > 2$; for $p = 2$, the only Farey fraction is $1/2$, but we are not interested in slopes greater than $1/2$; for $p = 1$, the case of $0 = 0/1$ is a bit special and has been considered in the beginning of this paragraph. So, summing up, we obtain that

$$f_{pairs}(n) = n(n + 1) + \frac{1}{2} \sum_{p=3}^{n} \varphi(p)(n^2 - p^2 + n + p). \qquad \square$$

Together with (3.3), this lemma already gives us an upper bound for $f(n + 1)$. However, to pass to a precise formula, we should classify the cases when different pairs of Sturmian words give the same rotation word.

We start from the following

Lemma 3.5. *If a rotation word r contains both factors 00 and 11, then it appears from only one pair of Sturmian words u and v of the same slope $\alpha < 1/2$. They can be found by the equalities $u_k = v_k = 0$ if $r_{k-1} = r_k = 1$ and $u_k = r_k$, $v_k = r_{k-1}$ otherwise.*

Proof. Let a be the symbol whose interval is shorter than the other one and thus not longer than $1/2$. The jump α of the moving point cannot be greater than the length of the longer interval, and thus, since aa appears in r, we see that in any pair of Sturmian words of the slope $\alpha < 1/2$, generating r, the slope α is less than the length of the interval corresponding to a, and all the more of the other interval. It means exactly that we can never have $u_k = v_k = 1$. So, we can uniquely reconstruct the words u and v: if $r_k = r_{k-1}$, then $u_k = v_k = 0$, if $r_k = 1$ and $r_{k-1} = 0$, then $u_k = 1$ and $v_k = 0$, and if $r_k = 0$ and $r_{k-1} = 1$, then $u_k = 0$ and $v_k = 1$. This is equivalent to the statement of the lemma. $\square$

So, to classify all pairs of Sturmian words of the same slope α, we must consider only those of them which contain consecutive occurrences of at most one symbol.

Due to the symmetry, we can suppose for a while that this symbol is 0, that is, that the rotation words considered do not contain the factor 11.

The words 0^{n+1} and 1^{n+1} have been already excluded from consideration and gave the addend 2 to the formula 3.3. So, in what follows we consider two cases: either r contains only one symbol 1, that is, $r = 0^i 10^{n-i}$ for some i, or r contains at least two 0 s and two 1 s. The proof will rely on properties of Sturmian words, so that we start from listing some of them.

4. PROPERTIES OF STURMIAN WORDS

Denote the set of all Sturmian words of the slope α by $St(\alpha)$; so, $St(\alpha, n) = \{0, 1\}^n \cap St(\alpha)$. As always, we say that a word u from a binary language L is left (right) *special* in L if both $0u$ and $1u$ (respectively, $u0$ and $u1$) are also in L. The mirror image $\tilde{u}$ of the word $u = u_1 \ldots u_m$ is the word $u_m \ldots u_1$.

The following statements are classical and can be found in any survey on Sturmian words (see, *e.g.*, [9]).

Claim 4.1. Each language $St(\alpha)$ contains exactly one left special word u and one right special word v of each length, and $u = \tilde{v}$. A shorter left special word from $St(\alpha)$ is always a prefix of a longer one; the same holds symmetrically on right special words and suffixes.

Claim 4.2. The total number of left (right) special Sturmian words of length n is $\sum_{p=1}^{n+1} \varphi(p)$.

Another family of known facts concerns the construction of Sturmian words with directive sequences, standard words and central words. The facts below can be found in [9].

Consider a *directive sequence* $(d_1, d_2, \ldots)$, where $d_1 \geq 0$, $d_i > 0$ for all $i > 1$, and construct a sequence of words

$$s_{-1} = 1, s_0 = 0, s_n = s_{n-1}^{d_n} s_{n-2} (n \geq 1).$$

The words s_n obtained are called *standard* words. All standard words are Sturmian, and for each Sturmian word w there exists a directive sequence such that all the standard words it generates are factors of w. So, the set of factors of w is uniquely determined by some directive sequence, which is directly related to the continuous fraction expansion $[0, 1 + d_1, d_2, \ldots]$ of the slope α of w. In what follows we denote this slope by α_d.

The slope α_d is not greater than $1/2$ if and only if $d_1 > 0$. Since in this paper we are interested in Sturmian words whose slope is less than $1/2$, from now on we assume for each directive sequence that $d_i > 0$ for all $i > 0$.

It can be easily checked that all standard words of length at least 2 end by 01 or 10. For each standard word $s_n = c_n ab$, where $a \neq b$, $a, b \in \{0, 1\}$, the word c_n is called a *central* word. In what follows we denote the word $c_n ba$ by s'_n.

The following facts on standard and central words can be easily proved. Once again, we refer to [9] for details.

Claim 4.3. All standard words from $St(\alpha_d)$ are left special in that language.

Claim 4.4. For each standard word $s_n \in St(\alpha_d)$, the word s'_n is also standard and belongs to $St(\alpha_d)$, but is not left special in that language.

Claim 4.5. If $s_n = s_{n-1}^{d_n} s_{n-2}$, then $s'_n = s_{n-1}^{d_n-1} s_{n-2} s_{n-1}$.

Claim 4.6. Central words c_n are bispecial in $St(\alpha_d)$.

Claim 4.7. A word c is bispecial in $St(\alpha_d)$ if and only if it is obtained by deleting two last symbols from some word $s_{n-1}^t s_{n-2}$, where $0 < t \leq d_n$. We shall denote this word by $c_{n,t}$; in particular, $c_{n,d_n} = c_n$.

Claim 4.8. The total number of central words of a length n is $\varphi(n+2)$.

Claim 4.9. The last two symbols of standard words alternate: if s_{n-1} ends by 01, then s_n ends by 10, and *vice versa*.

For each directive sequence $(d_1, d_2, \ldots)$, we denote the length of the nth standard word s_n by l_n.

The following lemma can be proved analogously to Theorem 2.2.31 from [9].

Lemma 4.10. *For all $n \geq 1$, if the standard word s_n in the language $St(\alpha_d)$ ends by a symbol a, then the word $a s_n^{d_{n+1}+2}$ is not a factor of $St(\alpha_d)$.*

5. The case of unique 1

Let r be a rotation word of the form $0^i 1 0^{n-i}$. Clearly, if $0 < i < n$, this word is generated by *all* pairs of Sturmian words of the form $(s10t, s01t)$ with $s10t, s01t \in S(n, \alpha)$ for some slope α, where $|s| = i-1, |t| = n-i-1$. In particular, s is a right special word in $St(i-1, \alpha)$, and t is a left special word in $St(n-i-1, \alpha)$. Due to Claim 4.1, the shorter of the words s and t is determined by the longer one.

Suppose first that $i \geq n - i$; then s is longer than t and determines the whole word $s10t$. So, the number of pairs of Sturmian words giving $0^i 1 0^{n-i}$ is equal to the number of (right) special Sturmian words of length $i-1$, that is, to $\sum_{p=1}^{i} \varphi(p)$ (see Claim 4.2). Symmetrically, if $i < n - i$, then this is t that defines the whole word $s10t$, and the number of such pairs is equal to the number of (left) special Sturmian words of length $n - i - 1$, that is, to $\sum_{p=1}^{n-i} \varphi(p)$.

Now if $i = 0$ and the rotation word is 10^n, the pairs of Sturmian words generating it are all pairs of the form 0 s, 1 s. So, s is a left special Sturmian word of length $n - 1$, and the number of such words is $\sum_{p=1}^{n} \varphi(p)$. Symmetrically, the number of pairs of Sturmian words generating the word 01^n is also the same.

Summing up all the numbers above, we see that the $n + 1$ rotation words of the form $0^i 1 0^{n-i}$, where $0 \leq i \leq n$, are generated by the following number of pairs of Sturmian words of length n:

$$f_1(n) = \begin{cases} 2 \sum_{i=k}^{2k} \sum_{p=1}^{i+1} \varphi(p), & \text{if } n = 2k + 1, \\ 2 \sum_{i=k}^{2k-1} \sum_{p=1}^{i+1} \varphi(p) + \sum_{p=1}^{k} \varphi(p), & \text{if } n = 2k. \end{cases} \tag{5.1}$$

Starting with $n = 3$, the sums in $f_1(n)$ involve only special words of length at least 1. Exactly a half of them, namely, the left special words starting with 0 and symmetrically the right special words ending with 0, correspond to slopes less than 1/2. So, the $n+1$ rotation words of the form $0^i 10^{n-i}$ are generated by $f_1(n)/2$ pairs of Sturmian words.

Exactly the same total number of pairs (in fact, the pairs (v, u), where (u, v) are the pairs considered above) generate the $n+1$ rotation words of the form $1^i 01^{n-i}$. Starting from $n = 3$, it gives us exactly $f_1(n)$ pairs generating $2(n + 1)$ rotation words.

6. The case of several 1s

Most of technical details of our result are hidden in the following

Theorem 6.1. *Suppose a rotation word w is generated by at least two different pairs of Sturmian words of slope less than 1/2, and w contains at least two 1s and at least two 0 s. Then $w = 0^i(10^l)^k 10^j$ or $w = 1^i(01^l)^k 01^j$ for some $i, j \geq 0$, $l \geq 2$, $k \geq 1$, and the number of pairs generating w is equal to $\varphi(l + 1)/2$ if $i, j \leq l$ and $\varphi(l + 1)$ otherwise.*

This section is devoted to its proof which is based on the theory of standard Sturmian words and their construction with directive sequences (see Sect. 4).

Lemma 6.2. *Suppose that a rotation word r contains a factor $010^{k_1}10^{k_2}1$, where $k_2 > k_1 > 0$, and is generated by a pair (u, v) of Sturmian words from some $St(\alpha_d)$. Then $k_1 = l_n - 1$ for some n and $(k_2 + 1) \bmod l_n = l_{n-1}$.*

Proof. Clearly, the pair (u, v) generating r contains some factors $u' = 10u_1 10u_2 1$ and $v' = 01u_1 01u_2 0$ for some u_1, u_2 with $|u_1| = k_1 - 1$ and $|u_2| = k_2 - 1$. The words u_1 and u_2 are bispecial in $St(\alpha_d)$. So, due to Claim 4.7, $u_1 = c_{n,d}$ for some $n > 0$ and $0 < d \leq d_n$. Note also that u_2 is left special in $St(\alpha_d)$ longer than s_{n-1}, so it starts from s_{n-1} since there is only one left special word of each length in $St(\alpha_d)$ (see Claim 4.1).

Without loss of generality, suppose that s_{n-2} ends by 10 (and thus s_{n-1} ends by 01 due to Claim 4.9); if the opposite holds, in all the arguments below we should just consider v' instead of u'.

Suppose that $d < d_n$; then $u' = 10 s_{n-1}^d s_{n-2} s_{n-1} u''$ since $u_1 10 = s_{n-1}^d s_{n-2}$ and u_2 starts with u_1 and thus with s_{n-1}. In particular, it means that the word $s_{n-1}^d s_{n-2} s_{n-1}$ can be extended to the left by 0. On the other hand, since $d < d_n$, the same word can clearly be extended to the left by s_{n-1}, and thus by its last symbol 1. We see that $s_{n-1}^d s_{n-2} s_{n-1}$ is left special; but it is not possible since due to Claim 4.5 $s_{n-2} s_{n-1}$ differs in two last symbols from $s_{n-1} s_{n-2}$, and thus $s_{n-1}^d s_{n-2} s_{n-1}$ is not equal to the prefix $s_{n-1}^{d+1} s_{n-2}$ of s_n of the same length. Since $s_{n-1}^{d+1} s_{n-2}$ is the only left special word of its length in $St(\alpha_d)$, we see that $0 s_{n-1}^d s_{n-2} s_{n-1}$ and thus u' are not in $St(\alpha_d)$, a contradiction. So, the case of $d < d_n$ is not possible, and thus $u_1 = c_n$ for some n, and $k_1 = l_n - 1$.

Now recall that u_2 is longer than u_1; so, it is equal to $c_{N,D}$ for some $N > n$ and $0 < D \leq d_N$. Suppose that $N > n + 1$; then u_2 starts with s_{n+1} and $u' = 10s_n s_{n+1} u''$. As above, the word $s_n s_{n+1}$ can be extended to the left by 1, which is the last symbol of s_{n+1}, and it is not left special since it is not equal to the only special word $s_{n+1} s_n$ of the same length, so, $0s_n s_{n+1}$ and thus u are not elements of $St(\alpha_d)$.

So, $N = n + 1$, $u_2 01 = s_n^d s_{n-1}$ for some $d \leq d_{n+1}$, and $(|u_2| + 2) \bmod l_n = l_{n-1}$, which was to be proved. $\qquad\square$

Lemma 6.3. *Consider two slopes α_d, $\alpha_{d'} < 1/2$ with corresponding directive sequences $(d_1, d_2, \ldots)$ and $(d_1', d_2', \ldots)$ and respective lengths l_n and l_n' of standard words. If $l_n = l_m'$ and $l_{n-1} = l_{m-1}'$ for some m and n, then $n = m$ and $d_i = d_i'$ for all $i = 1, \ldots n$.*

Proof. By the construction, for all i we have $l_{i-2} = l_i \bmod l_{i-1}$ and $d_i = \lfloor l_i : l_{i-1} \rfloor$. So, starting from $l_n = l_m'$ and $l_{n-1} = l_{m-1}'$, we can uniquely reconstruct $l_{n-2} = l_{m-2}'$, $l_{n-3} = l_{m-3}'$ *etc.* Note that here $d_1, d_1' > 0$ since both slopes are less than $1/2$. So, as soon as we get $l_{n-k} = l_{m-k}' = 1$, we immediately see that $n = m = k$ and $(d_1, \ldots, d_n) = (d_1', \ldots, d_n')$. $\qquad\square$

Lemma 6.4. *All rotation words arising from several pairs of Sturmian words of slope at most $1/2$ and not containing two consecutive 1s are of the form $0^i (10^l)^k 10^j$ for some $i, j \geq 0$, $k \geq 1$, $l \geq 2$.*

Proof. Let us prove that a word arising from two pairs of Sturmian words, of slopes $\alpha_d \neq \alpha_{d'}$, cannot contain a factor $w = 10^{m_1} 10^{m_2} 1$ with $m_1 \neq m_2$. Suppose it contains it. The proof is carried over for $m_2 > m_1$; the opposite case can be proved by the argument that the set of Sturmian words, the set of rotation words and the procedure generating a rotation word from two Sturmian words are symmetric under taking the mirror image.

First suppose that $1w$ is also a rotation word. Due to Lemma 3.5, $1w$ appears from only one pair of Sturmian words, and this pair is $(0^{m_1+1} 10^{m_2} 1, 010^{m_1} 10^{m_2})$. Since for a given slope, the number of consecutive 0 s between two 1 s in a Sturmian words takes only two consecutive values, these two words are Sturmian of the same slope only if $m_2 = m_1 + 1$. But in this case, we can apply Lemma 6.2 to the word $0w = 010^{m_1} 10^{m_1+1} 1$ and see that it is also generated by only one pair of Sturmian words: indeed, $m_1 = l_n - 1$ for some n, and $(m_1 + 2) \pmod{m_1 + 1} = l_{n-1}$, which means that $l_{n-1} = 1$ and $n = 1$ (since $d_1 > 0$). So, as it follows from the proof of Lemma 6.2, the word $0w$ can only be generated by the pair $(10^{m_1} 10^{m_1+1} 1, 010^{m_1} 10^{m_1+1})$. So, $1w$ and $0w$ are both generated only by pairs giving the pair $(0^{m_1} 10^{m_1+1} 1, 10^{m_1} 10^{m_1+1})$ after erasing the first symbol of each word. So, w is only generated by this latter pair, contradicting to our assumption.

So, if w arises from several pairs of Sturmian words of slope at most $1/2$, then so does $0w$. Since we suppose that $m_2 > m_1$, we can apply Lemma 6.2, according to which $m_1 = l_n - 1$ and $m_2 \bmod m_1 = l_{n-1} - 1$, where l_i are the lengths

coming from the directive sequence for the language of the Sturmian words involved. Due to Lemma 6.3, the values of l_n and l_{n-1} uniquely determine the sequence $(d_1, d_2, \ldots, d_n)$, its length n and thus the central words u_1 and u_2 such that $0w = 0u_1 10u_2 1$ and $v = 1u_1 01u_2 0$. So, we see that once again, $0w$ and thus w arise from only one pair of Sturmian words of slope less than $1/2$, a contradiction.

We have proved that our word is of the form $0^i (10^l)^k 10^j$ for some $i, j \geq 0$, $k \geq 1$ (since the case of a unique 1 is considered separately), and $l > 0$. It remains to consider the case of $l = 1$ and to see that the word $0^i (10)^k 10^j$ is generated by only one pair of Sturmian words of slope less than $1/2$ defined as follows: the central part $(10)^k 1$ is given by the pair $u = 0(10)^{k-1} 1$, $v = 1(01)^{k-1} 0$, and the prefix and suffix zeros correspond to the common prefix $\ldots 1010$ and the common suffix $0101 \ldots$ of the generating Sturmian words. No other pairs can give this word, and thus, we can state that $l \geq 2$. □

Lemma 6.5. *Each word $w = 0^i (10^l)^k 10^j$ with $i, j \leq l$, $l \geq 2$, $k \geq 1$ is generated by $\varphi(l+1)/2$ different pairs of Sturmian words of the same slope not greater than $1/2$.*

Proof. Suppose for simplicity that $i, j > 0$ and consider a pair (u, v) such that w is generated by it: we denote this fact by $w = r(u, v)$. Clearly, $(u, v) = (s10c_1 10 \ldots 10c_k 10p, s01c_1 01 \ldots 01c_k 01p)$ for some central words $c_1, \ldots, c_k$ of length $l - 1$ and some words s, p with $|s| = i - 1$, $|p| = j - 1$. Since each Sturmian language $St(\alpha)$ contains at most one central word of length $l - 1$, we have $c_1 = c_2 = \ldots = c_k = c$. Moreover, the word p is left special, and so it is a prefix of c, and the word s is right special, so it is a suffix of c. So, the pair (u, v) is uniquely determined by the central word c and the parameter i.

There exists $\varphi(l+1)$ central words of length $l-1$, and a half of them correspond to slopes less than $1/2$. So, it remains to prove that for each central word c and for each power k two words $u = s(10c)^k 10p$ and $v = s(01c)^k p$, where s is a suffix and p is a prefix of c, appear in some Sturmian language of a given slope.

Indeed, let $c10$ be equal to the standard word $s_n = s_{n-1}^{d_n} s_{n-2}$ in some language $St(\alpha_d)$ with $d = (d_1, \ldots, d_n, \ldots)$. Then u is a factor of s_n^{k+2}. At the same time, $c01 = s_n' = s_{n-1}^{d_n - 1} s_{n-2} s_{n-1}$, so that v is a factor of $(s')_n^{k+2} = s_{n-1}^{d_n - 1} s_{n-2} s_n^{k+1} s_{n-1}$, which is in its turn a factor of s_n^{k+3}. So, taking $d_{n+1} \geq k + 3$, we see that $u, v \in St(\alpha_d)$ for the directive sequence $d = (d_1, \ldots, d_n, k+3, \ldots)$, which was to be proved.

If $i = 0$, or $j = 0$, or/and the standard word s_n is equal to $c01$, not to $c10$, the proof is carried on similarly. □

Lemma 6.6. *Each word $w = 0^i (10^l)^k 10^j$ with $i > l$ or $j > l$, $l \geq 2$, $k \geq 1$ is generated by $\varphi(l+1)$ different pairs of Sturmian words of the same slope not greater than $1/2$.*

Proof. As in the previous lemma, if $w = r(u, v)$, then $u = s(10c)^k 10p$ and $v = s(01c)^k 01p$ for some central word c of length $l - 1$ and some words s, p with $|s| = i - 1$, $|p| = j - 1$. We have a choice which of the words $c01$ and $c10$ is

a standard word in the Sturmian language considered; suppose it is $c10 = s_n$. Suppose also that $j > l$. Clearly, p is left special and thus is a prefix of some standard word s_N, $N > n$. Suppose that p is not a prefix of some power s_n^d of s_n: it means that p contains as a prefix the word $s_n^{d_{n+1}}s_{n-1}s_n$, or, more precisely, the word obtained from it by erasing the last symbol, since $s_{n-1}s_n$ differs from $s_n s_{n-1}$, which is a prefix of s_n^2, by the two last symbols (see Claim 4.5). In particular, p starts by $s_n^{d_{n+1}+1}$, and thus the suffix $0c10p = 0s_n p$ of u starts with $0s_n^{d_{n+1}+2}$, which is not an element of $St(\alpha_d)$ due to Lemma 4.10, a contradiction. So, p is a prefix of s_n^d for some d. Symmetrically, if $i > l$, then s is a suffix of the mirror image of s_n^d for some d; by the way, this mirror image is equal to $01(c01)^{d-1}c$. For $s_n = c01$, we should just consider v instead of u to prove the similar statements.

Note that since at least one of the parameters i or j is indeed greater than l, the cases of $s_n = c10$ and of $s_n = c01$ are really different, which gives us $\varphi(l+1)$ cases: the total number of standard words of length $l+1$ is twice bigger than the number of central words of length $l-1$, that is, is equal to $2\varphi(l+1)$, but we are interested only in those of slope less than $1/2$.

It remains to mention that for each standard word $s = cab$ of length $l+1$, where $a, b \in \{0, 1\}$, $a \neq b$, and for all $d', d'', k \geq 0$, the words $u' = (cba)^{d'}(cab)^{k+d''}$ and $v' = (cba)^{d'+k}(cab)^{d''}$, so that w is a factor of $r(u', v')$, are factors of some language $St(\alpha_d)$. Indeed, let $s = s_n$ for the directive sequence $(d_1, \ldots, d_n)$; then $s = cab = s_{n-1}^{d_n}s_{n-2}$ and $cba = s_{n-1}^{d_n-1}s_{n-2}s_{n-1}$, so that $u' = (s_{n-1}^{d_n-1}s_{n-2}s_{n-1})^{d'}(s_{n-1}^{d_n}s_{n-2})^{k+d''} = s_{n-1}^{d_n-1}s_{n-2}s_n^{d'-1}s_{n-1}s_n^{k+d''}$, and $v' = s_{n-1}^{d_n-1}s_{n-2}s_n^{d'+k-1}s_{n-1}s_n^{d''}$. If we take $d_{n+1} = D = \max\{d'+k, d''+k\}$, we see that both u' and v' are factors of $s_n^D s_{n-1}s_n^D$. So, both u' and v', and thus the pair of words based on the standard word s of length $l+1$ and giving the rotation word w, are elements of the language $St(\alpha_d)$ for the directive sequence $d = (d_1, \ldots, d_n, D+2, \ldots)$, which completes the proof of the lemma. $\qquad\square$

This lemma, in its turn, completes the proof of Theorem 6.1.

7. Final computations

To find the precise formula for $f(n+1)$ for $n \geq 3$, we should subtract from the bound $f_{pairs}(n) + 2$, where $f_{pairs}(n)$ is found in Lemma 3.4, the number of pairs generating rotation words already obtained before.

As it was shown in Section 5, the $2(n+1)$ rotation words with only one symbol 1 or only one 0 are generated by $f_1(n)$ pairs (see (5.1)).

Now let us take into account the rotation words containing several 0 s and several 1 s. Their form is described in Theorem 6.1. Consider all rotation words of length $n+1$ of the form $0^i(10^l)^k10^j$, such that l and i are fixed and k and j are not. They are $\lfloor (n-i)/(l+1) \rfloor$; and taking all words with a given l together, we

see that they are

$$\sum_{i=0}^{n-(l+1)} \left\lfloor \frac{n-i}{l+1} \right\rfloor = \frac{1}{2} \left\lfloor \frac{n}{l+1} \right\rfloor (n - l + 1 + (n \bmod (l+1))).$$

In what follows, to make the formulas shorter, we will denote

$$g(n, l) = n - l + 1 + (n \bmod (l+1)),$$

so that the words of the form $0^i(10^l)^k10^j$ are $\frac{g(n,l)\lfloor n/(l+1)\rfloor}{2}$.

Each of these words is generated by $\varphi(l+1)$ pairs of Sturmian words, except for the $\min(l+1, n-l)$ words with $i, j \le l$ which are generated by $\varphi(l+1)/2$ pairs each. So, for each $l \ge 2$ we should subtract from the sum the following function:

$$f_2(n, l) = \left(\frac{1}{2} \left\lfloor \frac{n}{l+1} \right\rfloor g(n, l) - \min(l+1, n-l) \right) (\varphi(l+1) - 1)$$
$$+ \min(l+1, n-l) \left(\frac{\varphi(l+1)}{2} - 1 \right).$$

The same function $f_2(n, l)$ corresponds to the words of the form $1^i(01^l)^k01^j$. So, to take into account all rotation words arising from several pairs and containing at least two 0 s and at least two 1 s, we should subtract from the upper bound the sum $2\sum_{l=2}^{n-1} f_2(n, l)$.

Summarizing the above arguments, we see that

$$f(n + 1) = f_{pairs}(n) + 2 - f_1(n) + 2(n + 1) - 2 \sum_{l=2}^{n-1} f_2(n, l).$$

This is exactly the statement of Theorem 2.1 which was to be proved.

References

[1] P. Ambrož, A. Frid, Z. Masáková and E. Pelantová, On the number of factors in codings of three interval exchange, *Discr. Math. Theoret. Comput. Sci.* **13** (2011) 51–66.
[2] C.A. Berenstein, L.N. Kanal, D. Lavine and E.C. Olson, A geometric approach to subpixel registration accuracy. *Comput. Vision Graph.* **40** (1987) 334–360.
[3] J. Berstel and M. Pocchiola, A geometric proof of the enumeration formula for Sturmian words. *Int. J. Algebra Comput.* **3** (1993) 349–355.
[4] J. Berstel and M. Pocchiola, Random generation of finite Sturmian words. *Discr. Math.* **153** (1996) 29–39.
[5] J. Berstel and L. Vuillon, Coding rotations on intervals. *Theoret. Comput. Sci.* **281** (2002) 99–107.
[6] J. Cassaigne and A.E. Frid, On the arithmetical complexity of Sturmian words. *Theoret. Comput. Sci.* **380** (2007) 304–316.

[7] A. Frid, A lower bound for the arithmetical complexity of Sturmian words, *Siberian Electron. Math. Rep.* **2** (2005) 14–22 (in Russian, English abstract).

[8] E.P. Lipatov, A classification of binary collections and properties of homogeneity classes. *Problemy Kibernet.* **39** (1982) 67–84 (in Russian).

[9] M. Lothaire, *Algebraic Combinatorics on Words*. Cambridge University Press, Cambridge (2002).

[10] F. Mignosi, On the number of factors of Sturmian words. *Theoret. Comput. Sci.* **82** (1991) 71–84.

[11] G. Rote, Sequences with subword complexity 2n, *J. Number Theory* **46** (1994) 196–213.