JOURNAL DE THÉORIE DES NOMBRES DE BORDEAUX

ALEX HEINIS

On low-complexity bi-infinite words and their factors

Journal de Théorie des Nombres de Bordeaux, tome 13, nº 2 (2001), p. 421-442

<http://www.numdam.org/item?id=JTNB_2001__13_2_421_0>

Journal de Théorie des Nombres
de Bordeaux **13** (2001), 421–442

On low-complexity bi-infinite words and their factors

par Alex HEINIS

RÉSUMÉ. Dans cet article on étudie des mots bi-infinis sur deux symboles. On dit qu'un tel mot est de rigidité k si le nombre de facteurs différents de longueur n est egal à $n+k$ pour n grand. Un tel mot est appelé k-balancé si le nombre d'occurrences du symbole a dans deux facteurs quelconques de même longueur peuvent différer au plus de k. Dans cet article on donne une description complète de la classe des mots bi-infinis de rigidité k et on montre que le nombre de facteurs de longueur n de cette classe est de l'ordre de n^3. Dans le cas $k = 1$ on donne une formule exacte. On considère aussi la classe des mots bi-infinis k-balancés. Il est bien connu que le nombre de facteurs de longueur n est de l'ordre de n^3 si $k = 1$. En revanche, on montre que ce nombre est $\geq 2^{n/2}$ si $k \geq 2$.

ABSTRACT. In this paper we study bi-infinite words on two letters. We say that such a word has stiffness k if the number of different subwords of length n equals $n + k$ for all n sufficiently large. The word is called k-balanced if the numbers of occurrences of the symbol a in any two subwords of the same length differ by at most k. In the present paper we give a complete description of the class of bi-infinite words of stiffness k and show that the number of subwords of length n from this class has growth order n^3. In the case $k = 1$ we give an exact formula. We also consider the class of k-balanced bi-infinite words. It is well-known that the number of subwords of length n from this class has growth order n^3 if $k = 1$. In contrast, we show that the number is $\geq 2^{n/2}$ when $k \geq 2$.

1. Introduction: structure theorems.

In general, a word is defined to be a mapping $w : I \to \Sigma$ where I is an interval of integers and where Σ is a finite alphabet of symbols. In this paper we only consider words over the alphabet $\{a, b\}$. A *subword* of w is

Manuscrit reçu le 2 février 1999.

the restriction of w to some interval $J \subset I$. By abuse of notation we will write $x \subset w$ if x is a subword of w. We identify words which are translations of each other.

A **Z**-*word* is a word w with domain $\mathbf{Z}$. It is called *recurrent* if every finite subword appears more than once in w. It then appears infinitely often, but not a priori infinitely often in both directions. The word w is called *periodic* if there exists a $p \in \mathbf{N}^+$, the positive integers, such that $w_i = w_{i+p}$ for all i and then p is called a *period*. The minimal such p is called *the* period of w. Two finite words are called *conjugate* if they are cyclic permutations of each other and we write $x \sim y$. Now if p is the period for w then the conjugacy class of any subword of w of length p is called a *period cycle*. The word w is called *left periodic* if there exist numbers $p \in \mathbf{N}^+, N \in \mathbf{Z}$ such that $w_i = w_{i-p}$ if $i \leq N$. The minimal possible value for p is called the left period for w. A similar definition can be given with left replaced by right.

If a word w is finite we define the *content* of w as the number of a's in it, i.e., $|\{i \in I | w_i = a\}|$ and we write $c(w)$. A word w is called k-*balanced* if $|c(A) - c(B)| \leq k$ for every pair of subwords A, B of the same length. If $k = 1$ then w is called balanced.

For $n \geq 0$ we define $\mathcal{B}(w,n)$ as the collection of subwords of length n and we write $P(w,n) := |\mathcal{B}(w,n)|$ for its cardinality. It is called the *complexity function* of w. The word w is called k-*stiff* if $P(w,n) \leq n + k$ for all n. If $k = 1$ we will just write stiff. It is well-known, see [C, Theorems 2.06, 2.11, 2.14] that the following statements concerning a **Z**-word w are equivalent: (1) $P(w, n+1) = P(w,n)$ for some $n \geq 1$; (2) $P(w,n)$ is bounded; (3) $P(w,n) \leq n$ for some $n \geq 1$; (4) w is periodic. Hence, if w is k-stiff and not periodic, then $P(w,n)$ is strictly increasing in n and it follows that there exist integers $N \geq 0, k' \geq 1$ such that $P(w,n) = n + k'$ for $n \geq N$. Coven [C] uses the term *minimal block growth* to describe this and we will adopt this term. We call k' the *stiffness* of w and we write it as $k(w)$. In Section 3 we will recall some well-known facts about balanced and stiff words which we will need further on in the article.

The following two theorems describe the structure of non-recurrent k-stiff **Z**-words. Proofs can be found in [C, Section 3]. We note, however, that Theorem 3.10 in [C] is not entirely correct. To avoid confusion we will give our own proofs in Section 4 and afterwards we indicate what (minor) changes have to be made in Theorem 3.10 in order for it to be true.

Theorem A. *Let w be k-stiff ($k \geq 1$) and not recurrent. Then w is left periodic and right periodic.*

Now let w be an arbitrary **Z**-word which is left periodic and right periodic but not periodic, for instance $w = \cdots(ab)(ab)(aba)(aba)\cdots$ We denote the left and right period by s, r respectively and the corresponding maximal periodic domains by $(-\infty, \lambda], [\mu, \infty)$. We define the overlap of w to be $g = \lambda - \mu + 1$ (this overlap can be 0 or negative). Also we define $\chi = r = s$ if the period cycles are equal and $\chi = s + r$ otherwise.

Theorem B. *Let w be a $\mathbf{Z}$-word which is left periodic and right periodic but not periodic. Then w is not recurrent, it has minimal block growth and $k(w) = \chi - 1 - g$.*

Important for the recurrent case is the notion of a *substitution.* A substitution is a mapping from words on $\{a, b\}$ to words on $\{a, b\}$ which replaces every a by a finite word X and every b by a finite word Y. We write (X, Y) for this transformation. In this article we always assume that X begins with a and that Y begins with b. Let T be such a substitution. It might happen that T is of the form (X, ZX) or of the form (ZY, Y). Of course these cases are incompatible since $X \neq Y$. In the first case we define the *reduction* T^{red} of T as (X, Z) and in the second case as (Z, Y). If neither of these cases apply we call T *irreducible* and we just define $T^{\text{red}} = T$. Now let T be any substitution. After a finite number of reductions one obtains an irreducible substitution and we denote the result by T^{RED}. For example, if $T = (abba, bba)$ then $T^{\text{RED}} = (a, bb)$. Note that an irreducible substitution T is of the form $(A\sigma C, B\overline{\sigma}C)$ with uniquely determined A, B, C, σ. Conversely, every substitution of this form such that $A\sigma, B\overline{\sigma}$ have initial symbols a, b, respectively, is irreducible. Here σ is a symbol and $\overline{\sigma}$ its negation, i.e., the other member of the alphabet. We define the *stiffness* $k(T)$ of an irreducible substitution $T = (A\sigma C, B\overline{\sigma}C)$ by $k(T) = |ABC| + 1$. If T is reducible then we define $k(T) = k(T^{\text{RED}})$. The term "stiffness" will become clear from the next theorem and Lemma 7. Concerning the next theorem we note that recurrent, stiff, non-periodic $\mathbf{Z}$-words are called *Sturmian* $\mathbf{Z}$-words. See also Section 3.

Theorem 1. *Let w be a recurrent k-stiff $\mathbf{Z}$-word. Then there exist a recurrent stiff $\mathbf{Z}$-word σ and a substitution T such that $T\sigma = w$. Conversely, if σ is Sturmian, then $T\sigma$ is recurrent, has minimal block growth and $k(T\sigma) = k(T)$.*

The first assertion can be found in [A, Chapitre 3] for words with domain $\mathbf{N}^+$ over any finite alphabet Σ. The formula for the stiffness in the second assertion is new. Related results on recurrent k-stiff words can be found in Didier [D], Paul [P] and Coven [C]. The proof of Theorem 1 is given in Section 5.

In Theorem 2 we study the finite subwords of k-stiff $\mathbf{Z}$-words. Let $\mathcal{S}_k$ be the collection of $\mathbf{Z}$-words w such that $P(w, n) \leq n + k$ for all n and such that $P(w, n) = n + k$ for at least one n. We define

- $\mathcal{S}_k^{\text{per}} = \{w \in \mathcal{S}_k | w \text{ is periodic}\}$,
- $\mathcal{S}_k^{\text{rnp}} = \{w \in \mathcal{S}_k | w \text{ is recurrent but not periodic}\}$,

- $\mathcal{S}_k^{\mathrm{nr}} = \{w \in \mathcal{S}_k | w \text{ is non-recurrent}\}$.

It is obvious that $\mathcal{S}_k$ is the disjoint union of these three sets. Theorems 2a and 2b deal with the finite factors of these classes.

Theorem 2a. *The classes $\mathcal{S}_k^{\mathrm{per}}$ and $\mathcal{S}_k^{\mathrm{rnp}}$ contain the same n-words and all these n-words are contained in $\mathcal{S}_k^{\mathrm{nr}}$.*

Theorem 2b. *If $w \in \mathcal{S}_k^{\mathrm{nr}}$ has the same period cycle in both directions, then its n-factors are contained in $\mathcal{S}_k^{\mathrm{rnp}}$.*

Theorem 3. *If w is a k-stiff and recurrent $\mathbf{Z}$-word then w is k-balanced. The second k cannot be replaced by $k-1$.*

Remarks. Let $k \geq 1$ and consider the word $w = a^\infty(b^k a)^\infty$. (If x is a finite word then the notation x^∞ will be used to denote the left-infinite word $\cdots xxx$, the right-infinite word $xxx\cdots$ and the bi-infinite word $\cdots xxx \cdots$). If confusion is possible we will state the precise meaning, but this is not the case here. This word is in $\mathcal{S}_k^{\mathrm{nr}}$ (apply Theorem B) and the factor $a^{k+2}b^k ab$ is contained in no element of $\mathcal{S}_k^{\mathrm{rnp}}$ by Theorem 3. This shows that the inclusion in Theorem 2a is strict. At the same time it shows that the recurrency condition in Theorem 3 is necessary: in fact w is not k'-balanced for any k'. Finally we note that the $k=1$ case from Theorems 2 and 3 could be derived directly from the classification of stiff $\mathbf{Z}$-words, as given in Section 3. The proofs of Theorems 2a, 2b and 3 are given in Section 6.

2. Introduction: counting theorems.

First we discuss formulas for $\mathrm{bal}(n)$ and $\mathrm{st}(n)$, the number of balanced and stiff words of length n, respectively. In [D/GB, Conj. 6.4] it was conjectured that

$$\mathrm{bal}(n) = 1 + \sum_{i=1}^{n} (n+1-i)\phi(i).$$

This formula was afterwards proved in a number of ways, see [B/Po], [dL/Mi, Th. 7], [Mi]. We note that many of the authors above count finite factors of Sturmian words, rather than finite balanced words. As we will see in Section 3 this makes no difference. In this paper we prove a similar formula for the number of stiff, non-balanced words of length n.

Theorem 4. $\mathrm{st}(n) - \mathrm{bal}(n) = 4\sum_{2\leq i\leq n/2} (\frac{n+1}{2} - i)\phi(i)$.

With the asymptotic formula $\Phi(n) := \sum_{i=1}^{n} \phi(i) = \frac{3n^2}{\pi^2} + O(n \ln n)$, (see [H/W, Theorem 330]), Mignosi shows that $\mathrm{bal}(n) = \frac{n^3}{\pi^2} + O(n^{2+\delta})$ for any $\delta > 0$. In a similar way it will follow that $\mathrm{st}(n) = \frac{3n^3}{2\pi^2} + O(n^{2+\delta})$ for any $\delta > 0$.

As we have seen in Theorem 1 every $w \in \mathcal{S}_k^{\mathrm{rnp}}$ is of the form $w = T\sigma$ where σ is stiff and recurrent. Of course σ is not periodic, hence it is

Sturmian and $k(T) = k$. With the fact, to be shown in Lemma 10, that there exist only finitely many irreducible substitutions with given $k(T)$, we estimate from above the number of n-words in $\mathcal{S}_k^{\mathrm{rnp}}$. If $\mathcal{S}$ is a collection of words, we denote by $F_n(\mathcal{S})$ the collection of all n-factors of elements of $\mathcal{S}$.

Theorem 5a. *For every $k \in \mathbf{N}^+$ there exists a constant $\mathcal{C}_k$ such that $|F_n(\mathcal{S}_k^{\mathrm{rnp}})| \le \mathcal{C}_k n^3$.*

The next theorem estimates the number of words of length n which appear in a non-recurrent k-stiff $\mathbf{Z}$-word, but not in a recurrent $\mathbf{Z}$-word of stiffness at most k.

Theorem 5b. $|F_n(\mathcal{S}_k^{\mathrm{nr}}) \setminus \cup_{i=1}^{k} F_n(\mathcal{S}_i^{\mathrm{rnp}})| \le 2^k (n+k)^3$.

Corollary. $|F_n(\mathcal{S}_k)| \le \mathcal{D}_k n^3$ *for some* $\mathcal{D}_k$.

Remark. As shown in [H/T, Th. 4] there exist for every $k \ge 2$ finite k-stiff words which are not factors of k-stiff $\mathbf{Z}$-words. Such words are not counted in the above estimates.

The properties stiff and k-stiff do not seem to lie far apart. The situation is very different for balanced and k-balanced since $\mathrm{bal}_k(n)$, the number of k-balanced words of length n, is exponentially large in n when $k \ge 2$. We prove the following theorem.

Theorem 6. *For every $k \ge 2$ there exist positive constants c, d, C_k, D_k with $3^{1/3} \le C_k \le D_k < 2$ such that $c \cdot C_k^n \le \mathrm{bal}_k(n) \le d \cdot D_k^n$ for all n and such that $\lim_{k\to\infty} C_k = 2$.*

The proofs of Theorems 4-6 can be found in Section 7.

3. Prerequisites on balanced and stiff words

A. Every balanced word is stiff. See [C/H], Theorem 3.14, for a short proof. From Theorems 5 and 6 it follows that k-balanced does not imply k-stiff when $k \ge 2$.

B. Every finite k-balanced word is contained in some k-balanced $\mathbf{Z}$-word. The statement remains true when k-balanced is replaced by stiff but not when stiff is replaced by k-stiff, $k \ge 2$. See [H/T].

We now classify all stiff $\mathbf{Z}$-words; references for proofs can be found at the end of this section. Sometimes we describe a $\mathbf{Z}$-word w by the subset $W \subset \mathbf{Z}$, defined by $i \in W \iff w_i = a$.

C. The stiff $\mathbf{Z}$-words, defined modulo shift, are given by:

a) $w = a^\infty, w = b^\infty$ and $W = \{\lfloor \frac{in}{k} \rfloor\}_{i\in\mathbf{Z}}$ where $k, n \in \mathbf{N}^+$ satisfy $(k,n) = 1$ and $\frac{k}{n} \in (0,1)$.

b) $w = a^\infty b a^\infty, w = b^\infty a b^\infty$, $W = \{\lfloor \frac{in}{k} \rfloor\}_{-\infty}^{k-1} \cup \{\lfloor \frac{in-1}{k} \rfloor\}_1^\infty$ and $W = \{\lfloor \frac{in-1}{k} \rfloor\}_{-\infty}^{k-1} \cup \{\lfloor \frac{in}{k} \rfloor\}_1^\infty$, where $k, n \in \mathbf{N}^+$ satisfy $(k,n) = 1$ and $\frac{k}{n} \in (0,1)$.

c) $W=\{\lfloor\zeta i+\phi\rfloor\}, W=\{\lceil\zeta i+\phi\rceil\}$ where $\zeta>1$ is irrational and $0\leq\phi<\zeta$.
d) $w=a^{\infty}b^{\infty}, w=b^{\infty}a^{\infty}$,
$W=\{\lfloor\frac{is}{l}\rfloor\}_{-\infty}^{k+l-1}\cup\{\lfloor\frac{ir-1}{k}\rfloor\}_1^{\infty}$ and $W=\{\lfloor\frac{is-1}{l}\rfloor\}_{-\infty}^{k+l-1}\cup\{\lfloor\frac{ir}{k}\rfloor\}_1^{\infty}$.
Here k,l,r,s are integers with $0\leq l\leq s, 0\leq k\leq r$ and $lr-ks=1$ in the first case and $lr-ks=-1$ in the second case.

Remarks. Let w be a $\mathbf{Z}$-word and $x=(x_n)_{n=1}^{\infty}$ be a sequence of subwords of w where x_n has length n. If $\lim_{n\to\infty}\frac{c(x_n)}{n}$ exists for all x and is independent of x then the common value α is called the *density* of w. It turns out that (from the list above) only elements from **a)**, **b)**, **c)** have a well-defined density and that these classes exactly describe the balanced $\mathbf{Z}$-words. We elaborate a little on each of the classes.
a) describes the periodic balanced $\mathbf{Z}$-words. The periods are $1,1,n$, respectively and the corresponding densities are $1,0,\frac{k}{n}$. This class equals $\mathcal{S}_1^{\mathrm{per}}$.
b) In this case $\alpha=\frac{k}{n}$ and one verifies that w contains a unique n-word x with $c(x)\neq k$. The word w is called *skew*. If $c(x)=k-1$ then w is called skew of min-type and if $c(x)=k+1$ then w is called skew of max-type. The subword x is called its exceptional block. The reader may verify that the exceptional block in the last two cases of **b)** is situated at $[0,n-1]$. The classes **b)** and **d)** together constitute $\mathcal{S}_1^{\mathrm{nr}}$.
c) These $\mathbf{Z}$-words are called irrational Beatty sequences (the rational ones are given in **a)**) or Sturmian words. Given w, the number ζ is unique ($\zeta=\frac{1}{\alpha}$) and w has at most one representation of each type. In the special case that $\phi\in\mathbf{Z}\oplus\zeta\mathbf{Z}$ one speaks of *standard* Sturmian words and only in this case the type of the representation is fixed. The words with $\phi\in\frac{1}{2}+\mathbf{Z}\oplus\zeta\mathbf{Z}$ are known as Bernoulli words. Since Sturmian words have irrational density they are recurrent by Theorem A. The class of Sturmian $\mathbf{Z}$-words equals $\mathcal{S}_1^{\mathrm{rnp}}$.
d) It is easily seen that every word w in **d)** is left periodic with left period s, left density $\frac{l}{s}$ and right periodic with right period r, right density $\frac{k}{r}$. Since the fractions $\frac{l}{s}$ and $\frac{k}{r}$ are irreducible and unequal it follows that s,r are actually the minimal periods and that the period cycles are unequal. For the maximal periodic domains one finds $(-\infty,r+s-2]$ and $[1,\infty)$. Hence $g=r+s-2$ and $\chi-1-g=1$, in accordance with Theorem B. Words from class **d)** are called *infinite Hedlund words* and we write $\mathrm{PER}(s,r,\Delta)$ for the word with period pair (s,r) and $lr-ks=\Delta\in\{\pm1\}$. We call Δ the *signature* of w and the maximal overlap B of the periodic parts of w will be called its associated *finite Hedlund word.* We denote the class of finite Hedlund words by $\mathcal{H}$. Note that $\emptyset\in\mathcal{H}$. In terms of the parameters k,l,r,s one has $|B|=r+s-2, c(B)=k+l-1$ as can be verified directly from the formulas above. We summarize some well-known facts about B in the following two lemmata.

Lemma 1. *Let $w = \mathrm{PER}(s, r, \Delta)$ and B its associated finite Hedlund word. Then* (a) *B is a palindrome (invariant under reversal),* (b) *there is a unique symbol σ such that $\sigma B\overline{\sigma} \subset w$,* (c) *the word $\sigma B\overline{\sigma}$ appears only once in w, namely surrounding the overlap and finally we have* (d) $\sigma = a \iff \Delta = 1$.

Proof. All statements but (d) are contained in [C/H], Theorem 4.12. We note that from the statements $(3)(\rho), (\lambda)$ in that theorem it follows that the B appearing there is indeed our finite Hedlund word. Part (d) of our lemma is obtained by substituting $i = 0, k + l$ in the appropriate formulas. □

Lemma 2. *Let $B \in \mathcal{H}$ and $\Delta \in \{\pm 1\}$. Then there is exactly one infinite Hedlund word with signature Δ inducing B.*

Proof. Suppose first that two Hedlund words with period pairs $(s, r), (s', r')$ and the same Δ induce B. Writing $|B| + 2 = \alpha, c(B) + 1 = \beta$ we have $r + s = r' + s' = \alpha, k + l = k' + l' = \beta, lr - ks = l'r' - k's'$.
Then $l\alpha - s\beta = l(r+s) - s(k+l) = lr - ks = l'r' - k's' = l'\alpha - s'\beta \in \{\pm 1\}$. Hence α, β are coprime and $(l - l')\alpha = (s - s')\beta$. Since $s, s' < \alpha$ we have $s = s', r = r', l = l', k = k'$ and the infinite Hedlund words are indeed equal. Suppose now that $B \in \mathcal{H}$, then some $\mathrm{PER}(s, r, \Delta)$ induces B by definition. The word $\mathrm{PER}(r, s, -\Delta)$ is obtained from the previous one by mirroring and since B is a palindrome it follows that $\mathrm{PER}(r, s, -\Delta)$ also induces B. □

Literature. The classification of balanced words already appeared in [M/H], although their terminology is slightly different from ours. A construction of the infinite Hedlund words, completing the classification of stiff words, is given in [C/H]. A more recent discussion of the formulas above can be found in [T]. There is a vast literature on Sturmian words. It is not hard to show (using Theorems 2a,2b or explicitly with the formulas above) that the three classes of balanced $\mathbf{Z}$-words as described in **a**), **b**), **c**) induce the same collection of finite factors. In particular a finite word is balanced if and only if it is contained in a Sturmian word. For this reason finite balanced words are also known as finite Sturmian words and this is the name most often used. For relations with continued fractions, Christoffel words, Lyndon words and more we refer the reader to [B/dL], [Bo/L], [Br], [dL/Mi], [S]. In [S] a more detailed bibliography can be found.

4. The non-recurrent case

Proof of Theorem A. Assume that the subword w with domain $[1, n]$ does not occur elsewhere in w. Then every subword containing this one does not occur elsewhere in w either. Let $w' = w_2 w_3 \cdots$. For $N \geq n$ there exist $N - n + 1$ intervals of length N containing $[1, n]$. Hence $P(w', N) \leq (N + k) - (N - n + 1) = k + n - 1$. Since $P(w', N)$ is bounded it follows that w' and also w are right periodic. By a symmetry argument w is left periodic. $\square$

Proof of Theorem B. First we deal with $g < 0$. Hence we can write $w = A^\infty B C^\infty$ where $|A| = s, |B| = -g, |C| = r, A$ and C are primitive (not powers of a smaller word),$B \neq \emptyset, A$ and B have different initial symbols and B, C have different terminal symbols. If B starts with σ and ends in τ, then $A\sigma$ does not appear in A^∞ and τC does not appear in C^∞. It follows that $ABC = A\sigma \cdots \tau C$ appears only once in w, say in position $[1, n]$. In particular w is not recurrent.

Now let D be any subinterval of $\mathbf{Z}$ of length N and x the subword of w with domain D. We say that x has property $(*)$ if x has period cycle (A) after deleting the last $|BC| - 1$ symbols and we say that x has property $(**)$ if x has period cycle (C) after deleting the first $|AB| - 1$ symbols. If $D \subset (-\infty, n-1]$ then x has property $(*)$ and if $D \subset [2, \infty)$ then x has property $(**)$. The remaining intervals D are exactly those containing $[1, n]$.

Now suppose that x satisfies $(*)$ and $(**)$ simultaneously where

$$N \geq |A| \cdot |C| + |ABBC| - 2.$$

Performing both deletions we find a word x' with $|x'| \geq |A| \cdot |C|$ such that x' is contained in A^∞ and C^∞. Let x'' be a subword of x' of length $|A| \cdot |C|$. Then $x'' = \tilde{A}^{|C|} = \tilde{C}^{|A|}$ where $\tilde{A} \sim A, \tilde{C} \sim C$. So the $\mathbf{Z}$-words $\tilde{A}^\infty, \tilde{C}^\infty$ are equal, hence $A^\infty = C^\infty$ and by primitivity we have $A \sim C$. Now x is obtained by extending x'' with period $|A| = |C|$ in both directions. Therefore, if $N \geq |A| \cdot |C| + |ABBC| - 2$ then every word x satisfying $(*)$ and $(**)$ is contained in both periodic parts of w.

We note that for N large exactly χ words of length N are contained in some periodic part of w and that exactly $N - n + 1$ subwords of w contain ABC. We will now count the subwords x which do not contain ABC and are not contained in any of the two periodic parts. By the previous paragraph these subwords satisfy exactly one of the conditions $(*)$ and $(**)$. First suppose that x satisfies only $(*)$. Then any domain D of x satisfies $D \subset (-\infty, n-1], D \not\subset (-\infty, |A|]$, and there exist $|BC| - 1$ such intervals of length N. If D is such an interval, then the corresponding word x is not contained in the left periodic part (x contains $A\sigma$), not contained in the right periodic part (then x would satisfy $(*)$,$(**)$ simultaneously, hence be contained in the left periodic part), and does not contain ABC. Moreover, all such D yield different x, since the x's can be distinguished by the first appearance of $A\sigma$. Hence the number of words not contained in a periodic part, not containing ABC and satisfying only $(*)$ equals $|BC| - 1$. Similarly the number of such x satisfying only $(**)$ equals $|AB| - 1$. Hence

$$IP(w, N) = \chi + (N - |ABC| + 1) + |BC| - 1 + |AB| - 1 = N + \chi - 1 - g$$

Now we consider the case $g \geq 0$. Then we can write $w = A^\infty BC^\infty$ where $|A| = s, |B| = g, |C| = r$ where A, C are primitive and where the maximal periodic parts are given by $A^\infty B$ and BC^∞. In analogy with the previous argument we find that ABC appears only once in w and that $P(w, N) = \chi + (N - |ABC| + 1) + |C| - 1 + |A| - 1 = N + \chi - 1 - g$ for N large. □

Remark. A formula for $k(w)$ in Theorem B already appears in [C, Theorem 3.10]. Coven defines, in our terminology, that $\chi = r = s$ if w is asymptotically symmetric and $\chi = r + s$ otherwise. Here a $\mathbf{Z}$-word w is called asymptotically symmetric if there exist i, j such that $w_{i-k} = w_{j+k}$ for all $k \geq 0$. The reader should be aware that the condition of asymptotic symmetry appearing in [C] is *not* equivalent to our condition of equal period cycles on either side as the example $w = (aababb)^\infty(baabab)^\infty$ shows. If Coven's condition is replaced by ours, then the proof in [C] is correct.

Remark. Let w be as in Theorem B. Since w is not periodic we have $P(w, n) > n$ for all n, hence $k \geq 1$. This implies $g \leq \chi - 2$. The extreme cases with $g = \chi - 2$ correspond to the subclasses **b**), **d**) of the classification C, depending on whether or not the left and right period cycles coincide.

5. The recurrent case

Let $T = (X, Y)$ be a substitution. We associate with T a directed graph $G(T)$ each edge of which is labelled with a or b. The graph consists of two directed cycles α, β of lengths $|X|, |Y|$ whose only intersection is a vertex O, the origin. Also, if one follows α from O to itself the labels read X and if one follows β from O to itself the labels read Y. We call $G(T)$ the *representing graph* for T. An *acceptable path* in $G(T)$ is a directed path whose labels form the initial segment of a right-infinite word on $\{X, Y\}$. In other words, an acceptable path is a path whose labels can be obtained starting from O. If an acceptable path has label x, then there is a unique ξ of minimal length such that x is a left-factor of $T\xi$. We call ξ the *coding* for x. It indicates the order in which X, Y appear following the path. The choice is unique because X, Y start with different symbols.

Lemma 3. *Let T be an irreducible substitution and G its representing graph. From every vertex $P \neq O$ there is at most one acceptable path of given length. Moreover, there exist positive integers M, N such that every acceptable path with starting vertex $\neq O$ has a coding with period M after deleting the first N symbols.*

Proof. Let $P \neq O$ and first assume that there is an infinite acceptable path from P. The first symbol of the coding is fixed and following X, Y respectively in $G(T)$ we find a path $P \to Q$. The direction to take at O is completely determined by the next symbol. Suppose $Q = O$ and without loss of generality that the final edge of the path $P \to Q$ lies in α. Apparently the word you followed was Y, since otherwise $P = O$. If $|Y| \leq |X|$ then $X = ZY$ for some Z and if $|X| < |Y|$ then $Y = ZX$. Both are impossible because T was irreducible. Hence $Q \neq O$ and repeating the procedure we find a path $P \to Q \to R \to \cdots$ where each arrow has label X or Y. Since G has only $|XY| - 2$ vertices different from O, the first $|XY| - 1$ elements of the sequence $P, Q, R, \cdots$ cannot be distinct. Hence some vertex appears twice in the sequence and since every point in the sequence determines its successor uniquely it follows that the sequence, hence also the path, is eventually periodic. Let N be the maximal preperiodic part of the paths (taken over all P) and M the least common multiple of all the periods. These M, N will do. If there exist only finite acceptable paths from P then the uniqueness is shown as above and enlarging N sufficiently the second part of the Lemma becomes trivial for all such P. $\square$

We recall that a finite word x is called *primitive* if x is not a power of a strictly smaller word. Every finite word can be written uniquely as power of a primitive word. See [B/P, Proposition 3.1] for a proof.

Lemma 4. *Let T be an irreducible substitution. Then there exist only finitely many finite primitive words x such that Tx is not primitive.*

Proof. Suppose x is primitive and $Tx = \eta^n$ with η primitive and $n \geq 2$. Trace out η in $G(T)$, starting in O and find a path $O \to P$. If $P = O$ then $\eta = T(\xi)$, hence $Tx = T(\xi^n)$ and $x = \xi^n$, a contradiction. Here the injectivity of T is a direct consequence of the fact that X starts with a and Y with b. Thus $P \neq O$ and there is an acceptable path from P with label η^∞. But we know from the proof of Lemma 3 that η is determined up to conjugacy by P and hence the set of possible η is finite. Now suppose that x, y, η are primitive such that Tx, Ty are powers of η. Then we have $Tx = \eta^m, Ty = \eta^n$ for some $m, n \in \mathbf{N}^+$, hence $x^n = y^m$. From the Defect Theorem, (see [B/P, Theorem 2.8]), it follows that either $m = n$ or that x, y are powers of the same word. Hence x, y are powers of the same word and by primitivity we have $x = y$. We conclude that there are only finitely many η and each η yields at most one x. $\square$

Lemma 5. *Let T be an irreducible substitution. There are only finitely many x such that there exist y with $x \not\sim y$ but $Tx \sim Ty$.*

Proof. Suppose x, y are as in the lemma and write $Tx = \eta^m$ with η primitive and $m \geq 1$. Choose an admissible path γ from O with label $(Ty)^\infty$. Since $(Ty)^\infty$ has primitive period cycle η there is a $P \in \gamma$ such that the induced path from P onwards has label η^∞. If $P = O$ then $T(y^\infty) = T(zx^\infty)$ for some z hence $y^\infty = zx^\infty$. It follows that $x^\infty = y^\infty$ (where now these words are understood to be $\mathbf{Z}$-words), hence $x \sim v^k, y \sim v^l$ for some primitive v and $k, l \geq 1$. Then $T(v)^k \sim Tx \sim Ty \sim T(v)^l$ and by comparing lengths we find $k = l$ and $x \sim y$, a contradiction. Hence $P \neq O$, and since the infinite path from P onwards is admissible we find that η is determined up to conjugacy by P. Hence only finitely many η are possible and since every η yields at most one x we are done. □

Let $\mathcal{M}$ be the monoid of all substitutions generated by $(ab, b), (ba, b)$ and (b, a). In the next lemma we will show that every $T \in \mathcal{M}$ leaves the class of infinite Hedlund words h invariant. Next we show that $T\sigma$ has stiffness $k(T)$ when T is a fixed substitution and σ is a "generic" infinite Hedlund word.

Lemma 6. *Let $T \in \mathcal{M}$ where $\mathcal{M}$ is as above and σ a $\mathbf{Z}$-word. Then $T\sigma$ is infinite Hedlund when σ is infinite Hedlund.*

Proof. It suffices to take for T one of the generators above. The theorem is clear for $T = (b, a)$ and we need only consider $T = (ab, b)$ since the case $T = (ba, b)$ follows from this one by symmetry. Let $\sigma = \mathrm{PER}(s, r, \Delta)$ and write $w = T\sigma$. If x is the left period cycle for σ then Tx has length $l + s$ and content l. Similarly, if x is the right period cycle then Tx has length $r + k$ and content k. The fractions $\frac{l}{s+l}$ and $\frac{k}{k+r}$ are irreducible and unequal. It follows that w has left period $s + l$, right period $r + k$ and that the period cycles are unequal. By Lemma 1 the symbols surrounding the maximal overlap of σ are different. Hence in both cases $(\Delta = -1, \Delta = 1)$ we find $g(w) = r + s + k + l - 2$. Theorem B now gives us $k = \chi - 1 - g = r + s + k + l - 1 - (k + l + r + s - 2) = 1$. It follows that $w \in h$. □

Remark. Let $T \in \mathcal{M}$. Because every finite balanced word appears in the periodic part of some infinite Hedlund word it is clear from the previous lemma that $T : \mathrm{Bal} \to \mathrm{Bal}$ where Bal is the collection of all balanced words (finite and infinite). Assume now that σ is a Sturmian $\mathbf{Z}$-word with density α. Then $T\sigma$ is balanced and an easy calculation shows that $T\sigma$ has irrational density when T is one of the three given generators of $\mathcal{M}$. In fact these densities equal $\frac{\alpha}{\alpha+1}, \frac{\alpha}{\alpha+1}$ and $1 - \alpha$ respectively. It follows that $T\sigma$ is Sturmian for these T as well and then the same follows for all $T \in \mathcal{M}$. The fact that every $T \in \mathcal{M}$ maps Sturmian words into Sturmian words is well-known and the converse is also true; see [Mi/S]. For this reason members of $\mathcal{M}$ are sometimes called Sturmian transformations.

Lemma 7. *Let T be a substitution. Then there exists a finite set $V \subset \mathbf{Q}$ such that $P(Tw, n) = n + k(T)$ for n large whenever w is an infinite Hedlund word whose left- and right-density avoid V.*

Proof. If T is a substitution it is easy to see that $T = T^{\mathrm{RED}} \circ \Phi$ where $\Phi \in \mathcal{M}$. Since by Lemma 6 Φ induces an injection from h into itself it follows that we may assume without loss of generality that T is reduced.

Now choose $w = \mathrm{PER}(s, r, \Delta)$ such that the period cycles of length s, r are not in the exceptional sets of Lemmas 4 and 5. We write $|Ta| = \phi, |Tb| = \psi$ and A, B, C as in the definition of $k(T)$. Then Tw has minimal left-period $l\phi + (s-l)\psi$, minimal right-period $k\phi + (r-k)\psi$ and overlap $(k+l-1)\phi + (r+s-k-l-1)\psi + |C|$. The period cycles mentioned are not conjugate and applying Theorem B we have $k(Tw) = \chi - 1 - g = (k+l)\phi + (r+s-k-l)\psi - 1 - (k+l-1)\phi - (r+s-k-l-1)\psi - |C| = \phi + \psi - 1 - |C| = |ABC| + 1 = k(T)$. □

Remark. Let $T = (abba, bb)$ and $w = \mathrm{PER}(2,5,1) = (ab)^\infty(babab)^\infty$. The reader may verify that $k(T) = 5$ and that $k(Tw) = 2$. This shows that the restriction on w cannot be dropped.

Let w be a $\mathbf{Z}$-word and x a finite subword. We say that x has *multiple right extension* (MRE) in w if $xa, xb \subset w$ and we denote the set of these subwords by $\mathrm{MRE}(w)$. We note that some authors express this by saying that x is right-special in w. A similar definition can be given with left instead of right.

Now assume that w has minimal block growth, i.e., $P(w,n) = n + k$ for $n \geq N$. This means that for $n \geq N$ there is a unique word B_n of length n with MRE in w and also a unique word C_n of length n with MLE in w. It is obvious that B_n is a right-factor of B_{n+1} if $n \geq N$, hence there exists a unique left-infinite word B such that B_n equals the last n symbols of B for every $n \geq N$. Of course every right-factor from B has MRE and this allows us to define B_n for every $n \geq 0$. Similar definitions can be given for C and C_n. We note that for stiff words w we have $\mathrm{MRE}(w) = \{B_n\}, \mathrm{MRE}(w) = \{C_n\}$ and that for infinite Hedlund words B, C are equal to the maximal periodic tails. We say that w has a *jump* at $n \geq 0$ if $\mathrm{MRE}(w)$ has more than one element of length n or, equivalently, if $P(w, n+1) - P(w,n) > 1$. Of course $n < N$ for such n. The following lemma has been set apart since it will be used several times in the sequel.

Lemma 8. *Let w be a $\mathbf{Z}$-word with $P(w,n) = n+1$ for all n, T an irreducible substitution, γ the bi-infinite path in $G(T)$ induced by Tw and let $x \in \mathrm{MRE}(Tw)$ be finite. If all paths along γ with label x have the same endpoint, then x is a right-factor of the left-infinite word $T(B(w))$.*

Proof. If all paths in $G(T)$ with label x have the same endpoint, then this endpoint must be O. It follows that we can write $x = yT(z)$ where y is a strict right-factor of X or Y and where $z \subset w$. Here y, z are unique because T is irreducible. We have $z \in \mathrm{MRE}(w)$, hence $z = B_i(w)$ for some i and we have $B_{i+1}(w) = \lambda z$ for some symbol λ. If γ contains different paths with label x then y is a common right-factor of X and Y. In particular y is a right-factor of $T\lambda$ and x is a right-factor of $T(B(w))$, as stated. If there is a unique path along γ with label x, then y must be a right-factor of $T\lambda$ and we are done as well. $\square$

Lemma 9. *Let w be a $\mathbf{Z}$-word with $P(w,n) = n+1$ for all n, T an irreducible substitution, $\mu = \max(|Ta|, |Tb|)$ and M, N as in Lemma 3. If $x \in \mathrm{MRE}(Tw)$ has length τ and x is no right-factor of $T(B(w))$, then w contains an M-periodic word of length $\lceil \frac{\tau}{\mu} - 1 - N \rceil$.*

Proof. It follows from Lemma 8 that x has two representing paths $P_0 \cdots P_\tau$, $Q_0 \cdots Q_\tau$ along γ with $P_\tau \neq Q_\tau$ and, consequently, $P_0 \neq Q_0$. We may assume $\tau \geq \mu$ (otherwise the theorem is trivial), hence both paths pass through O. If neither of these paths starts at O then we delete the initial edge of both of them and we repeat until one of them does. The new length ν of both paths satisfies $\tau - \mu \leq \nu \leq \tau$ and both paths have label $\tilde{x}$, the final ν symbols of x. We denote the new paths, without loss of generality, by $OP'_1 \cdots P'_\nu$ and $Q'_0 \cdots Q'_\nu$. Note that $P'_\nu \neq Q'_\nu$ and $O \neq Q'_0$. Let ξ be the coding of these acceptable paths. From $OP'_1 \cdots P'_\nu \subset \gamma$ we deduce $\xi \subset w$ and from Lemma 3 applied to the acceptable path $Q'_0 \cdots Q'_\nu$ we deduce that ξ has period M after deleting the first N symbols. The result now follows from $|\xi|\mu \geq |T\xi| \geq \nu \geq \tau - \mu$. $\square$

Proof of Theorem 1. (first part) The first part of the theorem is trivial if w is periodic, hence we assume that this is not the case. For every $n \geq 1$ we define a directed graph G_n with vertex-set $\mathcal{B}(w,n)$ and edge-set $\mathcal{B}(w,n+1)$ in such a way that every $x \in \mathcal{B}(w,n+1)$ induces an arrow from its first n symbols to its last n symbols. Then every subword x of w of length $\geq n+1$ induces in a natural way a path in G_n, namely the path which has the successive n-factors of x as its successive vertices. This path has $x - n$ edges. The undirected graph underlying G_n is of course connected and for n large there is one point B_n of outdegree 2 and one point C_n of indegree 2. It follows that we have only three possibilities for the type of G_n for n large.

a) A loop from B_n to itself, a path of positive length from B_n to C_n and a loop from C_n to itself. (Apart from B_n, C_n the different loops and paths have to be disjoint).
b) If $B_n = C_n$ then two distinct loops from B_n to itself.
c) If $B_n \neq C_n$ then two paths from B_n to C_n and a single path from C_n to B_n.

If one is in case a) for some n then one is in case a) for all larger n and it is easy to see that the pathlength B_nC_n in G_n increases by 1 if n increases by 1. After at most one such step we find a point $P \neq B_n, C_n$ on this path and then there is no path in G_n from P to itself, contradicting the fact that w is recurrent. Hence only b) and c) occur. If one is in case c) then it is not hard to show either that the pathlength C_nB_n in G_n decreases by 1 if n increases by 1. See also [A/R, section 1]. It follows that case b) appears for infinitely many n and without loss of generality we take some n above all jumps such that case b) applies. The loops α, β correspond to words X, Y with different initial symbols. We suppose that X starts with a, that Y starts with b and we let $T = (X, Y)$. We define $\chi(a) = \alpha, \chi(b) = \beta$ and extend χ by concatenation to a mapping sending words to paths in G_n. It is immediate that $w = T(\sigma)$ for some $\mathbf{Z}$-word σ, that w induces a bi-infinite path γ in G_n and that $\gamma = \chi(\sigma)$. For all $p \geq 0$ we have a natural correspondence $\mathcal{B}(w, p+n) \leftrightarrow \{$ paths of length p in $\gamma\}$: a word $x = x_1 \cdots x_{n+p}$ corresponds to the path which has $x_i \cdots x_{i+n-1}$ as its i^{th} point. The cardinality of both sets equals $p+n+k$. It follows that for every $p \geq 0$ there is a unique path of length p with two right-extensions in γ. In particular, every two such paths are comparable, i.e., γ_p is a left-extension of γ_q if $p \geq q$. Now suppose that $x, \tilde{x} \in \mathrm{MRE}(\sigma)$. The words $\chi(x), \chi(\tilde{x})$ induce paths from B_n to itself with multiple right extension in γ and since they are comparable the same follows for $x, \tilde{x}$. Since σ is not periodic it follows that $\mathrm{MRE}(\sigma)$ contains exactly one word of length i for each i and hence that $P(\sigma, i) = i+1$ for all i. If σ is not recurrent then σ is left- and right-periodic with a finite overlap. The same is then true for w, but then Theorem B shows that w is not recurrent. Hence σ is recurrent and by the classification in C it is Sturmian.

(second part) Let σ be a Sturmian $\mathbf{Z}$-word. If $x \subset T(\sigma)$ then we can write $yxz = T\xi$ where $\xi \subset \sigma$. Since ξ appears infinitely often in σ the same holds for x in $T\sigma$. Hence $T(\sigma)$ is recurrent. We write $\mu = \max(|Ta|, |Tb|)$ as before. Choose $n \geq 1$ and let $\xi_n \subset \sigma$ be a finite subword such that ξ_n contains all $(n+2\mu)-$factors from σ and let $w = w(n)$ be an infinite Hedlund word containing ξ_n. Then $\mathcal{B}(\sigma, n+2\mu) = \mathcal{B}(w, n+2\mu)$, hence $\mathcal{B}(T\sigma, n) = \mathcal{B}(Tw, n)$ and $P(T\sigma, n) = P(Tw, n)$. For every $p \geq 1$ there exists a constant C_p such that σ contains no C_p-factors of period p. Otherwise, let ξ_n be a p-periodic subword of length n for every n. Then $d(\frac{c(\xi_n)}{n}, \frac{\mathbf{Z}}{p}) \to 0$ and $\frac{c(\xi_n)}{n} \to \alpha(s)$ as $n \to \infty$, a contradiction. It follows that the left and right density of $w = w(n)$ lie outside the set V from Lemma 7 for $n \geq n_1$. Taking $n \geq n_1$ we have $k(Tw) = k(T)$.

Suppose now that $\tau \geq n$ is such that $\mathrm{MRE}(Tw)$ contains more than one word of length τ. Then w contains an M-periodic word of length $\lceil \frac{n}{\mu} - 1 - N \rceil$ by Lemma 9 and since $\lceil \frac{n}{\mu} - 1 - N \rceil \leq n$ the same word appears in σ. Hence $\lceil \frac{n}{\mu} - 1 - N \rceil \leq C_M$ whence $n \leq \mu(N + C_M + 1)$. Taking n larger than this value it follows that Tw has no jumps at places $\tau \geq n$, hence $P(T\sigma, n) = P(Tw, n) = n + k(Tw) = n + k(T)$. □

Remark. Let w be a recurrent $\mathbf{Z}$-word of minimal block growth. It follows from the previous proof that $G_i(w)$ is strongly connected for large i. From this we conclude that $G_i(w)$ is strongly connected for all i.

6. Theorems 2 and 3

Proof of Theorem 2a. First let $w \in \mathcal{S}_k^{\text{per}}$ with primitive period cycle π and choose n maximal such that $P(w,n) = n + k$. Then $P(w, n-1) < P(w,n) = P(w,n+1)$. The graph G_n is a cycle and because $P(w,n-1) < P(w,n)$ it follows that $\mathrm{MRE}(w)$ contains an element A of length $n-1$. Choose a subword τA of w. Then only one of the words Aa, Ab succeeds it in G_n, say $A\tau'$. We let G_n' be the graph which is obtained by adding the edge $\tau A\overline{\tau'}$ to G_n. The two paths from B_n to itself in G_n' induce words which we designate by X, Y. Note that $\pi \in \{X, Y\}$ and without loss of generality we take $X = \pi$. Let σ be a $\mathbf{Z}$-word of stiffness 1 and $T(\sigma) = w'$ where $T = (X, Y)$. Then w' has minimal block growth.

We will show now that w' has no jumps at places $\geq n$, hence has stiffness k. Suppose that $x \in \mathrm{MRE}(w')$ satisfies $|x| \geq n$ and let $\Gamma : \mathbf{Z} \to G_n'$ and $\gamma : \mathbf{Z} \to G(T^{\text{RED}})$ be bi-infinite paths induced by w'. We normalize these bi-infinite paths by demanding that $\Gamma(0) = B_n, \gamma(0) = O$, and that the edge $\Gamma(t)\Gamma(t+1)$ has the same label as the edge $\gamma(t)\gamma(t+1)$ for all $t \in \mathbf{Z}$. For all $P \in G_n'$ there is a corresponding point $Q \in G(T^{\text{RED}})$ such that $\Gamma(t) = P$ implies $\gamma(t) = Q$. It is enough to show this for $P = B_n$ and $P = C_n$. For $P = B_n$ the statement is clearly true with $Q = O$, because X, Y can be written as words in $T^{\text{RED}}(a), T^{\text{RED}}(b)$. In G_n' we have two simple paths from B_n to C_n, exactly one of which has length 1. Denote this path by δ_1, the other path by δ_2, and let σ be the label of δ_1. By induction we have a Q for every $P \in \delta_2 \setminus \{C_n\}$, let P_0 be the vertex of δ_2 preceding C_n. If the subpath of δ_2 from B_n to P_0 has label X_1 and if the unique simple path from C_n to B_n (possibly empty) has label X_2, then $X = X_1\sigma X_2, Y = \sigma X_2$ and $T^{\text{red}} = (X_1, \sigma X_2)$. Therefore the point corresponding to P_0 is O, since X_1 can also be seen as a word in $T^{\text{RED}}(a), T^{\text{RED}}(b)$. From this the statement follows for $P = C_n$, hence for all $P \in G_n'$.

Now suppose that two finite subpaths of γ have label x. The corresponding paths along Γ have the same endpoint P, which corresponds to the last n symbols of x. But then the subpaths along γ both end in Q. We can now apply Lemma 8 with T^{RED} and $\tilde{\sigma}$ to find that $x = B_\tau(w')$. Hence w' has no jumps at any $\tau \geq n$ and w' has stiffness k, as claimed.

If x is a finite factor of w then x is also a factor of $w' = T(\sigma)$ whenever σ is a Sturmian word with a density sufficiently close to 1. Also, x is a factor of w' if σ is an infinite Hedlund word with one of its densities close enough to 1. This shows that $F_n(\mathcal{S}_k^{\text{per}})$ is contained in $F_n(\mathcal{S}_k^{\text{rnp}})$ and in $F_n(\mathcal{S}_k^{\text{nr}})$.

Now for the other part. Let $w \in \mathcal{S}_k^{\mathrm{rnp}}$ and $x \subset w$ a finite factor. We take $n \geq |x|$ above all jumps of w such that $B_n = C_n$, and if the loops in G_n have labels X, Y we set $w' := (XY)^\infty$ as a $\mathbf{Z}$-word and this w' contains x. The only n-factor of w' with MRE is B_n and in exactly one of the two loops we have that aB_{n-1} precedes B_n. These two facts imply that all $(n+1)$-factors of w' have unique right extension. Since $P(w', n) = P(w, n)$ this implies $w' \in \mathcal{S}_k^{\mathrm{per}}$. □

Proof of Theorem 2b. Let $w \in \mathcal{S}_k^{\mathrm{nr}}$ with equal period cycles in both directions and $x \subset w$ a finite factor. Again we take $n \geq |x|$ above all jumps of w. The common period cycle in w induces a cycle in G_n which contains B_n, C_n since B_n is contained in the left periodic part and C_n is contained in the right one. It follows that we are in case b) or c), as described in the proof of Theorem 1. Without loss of generality we assume that we are in case b). If the loops in G_n have label X, Y we set $w' := (XY)^\infty$ as before and this $w' \in \mathcal{S}_k^{\mathrm{per}}$ contains x. □

Proof of Theorem 3. Suppose that w is not k-balanced. We can find in w two subwords A, B with $|A| = |B| = n$ and $c(A) \geq c(B) + k + 1$. We take A, B such that n is minimal. Then $c(A) = c(B) + k + 1$. The graph $G_n(w)$ is strongly connected, as remarked after the proof of Theorem 1. Let α be a directed path from A to B of minimal length in G_n and β a directed path from B to A of minimal length. The path α has length $\geq n$ since otherwise a non-empty right-factor of A would equal a left-factor of B, contradicting the minimality of n. Likewise β has length $\geq n$. We define the *effect* of an arrow $x \to y$ in G_n as $c(y) - c(x)$, which is contained in $\{-1, 0, 1\}$. The effect of a set of edges is simply the sum of the individual effects. We denote the number of edges in $\alpha \setminus \beta$ by s and their total effect by σ. We denote the number of edges in $\beta \setminus \alpha$ by t and their effect by τ. We denote the number of common edges by f and their total effect by ϕ. Then

$$s + t + f \leq n + k + 1.$$

Also $\sigma + \phi = -(k+1), \tau + \phi = k + 1$, hence

$$s + t \geq |\sigma| + |\tau| = |\phi + k + 1| + |k + 1 - \phi| = 2\max(k+1, |\phi|) \geq 2(k+1).$$

Combining these inequalities with $s + f, t + f \geq n$ we find that equality holds everywhere. This implies that

- the path $\alpha\beta$ is induced by the word ABA;
- the number of $(n+1)-$words appearing in $\alpha\beta$ equals $n + k + 1$ and hence that these words form $\mathcal{B}(w, n+1)$;
- all edges in $\alpha \setminus \beta$ are of the form $a * b$ with $*$ a word of length $n - 1$;
- all edges in $\beta \setminus \alpha$ are of the form $b * a$.

Now suppose that an edge from α also appears in β. Let $P \to Q$ be the first such edge (following α). Then $P \neq A$ because β finishes when arriving in A. The arrow to P in α is of the form $a * b$. The arrow to P in β is of the form $b * a$. This is a contradiction, because the last symbol of P cannot be a and b at the same time. Hence $f = 0 = \phi, s = t = n$. But then all edges in α are of the form $a * b$, hence $A = a^n, B = b^n$. From $c(A) = c(B) + k + 1$ we read off $n = k+1$, but then $P(s, n) = |\{\text{vertices in } ABA\}| = 2n > n+k$, contradicting that w is k-stiff.

Now let σ be an arbitrary Sturmian word containing b^k and T the substitution with $T(a) = a^k, T(b) = b$. Then $T\sigma$ is k-stiff by Theorem 1, it is recurrent and obviously not $(k-1)-$balanced since $a^k, b^k \subset T\sigma$. Therefore the second k is indeed sharp. □

Corollary. *If σ is recurrent and stiff, then $T\sigma$ is $k(T)$-balanced.*

Proof. Indeed, every factor of σ is contained in a Sturmian word σ' by Theorem 2a. Now $T\sigma'$ is recurrent and by Theorem 1 it is $k(T)$-stiff. Applying Theorem 3 we find that $T\sigma'$ is $k(T)$-balanced. □

7. The counting theorems

We now turn to Theorem 4. Let x be a finite stiff word which is not balanced. By Section 3, part B, x is contained in a stiff $\mathbf{Z}$-word w. This w is not balanced, hence it is an infinite Hedlund word. In Lemma 10 we identify the finite subwords of a given w which are not balanced and we show that each of these subwords determines w completely. From this our formula will follow quite easily.

Lemma 10. *Let $w = \mathrm{PER}(s, r, \Delta)$ be normalised such that its finite Hedlund block B is situated at $[1, n]$. Let $S = [-s, n + r + 1]$. Then a finite subword $x \subset w$ is not balanced if and only if it has a domain containing S. Also, every non-balanced subword determines w completely.*

Proof. Suppose that $x \subset w$ is finite and not balanced. A standard argument, as in [C/H, Lemma 3.06], shows that there exists a word X such that $aXa, bXb \subset x$. Since X has MRE and MLE in w it follows that $X = B_m = C_m$ where $m = |X|$. Hence the word X has periods s and r simultaneously and it is well-known, see for instance [T, Section 4] or Lemma 12 in this article, that this implies $|X| \leq r + s - 2 = |B|$. Suppose that $|X| < |B|$. Since B has MRE and MLE too, it follows that $B = PX = XQ$ with non-empty P and Q. Let τ be the final symbol of P. Using the fact that B is a palindrome, see Lemma 1, we find that Q begins with τ. Then $aB, bB \subset w$ implies $aX\tau, bX\tau \subset w$ hence $X\tau = C_{m+1}$. Since $B\overline{\tau} \subset w$ we have $\tau X\overline{\tau} \subset w$ and together with $\overline{\tau}X\overline{\tau} \subset w$ this implies $X\overline{\tau} = C_{m+1}$. The contradiction implies $|X| = |B|, X = B$ and $aBa, bBb \subset x$. This shows that B is determined by x as the only solution X of $aXa, bXb \subset x$. The word $\sigma B\overline{\sigma}$ from Lemma 1 appears in x, for otherwise x would be contained in a periodic part of w and be balanced.

From Lemma 2 we deduce that w is unique, which is the second part of the lemma.

If we look at s places to the left of the σB in $\sigma B\overline{\sigma} \subset w$ we find another σB, now followed by σ. Furthermore, there is no σB in between because s is the minimal left period and $|\sigma B| \geq s$. This implies that the right-most copy of $\sigma B\sigma$ appears at s places to the left of $\sigma B\overline{\sigma}$, hence at $[-s, r-1]$. Similarly, the left-most copy of $\overline{\sigma}B\overline{\sigma}$ is at $[r, n+1+r]$. The first part of the lemma now follows. □

Proof of Theorem 4. Note that in the previous lemma $|S| = 2(r+s)$. We want to count the stiff words x of length n, which are not balanced. To construct such an x, one chooses coprime positive integers s, r with $2(r+s) \leq n$, one chooses $\Delta \in \{\pm 1\}$ and then x can be any subword of $w := \mathrm{PER}(s, r, \Delta)$ containing S. It follows from Lemma 10 that no different choices for (s, r, Δ) yield the same x and writing $i = r + s$ we have

$$\mathrm{st}(n) - \mathrm{bal}(n) = 2 \sum_{2 \leq i \leq n/2} (n+1-2i)\phi(i) = 4 \sum_{2 \leq i \leq n/2} (\frac{n+1}{2} - i)\phi(i) \quad \square$$

Remark. Some elementary partial summation shows that the previous formula can be rewritten as

$$\mathrm{st}(n) - \mathrm{bal}(n) = 4 \sum_{i=1}^{m} \Phi(i) - 2(n-1) + 2(n-1-2m)\Phi(m)$$

where $m = \lfloor \frac{n}{2} \rfloor$. Together with the asymptotic formulas for $\Phi(n)$ and $\mathrm{bal}(n)$ as given in Section 2 this yields the asymptotic formula for $\mathrm{st}(n)$.

Lemma 11. *For $k \geq 1$ there exist exactly $\nu := (k^2 + k + 2)2^{k-3}$ irreducible substitutions with $k(T) = k$.*

Proof. Write $Ta = A\sigma C, Tb = B\overline{\sigma}C$ as before. Then $|ABC| = k - 1$. It immediately follows that there exist only finitely many T and for the explicit calculation we distinguish between four cases. If $A, B = \emptyset$ then $Ta = aC, Tb = bC$ where $|C| = k-1$ and we have 2^{k-1} choices. If $A = \emptyset \neq B$ then $Ta = aC, Tb = b\tilde{B}bC$ where $|\tilde{B}C| = k-2$ and we have $2^{k-2}(k-1)$ choices. The case $B = \emptyset \neq A$ is similar and gives the same number. If $A, B \neq \emptyset$ then $Ta = a\tilde{A}\sigma C, Tb = b\tilde{B}\overline{\sigma}C$ where $|\tilde{A}\tilde{B}C| = k-3$ and we have $2^{k-2}\binom{k-1}{2}$ choices. The total number then becomes $2^{k-1} + 2^{k-1}(k-1) + 2^{k-2}\binom{k-1}{2} = (k^2+k+2)2^{k-3}$. □

Proof of Theorem 5a. Let $w \in \mathcal{S}_k^{\mathrm{rnp}}$ and $x \in F_n(w)$. We know by Theorem 1 that $w = T(\sigma)$ where σ is Sturmian, T is irreducible and $k(T) = k$. We denote the set of these T by $\{T_1, \cdots, T_\nu\}$, where ν is defined in

Lemma 11. There exist y, z, ξ such that $xy = zT(\xi)$ where z is a strict right-factor of X or Y and $\xi \subset \sigma$ (possibly empty). Then $x = [zT(\xi)]_n$ where $[\cdots]_n$ stands for taking the first n symbols only. We will write Z for the set of possible z and define $\phi : Z \times \{T_i\} \times \mathrm{Bal}(n) \to \{a,b\}^n$ by $\phi(z,T,\xi) = [zT(\xi)]_n$. The image contains all words in $F_n(\mathcal{S}_k^{\mathrm{rnp}})$. Therefore $|F_n(\mathcal{S}_k^{\mathrm{rnp}})| \leq \nu|Z|\mathrm{bal}(n)$ and the result now follows from the result of Mignosi mentioned after Theorem 4. $\square$

Example. If $k = 2$ then $T \in \{(aa,b),(a,bb),(ab,bb),(aa,ba)\}, Z = \{\emptyset, a, b\}$ and $|F_n(\mathcal{S}_2^{\mathrm{rec}})| \leq 12\mathrm{bal}(n)$.

Lemma 12. *Let $r, s \in \mathbf{N}^+, d = (s,r), \phi = r + s - 2$. The total number of $\mathbf{Z}$-words w with left-period s, right-period r (not per se minimal) and exact overlap g equals $\max(2^d, 2^{\phi-g})$ if $g \leq \phi + 1 - d$ and 0 otherwise.*

Proof. We first deal with $g < 0$. Then $w = A^\infty BC^\infty$ with $|A| = s, |B| = -g, |C| = r$, A, B have different initial symbols and B, C have different terminal symbols. Choose the symbols in B arbitrarily. This yields only 2 letter restrictions on A, C hence we find $2^{|B|+(r-1)+(s-1)} = 2^{\phi-g}$ for the number of possible w. Note that $\phi - g \geq d$.

Now for $g \geq 0$, we first consider the case $d = 1$. The proof is basically a generalisation of Tijdeman's proof in [T, Section 4] when $d = 1$ and $g = \phi$. We will use Tijdeman's result that if $x_0 = 0$ and x_n is inductively defined by

$$x_{n+1} = \begin{cases} x_n + s \text{ if } x_n < r \\ x_n - r \text{ if } x_n \geq r \end{cases},$$

then $\{x_0, \cdots, x_{\phi+1}\} = \{0, \cdots, \phi+1\}$ and also $x_{\phi+1} = r$.

Now first suppose that $g > \phi$ and without loss of generality that $[1, \phi+1]$ lies in the intersection of the periodic parts. If we define $(x_n)_0^\infty$ as above and write $\sigma_n = w_{x_n}$, then $\sigma_n = \sigma_{n+1}$ for $n \leq \phi$, hence w is constant on $[1, \phi+1]$. Since $\phi+1 = r+s-1 \geq r, s$ we see that the $\mathbf{Z}$-word w is constant, a contradiction.

Now let $g \leq \phi$ and assume, without loss of generality, that the intersection of the periodic parts of w is situated at $[1, g]$. At first we have to fill in $(\sigma_n)_0^{\phi+1}$ subject to the condition that $\sigma_n = \sigma_{n+1}$ when $x_{n+1} \leq x_n$ or $x_{n+1} \leq g$. Then $\sigma_n \neq \sigma_{n+1}$ is only possible when $x_{n+1} > x_n, g$ hence if

$$x_n \in [\max(g - s + 1, 0), r - 1].$$

Note that this interval is not empty because $g \leq \phi$. This yields $1 + |[\max(g - s + 1, 0), r - 1]| = r + 1 - \max(0, g - s + 1)$ choices. Next we have to fill in the $\max(0, s - (g+1))$ symbols to the left of the 0-position to complete the left period cycle. The total number of choices is then $r + 1 - \max(0, g - s + 1) + \max(0, s - g - 1) = r + 1 + s - g - 1 = \phi + 2 - g$, which is apparently the number of $\mathbf{Z}$-words with left-period s, right-period r (we will abbreviate this as s-r $\mathbf{Z}$-words) with $[1,g]$ in the overlap. The number of s-r $\mathbf{Z}$-words with $[1,g]$ in the overlap but not $[1,g+1]$ then equals

$$\begin{cases} 2^{\phi+2-g} - 2^{\phi+1-g} = 2^{\phi+1-g}, & \text{if } g < \phi; \\ 2^{\phi+2-g} - 2 = 2, & \text{if } g = \phi. \end{cases}$$

The second formula follows from the fact just proved that w is constant if $g > \phi$. Hence the number of s-r $\mathbf{Z}$-words with exact overlap $[1,g]$ equals

$$\begin{cases} 2^{\phi+1-g} - 2^{\phi-g} = 2^{\phi-g}, & \text{if } g < \phi; \\ 2, & \text{if } g = \phi. \end{cases}$$

This proves our theorem when $(r,s) = 1$.

Now for the case $(r,s) = d > 1$. We write $r = d\rho$, $s = d\sigma$ and $t = \rho+\sigma-2$. If an s-r $\mathbf{Z}$-word has finite overlap then the overlap contains at most t elements from a certain residue-class λ mod d, for otherwise the $\mathbf{Z}$-word would be constant on every residue class modulo d, and this would imply periodicity. Hence $g \leq t+(t+1)(d-1) = td+d-1 = r+s-d-1 = \phi+1-d$. Now suppose that an s-r $\mathbf{Z}$-word has finite overlap containing $[1,g]$ where $g \leq \phi + 1 - d$ and write $c(\lambda) = |(\lambda \bmod d) \cap [1,g]|$ for $\lambda \in \mathbf{Z}$. The number of choices is then $\sum_{\lambda=1}^{d} (t + 2 - c(\lambda)) = d(t+2) - g = \phi + 2 - g$ and the rest is similar to the discussion when $d = 1$. □

Proof of Theorem 5b. Let $x \in F_n(\mathcal{S}_k^{\mathrm{nr}}) \setminus \cup_{i=1}^{k} F_n(\mathcal{S}_i^{\mathrm{rnp}})$ and w a non-recurrent $\mathbf{Z}$-word of stiffness k containing x. Write s,r as usual for the minimal periods of w. Without loss of generality we write $(-\infty, g], [1,\infty)$ for its maximal periodic domains. By Theorem 2b the period cycles are not conjugate. If $[\lambda, \lambda+n-1]$ is a domain for x in w, then $g+2-n \leq \lambda \leq 0$, for x is contained in no periodic part of w. Hence for given w it follows that the number of possible x is at most $n-1-g = n+k-\chi \leq n+k$. From the above we also conclude $g \leq n-2$, hence $r,s \leq r+s = \chi = k+1+g \leq n+k$. For fixed s,r the number g is determined by $g = r+s-1-k$ and we have at most $\max(2^d, 2^{\phi-g})$ possiblities for w by Lemma 12. Now $g \leq \phi+1-d$, again by Lemma 12, and substituting $g = \phi+1-k$ this gives $d \leq k$. Also $\phi-g = k-1$ and therefore each choice s,r yields at most 2^k possible w. Combining all previous inequalities we have $|F_n(\mathcal{S}_k^{\mathrm{nr}}) \setminus \cup_{i=1}^{k} F_n(\mathcal{S}_i^{\mathrm{rnp}})| \leq 2^k(n+k)^3$. □

Lemma 13. *Let N,k,ϕ be positive integers with $2\phi \leq k, N - \phi \leq k$ and $\mathcal{A} = \{w |\, |w| = N, c(w) = \phi\}$. Then every element of $\mathcal{A}^\infty$ is k-balanced.*

Proof. For any $w \in \mathcal{A}^\infty$ there is a partition of $\mathbf{Z}$ into intervals $I_i = p + [Ni, Ni+N-1]$ such that the word with domain I_i is contained in $\mathcal{A}$ for all i. Now let us assume that $w \in \mathcal{A}^\infty$ is not k-balanced. Choose subwords $A, B \subset w$ of equal length n such that $|c(A) - c(B)| > k$ and choose domains $D(A), D(B)$ for A and B. We can write $D(A) = PQR, D(B) = STU$ where Q,T are unions of I_i's and where P,R,S,U are strict subintervals of some

I_i. We will identify P, Q, R, S, T, U with the subwords they induce and we write $|Q| = \lambda N, |T| = \mu N$ where $\lambda, \mu \geq 0$. Comparing lengths and contents we find

$$\begin{cases} |PR| + \lambda N = |SU| + \mu N \\ |(c(PR) + \lambda\phi) - (c(SU) + \mu\phi)| > k \end{cases}$$

We have $|\lambda - \mu|N = ||PR| - |SU|| \leq \max(|PR|, |SU|) < 2N$ hence $|\lambda - \mu| \leq 1$. If $\lambda = \mu$ then the second inequality would imply $k < |c(PR) - c(SU)| \leq \max(c(PR), c(SU)) \leq 2\phi \leq k$, a contradiction. It follows that it is safe to assume that $\lambda = \mu + 1$, interchanging A and B if necessary. Counting the number of b's in SU we find $|SU| - c(SU) \leq 2(N - \phi)$. Hence

$$|PR| \geq c(PR) > k - \phi + c(SU) \geq |SU| + k + \phi - 2N = |PR| + k + \phi - N \geq |PR|$$

This contradiction completes the proof. □

Proof of Theorem 6. Let $k \geq 2$ and choose $\phi, N \in \mathbf{N}^+$ such that $2\phi, N - \phi \leq k$. For instance $\phi = \lfloor \frac{k}{2} \rfloor, N = \phi + k$ will do. Now $\mathrm{bal}_k(i)$ is increasing in i (this follows from Section 3, part B) and applying the previous lemma we find

$$\mathrm{bal}_k(n) \geq \mathrm{bal}_k(N\lfloor \frac{n}{N} \rfloor) \geq \binom{N}{\phi}^{\lfloor n/N \rfloor} \geq \binom{N}{\phi}^{-1} \binom{N}{\phi}^{n/N} =: c \cdot C_k^n$$

In particular we can take $C_2 = 3^{\frac{1}{3}}$ and because $\mathrm{Bal}_2(n) \subset \mathrm{Bal}_k(n)$ for $k \geq 2$ we can take all $C_k \geq 3^{\frac{1}{3}}$. The choice $\phi = \lfloor \frac{k}{2} \rfloor, N = k$ gives $C_k = \binom{k}{[k/2]}^{1/k}$. The inequalities $2^{2\kappa} \leq (2\kappa + 1)\binom{2\kappa}{\kappa}$ and $2^{2\kappa+1} \leq (2\kappa + 2)\binom{2\kappa+1}{\kappa}$ then show that $C_k \geq \frac{2}{\sqrt[k]{k+1}}$, hence $\lim_{k\to\infty} C_k = 2$. Now for the upper bound, let $N = 2k + 2$. Then $\mathrm{bal}_k(N) =: \lambda < 2^N$, hence

$$\mathrm{bal}_k(n) \leq \mathrm{bal}_k(N\lceil \frac{n}{N} \rceil) \leq \lambda^{\lceil n/N \rceil} \leq \lambda \cdot (\lambda^{1/N})^n =: d \cdot D_k^n$$

with $D_k = \lambda^{1/N} < 2$. □

Acknowledgements. I would like to thank my promotor, Professor Tijdeman, for suggesting some of the problems that I deal with in this paper.

References

[A] P. ALESSANDRI, *Codages de rotations et de basses complexités.* Thèse de Doctorat, Université d'Aix-Marseille II, 1996.

[A/R] P. ARNOUX, G. RAUZY, *Réprésentation géométrique des suites de complexité* $2n+1$. Bull. Soc. Math. France **119** (1991), 199–215.

[B/dL] J. BERSTEL, A. DE LUCA, *Sturmian words, Lyndon words and trees.* Theoret. Comput. Sc. **178** (1993), 171–203.

[B/P] J. BERSTEL, D. PERRIN, *Theory of Codes.* Academic Press, 1985.

[B/Po] J. BERSTEL, M. POCCHIOLA, *A geometric proof of the enumeration formula for Sturmian words.* J. Algebra Comput. **3** (1993), 349–355.

[Bo/L] J.-P. BOREL, F. LAUBIE, *Quelques mots sur la droite projective réelle.* J. Théor. Nombres Bordeaux **5** (1993), 23–51.

[Br] T. C. BROWN, *Descriptions of the characteristic sequence of an irrational.* Canad. Math. Bull. **36** (1993), 15–21.

[C] E. M. COVEN, *Sequences with minimal block growth II.* Math. Systems Th. **8** (1975), 376–382.

[C/H] E. M. COVEN, G. A. HEDLUND, *Sequences with minimal block growth.* Math. Systems Th. **7** (1971), 138–153.

[D] G. DIDIER, *Caractérisation des $N-$écritures et application à l'étude des suites de complexité $n + c^{ste}$.* Theoret. Comput. Sc. **215** (1999), 31–49.

[D/GB] S. DULUCQ, D. GOUYOU-BEAUCHAMPS, *Sur les facteurs des suites de Sturm.* Theoret. Comput. Sc. **71** (1990), 381–400.

[H/W] G. H. HARDY, E. M. WRIGHT, *An Introduction to the Theory of Numbers.* Oxford Univ. Press, Third Edition, 1954.

[H/T] A. HEINIS, R. TIJDEMAN, *Extending balanced and stiff words.* Report no. W97-15, University of Leiden, 1997.

[dL/Mi] A. DE LUCA, F. MIGNOSI, *Some combinatorial properties of Sturmian words.* Theoret. Comput. Sc. **136** (1994), 361–385.

[H/H] M. MORSE, G. A. HEDLUND, *Symbolic dynamics II - Sturmian trajectories.* Amer. J. Math. **62** (1940), 1–42.

[Mi] F. MIGNOSI, *On the number of factors of Sturmian words.* Theoret. Comput. Sc. **82** (1991), 71–84.

[Mi/S] F. MIGNOSI, P. SÉÉBOLD, *Morphismes sturmiens et règles de Rauzy.* J. Théor. Nombres Bordeaux **5** (1993), 211–233.

[P] M. E. PAUL, *Minimal symbolic flows having minimal block growth.* Math. Systems Th. **8** (1975), 309–315.

[S] K. B. STOLARSKY, *Beatty sequences, continued fractions and certain shift operators.* Canad. Math. Bull. **19** (1976), 473–482.

[T] R. TIJDEMAN, *Intertwinings of periodic sequences.* Indag. Math. **9** (1998), 113–122.

Alex HEINIS
University of Leiden
The Netherlands
E-mail : `heinis@math.leidenuniv.nl`