

ON THE GROWTH RATES OF COMPLEXITY OF THRESHOLD LANGUAGES

ARSENY M. SHUR¹ AND IRINA A. GORBUNOVA¹

Abstract. Threshold languages, which are the $(k/(k-1))^+$ -free languages over k -letter alphabets with $k \geq 5$, are the minimal infinite power-free languages according to Dejean's conjecture, which is now proved for all alphabets. We study the growth properties of these languages. On the base of obtained structural properties and computer-assisted studies we conjecture that the growth rate of complexity of the threshold language over k letters tends to a constant $\hat{\alpha} \approx 1.242$ as k tends to infinity.

Mathematics Subject Classification. 68Q70, 68R15.

INTRODUCTION

The study of words and languages avoiding repetitions is one of the central topics in combinatorics of words since the pioneering work of Thue [18]. A repetition is called *avoidable* on a given alphabet, if there exists an infinite word over this alphabet (or, equivalently, an infinite set of finite words) without this repetition. Thue proved that squares are avoidable on the ternary alphabet, while cubes and overlaps are avoidable already over two letters. Integral powers, which are certainly the simplest repetitions, can be generalized in several ways. Among such generalizations we mention patterns, Abelian powers, relational powers, and, of course, *fractional powers*, which are expressed numerically by *exponents*. An exponent of a word is the ratio between its length and its minimal period. If $\beta > 1$ is a rational number, then a word is called β -free (β^+ -free) if all its factors have exponents less than β (respectively, at most β).

The most natural and challenging problem on the avoidability of exponents is to determine the *repetitive threshold*, which is the exact border between avoidable and unavoidable exponents for any given finite alphabet. For the binary alphabet,

Keywords and phrases. Power-free languages, Dejean's conjecture, threshold languages, combinatorial complexity, growth rate.

¹ Ural State University, Ekaterinburg, Russia; Arseny.Shur@usu.ru

this border is known from Thue's works: the exponent 2 is unavoidable, while there exist infinitely many 2^+ -free (= overlap-free) words. For three letters, this border is between $7/4$ and $(7/4)^+$, as was shown by Dejean [7]. Dejean also conjectured that $7/5$ gives the threshold for a 4-letter alphabet, and $k/(k-1)$ gives the threshold for any k -letter alphabet with $k \geq 5$. By the efforts of different authors, see [2, 10–12], this long-standing conjecture was proved except for the gap $15 \leq k \leq 32$. The paper of Carpi [2] actually left no reason to doubt that Dejean's conjecture is true for all alphabets. Quite recently, the remaining gap was fully removed through extensive computations. Only preliminary versions of the papers with this final result are available now, see [5, 6], and also [13] with a different approach.

We use the term *threshold language* for the minimal by inclusion infinite power-free language over a given alphabet. By Dejean's conjecture, the $(k/(k-1))^+$ -free languages over k -letter alphabets with $k \geq 5$ are threshold languages.

The properties of the minimal language avoiding a given repetition are certainly interesting. For any *minimal* language $L \subseteq \Sigma^*$ its growth properties surely need to be studied. Such properties are represented by the *combinatorial complexity* (or *counting function*) $C_L(n) = |L \cap \Sigma^n|$ and the *growth rate* $\alpha(L) = \limsup_{n \rightarrow \infty} (C_L(n))^{1/n}$. Note that the growth properties of repetition-free languages are intensively studied, starting with the paper of Brandenburg [1].

In the case of threshold languages there is one more specific point of interest. We mean the asymptotic behaviour of properties when the size of the alphabet tends to infinity. In this paper we try to conjecture this asymptotic behaviour of the growth rate of threshold languages. The conjecture is made on the base of extensive study, both purely theoretic and computer-assisted, of different particular cases. The computer-assisted part of studies is based on a fast algorithm determining the growth rate of a regular language. This algorithm is due to one of the authors and is shortly described in [15].

After necessary preliminaries, we introduce a convenient two-dimensional representation of words, which is used for elements of threshold languages and forbidden repetitions. Using this representation, we clarify the structure of "short" forbidden repetitions. These structural results (Thms. 3.1, 3.4, 3.5, 3.6) allows us to obtain nontrivial upper bounds for the growth rates of threshold languages up to the 60-letter alphabet and to compare the contribution to the growth rate of different "short" forbidden repetitions. Summarizing these results and computer-assisted results on "longer" repetitions for relatively small alphabets (5 to 10 letters), we formulate the conjecture mentioned in the abstract.

1. PRELIMINARIES

1.1. WORDS AND LANGUAGES

We recall only necessary notions. See [3, 9] for more background.

An *alphabet* is a nonempty finite set, the elements of which are called *letters*. *Words* are finite sequences of letters. The length of the word w is denoted by $|w|$,

and the positions of letters in w are indexed by the numbers $1, \dots, |w|$. We also consider Z -words (or double-infinite words), in which the positions are indexed by the set of all integers. The *distance* between two occurrences of letters in a finite or infinite word is the (positive) difference between the numbers of their positions. A word u is a *factor* (respectively *prefix*, *suffix*) of the word w if w can be represented as $\bar{v}u\hat{v}$ (respectively $u\hat{v}$, $\bar{v}u$) for some (possibly empty) words \bar{v} and \hat{v} . A *factor* (respectively *prefix*, *suffix*) of w is called *proper*, if it does not coincide with w .

As usual, we write Σ^* ($\Sigma^{\mathbb{Z}}$) for the set of all words (respectively, all Z -words) over a fixed alphabet Σ . The subsets of Σ^* are called *languages* (over Σ). A language is *factorial*, if it is closed under taking factors of its words, and *antifactorial*, if any of its words is not a factor of any other one. A word w is *forbidden* for the language L if it is not a factor of any element of L (alternatively, one can say that words of L *avoid* w). The set of all minimal (with respect to the factorization order) forbidden words for L is called the *antidictionary* of L . A factorial language is regular if and only if its antidictionary is regular (in particular, finite). An antidictionary is always an antifactorial language. Moreover, any antifactorial language is the antidictionary of some factorial language.

A word $w \in \Sigma^*$ of length n can be viewed as a function $\{1, \dots, n\} \rightarrow \Sigma$. Then a *period* of w is any period of this function. The *exponent* of w is the ratio between its length and its minimal period; if this ratio is greater than 1, then w is a *fractional power*. If $\beta > 1$ is a rational number, then w is called β -free (β^+ -free) if all its factors have exponents less than β (respectively, at most β). By β -free (β^+ -free) languages we mean the languages of *all* β -free (respectively β^+ -free) words over a given alphabet. These languages are obviously factorial and are also called *power-free* languages.

In this paper we study β^+ -free words and languages over the alphabets Σ_k of size $k \geq 5$. Following Dejean's conjecture [7], the minimal infinite power-free (or *threshold*) languages over these alphabets are the $(k/(k-1))^+$ -free languages. The threshold language over Σ_k is denoted below by T_k . In the theoretic studies of this paper we assume that the number k is fixed.

We denote the antidictionary of T_k by A_k . A word $u \in A_k$ can be factorized as $u = yzy$, where $|yz|$ is the minimal period of u , $|u|/|yz| > k/(k-1)$, and all proper factors of u have the exponent at most $k/(k-1)$. If $|y| = m$, we call such a word an *m-repetition*. The finite set $A_k^{(m)} \subset A_k$ consists of all r -repetitions with $r \leq m$. The notation $T_k^{(m)}$ is used for the (regular) language with the antidictionary $A_k^{(m)}$. Then, $T_k \subseteq T_k^{(m)}$. Since an infinite regular language contains arbitrary powers of some word, one has $T_k \subset T_k^{(m)}$.

1.2. GROWTH RATES. REDUCTION TO EXTENDABLE LANGUAGES

The *combinatorial complexity* of a language L is a function $C_L(n)$ which returns the number of words in L of length n . This function serves as a natural quantitative measure of L . For factorial languages, the combinatorial complexity is either bounded by a constant or strictly increasing [8]. Increasing complexity functions

can grow exponentially “fast” or subexponentially “slow”. The behaviour of a “fast” complexity can be described by the *growth rate* $\alpha(L) = \lim_{n \rightarrow \infty} (C_L(n))^{1/n}$ (the value $\alpha(L) = 1$ indicates a “slow” complexity function). Note that one should replace \lim by \limsup to extend the definition of growth rate to arbitrary languages.

The growth rate of $T_k^{(m)}$ approximates the growth rate of T_k from above. Moreover, it is easy to prove that $\lim_{m \rightarrow \infty} \alpha(T_k^{(m)}) = \alpha(T_k)$ (see, *e.g.*, [14]). As we will see, this approximation can be good enough even if m is small.

It is well known that the growth rate of a regular language L can be calculated using the graph of a deterministic finite automaton recognizing L . More precisely, this rate is a root of the characteristic polynomial of the adjacency matrix of this automaton. Hence, in the general case the growth rate can not be found exactly, but can be approximated with any prescribed absolute error (the admissible error is considered as a part of input). An efficient practical algorithm for the calculation of such growth rates is described in [15,16]. Its time complexity is $\Theta(n \log(1/\delta))$, where n is the number of states of the automaton and δ is the admissible error of the result. We made an extensive use of this algorithm to obtain the growth rates of the languages $T_k^{(m)}$ for different values of k and m (see Tabs. 1, 2). An important note is that if a regular language is given by a finite antidictionary, then the recognizing automaton can be efficiently constructed by the algorithm of [4].

A word $w \in L$ is said to be (*two-sided*) *extendable* in the language L , if there exist arbitrarily long words u, v such that $u w v \in L$. By $e(L)$ we denote the set of all extendable words of the language L . We use the following result of [17].

Theorem 1.1. *For any factorial language L , $\alpha(e(L)) = \alpha(L)$.*

The antidictionary of $e(L)$ often admits more compact and handy representation than the antidictionary of L . So, in this paper we describe extendable words in the languages $T_k^{(m)}$.

1.3. PANSIOT WORDS AND BINARY ENCODING

In [12], Pansiot showed how to encode the words of threshold languages over any alphabet with “characteristic” words over the binary alphabet $B = \{0, 1\}$. This idea proved very fruitful and was used in the papers [2,5,6,10,11,13]. Below we describe Pansiot’s construction.

Any $(k/(k-1))^+$ -free word over Σ_k avoids 1- and 2-repetitions, and hence satisfies two local conditions:

- (1) any $k-1$ consecutive letters are different;
- (2) two closest occurrences of a letter are followed by different letters.

We define a *Pansiot word* to be any word or Z-word satisfying (1), (2). Thus, Pansiot words are exactly the elements of $T_k^{(2)}$.

Observation 1.2. By (1) and (2), the distance between two closest occurrences of a letter in a Pansiot word is $k-1$, k , or $k+1$.

Suppose that the letter a in a Pansiot word is preceded by the factor $a_1 \dots a_{k-1}$, and $a_k \notin \{a_1, \dots, a_{k-1}\}$. By (1), either $a = a_1$ or $a = a_k$. Let us encode a by 0 in the first case and by 1 in the second case. Using this rule, we construct a binary *codeword* which encodes the original Pansiot word up to the first $k-1$ letters:

$$\begin{array}{lcl} \text{Pansiot word over 5 letters} & \rightarrow & a b c d a e c b d e a b \dots \\ \text{codeword} & \rightarrow & 0 1 0 1 1 0 1 0 \dots \end{array}$$

More precisely, a codeword of length $n-k+1$ determines $k!$ Pansiot words of length n ; all these Pansiot words can be obtained from each other by permutations of the alphabet. We denote the codeword of a Pansiot word w by $\text{Bin}(w)$, and the set of all possible codewords by P . The definition of the codeword is extended to Z -words in an obvious way.

The following two properties of codewords are easy but very important. The second one follows from condition (2).

Observation 1.3. The growth rate of a set of codewords coincides with the growth rate of the set of all Pansiot words represented by those codewords.

Observation 1.4. Independently of k , P is the language with the finite antidictionary $\{00, 111\}$.

Using Observations 1.3 and 1.4, we easily get $\alpha(T_k^{(2)}) = \alpha(P) \approx 1.324718$.

2. TYPES OF REPETITIONS

2.1. CYLINDRIC REPRESENTATION

To make our considerations visual, we give a graphical representation of Pansiot words on an infinite cylinder. Imagine the word (finite or infinite) as a rope with knots, which are representing letters. This rope is wound around a cylinder such that the knots at distance k are placed one under another (Fig. 1a). By Observation 1.2, the knots labeled by two closest occurrences of the same letter are situated on two consecutive winds of the rope one under another or shifted by one knot (Fig. 1b). If we connect these closest occurrences by “sticks”, we get three types of such sticks: vertical, left-slanted, and right-slanted (Fig. 1b). We refer to this two-dimensional construction, which is a graph on an infinite cylinder, as the *cylindric representation* of a Pansiot word.

The following observation immediately follows from condition (2).

Observation 2.1. Following the wind, any two consecutive sticks in the cylinder representation of a Pansiot word are different.

It is easy to observe that the sequence of sticks, corresponding to a given Pansiot word, is a convenient way to represent the codeword of this Pansiot word. Indeed, the stick going up from a given knot represents the code of the letter in this position according to the following observation, which follows easily from Observation 2.1.

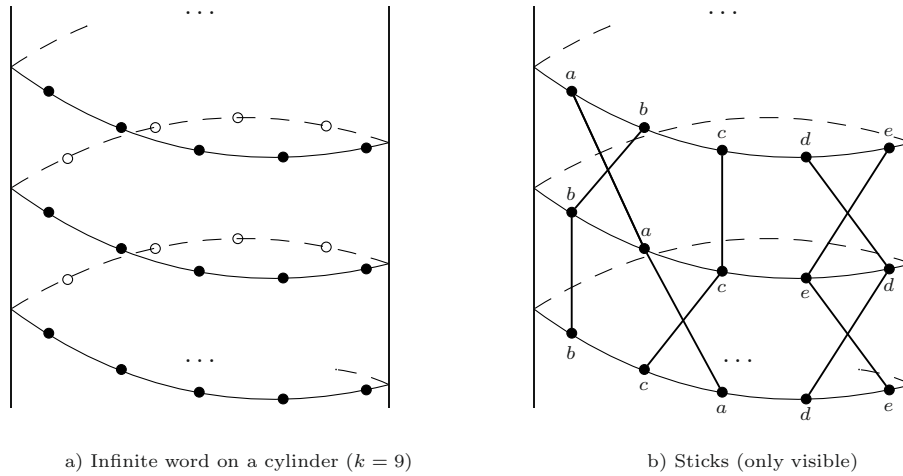


FIGURE 1. Cylindric representation of Pansiot words.

Observation 2.2. The right-slanted stick represents 0, the left-slanted stick represents 1 which follows 0 in the codeword, and the vertical stick represents 1 which follows another 1 in the codeword.

On the other hand, the cylindric representation gives us new possibilities: we can now obtain handy and visual description of repetitions in terms of 2-dimensional patterns.

Note that the sticks constitute k polylines, one for each letter. We call these polylines the *traces* of letters.

2.2. SHORT AND KERNEL REPETITIONS

The partition of m -repetitions into two natural classes was proposed in [11]. These two classes are formed by *short* repetitions ($m < k$) and *kernel* repetitions ($m \geq k$). To see the main difference between these classes, note that the codeword of the repetition $yzzy$ also has the period $|yz|$. In the case of a short repetition, this period is greater than or equal to the length of the codeword. Thus, the codeword has no “global” repetitiveness. For the codewords of kernel repetitions, the period $|yz|$ is proper, so these words are fractional powers.

The termin “kernel” is due to the group-theoretic view on the codewords. In a natural way, 0 and 1 can be considered as the permutations on k letters, thus making each codeword an element of the symmetric group S_k . Then, all factors of length $|yz|$ of the codeword of a kernel repetition are equal to the unit of this group. For details on kernel repetitions the reader is referred to [2,11].

The properties of kernel repetitions are the properties of words in finite symmetric groups. It seems reasonable that the properties of such words do not heavily depend on k , especially if k is big enough. In particular, one can prove that 0 and 1 generate the whole symmetric group for any $k \geq 5$.

In this paper we study structural properties only for short repetitions. Nevertheless, in our numerical results for relatively small alphabets the kernel repetitions will appear as well.

2.3. TYPES OF SHORT REPETITIONS

Let zyz be a short m -repetition. Then all letters of the word y are distinct by condition (1); we refer to these letters as Y-letters.

Lemma 2.3. *Let zyz be an m -repetition. Then (i) $(m-1)k+1 \leq |zyz| \leq mk-1$, (ii) each Y-letter occurs in zyz at most m times.*

Proof. An m -repetition is a minimal forbidden word for the threshold language T_k by definition. If $|zyz| \geq mk$, then the exponent of zyz is at most $k/(k-1)$, while in the case $|zyz| \leq (m-1)k$ the longest proper prefix of zyz is an $(m-1)$ -repetition by definition. In both cases zyz is not a minimal forbidden word, whence (i).

Now let a be an Y-letter. All a 's in the word zyz occur in a factor starting in the position of a in the left y and finishing in the position of a in the right y . We denote this factor by v ; its length is $n = |zyz| - m + 1$. Since the distance between equal letters in a Pansiot word is at least $k-1$ by Observation 1.2, v contains at most $\lceil n/(k-1) \rceil$ occurrences of a . But $n = |zyz| - m + 1 \leq m(k-1)$, whence this number of occurrences is at most m . The lemma is proved. \square

We say that zyz is *uniform*, if all Y-letters have the same number of occurrences in zyz . The following lemma shows that all ‘‘really short’’ repetitions are uniform.

Lemma 2.4. *Let zyz be an m -repetition such that $m < (k+3)/2$. Then each Y-letter occurs in zyz exactly m times.*

Proof. First note that zyz is a short repetition, since $(k+3)/2 \leq k-1$.

Let a be an Y-letter, v be the shortest factor of zyz containing all a 's, $n = |v| = |zyz| - m + 1$. By Lemma 2.3, a occurs at most m times in v . Suppose that v contains less than m a 's. Since by Observation 1.2 the distance between two nearest a 's is at most $k+1$, we have $n \leq (m-2)(k+1) + 1$. On the other hand, $n = |zyz| - m + 1 \geq (m-1)(k-1) + 1$ by Lemma 2.3. Comparing these bounds for $m < (k+3)/2$, we get a contradiction:

$$((m-1)(k-1) + 1) - ((m-2)(k+1) + 1) = k + 3 - 2m > 0.$$

Hence, a occurs exactly m times in zyz . The lemma is proved. \square

Remark 2.5. The bound in the previous lemma is sharp, because the existence of non-uniform $\lceil (k+3)/2 \rceil$ -repetitions can be proved for any $k \geq 5$. For odd k , such a repetition of length $(k+1)k/2 + 1$ has the cylindrical representation consisting of slanted sticks only.

The next lemma establishes the length property of uniform repetitions.

Lemma 2.6. *Let zyz be a uniform short m -repetition. Then all k letters occur in the word yz exactly $m-1$ times, implying $|yz| = (m-1)k$.*

Proof. First we count the occurrences of Y-letters. By Lemma 2.3, an Y-letter occurs at most m times in $yzzy$. Suppose that Y-letters occur in $yzzy$ less than m times. Then, z contains at most $m-3$ occurrences of each of m Y-letters. By Lemma 2.3 (i), there is a letter c occurring in z at least m times. By condition (1), at least $k-2$ different letters appear between two consecutive occurrences of c . At least $m-1$ of those letters are Y-letters. Thus, z contains at least $(m-1)(m-1)$ occurrences of Y-letters, which is greater than $m(m-3)$, a contradiction. So, each Y-letter occurs m times in $yzzy$ and $m-1$ times in yz .

Now consider a letter c which is not an Y-letter and estimate the number of its occurrences in yz . For convenience, we assume that the occurrences of any letter are numbered from left to right.

Suppose that c has at least m occurrences. The first occurrences of all m Y-letters are on the left of the first occurrence of c . Now look at the cylinder representation of $yzzy$. On every wind of the rope the letter c can swap the places with at most one letter. Hence, the second occurrences of at least $(m-1)$ Y-letters are on the left of the second occurrence of c , and so on. Finally, the m th occurrence of some Y-letter a is certainly on the left of the m th occurrence of c . This is a contradiction, because the m th occurrence of a belongs to the right y , while all occurrences of c belongs to z .

Now suppose that c has at most $m-2$ occurrences. From the definition of Pansiot's words it follows that the second occurrences of at least $(m-1)$ Y-letters are on the right of the first occurrence of c . Similar to the above, we obtain that the $(m-1)$ th occurrences of at least two Y-letters are on the right of the last occurrence of c . The $(m-1)$ th and m th occurrences of each of these two letters are at distance at least $k-1$. Hence, $yzzy$ has a suffix of length at least $(k+1)$ which does not contain the letter c in contradiction with the definition of Pansiot words.

Therefore, c occurs in $yzzy$ exactly $m-1$ times, whence the result. \square

Corollary 2.7. *The trace of an Y-letter in a uniform repetition $yzzy$ consists of $m-1$ sticks and contains equal number of left-slanted and right-slanted sticks.*

Proof. Such a trace connects the positions of a letter a in the left and the right y 's. The distance between these positions is $(m-1)k$, and the total number of occurrences of a is m . The required statement follows from this. \square

Observation 2.8. In non-uniform short repetitions the traces of some Y-letters have different lengths. Hence, for any two such letters one trace goes i knots to the right, while another one goes $(k-i)$ knots to the left. In particular, these traces intersect odd number of times.

3. UNIFORM REPETITIONS

A natural property of antidictionaries is that by adding a short word to the antidictionary of a given language we affect the growth rate of this language much stronger than by adding a long word. So, the study of short forbidden words is of utmost interest for estimating growth rates. From the previous section we

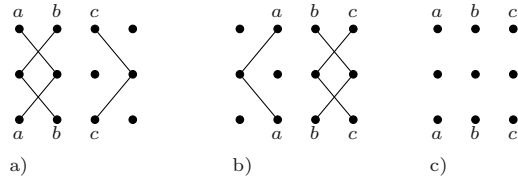


FIGURE 2. Cylindric representations for 3-repetitions.

learn that in Pansiot words all repetitions of length $O(k)$ are uniform, and all other repetitions have the length $\Omega(k^2)$. So, here we study particular uniform repetitions in more detail. In view of Theorem 1.1, we may suppose that all considered repetitions are factors of Pansiot Z-words.

3.1. 3-REPETITIONS

Note that by Lemma 2.4 3-repetitions are uniform for any $k \geq 5$.

Theorem 3.1. *The following conditions are equivalent for a Pansiot Z-word $W \in \Sigma_k^Z$:*

- 1) *W contains a 3-repetition;*
- 2) *the cylindric representation of W contains one of the patterns $\updownarrow, \times \times$;*
- 3) *$\text{Bin}(W)$ has a factor of the form $1101x1011$ or $0101y1010$, where $|x| = k-6, |y| = k-5$ ¹.*

Proof. $1 \Rightarrow 2$. Let $abc z abc$ be a 3-repetition and a factor of W . By Corollary 2.7, the trace of each of the letters a, b , and c consists either of two vertical sticks (the pattern \updownarrow), or of two different slanted sticks. If all three traces consist of slanted sticks, then by Observation 2.1 we have the pattern from Figure 2a or 2b. In both cases the cylindric representation of W contains the pattern $\times \times$.

$2 \Rightarrow 1$. By Lemma 2.6 we see that for a 3-repetition the occurrences of three letters should appear in the cylindric representation as is shown in Figure 2c. The pattern $\times \times$ readily provides such letters, while the pattern \updownarrow is preceded (and followed) by crossed sticks by Observation 2.1, thus providing the required letters as well.

$2 \Leftrightarrow 3$. We make use of Observation 2.2. The pattern \updownarrow indicates that the codeword contains two occurrences of the factor 11 at distance k (and *vice versa* – such occurrences give two vertical sticks one under another). Since 11 in the codeword of a Pansiot word is always preceded by 10 and followed by 01, we get the required factor $1101x1011$. Similarly, the pattern $\times \times$ occurs in the cylindric representation if and only if the codeword contains two occurrences of the factor 0101 at distance k . Since 0 in the codeword of a Pansiot word is always preceded by 1, we obtain the factor $0101y10101$, which contains the required one.

¹If $k = 5$, then the first word is 1101011.

The theorem is proved. \square

The language

$$A'_k = \{1101x1011 \mid |x| = k-6\} \cup \{0101y1010 \mid |y| = k-5\} \cup \{00, 111\},$$

where the first two sets consist of the words without factors 00 and 111, is antifactorial, and hence is the antidictionary of some factorial binary language $P'_k \subset P$. This language can be used to calculate the growth rate of the language $T_k^{(3)}$, as the following proposition shows.

Proposition 3.2. *The growth rates of the languages $T_k^{(3)}$ and P'_k coincide.*

Proof. By Theorem 1.1, the languages $T_k^{(3)}$ and $e(T_k^{(3)})$ have the same growth rate. The language $T_k^{(3)}$ consists of all Pansiot words avoiding 3-repetitions. Then, the language $e(T_k^{(3)})$ consists of all finite factors of Pansiot Z-words avoiding 3-repetitions. By Theorem 3.1, a word belongs to $e(T_k^{(3)})$ if and only if its codeword belongs to P'_k . The growth rate of a set of Pansiot words coincides with the growth rate of the set of their codewords by Observation 1.3, whence the result. \square

The antidictionary A'_k of the language P'_k is easy to calculate. So, we apply the algorithm of [4] to construct the automata and the algorithm of [15] to obtain the growth rates of languages P'_k for $k = 5, 6, \dots, 60$ (the restriction from above is due to the memory constraints of the personal computer). Some of our results are presented in the second column of Table 1, p. 190.

This table gives a strong evidence that the sequence of growth rates of the languages avoiding 3-repetitions converges to a limit $\bar{\alpha} \approx 1.242096777\dots$, although we have no analytic proof of this fact. Another interesting feature is that the obtained growth rates suit very well to the curve of damped oscillations (see Fig. 3). The oscillating function in this graph is $y = \bar{\alpha} + 0.125 \times 1.425^{-x} \cos(2.412x - 2.7)$.

3.2. LACK OF 4- AND 5-REPETITIONS

The following lemma is used many times in the current and the next subsections.

Lemma 3.3. *Suppose that $yzzy$ is a uniform m -repetition, and a is an Y -letter. Then the trace of any letter $b \neq a$ intersects the trace of a even number of times².*

Proof. By Corollary 2.7 the trace of a consists of $(m-1)$ sticks and connects the positions i and $i + (m-1)k$ for some i . If the trace of b intersects the trace of a odd number of times, then we can assume that the trace of b connects the positions j_1 and $j_2 + (m-1)k$ for some j_1, j_2 such that $j_1 < i < j_2$ or $j_2 < i < j_1$. Note that

²Recall that $yzzy$ is considered as a factor of some Pansiot Z-word. For example, if a occupies the 1st and k th positions in $yzzy$, some letter b first occurs in $yzzy$ in $(k+1)$ th position. But if we extend $yzzy$ to the left, b will also occupy the position 0. Hence, we consider the stick $(0, k+1)$ as a part of the trace of b .

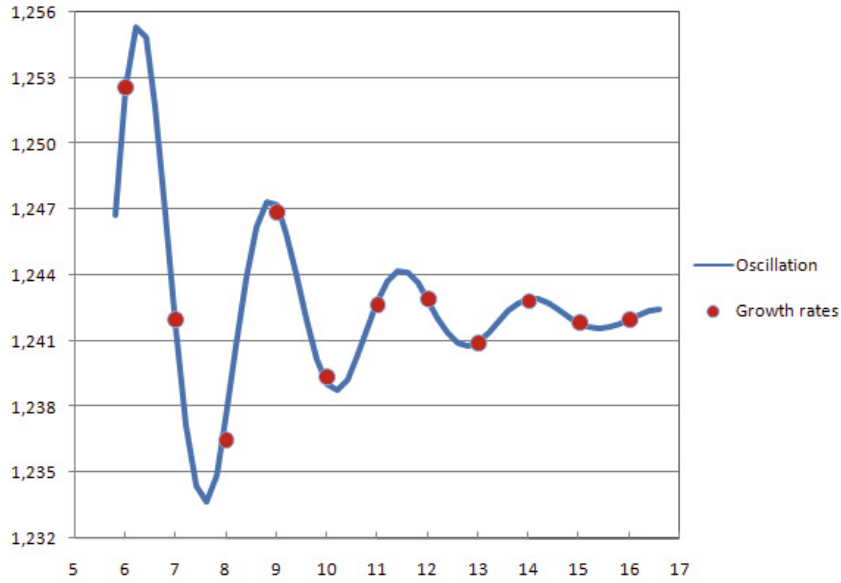


FIGURE 3. (Color online) Growth rates and dumped oscillations.

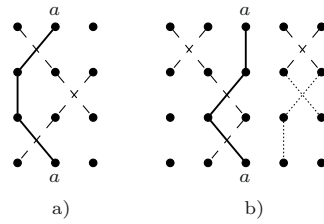


FIGURE 4. Cylindric representations for 4-repetitions. The dashed sticks are determined by the trace of the letter a .

$|j_2 - j_1| \leq m-1$, because the distance between two consecutive occurrences of b is at least $k-1$ and at most $k+1$. Since the positions j_1 and j_2 can not belong to y by Corollary 2.7, we have $|y| < m$ in contradiction with the definition of m -repetition. \square

Theorem 3.4. *There exist no uniform 4- or 5-repetitions.*

Proof. In view of Corollary 2.7, the trace of an Y-letter a in a 4-repetition must look (up to the symmetry) like in Figures 4a, 4b. In case (a) there are traces intersecting the trace of a only once. Hence, no repetition occurs by Lemma 3.3. In case (b) a must be the first letter of the repetition. The second letter returns to its place only if the dotted sticks belong to the cylindric representation of this

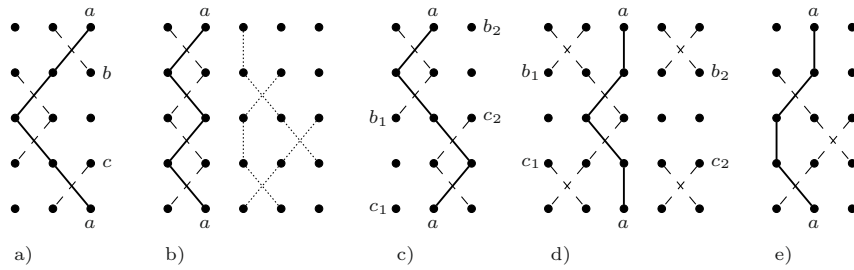


FIGURE 5. Cylindric representations for 5-repetitions. The dashed sticks are determined by the trace of the letter a .

repetition. But then the third letter can not return to its place. So, no uniform 4-repetition can exist.

Similarly, there are five possibilities for the trace of an Y-letter in a 5-repetition, see Figures 5a–5e. If the knots labeled by b and c in case (a) are not connected, then no 5-repetition occur by Lemma 3.3. But if they are connected, the pattern \dagger or $\times \times$ appears, indicating a 3-repetition by Theorem 3.1. Hence, no 5-repetition occur by definition.

In case (b) the dotted sticks show the only way (up to reversing upside down) to avoid the patterns \dagger and $\times \times$. We see that the letter following a does not return to its place. By a similar argument, the letter which is initially two knots left from a does not return to its place also, so no repetition occurs.

To obtain a repetition in cases (c) and (d), one should connect the knots labeled by b_1 and c_1 or the knots labeled by b_2 and c_2 . In both cases the pattern \dagger or $\times \times$ will be obtained.

In case (e) no 5-repetition can occur by Lemma 3.3. Thus, we have examined all possibilities for the trace of an Y-letter in a 5-repetition. The theorem is proved. \square

3.3. 6-REPETITIONS

Figure 6 shows five traces, which we denote by $\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}$. The mirror images of these traces (under the vertical axis) will be referred to as the traces $\mathbf{1}', \mathbf{2}', \mathbf{3}', \mathbf{4}',$ and $\mathbf{5}'$ respectively. The pattern consisting of several traces will be denoted by the sequence of traces in the order in which the traces appear in the top row of the pattern. Thus, the pattern in Figure 6 is $\mathbf{12345}$.

Theorem 3.5. *Let $W \in \Sigma_k^Z$ ($k \geq 7$) be a Pansiot Z -word without 3-repetitions. Then W contains a uniform 6-repetition if and only if the cylindric representation of W contains one of the patterns $\mathbf{123451}, \mathbf{234512}, \mathbf{345123}, \mathbf{451234}, \mathbf{512345}, \mathbf{1'5'4'3'2'1'}, \mathbf{2'1'5'4'3'2'}, \mathbf{3'2'1'5'4'3'}, \mathbf{4'3'2'1'5'4'}, \mathbf{5'4'3'2'1'5'}$.*

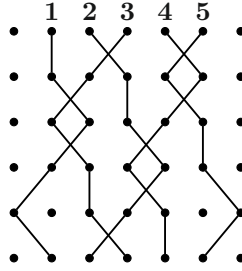


FIGURE 6. Admissible traces for 6-repetitions.

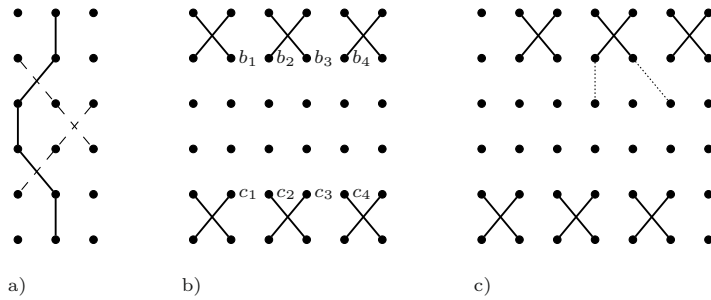


FIGURE 7. Impossible patterns for 6-repetitions.

Proof. First we prove necessity. Let zyz be a uniform 6-repetition and a factor of W . The trace of an Y-letter contains equal number of left and right sticks by Corollary 2.7 and no \uparrow pattern by Theorem 3.1. So, if such a trace contains three vertical sticks, it looks (up to symmetry) like in Figure 7a. We see that in this case no uniform 6-repetition occur by Lemma 3.3. Hence, the trace of an Y-letter contains only one vertical stick. If no one of these traces begins or ends by a vertical stick, then we have (up to symmetry) the pattern in Figure 7b or 7c. To obtain the required repetition from the pattern in Figure 7b, one should connect b_i to c_i for $i = 1, 2, 3, 4$, thus getting a uniform 4-repetition, which is impossible by Theorem 3.4. To get the repetition from the pattern in Figure 7c, one should insert into each trace one vertical stick and two slanted sticks, which differ from the first stick. If we continue one trace with a vertical stick, its neighbor will get a wrong slanted stick (dotted sticks in Fig. 7c). But if we use only slanted sticks in the second row, we will get the pattern $\begin{matrix} \times & \times & \times \\ \times & \times & \times \end{matrix}$, and hence a 3-repetition by Theorem 3.1. Therefore, the considered cases are impossible, and some trace has a marginal vertical stick. Without loss of generality we suppose that this stick is the first one, and examine possible cases.

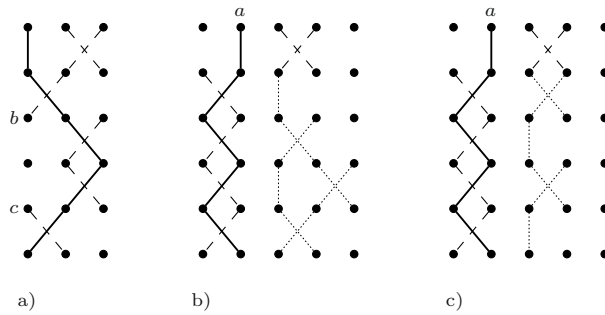


FIGURE 8. More impossible patterns for 6-repetitions.

Up to symmetry, there are three possible traces of an Y-letter beginning with a vertical stick. First one is shown in Figure 8a. To satisfy Lemma 3.3, we should connect the knots b and c , but then we can not avoid the \uparrow or $\begin{smallmatrix} \times & \times \\ \times & \times \end{smallmatrix}$ pattern. The second case is represented in Figures 8b, 8c. The letter a must be the first letter of the repetition, and there are two ways to place sticks on the right of the trace of a (dotted sticks in Figs. 7b and 7c). We see that the two letters, following a , can not belong to the repetition simultaneously. The third case is the trace **1** (Fig. 6). In this case, if the next letter is in y , its trace is **2** (otherwise we obtain a pattern of 3-repetition). Similarly, after **2** only the trace **3** can appear, then **4**, **5**, and another **1**. Starting with traces **2**, **3**, **4**, and **5** we get four more required patterns. Symmetrically, the remaining five patterns can be obtained. Note that trace **4** ends by a vertical stick. Hence, if we consider the possible traces ending by a vertical stick instead of traces beginning with such a stick, we will obtain the same set of patterns. The necessity is thus proved.

Now check sufficiency. Each of listed patterns indicates a factor $yzzy$ in W , where $|yz| = 5k$, $|y| = 6$ (the traces in the pattern are exactly the ones of Y-letters). So, the exponent of $yzzy$ is strictly greater than $k/(k-1)$. In addition, we must show that $yzzy$ contains no repetition as proper factor. W contains no 1- and 2-repetitions by definition of Pansiot words, no 3-repetitions by conditions of theorem, no uniform 4- and 5-repetitions by Theorem 3.4, and no non-uniform 4-repetitions by Lemma 2.4. It remains to check that $yzzy$ does not contain non-uniform 5-repetitions for $k = 7$ and non-uniform 6-repetitions for $k = 7, 8, 9$ (the restrictions on k follow from Lem. 2.4). This can be easily done using Observation 2.8. \square

Similar to Theorem 3.1, we have an equivalent description of uniform 6-repetitions in terms of forbidden codewords. But this description is bulky (ten patterns should be described by codewords with four variable factors in each), so we omit it here. Nevertheless, we use this description to obtain the growth rate of the languages $T_k^{(6)}$ for $k = 10, \dots, 22$. (For $k \leq 9$ there exist non-uniform repetitions which are shorter than uniform 6-repetitions.) The results are given in Table 1.

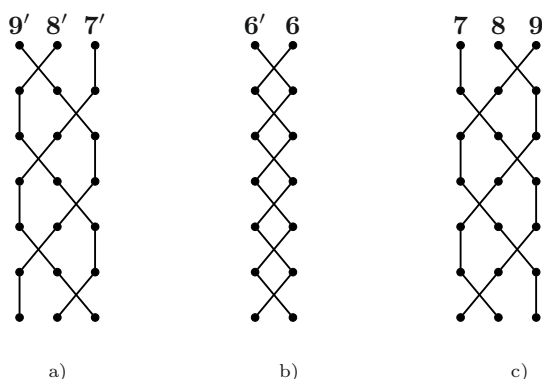


FIGURE 9. Admissible traces for 7-repetitions.

3.4. LONGER REPETITIONS AND NUMERICAL RESULTS

We see that, for a fixed m , all uniform short m -repetitions can be described by a finite set of two-dimensional patterns, independent of k . So, in principle, the analogs of Theorem 3.5 can be proved for any fixed m . The question is whether we can give some inductive description of these patterns for m -repetitions from smaller cases. We suggest the negative answer. At least, no similarity can be seen between 6- and 7-repetitions, as the following theorem shows. We omit the proof, because it is yet another pretty long case examination. The admissible traces for 7-repetitions are presented in Figure 9.

Theorem 3.6. *Let $W \in \Sigma_k^Z$ ($k \geq 8$) be a Pansiot Z -word without 3- and 6-repetitions. Then W contains a uniform 7-repetition if and only if the cylindric representation of W contains seven consecutive letters the traces of which are among $\mathbf{6} - \mathbf{9}$, $\mathbf{6}' - \mathbf{9}'$.*

Once again, the two-dimensional patterns can be translated into forbidden binary codewords to get the growth rates of the languages $T_k^{(7)}$. The results for $k = 12, \dots, 20$ are given in Table 1. (For $k \leq 11$ there exist non-uniform repetitions which are shorter than uniform 7-repetitions.)

Now compare the figures in Table 1 to estimate the contribution of uniform 6- and 7-repetitions to the growth rate of threshold languages. We see that this contribution is very small with respect to the contribution of 3-repetitions. On the other hand, this contribution does not tend to zero as k increases. Moreover, the contribution of 6-repetitions (respectively, 7-repetitions) seems to admit a limit close to 4.3×10^{-5} (respectively, 3.6×10^{-6}).

TABLE 1. Growth rates of the languages $T_k^{(3)}$, $T_k^{(6)}$ and $T_k^{(7)}$.

k	$\alpha(T_k^{(3)})$	$\alpha(T_k^{(6)})$	Δ_{36}	$\alpha(T_k^{(7)})$	Δ_{67}
5	1.2149529259				
6	1.2525850156				
7	1.2419872870				
8	1.2365057034				
9	1.2468689328				
10	1.2393804408	1.2393569373	2.35×10^{-5}		
11	1.2426566183	1.2426131641	4.35×10^{-5}		
12	1.2429286087	1.2428815613	4.70×10^{-5}	1.2428793902	2.17×10^{-6}
13	1.2409226614	1.2408787655	4.39×10^{-5}	1.2408729371	5.83×10^{-6}
14	1.2428289279	1.2427804533	4.85×10^{-5}	1.2427767435	3.71×10^{-6}
15	1.2418774954	1.2418379960	3.95×10^{-5}	1.2418340134	3.98×10^{-6}
16	1.2420000807	1.2419572202	4.29×10^{-5}	1.2419536019	3.62×10^{-6}
17	1.2423240470	1.2422793681	4.47×10^{-5}	1.2422759823	3.39×10^{-6}
18	1.2418973834	1.2418553246	4.21×10^{-5}	1.2418514866	3.84×10^{-6}
19	1.2421895750	1.2421451742	4.44×10^{-5}	1.2421416402	3.53×10^{-6}
20	1.2420949436	1.2420520022	4.29×10^{-5}	1.2420483333	3.67×10^{-6}
21	1.2420552103	1.2420123323	4.29×10^{-5}		
22	1.2421449456	1.2421012197	4.37×10^{-5}		
...	...				
58	1.2420967776				
59	1.2420967762				
60	1.2420967771				

4. MORE NUMERICAL RESULTS AND MAIN CONJECTURE

The results of previous section strongly suggest the idea that the contribution of uniform m -repetitions to the growth rate of threshold languages is almost independent of k if k is big enough. Indeed, the two-dimensional description of such repetitions consists of patterns of approximately $m \times m$ size and is independent of k . From the numerical results we also suspect that this contribution quickly decreases as m grows.

TABLE 2. Bounds of the growth rates of the threshold languages: n is the length of the longest word in the antidictionary, $\#$ is the number of different codewords in the antidictionary, α is the growth rate.

k	$T_k^{(3)}$			$T_k^{(k-1)}$			$T_k^{(2k-1)}$			Best obtained bound		
	n	$\#$	α	n	$\#$	α	n	$\#$	α	n	$\#$	α
5	13	10	1.214953	16	12	1.186362	41	60	1.164888	134	570891	1.158057
6	15	11	1.252585	26	23	1.232128	65	1170	1.225386	103	296961	1.224784
7	17	14	1.241987	41	32	1.238146	90	16378	1.236972	97	54114	1.236948
8	19	18	1.236506	55	74	1.234967			n/a	98	14182	1.234857
9	21	21	1.246869	71	142	1.246694			n/a	94	3055	1.246682
10	23	23	1.239380	89	459	1.239310			n/a	95	1184	1.239309

The length of non-uniform short repetitions and kernel repetitions is at least quadratic in k . So, the two-dimensional descriptions of such repetitions will depend on k . Then the contribution of such repetitions to the growth rate is likely to depend on k also. We can suggest the type of this dependence studying threshold languages over relatively small alphabets, where the antidictionaries $A_k^{(m)}$ can be constructed by brute force. Table 2 contains our results for the alphabets of 5 to 10 letters.

First, note the difference between the growth rates of $T_k^{(3)}$ and $T_k^{(k-1)}$. Surprisingly, it greatly (and monotonely) decreases with the increase of k . For $k = 5$, we have the difference about 0,03 with only two additional words in the antidictionary, while for $k = 10$ the difference is only 0,00007 (with several hundred additional words). So, it seems that the total contribution of non-uniform short repetitions to the growth rate tends to zero as k approaches infinity.

Second, the contribution of “relatively short” kernel repetitions ($k \leq m < 2k$) is less than the contribution of non-uniform short repetitions in all three cases. It is hard to be definite about “long” kernel repetitions, but for $k = 5$ more than 500 000 shortest kernel repetitions have approximately the same total contribution as two non-uniform short repetitions. We summarize the above considerations in our main conjecture:

Conjecture 4.1. The sequence $\{\alpha(T_k)\}$ of the growth rates of threshold languages converges to a limit $\hat{\alpha} \approx 1.242$ as k tends to infinity.

Acknowledgements. The authors heartly thank the referees for their valuable comments and remarks.

REFERENCES

- [1] F.-J. Brandenburg, Uniformly growing k -th power free homomorphisms. *Theoret. Comput. Sci.* **23** (1983) 69–82.
- [2] A. Carpi, On Dejean’s conjecture over large alphabets. *Theoret. Comput. Sci.* **385** (2007) 137–151.
- [3] C. Choffrut, J. Karhumäki, *Combinatorics of words*, edited by G. Rosenberg and A. Salomaa. Handbook of formal languages, Vol. 1, Chap. 6. Springer, Berlin (1997) 329–438.
- [4] M. Crochemore, F. Mignosi and A. Restivo, Automata and forbidden words. *Inform. Process. Lett.* **67** (1998) 111–117.
- [5] J.D. Currie, N. Rampersad, Dejean’s conjecture holds for $n \geq 27$. *RAIRO-Theor. Inf. Appl.* **43** (2009) 775–778.
- [6] J.D. Currie, N. Rampersad, A proof of Dejean’s conjecture, <http://arxiv.org/PScache/arxiv/pdf/0905/0905.1129v3.pdf>
- [7] F. Dejean, Sur un Theoreme de Thue. *J. Combin. Theory Ser. A* **13** (1972) 90–99.
- [8] A. Ehrenfeucht and G. Rozenberg, On subword complexities of homomorphic images of languages. *RAIRO Inform. Theor.* **16** (1982) 303–316.
- [9] M. Lothaire, *Combinatorics on words*. Addison-Wesley (1983).
- [10] M. Mohammad-Noori and J.D. Currie, Dejean’s conjecture and Sturmian words. *Eur. J. Combin.* **28** (2007) 876–890.
- [11] J. Moulin-Ollagnier, Proof of Dejean’s Conjecture for Alphabets with 5, 6, 7, 8, 9, 10 and 11 Letters. *Theoret. Comput. Sci.* **95** (1992) 187–205.
- [12] J.-J. Pansiot, À propos d’une conjecture de F. Dejean sur les répétitions dans les mots. *Discrete Appl. Math.* **7** (1984) 297–311.
- [13] M. Rao, Last Cases of Dejean’s Conjecture, accepted to WORDS’2009.
- [14] A.M. Shur, Rational approximations of polynomial factorial languages. *Int. J. Found. Comput. Sci.* **18** (2007) 655–665.
- [15] A.M. Shur, Combinatorial complexity of regular languages, *Proceedings of CSR’2008. Lect. Notes Comput. Sci.* **5010** (2008) 289–301.
- [16] A.M. Shur, Growth rates of complexity of power-free languages. Submitted to *Theoret. Comput. Sci.* (2008).
- [17] A.M. Shur, Comparing complexity functions of a language and its extendable part. *RAIRO-Theor. Inf. Appl.* **42** (2008) 647–655.
- [18] A. Thue, *Über unendliche Zeichenreihen*, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl., Christiania* **7** (1906) 1–22.