Journal of Integer Sequences, Vol. 24 (2021),
Article 21.8.7

Squarefree Extensions of Words

Jarosław Grytczuk and Hubert Kordulewski
Faculty of Mathematics and Information Science
Warsaw University of Technology
00-662 Warsaw
Poland
j.grytczuk@mini.pw.edu.pl
apostatajulian@gmail.com

Bartłomiej Pawlik
Institute of Mathematics
Silesian University of Technology
44-100 Gliwice
Poland
bpawlik@polsl.pl

Abstract

A word is *squarefree* if it does not contain nonempty factors of the form XX. In 1906 Thue proved that there exist arbitrarily long squarefree words over a 3-letter alphabet. It was proved recently that among these words there are infinitely many *extremal* ones, that is, having a square in every single-letter extension.

We study diverse problems concerning extensions of words preserving the property of avoiding squares. Our main motivation is the conjecture stating that there are no extremal words over a 4-letter alphabet. We also investigate a natural recursive procedure of generating squarefree words by a single-letter rightmost extension. We present the results of computer experiments supporting a supposition that this procedure gives an infinite squarefree word over any alphabet of size at least three.

1 Introduction

A *square* is a finite nonempty word of the form XX. For instance, the word `hotshots` is a square with $X = \texttt{hots}$. A word W *contains* a square if it can be written as $W = UXXV$ for some words U, V, and a nonempty word X. A word is *squarefree* if it does not contain any squares. For instance, the word `repetition` contains the square `titi`, while `recreation` is squarefree.

It is easy to check that there are no binary squarefree words of length greater than 3. However, there exist ternary squarefree words of any length, as proved by Thue [16, 3]. This result is the starting point of combinatorics on words, a wide discipline with many exciting problems, deep results, and important applications [1, 2, 4, 6, 9, 10].

In this paper we study problems concerning *squarefree extensions* of words, a concept introduced recently by Grytczuk, Kordulewski, and Niewiadomski [7]. Let $\mathbb{A}$ be a fixed alphabet and let W be a finite word over $\mathbb{A}$. The set of all finite words over $\mathbb{A}$ is denoted by $\mathbb{A}^*$. An *extension* of W over $\mathbb{A}$ is any word of the form $W'\mathtt{x}W''$, where $\mathtt{x}$ is any letter from $\mathbb{A}$ and $W = W'W''$ (words W' and W'' are possibly empty). For instance, the word `bear` is an extension of the word `bar` by inserting the letter `e` between letters `b` and `a`. A squarefree word $W \in \mathbb{A}^*$ is *maximal* if for every $\mathtt{x} \in \mathbb{A}$ the extensions $\mathtt{x}W$ and $W\mathtt{x}$ contain a square. A squarefree word W is called *extremal* over $\mathbb{A}$ if there is no squarefree extension of W. For instance, the word

$$H = \texttt{123121323123212312132312 3}$$

is the shortest extremal word over the alphabet $\{\mathtt{1}, \mathtt{2}, \mathtt{3}\}$. This means that inserting any letter from the alphabet $\{\mathtt{1}, \mathtt{2}, \mathtt{3}\}$ at any position in the word H, including the beginning as well as the end of H, results with a square.

A natural intuition is that extremal words should be rare or even should have bounded length. However, in the case of a 3-letter alphabet, this intuition turned out to be faulty.

Theorem 1 (Grytczuk, Kordulewski, and Niewiadomski [7])**.** *There exist infinitely many extremal squarefree words over a* 3*-letter alphabet.*

The proof of this theorem is by a recursive construction whose validity is partially based on computer verification. Mol and Rampersad [11] determined all positive integers k for which there exist extremal ternary words of length exactly k. In particular, they proved that for every $k \geq 87$ there exists an extremal ternary word of length k.

One may naturally wonder what the case is for larger alphabets. Actually, we do not know if there are any extremal words over a 4-letter alphabet. Computational experiments prompted us to state the following conjecture.

Conjecture 2 (Grytczuk, Kordulewski, and Niewiadomski [7])**.** Every squarefree word over a 4-letter alphabet can be extended to a squarefree word.

A weaker version of this statement was recently confirmed by Hong and Zhang [8] who proved that there are no extremal squarefree words over an alphabet of size 15.

In the forthcoming sections we shall present some results and effects of computer experiments inspired by Conjecture 2.

2 Nonchalant words

2.1 The main conjectures

The problem of extremal squarefree words is connected to the following recursive construction.

Given a fixed ordered alphabet $\mathbb{A}$, we start with the first letter from $\mathbb{A}$ and continue by inserting the earliest possible letter at the rightmost position of the previous word so that the new word is squarefree. For instance, for the alphabet $\{\mathtt{1},\mathtt{2},\mathtt{3}\}$ this greedy procedure starts with the following sequence of squarefree words:

$$\mathtt{1}, \mathtt{12}, \mathtt{121}, \mathtt{1213}, \mathtt{12131}, \mathtt{121312}, \mathtt{1213121}, \mathtt{12131231}.$$

The last word was obtained by inserting $\mathtt{3}$ at the penultimate position of the previous word.

We conjecture that the aforementioned procedure never stops. To state it formally, let us define recursively a sequence of *nonchalant words* N_i over the alphabet $\mathbb{A}_n = \{\mathtt{1},\mathtt{2},\ldots,\mathtt{n}\}$ by putting $N_1 = \mathtt{1}$, and letting $N_{i+1} = N_i'\mathtt{x}N_i''$ to be a squarefree extension of N_i such that in the first instance N_i'' is the shortest possible suffix of N_i and then $\mathtt{x} \in \mathbb{A}_n$ is the earliest possible letter.

Conjecture 3 (Grytczuk, Kordulewski, and Niewiadomski [7])**.** The sequence of nonchalant words over $\mathbb{A}_n$ is infinite for every $n \geq 3$.

In other words, we believe that the nonchalant algorithm never produces an extremal word. The results of computer experiments support this conjecture. For instance, for $n = 3$ a nonchalant word of length 10000 was obtained. Moreover, the new letter was never inserted more than 20 positions from the end of the previous word (Appendix A contains more details). Therefore the following conjecture also seems plausible.

Conjecture 4 (Grytczuk, Kordulewski, and Niewiadomski [7])**.** The sequence of nonchalant words over $\mathbb{A}_n$ converges to an infinite word $\mathcal{N}_n$ for every $n \geq 3$.

Here are the first 70 terms of the presumably infinite limit word for $n = 3$:

$\mathcal{N}_3 = \mathtt{12131231321231213123132312131232123121312313212312131232123121321231}\mathtt{32}...$

N_1	0	1	2	3	4	7	9	12	14	15	20
1	9457	310	184	1	33	11	1	0	0	1	2
2	9457	309	186	1	33	11	0	0	1	0	2
3	9457	307	185	0	34	13	1	0	1	0	2
13	9454	310	185	0	34	13	1	0	1	0	2
23	9458	307	185	1	34	11	0	1	1	0	2
32	9458	309	185	1	33	11	0	0	1	0	2
3213	9455	309	185	0	34	13	1	0	1	0	2
2313213	9455	309	185	0	34	13	1	0	1	0	2
32132313213	9455	309	185	0	34	13	1	0	1	0	2

Table 1: Number of steps in which a new letter was inserted before the suffix of given length (10000 iterations).

2.2 Playing with initial words

The above version of the nonchalant algorithm with the two corresponding conjectures were stated by Grytczuk, Kordulewski, and Niewiadomski [7]. Our numerical experiments led us to introduce a more general approach.

Starting with the word 1 is the most natural approach, because such word can be considered as a result of performing the nonchalant procedure on an empty word. Let us consider starting the nonchalant algorithm with a different initial word. Namely, let the nonchalant word N_1 be some squarefree word over a fixed alphabet. From now on N_1 will be called the *initial word* of the nonchalant algorithm. The results of testing 10000 iterations of the nonchalant algorithm over the ternary alphabet for various initial words are presented in Table 1. The first column contains the initial words, while the other columns show how many times the procedure moved back by the given number of positions (the column initialized by 0 shows how many times a new letter is inserted at the rightmost position, by 1 - at the penultimate position, etc.). These experimental results suggest that sequences of nonchalant words still bear many similarities. For example, for each considered initial word, a new letter is inserted 33 to 36 times right before the suffix of length 4. More data from our experiments is presented in Appendix A.

2.3 Nonchalant words over four letters

In the case of a 4-letter alphabet the situation looks even more exciting. In our experiments for the initial word 1 the procedure never moved back by more than one position. In fact, through 50000 iterations the word was extended on the penultimate position only 33 times (in other cases, the word was extended at the last position).

The squarefree word W over given alphabet $\mathbb{A}$ is *almost extremal* if for all nonempty words W' and W'' such that $W = W'W''$, the word $W'\mathtt{x}W''$ contains a square for every $\mathtt{x} \in \mathbb{A}$. Given that, let us consider another variant of the nonchalant algorithm. Namely,

let 12 be the initial word of the nonchalant algorithm over $\mathbb{A}_4$ and let us allow for word extensions only at internal positions (extension of a word at the end or at the beginning is forbidden). This procedure starts with the following sequence of squarefree words:

$$12, 132, 1312, 13142, 131412, 1314132, 13141312, 131413212.$$

The last word was obtained by inserting 2 right before the suffix 12. Through 50000 iterations, the procedure never moved back by more than two positions (in this case the number of iterations in which procedure moved back by two positions is approximately equal to 10% of all iterations).

2.4 Extensions close to the ends

The above experiments led us to the following two suppositions: (1) every quaternary squarefree word can be extended at the end or at the penultimate position, and (2) every quaternary squarefree word (of length at least 3) can be extended at one of the two rightmost internal positions. However, both suppositions turned out not to be true.

Proposition 5. *There exists a quaternary squarefree word S which cannot be extended, neither at the last, nor at the penultimate position.*

Proof. Let $A = 1213121$ and $B = 121312$. Next, let $Y = 3B4$ and $Z = 41YA4YB341$. Finally, put $S = ZYA4YA$, which gives the word

$$S = 423121312412131214312131241213123142312131241213121431213124121312143121312412131214312131241213121.$$

It can be verified (by a computer) that S is indeed squarefree. Now, $A\mathtt{x}$ contains a square for every $\mathtt{x} \in \{1, 2, 3\}$. Also $S4 = Z(YA4)(YA4)$ is not squarefree. For the penultimate position it suffices to check only letters 3 and 4. So, the suffix A in S will turn to one of the forms, $B31$ or $B41$, respectively. In the latter case we get the word

$$ZYA4YB41 = ZYA43(B4)(B4)1. \tag{1}$$

In the former case we obtain

$$ZYA4YB31 = (41YA4YB3)(41YA4YB3)1. \tag{2}$$

The assertion is proved. □

The proof of Proposition 5 could be verified entirely with the use of a computer. However, we decided to include some explanations to give a better insight into the structure of the word S.

In the case of the second supposition, we will present a more general result. We use the well known *Zimin words* Z_n, defined recursively over an infinite alphabet of variables $\{1, 2, 3, \ldots\}$ by $Z_1 = 1$ and $Z_n = Z_{n-1}\mathtt{n}Z_{n-1}$ for every $n \geq 2$. Let us notice that the Zimin word Z_n is maximal over the alphabet $\{1, 2, \ldots, \mathtt{n}\}$. In the following proposition, the construction in the proof is more clear when we analyze the *leftmost* internal positions instead of the *rightmost* ones. Obviously, the result holds in the latter case as well.

Theorem 6. *For all natural numbers n and t, with $n \geq 4$ and $1 \leq t < n$, there exists a squarefree word W over the alphabet $\mathbb{A}_n$, which cannot be extended at any of its t leftmost internal positions.*

Proof. Let $A = \mathtt{12}\cdots\mathtt{n}$ be a word over alphabet $\mathbb{A}_n$. This word has exactly $N = (n-1)(n-2)$ distinct internal squarefree extensions. Let Z_N be the Zimin word over the alphabet $\mathbb{A}_N$. Consider the homomorphism $\varphi : \mathbb{A}_N^* \to \mathbb{A}_n^*$ defined so that the image of every letter in $\mathbb{A}_N$ is a unique internal extension of the word A assigned in a natural way as follows:

$$\begin{aligned}
\varphi(\mathtt{1}) &= \mathtt{1}\mathbf{3}\mathtt{23}\cdots\mathtt{n},\\
\varphi(\mathtt{2}) &= \mathtt{1}\mathbf{4}\mathtt{23}\cdots\mathtt{n},\\
&\vdots\\
\varphi(\mathtt{n}-\mathtt{2}) &= \mathtt{1}\mathbf{n}\mathtt{23}\cdots\mathtt{n},\\
\varphi(\mathtt{n}-\mathtt{1}) &= \mathtt{12}\mathbf{1}\mathtt{3}\cdots\mathtt{n},\\
\varphi(\mathtt{n}) &= \mathtt{12}\mathbf{4}\mathtt{3}\cdots\mathtt{n},\\
&\vdots\\
\varphi(\mathtt{N}) &= \mathtt{123}\cdots(\mathtt{n}-\mathtt{1})(\mathbf{n}-\mathbf{2})\mathtt{n}.
\end{aligned}$$

For every $\mathtt{i} \in \mathbb{A}_N$ the word $\varphi(\mathtt{i})$ has a unique factor $\mathtt{je(j+1)}$, where $\mathtt{e}$ is the inserted letter that extended the word A. Moreover, if $\mathtt{j} \neq \mathtt{1}$, then $\varphi(\mathtt{i})$ has also a unique factor $\mathtt{(j-1)je}$ and if $\mathtt{j+1} \neq \mathtt{n}$, then $\varphi(\mathtt{i})$ has a unique factor $\mathtt{e(j+1)(j+2)}$.

Let us assume that the word $\varphi(Z_N)$ contains a square XX. It is not hard to verify that for any $\mathtt{x}, \mathtt{y} \in \mathbb{A}_N\backslash\{\mathtt{1}\}$, $\mathtt{x} \neq \mathtt{y}$, the words $\varphi(\mathtt{1x})$, $\varphi(\mathtt{x1})$, $\varphi(\mathtt{1x1})$, $\varphi(\mathtt{x1y})$, $\varphi(\mathtt{1x1y})$ and $\varphi(\mathtt{x1y1})$ are squarefree. It follows that the length of XX has to be greater than $3n+2$ and XX has to contain a block $\varphi(\mathtt{u})$ for possibly the largest $\mathtt{u} \neq \mathtt{1}$. Moreover, this block is unique in XX since every factor of a Zimin word contains a unique single letter of the greatest value (in that factor). In consequence, the square XX contains a unique factor $\mathtt{je(j+1)}$, which must occupy the middle of the word XX. This fact gives us two possible cases for the form of the word X, namely

$$X = \mathtt{(j+1)}Y\mathtt{je} \quad \text{or} \quad X = \mathtt{e(j+1)}Y\mathtt{j},$$

for some nonempty word Y. We may also assume that $\mathtt{j} = 1$ or $\mathtt{j+1} = \mathtt{n}$, since otherwise one of the parts of the square XX would have to contain another unique factor, which is clearly impossible.

Let us consider the case $\mathtt{j} = \mathtt{1}$. Then we have

$$\varphi(\mathtt{u}) = \mathtt{1e23}\cdots\mathtt{n}.$$

To avoid a second unique factor, the word X has to be of the form

$$X = \mathtt{e23}\cdots\mathtt{n}B\mathtt{1}$$

for some nonempty word B, which gives

$$XX = \mathtt{e23}\cdots\mathtt{n}B\varphi(\mathtt{u})B\mathtt{1}.$$

Let us notice that XX is a factor of the word

$$\varphi(\mathtt{u})B\varphi(\mathtt{u})B\mathtt{1},$$

since the only word of form $\varphi(\mathtt{y})$ with the suffix $\mathtt{e23}\cdots(\mathtt{n}-\mathtt{1})\mathtt{n}$ is $\varphi(\mathtt{u})$. Between any two occurrences of the same letter in the Zimin word there is a letter of greater value, so the word XX contains a factor $\varphi(\mathtt{z})$ for some $\mathtt{z}$ greater than $\mathtt{u}$, and this fact creates a contradiction.

The reasoning in the case $\mathtt{j}+\mathtt{1}=\mathtt{n}$ goes analogously.

Thus we have proved that the word $\varphi(Z_N)$ is squarefree. In a similar way one may prove that the word $A\varphi(Z_N)$ is also squarefree. Inserting a single letter at one of the internal positions of the prefix A in the word $A\varphi(Z_N)$ generates a square (by the fact that Zimin words are maximal). □

The above results lead naturally to the following question.

Question 7. Is it true that there is some constant $t \geq 3$ such that every quaternary squarefree word can be extended at one of its rightmost t positions?

We can only prove that the answer is negative over a 5-letter alphabet for sufficiently large t and provided that we omit the very last position in the process of extension.

Proposition 8. *For every $t \geq 87$ there exists a squarefree word over a 5-letter alphabet which cannot be extended at any of its t rightmost internal positions.*

Proof. Let $A = \mathtt{a_1a_2}\cdots\mathtt{a_t}$ be any extremal word of length t over the alphabet $\{\mathtt{1},\mathtt{2},\mathtt{3}\}$. Consider the morphism $\alpha : \mathbb{A}_{2t-2}^* \to \mathbb{A}_5^*$ defined similarly as in the previous proof, that is, its blocks are all possible extensions of the word A by the letters $\mathtt{4}$ or $\mathtt{5}$ at the internal positions of A. Let us denote $A_i = \mathtt{4}\alpha(\mathtt{i})\mathtt{5}$ for all $i = 1, 2, \ldots, 2t-2$, and additionally $A_{2t-1} = \mathtt{4}A\mathtt{45}$. Now, consider the word W obtained as an effect of a substitution $\mathtt{i} \to A_i$ of words A_i for the corresponding letters of the Zimin word Z_{2t-1}. Finally, let us denote $S = \mathtt{4}A\mathtt{5}$ and let $P = WS$.

We claim that the word P satisfies the assertion of the proposition. Indeed, consider any extension of P at any of its t final internal positions. If the inserted letter is from the alphabet $\{\mathtt{1},\mathtt{2},\mathtt{3}\}$, then we get a square by the extremality of A. Otherwise, if the inserted letter is from $\{\mathtt{4},\mathtt{5}\}$, then the suffix S of P becomes one of the words A_i and we get a square by the structure of the Zimin word Z_{2t-1}.

It is also not hard to demonstrate that the word P is indeed squarefree, by a reasoning similar to the one in the proof of Theorem 6. □

3 The number of squarefree extensions

3.1 The squarefree potential

The problem of squarefree extensions leads to some naturally defined functions on words. For instance, given a squarefree word W over alphabet $\mathbb{A}$, let $Æ(W)$ and $æ(W)$ denote[1], respectively, the number of different squarefree extensions and the number of different *internal* squarefree extensions of W. The quantities $Æ(W)$ and $æ(W)$ will be called a *squarefree potential* and an *internal squarefree potential* of the word W, respectively. We can rephrase some definitions in the terms of squarefree potentials: the squarefree word W is *extremal* if $Æ(W) = 0$, *almost extremal* if $æ(W) = 0$, and *maximal* if $Æ(W) = æ(W)$.

Let us notice that for every squarefree word W over the alphabet $\mathbb{A}_n$, the inequality

$$Æ(W) \leq æ(W) + 2(n-1)$$

holds.

As we already know, $Æ(W) = 0$ for infinitely many squarefree ternary words. But how large can values of this function be for words of length n?

Let $\mathcal{S}_n$ denote the set of all finite squarefree words over the alphabet $\mathbb{A}_n$. Let $Æ_n(k)$ and $æ_n(k)$ be the maximum values of $Æ(W)$ and $æ(W)$ for words of length k in $\mathcal{S}_n$. Clearly, $Æ_3(k) \leq k+3$ and $æ_3(k) \leq k-1$ for all $k \geq 1$, by definition (every ternary squarefree word can be potentially extended at every internal position by just one letter, and at any of the border positions by at most two distinct letters). However, notice that the number of internal positions where such word may be extended is limited by the number of palindromic factors of the form $\mathtt{xyx}$, with $\mathtt{x}, \mathtt{y} \in \mathbb{A}$. Indeed, if $W = U\mathtt{xyx}V$ is a squarefree ternary word, then it cannot be extended at the bordering positions of $\mathtt{xyx}$, unless U or V is the empty word. Such palindromic factors occur very often in squarefree ternary words, at least once in every factor of length 7, which gives the following upper bound on the function $Æ_3(k)$.

Proposition 9. *For every sufficiently large k, we have $Æ_3(k) \leq \frac{5}{7}k + 2$.*

This upper bound can be easily improved by a more careful counting of palindromic factors in words of $\mathcal{S}_3$. For instance, by using structural observations made by Shur [14], one may improve the multiplicative constant to at least $4/7$. We omit the details since we feel that it is still far from the optimum, as suggested by Table 2. Also, any non-trivial lower bound for $Æ_3(k)$ would be interesting.

Conjecture 10. There exists a constant $\eta > 0$ such that $Æ_3(k) \geq \eta k$ holds for all sufficiently large k.

A position in a squarefree word W at which it can be extended will be called *extensible*. We will also say that W is *extensible* at that position. Let us consider the following squarefree

[1]Such designation of the functions, borrowed from the Norwegian alphabet, was chosen in order to honor Axel Thue.

k	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$æ_3(k)$	2	3	4	3	2	3	3	3	4	4	4	5	5	5	6	6
$Æ_3(k)$	6	7	6	6	6	6	6	6	6	7	7	7	8	8	8	9
k	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
$æ_3(k)$	6	7	7	7	8	8	8	9	9	9	10	10	10	11	11	11
$Æ_3(k)$	9	9	10	10	10	11	11	11	12	12	12	13	13	13	14	14
k	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
$æ_3(k)$	12	11	11	11	12	12	12	12	12	13	13	13	14	14	14	15
$Æ_3(k)$	14	15	14	15	16	15	14	15	15	15	16	16	16	17	17	17

Table 2: Values of the functions $æ_3(k)$ and $Æ_3(k)$ for $k \leq 50$.

i	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$æ(N_i)$	1	2	2	2	2	2	3	3	3	3	3	2	2	1	1	1	2	2	2
i	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
$æ(N_i)$	3	3	3	3	3	4	4	4	4	4	4	5	4	4	4	4	5	5	5

Table 3: Values of the function æ for nonchalant words N_i for $i \leq 39$.

word M of length 35

$$M = \texttt{1_213_123_132_312_321_231_213_123_132_312_321_2},$$

where the symbol "_" stands for an extensible position in M. It is worth noticing that for every $k \in \{7, 8, \ldots, 35\}$, the prefix M_k of length k of the word M satisfies $æ(M_k) = æ_3(k)$. Moreover, the internal positions at which the words M_k are extensible coincide with internal extensible positions of the word M.

Analogously, one may consider the minimum values of the functions æ and Æ for words of given length k. Theorem 1 implies that over the alphabet $\{\texttt{1}, \texttt{2}, \texttt{3}\}$ these values are equal to 0 for infinitely many k. In the case of a 4-letter alphabet, minimum values for words of small lengths seem to support Conjecture 2.

It is also interesting to look for the values of the function Æ or æ for the nonchalant words N_i. For instance, our numerical experiments shows that in the first 1000 iterations of the procedure, the values of $æ(N_i)$ are "gradually increasing" in the sense that when the new maximal value is obtained for the nonchalant word N_i, then the value of æ for the following words is never less than $æ(N_i) - 2$ (more details are presented in Tables 3 and 4).

This fact led us to the following version of Conjecture 3.

Conjecture 11. If $æ(N_1) > 2$ for the starting word N_1 of the nonchalant algorithm, then the respective sequence of nonchalant words is infinite.

i	2	3	8	26	32	40	46	64	79	100	108	111	117	135	172
æ(N_i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
i	175	183	189	222	243	251	254	260	279	286	314	338	346	352	370
æ(N_i)	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
i	385	406	414	417	423	445	469	477	489	496	524	548	556	562	580
æ(N_i)	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
i	595	616	624	627	633	655	687	706	712	737	740	743	764	779	800
æ(N_i)	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
i	808	811	817	835	850	872	875	878	881	902	917	938	967	973	997
æ(N_i)	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75

Table 4: Indices $i < 1000$ for which the nonchalant procedure gave new maximum values of æ(N_i).

3.2 The squarefree potential of Zimin words

Let us consider the squarefree potential of Zimin words Z_n. Since Zimin words are non-extensible on the external positions, we have $Æ(Z_n) = æ(Z_n)$. It is easy to verify that $Æ(Z_1) = Æ(Z_2) = 0$ and $Æ(Z_3) = 2$. Let $n \geq 4$. Since $Z_n = Z_{n-1}\mathtt{n}Z_{n-1}$, we get that the only extensions of Z_n by the letters $\mathtt{1}, \mathtt{2}, \ldots, \mathtt{n-1}$ are those induced by the prefix or suffix Z_{n-1} (the letter $\mathtt{n}$ in the center of the word Z_n occurs only once). Thus

$$Æ(Z_n) = 2 \cdot Æ(Z_{n-1}) + t(n),$$

where $t(n)$ stands for the number of different squarefree extensions of the word Z_n by inserting the letter $\mathtt{n}$. Let us notice that the extension of Z_n by the letter $\mathtt{n}$ would generate a square if and only if we insert this letter right after the last appearance of any other letter in the prefix $Z_{n-1}\mathtt{n}$ or, analogously, right before the first appearance of any other letter in the suffix $\mathtt{n}Z_{n-1}$. Since there are $2^n - 2$ internal positions in the word Z_n, we have

$$t(n) = (2^n - 2) - 2(n-1).$$

This leads us to the following statement.

Proposition 12. *Let Z_n be a Zimin word over the alphabet $\mathbb{A}_n$. Then $Æ(Z_1) = Æ(Z_2) = 0$, $Æ(Z_3) = 2$, and*

$$Æ(Z_n) = 2^n - 2n + 2 \cdot Æ(Z_{n-1}), \tag{3}$$

for all $n \geq 4$.

Before we came up with the presented formula, our computer calculations gave us the following sequence of the squarefree potentials of Zimin words:

$$0, 0, 2, 12, 46, 144, 402.$$

Excluding the first two 0's, there is only one sequence A006742 in Sloane's *OEIS* [15] which is initialized by such integers. The sequence is described as *series for second perpendicular moment of hexagonal lattice*, which sounds rather distant from combinatorics on words. Needless to say, the next term of the *OEIS* sequence is equal to 1040 and, as it turned out, the next term of our sequence is equal to 1044. Thus, for a brief moment, the authors became victims of the well known *Strong Law of Small Numbers*.

4 Final discussion

Let us conclude the paper with some remarks and more general open problems.

First notice that one may consider extremal words and nonchalant words with respect to any *avoidable pattern*. For instance, a close relative to the notion of the square is that of the *overlap*, which is a word consisting of two identical intersecting factors. Equivalently, this is any word of the form $\mathtt{x}B\mathtt{x}B\mathtt{x}$, where $\mathtt{x}$ is a single letter and B is an arbitrary word. A word is *overlap-free* if none of its factors is an overlap. Such word is called *extremal* (over a fixed alphabet) if any of its extensions (by a single letter) contains an overlap.

A classical result of Thue [17] asserts that there exist arbitrarily long overlap-free binary words. Recently, Mol, Rampersad, and Shallit [12] proved that among them there are infinitely many extremal ones. Moreover, they determined exactly the set of possible lengths of such extremal overlap-free words, which, unlike in the case of squares, does not contain all sufficiently large integers. In analogy to the case of squarefree words one may ask the following questions.

Question 13. Is there any extremal overlap-free word over a 3-letter alphabet?

Question 14. Is the sequence of nonchalant overlap-free binary words infinite?

Another generalization of squares is that of *k-powers*, which are words of the form $XX\cdots X$ consisting of k copies of any nonempty word X. By the mentioned result of Thue [17], there exist infinitely many *cube-free* words over a 2-letter alphabet. Is the sequence of cube-free nonchalant *binary* words infinite? Is it true that every cube-free *ternary* word is extensible?

To state a general conjecture let us recall briefly some basic notions of pattern avoidance as introduced independently by Bean, Ehrenfeucht, and McNulty [2] and Zimin [18]. Let $\mathbb{V}$ be an alphabet of variables. A *pattern* $P = p_1p_2\cdots p_r$, with $p_i \in \mathbb{V}$, is any nonempty word over $\mathbb{V}$. A word W *realizes* a pattern P if it can be split into nonempty factors $W = W_1W_2\cdots W_r$ so that $W_i = W_j$ if and only if $p_i = p_j$, for all $i, j = 1, 2, \ldots, r$. A word W *avoids* a pattern P if no factor of W realizes P. For instance, a squarefree word avoids a pattern $P = xx$. The pattern P is *avoidable* if there exist arbitrarily long words avoiding P over some finite alphabet. The least size of such an alphabet is denoted by $\mu(P)$ and called the *avoidability index* of P. For instance, $\mu(xx) = 3$. A complete characterization of avoidable patterns was provided independently by Zimin [18] and Bean, Ehrenfeucht, and McNulty [2].

Now, given a fixed pattern P, we may define *extremal P-avoiding* words and *P-nonchalant words* analogously as in the case of squares. The following conjectures are worth consideration.

Conjecture 15. For every avoidable pattern P, there are no extremal P-avoiding words over an $(\mu(P)+1)$-letter alphabet.

Conjecture 16. For every avoidable pattern P and any integer $n \geq \mu(P)$, the sequence of P-nonchalant words over $\mathbb{A}_n$ is infinite and converges to a unique infinite word $\mathcal{N}_n^{(P)}$.

The avoidability index $\mu(P)$ is quite a mysterious parameter. A pattern P with $\mu(P) = 4$ was found already by Bean, Ehrenfeucht, and McNulty in [2]. A current record is only slightly better and belongs to Clark [5], who found a pattern P with $\mu(P) = 5$. It is not known if there exist patterns with avoidability index six or more. Is it possible that every avoidable pattern can be avoided over just 5-letter alphabet?

Actually, one may consider similar problems with respect to any reasonable "property" of words. For instance, Ter-Saakov and Zhang [13] considered extremal words avoiding *abelian squares* (these are words of the form XY, where Y is a permutation of X). They conjecture that there are infinitely many such words over a 4-letter alphabet. It is however not true that every abelian squarefree word over a sufficiently large alphabet is extensible. Indeed, every Zimin word Z_n is extremal. This implies that the nonchalant procedure (starting from the empty word) stops over any finite alphabet.

5 Acknowledgments

We would like to thank the anonymous referee for careful reading of the manuscript and many valuable comments and suggestions.

A 10000 iterations of the nonchalant procedure

We consider a 3-letter alphabet. Let p be the number of positions that the procedure moved back in the i-th iteration. Tables 5 and 6 contain information about the first 10000 iterations of the procedure with initial word `1`. For example, in the seventh iteration, the 8-letter long word was obtained by inserting a single letter in the penultimate position of the previous, 7-letter long, word. The three iterations with the biggest number of positions moved back are bolded (in fact, these are the only iterations in which the procedure moved back more than 9 positions (cf. Table 1).

Table 7 contains an example of more detailed common results for various initial words. Let us focus on the distances between consecutive iterations, in which the procedure moved back by exactly four positions (the number 4 was chosen arbitrarily). We present the number of occurrences of such distances with respect to various initial words (for each initial word we, again, analyze the first 10000 iterations of the procedure). For example, for initial word `1`,

procedure moves back by four positions in iteration 480. It happens again 659 iterations later, i.e., in iteration 1139 (Table 5). Such number of steps between two consecutive iterations does not happen again, so for the initial word 1 we have a number 1 in the column started by 659. As we can see, the most common distances among considered iterations are about 210-211 steps.

i	p	i	p	i	p	i	p	i	p	i	p	i	p	i	p
7	1	640	1	1307	2	1965	1	2592	4	3253	1	3861	1	**4436**	**20**
25	2	648	2	1338	1	1986	1	2625	1	3256	2	3883	2	4453	1
32	1	676	1	1349	7	1994	2	2657	1	3279	4	3890	2	4485	1
64	1	698	2	1382	1	2025	1	2662	1	3312	1	3921	1	4490	1
69	1	705	2	1387	1	2036	4	2665	2	3344	1	3953	1	4493	2
72	2	764	1	1390	2	2040	7	2696	1	3349	1	3958	1	4524	1
103	1	769	1	1421	1	2067	2	2728	1	3352	2	3961	2	4556	1
135	1	772	2	1453	1	2074	2	2730	2	3383	1	3992	1	4558	2
140	1	803	1	1458	1	2105	1	2732	2	3415	1	4013	1	4560	2
143	**15**	835	1	1461	2	2137	1	2761	2	3417	9	4021	2	4589	2
144	2	840	1	1484	4	2142	1	2792	1	3438	2	4052	1	4620	1
175	1	843	2	1517	1	2145	2	2824	1	3445	2	4063	4	4652	1
207	1	902	1	1549	1	2176	1	2829	1	3476	1	4093	2	4657	1
212	1	907	1	1554	1	2197	1	2832	2	3508	1	4100	2	4660	2
215	2	910	2	1557	2	2205	2	2863	1	3513	1	4131	1	4691	1
246	1	931	2	1588	1	2236	1	2884	1	3516	2	4163	1	4712	1
270	4	959	2	1620	1	2247	4	2892	2	3547	1	4168	1	4720	2
300	2	966	2	1622	2	2277	2	2923	1	3568	1	4171	2	4751	1
307	2	997	1	1624	2	2284	2	2934	4	3576	2	4202	1	4762	4
338	1	1029	1	1653	2	2315	1	2964	2	3607	1	4223	1	4792	2
370	1	1034	1	1684	1	2347	1	2971	2	3618	4	4231	2	4799	2
375	1	1037	2	1716	1	2352	1	3002	1	3648	2	4262	1	4830	1
378	2	1068	1	1721	1	2355	2	3034	1	3655	2	4273	4	4862	1
409	1	1089	1	1724	2	2386	1	3039	1	3686	1	4274	1	4867	1
430	1	1097	2	1755	1	2407	1	3042	2	3718	1	4275	1	4870	2
438	2	1128	1	1776	1	2415	2	3073	1	3723	1	4278	2	4901	1
469	1	1139	4	1784	2	2446	1	3094	1	3726	2	4296	2	4922	1
480	4	1169	2	1815	1	2457	7	3102	2	3757	1	4303	2	4930	2
510	2	1176	2	1826	4	2490	1	3133	1	3778	1	4342	1	4961	1
517	2	1207	1	1856	2	2495	1	3144	7	3786	2	4347	1	4972	4
548	1	1239	1	1863	2	2498	2	3177	1	3817	1	4350	2	4976	7
580	1	1244	1	1894	1	2529	1	3182	1	3828	4	4381	1		
585	1	1247	2	1926	1	2561	1	3185	2	3829	1	4413	1		
588	2	1278	1	1931	1	2566	1	3216	1	3830	1	4418	1		
619	1	1299	1	1934	2	2569	2	3248	1	3833	2	4421	2		

Table 5: Nonzero positions moved back by the nonchalant procedure in iterations 1-5000.

i	p	i	p	i	p	i	p	i	p	i	p	i	p	i	p
5003	2	5632	1	6249	2	6852	2	7483	2	8144	1	8835	1	9485	2
5010	2	5664	1	6280	1	6883	1	7514	1	8147	2	8843	2	9516	1
5041	1	5666	2	6291	4	6904	1	7535	1	8178	1	8874	1	9527	4
5073	1	5668	2	6321	2	6912	2	7543	2	8205	2	8885	4	9557	2
5078	1	5697	2	6328	2	6943	1	7574	1	8238	1	8915	2	9564	2
5081	2	5728	1	6359	1	6954	4	7585	7	8261	3	8922	2	9595	1
5112	1	5760	1	6391	1	6984	2	7618	1	8267	2	8953	1	9627	1
5133	1	5765	1	6396	1	6991	2	7623	1	8298	1	8985	1	9632	1
5141	2	5768	2	6399	2	7022	1	7626	2	8330	1	8990	1	9635	2
5172	1	5799	1	6430	1	7054	1	7657	1	8335	1	8993	2	9666	1
5183	4	5820	1	6451	1	7059	1	7689	4	8338	2	9024	1	9687	1
5213	2	5828	2	6459	2	7062	2	7716	2	8369	1	9045	1	9695	2
5220	2	5859	1	6490	1	7093	1	7723	2	8390	1	9053	2	9726	1
5251	1	5870	4	6501	7	7114	1	7754	1	8398	2	9084	1	9737	4
5283	1	5900	2	6534	1	7122	2	7786	1	8429	1	9123	2	9738	1
5288	1	5907	2	6539	1	7153	1	7791	1	8440	4	9130	2	9739	1
5291	2	5938	1	6542	2	7164	4	7794	2	8470	2	9161	1	9742	2
5322	1	5970	1	6573	1	7168	7	7825	1	8477	2	9193	1	9770	1
5343	1	5975	1	6605	1	7195	2	7846	1	8508	1	9198	1	9792	2
5351	2	5978	2	6610	1	7202	2	7854	2	8540	1	9201	2	9799	2
5382	1	6009	1	6613	2	7233	1	7885	1	8545	1	9232	1	9830	1
5393	7	6030	1	**6628**	**20**	7265	1	7896	4	8548	2	9253	1	9862	1
5426	1	6038	2	6645	1	7270	1	7926	2	8579	1	9261	2	9867	1
5431	1	6069	1	6677	1	7273	2	7933	2	8600	1	9292	1	9870	2
5434	2	6080	4	6682	1	7304	1	7964	1	8608	2	9294	1	9901	1
5465	1	6084	7	6685	2	7325	1	7996	1	8639	1	9328	4	9922	1
5497	1	6111	2	6716	1	7333	2	8001	1	8650	4	9347	2	9930	2
5502	1	6118	2	6748	1	7364	1	8004	2	8705	2	9354	2	9961	1
5505	2	6149	1	6750	2	7375	4	8035	1	8712	2	9385	1	9972	4
5528	4	6181	1	6752	2	7405	2	8056	1	8743	1	9417	1		
5561	1	6186	1	6781	2	7412	2	8064	2	8775	1	9422	1		
5593	1	6189	2	6812	1	7443	1	8095	1	8780	1	9425	2		
5598	1	6220	1	6844	1	7475	1	8106	7	8783	2	9456	1		
5601	2	6241	1	6849	1	7480	1	8139	1	8814	1	9477	1		

Table 6: Nonzero positions moved back by the nonchalant procedure in iterations 5001-10000.

N_1	199	207	210	211	233	235	314	339	342	345	443	460
1	1	1	9	4	0	3	1	1	3	4	1	0
2	2	1	8	5	0	3	1	0	3	4	1	1
3	1	1	8	6	0	1	1	1	5	6	0	1
13	1	1	8	6	0	1	1	1	5	6	0	1
23	2	1	9	5	0	3	1	0	3	4	1	1
32	2	1	8	5	0	3	1	0	3	4	1	1
3213	1	1	8	6	0	1	1	1	5	6	0	1
2313213	1	1	8	6	0	1	1	1	5	6	0	1
32132313213	1	1	8	6	0	1	1	1	5	6	0	1

N_1	489	544	659	663	688	806
1	1	1	1	1	0	0
2	0	1	0	1	1	0
3	1	0	0	1	0	0
13	1	0	0	1	0	0
23	1	1	0	1	0	0
32	0	1	0	1	1	0
3213	1	0	0	1	0	0
2313213	1	0	0	1	0	0
32132313213	1	0	0	1	0	0

Table 7: Numbers of occurrences of distances (first row) between two consecutive iterations in which the nonchalant procedure moved back by exactly four positions, for various initial words (for 10000 iterations of the procedure).

References

[1] J.-P. Allouche, J. Shallit, *Automatic Sequences: Theory, Applications, Generalizations*, Cambridge University Press, 2003.

[2] D. R. Bean, A. Ehrenfeucht, and G. F. McNulty, Avoidable patterns in strings of symbols, *Pacific J. Math.* **85** (1979), 261–294.

[3] J. Berstel, Axel Thue's papers on repetitions in words: a translation, *Publications du LaCIM* **20**, Université du Québec à Montréal, 1995.

[4] J. Berstel, D. Perrin, The origins of combinatorics on words, *Europ. J. Combin.* **28** (2007), 996–1022.

[5] R. J. Clark, The existence of a pattern which is 5-avoidable but 4-unavoidable, *Int. J. Alg. Comput.* **16** (2006), 351–367.

[6] J. Currie, Pattern avoidance: themes and variations, *Theoret. Comput. Sci.* **339** (2005), 7–18.

[7] J. Grytczuk, H. Kordulewski, and A. Niewiadomski, Extremal squarefree words, *Electron. J. Combin.* **27** (2020), Paper P1.48.

[8] L. Hong and S. Zhang, No extremal squarefree words over large alphabets, preprint, September 2021. Available at https://arxiv.org/abs/2107.13123.

[9] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, 1983.

[10] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press, 2002.

[11] L. Mol and N. Rampersad, Lengths of extremal squarefree ternary words, *Contrib. Discrete Math.* **16** (2020), 8–19.

[12] L. Mol, N. Rampersad, and J. Shallit, Extremal overlap-free and extremal β-free binary words, *Electron. J. Combin.* **27** (2020), Paper P4.42.

[13] N. Ter-Saakov and E. Zhang, Extremal pattern-avoiding words, preprint, 2020. Available at https://arxiv.org/abs/2009.10186.

[14] A. M. Shur, On ternary square-free circular words, *Electron. J. Comb.* **17** (2010), Paper R1.40.

[15] N. J. A. Sloane et al., *The On-Line Encyclopedia of Integer Sequences*, published electronically at https://oeis.org.

[16] A. Thue, Über unendliche Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.

[17] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 413–478.

[18] A. I. Zimin, Blocking sets of terms, *Mat. Sb.* **119** (1982), 363–375. Translated in *Sb. Math.* **47** (1984), 353–364.

2020 *Mathematics Subject Classification*: 68R15.
Keywords: squarefree word, extremal word, Zimin word, word extension.

(Concerned with sequence A006742.)

Received May 1 2021; revised versions received May 4 2021; August 17 2021; August 30 2021. Published in *Journal of Integer Sequences*, September 23 2021.

Return to Journal of Integer Sequences home page.