On Number of Rich Words

Josef Rukavicka*

January 25, 2017
Mathematics Subject Classification: 68R15

Abstract

Any finite word w of length n contains at most $n+1$ distinct palindromic factors. If the bound $n+1$ is reached, the word w is called rich. The number of rich words of length n over an alphabet of cardinality q is denoted $R_n(q)$. For binary alphabet, Rubinchik and Shur deduced that $R_n(2) \leq c1.605^n$ for some constant c. We prove that $\lim_{n\to\infty} \sqrt[n]{R_n(q)} = 1$ for any q, i.e. $R_n(q)$ has a subexponential growth on any alphabet.

1 Introduction

The study of palindromes is a frequent topic and many diverse results may be found. In recent years, some of the papers deal with so-called *rich* words, or also words having *palindromic defect* 0. They are words that have the maximum number of palindromic factors. As noted by [6], a finite word w can contains at most $|w|+1$ distinct palindromic factors with $|w|$ being the length of w. The rich words are exactly those that attain this bound. It is known that on binary alphabet the set of rich words contains factors of Sturmian words, factors of complementary symmetric Rote words, factors of the period-doubling word, etc., see [6, 4, 1, 13]. On multiliteral alphabet, the set of rich words contains for example factors of Arnoux–Rauzy words and factors of words coding symmetric interval exchange.

*Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CZECH TECHNICAL UNIVERSITY IN PRAGUE (josef.rukavicka@seznam.cz).

Rich words can be characterized using various properties, see for instance [8, 5, 2]. The concept of rich words can also be generalized to respect so-called pseudopalindromes, see [10]. In this paper we focus on an unsolved question of computing the number of rich words of length n over an alphabet with $q > 1$ letters. This number is denoted $R_n(q)$.

This question is investigated in [15], where J. Vesti gives a recursive lower bound on the number of rich words of length n, and an upper bound on the number of binary rich words. Both these estimates seem to be very rough. In [9], C. Guo, J. Shallit and A.M. Shur constructed for each n a large set of rich words of length n. Their construction gives, currently, the best lower bound on the number of binary rich words, namely $R_n(2) \geq \frac{C^{\sqrt{n}}}{p(n)}$, where $p(n)$ is a polynomial and the constant $C \approx 37$. On the other hand, the best known upper bound is exponential. As mentioned in [9], calculation performed recently by M. Rubinchik provides the upper bound $R_n(2) \leq c1.605^n$ for some constant c, see [11].

Our main result stated as Theorem 4.3 shows that $R_n(q)$ has a subexponential growth on any alphabet. More precisely, we prove

$$\lim_{n\to\infty} \sqrt[n]{R_n(q)} = 1\,.$$

In [14], Shur calls languages with the above property small. Our result is an argument in favor of a conjecture formulated in [9] saying that for some infinitely growing function $g(n)$ the following holds true $R_n(2) = \mathcal{O}\Big(\frac{n}{g(n)}\Big)^{\sqrt{n}}$.

To derive our result we consider a specific factorization of a rich word into distinct rich palindromes, here called UPS-factorization (Unioccurrent Palindromic Suffix factorization), see Definition 3.2. Let us mention that another palindromic factorizations have already been studied, see [3, 7]: Minimal (minimal number of palindromes), maximal (every palindrome cannot be extended on the given position) and diverse (all palindromes are distinct). Note that only the minimal palindromic factorization has to exist for every word.

The article is organized as follows: Section 2 recalls notation and known results. In Section 3 we study a relevant property of UPS-factorization. The last section is devoted to the proof of our main result.

2 Preliminaries

Let us start with a couple of definitions: Let A be an alphabet of q letters, where $q > 1$ and $q \in \mathbb{N}$ ($\mathbb{N}$ denotes the set of nonnegative integers). A finite sequence $u_1u_2\cdots u_n$ with $u_i \in A$ is a *finite word*. Its length is n and is denoted $|u_1u_2\cdots u_n| = n$. Let A^n denote the set of words of length n. We define that A^0 contains just the empty word. It is clear that the size of A^n is equal to q^n.
Given $u = u_1u_2\cdots u_n \in A^n$ and $v = v_1v_2\cdots v_k \in A^k$ with $0 \le k \le n$, we say that v is a *factor* of u if there exists i such that $0 < i$, $i + k \le n$ and $u_i = v_1$, $u_{i+1} = v_2$, ..., $u_{i+k-1} = v_k$.
A word $u = u_1u_2\cdots u_n$ is called a *palindrome* if $u_1u_2\cdots u_n = u_nu_{n-1}\cdots u_1$. The empty word is considered to be a palindrome and a factor of any word.

A word u of length n is called *rich* if u has $n + 1$ distinct palindromic factors. Clearly, $u = u_1u_2\cdots u_n$ is rich if and only if its *reversal* $u_nu_{n-1}\cdots u_1$ is rich as well.

Any factor of a rich word is rich as well, see [8]. In other words, the language of rich words is factorial. In particular it means that $R_n(q)R_m(q) \le R_{n+m}(q)$ for any $m, n, q \in \mathbb{N}$. Therefore, the Fekete's lemma implies existence of the limit of $\sqrt[n]{R_n(q)}$ and moreover

$$\lim_{n\to\infty} \sqrt[n]{R_n(q)} = \inf\left\{ \sqrt[n]{R_n(q)} \colon n \in \mathbb{N}\right\}.$$

For a fixed n_0, one can find the number of all rich words of length n_0 and obtain an upper bound on the limit. Using computer Rubinchik counted $R_n(2)$ for $n \le 60$, (see the sequence A216264 in OEIS). As $\sqrt[60]{R_{60}(2)} < 1.605$, he obtained the upper bound given in Introduction.

As shown in [8], any rich word u over alphabet A is richly prolongable, i.e., there exist letters $a, b \in A$ such that aub is also rich. Thus a rich word is a factor of an arbitrarily long rich word. But the question whether two rich words can appear simultaneously as factors of a longer rich word may have negative answer. It means that the language of rich words is not recurrent. This fact makes enumeration of rich words hard.

3 Factorization of rich words into rich palindromes

Let us recall one important property of rich words [6, Definition 4 and Proposition 3]: the longest palindromic suffix of a rich word w has exactly one occurrence in w (we say that the longest palindromic suffix of w is *unioccurrent* in w). It implies that $w = w^{(1)}w_1$, where w_1 is a palindrome which is not a factor of $w^{(1)}$. Since every factor of a rich word is a rich word as well, it follows that $w^{(1)}$ is a rich word and thus $w^{(1)} = w^{(2)}w_2$, where w_2 is a palindrome which is not a factor of $w^{(2)}$. Obviously $w_1 \neq w_2$. We can repeat the process until $w^{(p)}$ is the empty word for some $p \in \mathbb{N}$, $p \geq 1$. We express these ideas by the following lemma:

Lemma 3.1. *Let w be a rich word. There exist distinct non-empty palindromes $w_1, w_2, \ldots, w_p$ such that*

$$w = w_p w_{p-1} \cdots w_2 w_1 \text{ and } w_i \text{ is the longest palindromic suffix of } w_p w_{p-1} \cdots w_i \text{ for } i = 1, 2, \ldots, p. \quad (1)$$

Definition 3.2. *We define UPS-factorization (Unioccurrent Palindromic Suffix factorization) to be the factorization of a rich word w into the form (1).*

Since w_i in the factorization (1) are non-empty, it is clear that $p \leq n = |w|$. From the fact that the palindromes w_i in the factorization (1) are distinct we can derive a better upper bound for p. The aim of this section is to prove the following theorem:

Theorem 3.3. *There is a constant $c > 1$ such that for any rich word w of length n the number of palindromes in the UPS-factorization of $w = w_p w_{p-1} \cdots w_2 w_1$ satisfies*

$$p \leq c\frac{n}{\ln n}. \quad (2)$$

Before proving the theorem, we show two auxiliary lemmas:

Lemma 3.4. *Let $q, n, t \in \mathbb{N}$ such that*

$$\sum_{i=1}^{t} i q^{\lceil \frac{i}{2} \rceil} \geq n. \quad (3)$$

The number p of palindromes in the UPS-factorization $w = w_p w_{p-1} \dots w_2 w_1$ of any rich word w with $n = |w|$ satisfies

$$p \leq \sum_{i=1}^{t} q^{\lceil \frac{i}{2} \rceil}. \tag{4}$$

Proof. Let $f_1, f_2, f_3, \dots$ be an infinite sequence of all non-empty palindromes over an alphabet A with $q = |A|$ letters, where the palindromes are ordered in such a way that $i < j$ implies that $|f_i| \leq |f_j|$. In consequence $f_1, \dots, f_q$ are palindromes of length 1, $f_{q+1}, \dots, f_{2q}$ are palindromes of length 2, etc. Since $w_1, \dots, w_p$ are distinct non-empty palindromes we have $\sum_{i=1}^{p} |f_i| \leq \sum_{i=1}^{p} |w_i| = n$. The number of palindromes of length i over the alphabet A with q letters is equal to $q^{\lceil \frac{i}{2} \rceil}$ (just consider that that the "first half" of a palindrome determines the second half). The number $\sum_{i=1}^{t} i q^{\lceil \frac{i}{2} \rceil}$ equals the length of a word concatenated from all palindromes of length less than or equal to t. Since $\sum_{i=1}^{p} |f_i| \leq n \leq \sum_{i=1}^{t} i q^{\lceil \frac{i}{2} \rceil}$, it follows that the number of palindromes p is less than or equal to the number of all palindromes of length at most t; this explains the inequality (4). □

Lemma 3.5. *Let $N \in \mathbb{N}$, $x \in \mathbb{R}$, $x > 1$ such that $N(x-1) \geq 2$. We have*

$$\frac{N x^N}{2(x-1)} \leq \sum_{i=1}^{N} i x^{i-1} \leq \frac{N x^N}{(x-1)}. \tag{5}$$

Proof. The sum of the first N terms of a geometric series with the quotient x is equal to $\sum_{i=1}^{N} x^i = \frac{x^{N+1}-x}{x-1}$. Taking the derivative of this formula with respect to x with $x > 1$ we obtain: $\sum_{i=1}^{N} i x^{i-1} = \frac{x^N(N(x-1)-1)+1}{(x-1)^2} = \frac{N x^N}{x-1} + \frac{1-x^N}{(x-1)^2}$. It follows that the right inequality of (5) holds for all $N \in \mathbb{N}$ and $x > 1$. The condition $N(x-1) \geq 2$ implies that $\frac{1}{2} N(x-1) \leq N(x-1) - 1$, which explains the left inequality of (5). □

We can start the proof of Theorem 3.3:

Proof of Theorem 3.3. Let $t \in \mathbb{N}$ be a minimal nonnegative integer such that the inequality (3) in Lemma 3.4 holds. It means that:

$$n > \sum_{i=1}^{t-1} i q^{\lceil \frac{i}{2} \rceil} \geq \sum_{i=1}^{t-1} i q^{\frac{i}{2}} = q^{\frac{1}{2}} \sum_{i=1}^{t-1} i q^{\frac{i-1}{2}} \geq \frac{(t-1) q^{\frac{t}{2}}}{2(q^{\frac{1}{2}} - 1)}, \tag{6}$$

where for the last inequality we exploited (5) with $N = t-1$ and $x = q^{\frac{1}{2}}$. If $q \geq 9$, then the condition $N(x-1) = (t-1)(q^{\frac{1}{2}}-1) \geq 2$ is fulfilled (it is the condition from Lemma 3.5) for any $t \geq 2$. Hence let us suppose that $q \geq 9$ and $t \geq 2$. From (6) we obtain:

$$\frac{q^{\frac{t}{2}}}{q^{\frac{1}{2}}-1} \leq \frac{2n}{t-1} \leq \frac{4n}{t}. \tag{7}$$

Since t is such that the inequality (3) holds and $i \leq q^{\frac{i+1}{2}}$ for any $i \in \mathbb{N}$ and $q \geq 2$, we can write:

$$n \leq \sum_{i=1}^{t} i q^{\frac{i+1}{2}} \leq \sum_{i=1}^{t} q^{i+1} = q^2 \frac{q^t-1}{q-1} \leq \frac{q^2}{q-1} q^t \leq q^{2t}. \tag{8}$$

We apply a logarithm on the previous inequality:

$$\ln n \leq 2t \ln q. \tag{9}$$

An upper bound for the number of palindromes p in UPS-factorization follows from (4), (7), and (9):

$$p \leq \sum_{i=1}^{t} q^{\lceil \frac{i}{2} \rceil} \leq \sum_{i=1}^{t} q^{\frac{i+1}{2}} \leq q^{\frac{3}{2}} \frac{q^{\frac{t}{2}}}{q^{\frac{1}{2}}-1} \leq q^{\frac{3}{2}} \frac{4n}{t} \leq q^{\frac{3}{2}} 8 \ln q \frac{n}{\ln n}. \tag{10}$$

The previous inequality supposes that $q \geq 9$ and $t \geq 2$. If $t = 1$ then we can easily derive from (3) that $n \leq q$ and consequently $p \leq n \leq q$. Thus the inequality $p \leq q^{\frac{3}{2}} 8 \ln q \frac{n}{\ln n}$ holds as well for this case. Since every rich word over an alphabet with the cardinality $q < 9$ is also a rich word over the alphabet with the cardinality 9, the estimate (2) in Theorem 3.3 holds if we set the constant c as follows: $c = \max\{8q^{\frac{3}{2}} \ln q, 8 \cdot 9^{\frac{3}{2}} \ln 9\}$. □

Remark 3.6. Theorem 3.3 implies that average length of a palindrome of UPS-factorization of a rich word of length n is $\mathcal{O}(\ln(n))$. Note that in [12] it is shown that most of palindromic factors of a random word of length n are of length close to $\ln(n)$.

4 Rich words form a small language

The aim of this section is to show that the set of rich words forms a small language, see Theorem 4.3.

We present a recurrent inequality for $R_n(q)$. To ease our notation we omit the specification of the cardinality of alphabet and write R_n instead of $R_n(q)$.

Denote $\kappa_n = \left\lceil c\frac{n}{\ln n}\right\rceil$, where c is the constant from Theorem 3.3 and $n \geq 2$.

Theorem 4.1. *Let $n \geq 2$, then*

$$R_n \leq \sum_{p=1}^{\kappa_n} \sum_{\substack{n_1+n_2+\cdots+n_p=n \\ n_1,n_2,\ldots,n_p\geq 1}} R_{\lceil \frac{n_1}{2}\rceil} R_{\lceil \frac{n_2}{2}\rceil} \ldots R_{\lceil \frac{n_p}{2}\rceil}. \tag{11}$$

Proof. Given $p, n_1, n_2, \ldots, n_p$, let $R(n_1, n_2, \ldots, n_p)$ denote the number of rich words with UPS-factorization $w = w_p w_{p-1} \ldots w_1$, where $|w_i| = n_i$ for $i = 1, 2, \ldots, p$. Note that any palindrome w_i is uniquely determined by its prefix of length $\lceil \frac{n_i}{2}\rceil$; obviously this prefix is rich. Hence the number of words that appears in UPS-factorization as w_i cannot be larger than $R_{\lceil \frac{n_i}{2}\rceil}$. It follows that $R(n_p, n_{p-1}, \ldots, n_1) \leq R_{\lceil \frac{n_1}{2}\rceil} R_{\lceil \frac{n_2}{2}\rceil} \ldots R_{\lceil \frac{n_p}{2}\rceil}$. The sum of this result over all possible p (see Theorem 3.3) and $n_1, n_2, \ldots, n_p$ completes the proof. □

Proposition 4.2. *If $h > 1, K \geq 1$ such that $R_n \leq Kh^n$ for all n, then $\lim\limits_{n\to\infty} \sqrt[n]{R_n} \leq \sqrt{h}$.*

Proof. For any integers $p, n_1, \ldots, n_p \geq 1$, the assumption implies that $R_{\lceil \frac{n_1}{2}\rceil} R_{\lceil \frac{n_2}{2}\rceil} \ldots R_{\lceil \frac{n_p}{2}\rceil} \leq K^p h^{\frac{n_1+1}{2}} h^{\frac{n_2+1}{2}} \ldots h^{\frac{n_p+1}{2}} \leq K^p h^{\frac{n+p}{2}}$. Exploiting (11) we obtain:

$$R_n \leq K^{\kappa_n} h^{\frac{n+\kappa_n}{2}} \sum_{p=1}^{\kappa_n} \sum_{\substack{n_1+n_2+\cdots+n_p=n \\ n_1,n_2,\ldots,n_p\geq 1}} 1. \tag{12}$$

The sum

$$S_n = \sum_{\substack{n_1+n_2+\cdots+n_p=n \\ n_1,n_2,\ldots,n_p\geq 1}} 1$$

can be interpreted as the number of ways how to distribute n coins between p people in such a way that everyone has at least one coin. That is why

$S_n = \binom{n-1}{p-1}$.
It is known (see Appendix for the proof) that

$$\sum_{i=0}^{L} \binom{N}{i} \leq \left(\frac{eN}{L}\right)^L, \text{ for any } L, N \in \mathbb{N} \text{ and } L \leq N. \tag{13}$$

From (12) we can write: $R_n \leq K^{\kappa_n} h^{\frac{n+\kappa_n}{2}} \binom{en}{\kappa_n}^{\kappa_n}$. To evaluate $\sqrt[n]{R_n}$, just recall that $\lim\limits_{n\to\infty} (const)^{\frac{\kappa_n}{n}} = \lim\limits_{n\to\infty} (const)^{\frac{c}{\ln n}} = 1$ for any constant *const* and moreover $\lim\limits_{n\to\infty} \left(\frac{n}{\kappa_n}\right)^{\frac{\kappa_n}{n}} = \lim\limits_{n\to\infty} (c \ln n)^{\frac{1}{c \ln n}} = 1$. □

The main theorem of this paper is a simple consequence of the previous proposition.

Theorem 4.3. *Let R_n denote the number of rich words of length n over an alphabet with q letters. We have $\lim\limits_{n\to\infty} \sqrt[n]{R_n} = 1$.*

Proof. Let us suppose that $\lim_{n\to\infty} \sqrt[n]{R_n} = \lambda > 1$. We are going to find $\epsilon > 0$ such that $\lambda + \epsilon < \lambda^2$. The definition of a limit implies that there is n_0 such that $\sqrt[n]{R_n} < \lambda + \epsilon$ for any $n > n_0$, i.e. $R_n < (\lambda + \epsilon)^n$. Let $K = \max\{R_1, R_2, \ldots, R_{n_0}\}$. It holds for any $n \in \mathbb{N}$ that $R_n \leq K(\lambda + \epsilon)^n$. Using Proposition 4.2 we obtain $\lim\limits_{n\to\infty} \sqrt[n]{R_n} \leq \sqrt{\lambda + \epsilon} < \lambda$, and this is a contradiction to our assumption that $\lim\limits_{n\to\infty} \sqrt[n]{R_n} = \lambda > 1$. □

5 Appendix

For the reader's convenience, we provide a proof of the well-known inequality we used the proof of Proposition 4.2.

Lemma 5.1. $\sum_{k=0}^{L} \binom{N}{k} \leq \left(\frac{eN}{L}\right)^L$, *where $L \leq N$ and $L, N \in \mathbb{N}$.*

Proof. Consider $x \in (0, 1]$. The binomial theorem states that

$$(1 + x)^N = \sum_{k=0}^{N} \binom{N}{k} x^k \geq \sum_{k=0}^{L} \binom{N}{k} x^k.$$

By dividing by the factor x^L we obtain

$$\sum_{k=0}^{L} \binom{N}{k} x^{k-L} \leq \frac{(1 + x)^N}{x^L}.$$

Since $x \in (0,1]$ and $k - L \leq 0$, then $x^{k-L} \geq 1$, it follows that

$$\sum_{k=0}^{L} \binom{N}{k} \leq \frac{(1+x)^N}{x^L}.$$

Let us substitute $x = \frac{L}{N} \in (0,1]$ and let us exploit the inequality $1 + x < e^x$, that holds for all $x > 0$:

$$\frac{(1+x)^N}{x^L} \leq \frac{e^{xN}}{x^L} = \frac{e^{\frac{L}{N}N}}{(\frac{L}{N})^L} = \left(\frac{eN}{L}\right)^L.$$

□

Acknowledgments

The author wishes to thank Edita Pelantová and Štěpán Starosta for their useful comments. The authors acknowledges support by the Czech Science Foundation grant GAČR 13-03538S and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS14/205/OHK4/3T/14.

References

[1] L. Balková, *Beta-integers and Quasicrystals*, PhD thesis, Czech Technical University in Prague and Université Paris Diderot-Paris 7, 2008.

[2] L. Balková, E. Pelantová, and Š. Starosta, *Sturmian jungle (or garden?) on multiliteral alphabets*, RAIRO-Theor. Inf. Appl., 44 (2010), pp. 443–470.

[3] H. Bannai, T. Gagie, S. Inenaga, J. Kärkkäinen, D. Kempa, M. Piątkowski, S. J. Puglisi, and S. Sugimoto, *Diverse palindromic factorization is NP-complete*, in Developments in Language Theory: 19th International Conference, DLT 2015, Liverpool, UK, July 27-30, 2015, Proceedings., I. Potapov, ed., Springer International Publishing, 2015, pp. 85–96.

[4] A. Blondin Massé, S. Brlek, S. Labbé, and L. Vuillon, *Palindromic complexity of codings of rotations*, Theor. Comput. Sci., 412 (2011), pp. 6455–6463.

[5] M. Bucci, A. De Luca, A. Glen, and L. Q. Zamboni, *A new characteristic property of rich words*, Theor. Comput. Sci., 410 (2009), pp. 2860–2863.

[6] X. Droubay, J. Justin, and G. Pirillo, *Episturmian words and some constructions of de Luca and Rauzy*, Theor. Comput. Sci., 255 (2001), pp. 539–553.

[7] A. Frid, S. Puzynina, and L. Zamboni, *On palindromic factorization of words*, Adv. Appl. Math., 50 (2013), pp. 737–748.

[8] A. Glen, J. Justin, S. Widmer, and L. Q. Zamboni, *Palindromic richness*, Eur. J. Combin., 30 (2009), pp. 510–531.

[9] C. Guo, J. Shallit, and A. M. Shur, *Palindromic rich words and run-length encodings*, Inform. Process. Lett., 116 (2016), pp. 735–738.

[10] E. Pelantová and Š. Starosta, *Palindromic richness for languages invariant under more symmetries*, Theor. Comput. Sci, 518 (2014), pp. 42–63.

[11] M. Rubinchik and A. M. Shur, *EERTREE: An Efficient Data Structure for Processing Palindromes in Strings*, Springer International Publishing, Cham, 2016, pp. 321–333.

[12] M. Rubinchik and A. M. Shur, *The number of distinct subpalindromes in random words*, Fund. Inform., 145 (2016), pp. 371–384.

[13] L. Schaeffer and J. Shallit, *Closed, palindromic, rich, privileged, trapezoidal, and balanced words in automatic sequences*, Electr. J. Comb., 23 (2016), p. P1.25.

[14] A. M. Shur, *Growth properties of power-free languages*, Computer Science Review, 6 (2012), pp. 187–208.

[15] J. Vesti, *Extensions of rich words*, Theor. Comput. Sci., 548 (2014), pp. 14–24.