České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering

Ing. Ľubomíra Dvořáková, Ph.D.

Palindromy v jazyce zajímají nejen lingvisty

Palindromes in the language are interesting not only for linguists

Summary

The habilitation thesis "Combinatorics on Words and Applications" combines two fields of mathematics – the main one is Combinatorics on Words and combinatorial results are applied in Random Number Generation [2, 3]. In the habilitation lecture, we focus on the first field – Combinatorics on Words.

We start the lecture with a short look inside the history of Combinatorics on Words and with the most famous examples of infinite words: the Thue–Morse word and the Fibonacci word that belongs to the class of Sturmian words (aperiodic infinite words with the lowest possible complexity). These examples are then useful in the sequel. We subsequently pass to the main topic of this lecture: palindromes. A palindrome is a word that stays the same when read backwards. It is a popular linguistic game to look for the longest palindromes in natural languages. Even more interesting are palindromic sentences. However, according to the structure of natural languages, nobody can be surprised that there are no especially long palindromes in such languages. A much more interesting situation comes to light in the world of infinite words. Palindromes of any length can occur here. However, any anarchy is not reigning here neither. We devote the rest of the lecture to the study of palindromes and notions related to palindromes (for instance palindromic complexity, rich words, defect, almost rich words) in infinite words. Our results that are related to palindromes and that will be presented here in a nutshell are part of the following papers:

1. *Sturmian Jungle (or Garden?) on Multiliteral Alphabets* [4];

2. *Infinite Words with Finite Defect* [5];

3. *On the Brlek–Reutenauer Conjecture* [6];

4. *Proof of the Brlek–Reutenauer Conjecture* [7].

Souhrn

Habilitační práce “Kombinatorika na slovech a aplikace” se věnuje dvěma oblastem matematiky – hlavní oblastí je kombinatorika na slovech a kombinatorické výsledky jsou aplikovány na generování náhodných čísel [2, 3]. V habilitační přednášce se budeme zabývat první oblastí – kombinatorikou na slovech.

Přednášku zahájíme nahlédnutím do historie kombinatoriky na slovech a představením nejslavnějších příkladů nekonečných slov: Thueova–Morseova slova a Fibonacciho slova, které patří do třídy sturmovských slov (aperiodická slova s nejnižší možnou komplexitou). Tyto příklady se nám budou následně hodit. Posléze přejdeme k hlavnímu tématu přednášky: palindromům. Palindrom je slovo, které zůstane stejné, když ho přečteme pozpátku. Známou lingvistickou hříčkou je hledání dlouhých palindromů v přirozených jazycích. Ještě zábavnější jsou palindromické věty. Ovšem vzhledem ke struktuře přirozených jazyků nás nepřekvapí, že žádné převratně dlouhé palindromy přirozené jazyky neobsahují. Mnohem zajímavější situace nastává v případě nekonečných slov. Taková slova mohou obsahovat palindromy libovolné délky. Nicméně žádná anarchie nepanuje ani tady. Zbytek přednášky věnujeme právě studiu palindromů a pojmů souvisejících s palindromy (např. palindromická komplexita, slova bohatá na palindromy, defekt, slova s konečným defektem) v nekonečných slovech. Vlastní výsledky týkající se palindromů, které v přednášce v krátkosti představíme, byly publikovány v následujících článcích:

1. *Sturmian Jungle (or Garden?) on Multiliteral Alphabets* [4];
2. *Infinite Words with Finite Defect* [5];
3. *On the Brlek–Reutenauer Conjecture* [6];
4. *Proof of the Brlek–Reutenauer Conjecture* [7].

Klíčová slova

palindrom
palindromická komplexita
palindromická bohatost
defekt
slova s konečným defektem
Brlekova–Reutenauerova hypotéza

Keywords

palindrome
palindromic complexity
richness
defect
almost rich words
the Brlek–Reutenauer conjecture

Obsah

1 Combinatorics on Words

Let us start our look inside the subject of Combinatorics on Words in a gentle manner. In any lecture devoted to Combinatorics on Words, either the Thue–Morse word or the Fibonacci word should not be missing. It is even better if one speaks about Sturmian words, a class that contains the Fibonacci word. When studying palindromes in the sequel, we will see that all of the above famous examples are worth mentioning.

1.1 The Thue–Morse Word

The Norwegian number theorist and logician Axel Thue known in particular for his results in Diophantine approximations published in 1906 in an obscure Norwegian journal [18] answers to the following questions:
Question 1: “Does there exist an infinite binary word that does not contain cubes?” Let us illustrate on a concrete example the notions from his question. An infinite binary word is an infinite sequence consisting of only two symbols, say $a,\ b$. Take for simplicity a periodic sequence:

$$abbabbabbabbabbabbabbabbabb\dots. \tag{1.1}$$

It is thus an infinite repetition of the chain abb, which we usually denote $(abb)^\omega$. Such a word contains a cube because for instance the chain abb occurs three times in a row in this word:

$$\underline{abbabbabb}abbabbabbabbabbabb\dots.$$

Question 2: “Does there exist an infinite ternary word (consisting of three symbols) that does not contain squares?” A square is a double repetition of a chain.

Both of these questions were answered by Thue confirmatively. Moreover he explained that he had no specific application in his mind, but he studied these questions since he found them interesting.

The Thue–Morse word provided by Thue as an example of an infinite word without cubes consisting of two symbols 0 and 1 – that answers thus positively **Question 1** – can be constructed in the following way. Let us define a morphism $\varphi_{\rm TM}\colon \{0,1\}^* \to \{0,1\}^*$ by

$$\varphi_{\rm TM}(0) = 01 \quad \text{and} \quad \varphi_{\rm TM}(1) = 10,$$

where $\{0,1\}^*$ denotes the set of all finite sequences consisting of symbols 0 and 1. Then the Thue–Morse word $\mathbf{u}_{\rm TM}$ is the so-called fixed point of $\varphi_{\rm TM}$ starting in 0. In order to get the fixed point, we apply the morphism repeatedly

$$\begin{array}{rcl}\varphi^0_{\rm TM}(0) & = & 0\\ \varphi^1_{\rm TM}(0) & = & 01\\ \varphi^2_{\rm TM}(0) & = & 0110\\ \varphi^3_{\rm TM}(0) & = & 0110100110010110\\ \varphi^4_{\rm TM}(0) & = & 01101001100101101001011001101001.\end{array} \tag{1.2}$$

Since every iteration is the prefix of the next iteration and their lengths are strictly growing, it is possible to find an infinite word such that for all n the word $\varphi^n_{\rm TM}(0)$ is its prefix. This infinite word is the unique fixed point $\mathbf{u}_{\rm TM}$. We use the symbolic notation $\mathbf{u}_{\rm TM} = \lim_{n\to+\infty} \varphi^n_{\rm TM}(0)$.

Thue further constructed an infinite word $\mathbf{v}$ over $\{0,1,2\}$ that does not contain squares and responded hence positively also to **Question 2**. For $n \geq 1$ he denoted by v_n the number of ones between the n-th and $(n+1)$-th occurrence of zero in the Thue–Morse word. The desired word was then $\mathbf{v} = v_1v_2v_3\ldots$, i.e.,

$$\underbrace{\mathbf{u}_{\mathrm{TM}}}_{\mathbf{v}} = 0\underbrace{11}_{2}0\underbrace{1}_{1}0\underbrace{\quad}_{0}0\underbrace{11}_{2}0\underbrace{\quad}_{0}0\underbrace{1}_{1}0\underbrace{11}_{2}0\ldots$$

Since the result of Thue did not become known, Marston Morse rediscovered the Thue–Morse word in 1921 when he was studying differential geometry [16].

1.2 Dictionary of Combinatorics on Words

In order to describe the second moment considered as the birth of Combinatorics on Words and in order to present our results in the sequel, we have to introduce basic notions from Combinatorics on Words. By $\mathcal{A}$ we denote a finite set of symbols, usually called **letters**. The set $\mathcal{A}$ is therefore called an **alphabet**. A finite string $w = w_0w_1\ldots w_{n-1}$ of letters of $\mathcal{A}$ is said to be a **finite word**, its length is denoted by $|w| = n$. Finite words over $\mathcal{A}$ together with the operation of concatenation and the empty word ε as the neutral element form a monoid $\mathcal{A}^*$. The map

$$w = w_0w_1\ldots w_{n-1} \quad \mapsto \quad \overline{w} = w_{n-1}w_{n-2}\ldots w_0$$

is a bijection on $\mathcal{A}^*$ and the word $\overline{w}$ is called the **reversal** or the **mirror image** of w. A word w which coincides with its mirror image is a **palindrome**.

Under an **infinite word** we understand an infinite string $\mathbf{u} = u_0u_1u_2\ldots$ of letters from $\mathcal{A}$. A finite word w is a **factor** of a word v (finite or infinite) if there exist words p and s such that $v = pws$. If $p = \varepsilon$, then w is said to be a **prefix** of v. If $s = \varepsilon$, then w is a **suffix** of v.

The **language** $\mathcal{L}(\mathbf{u})$ of an infinite word $\mathbf{u}$ is the set of all its factors. Factors of $\mathbf{u}$ of length n form the set denoted by $\mathcal{L}_n(\mathbf{u})$. Clearly, $\mathcal{L}(\mathbf{u}) = \cup_{n\in\mathbb{N}}\mathcal{L}_n(\mathbf{u})$. We say that the language $\mathcal{L}(\mathbf{u})$ is **closed under reversal** if $\mathcal{L}(\mathbf{u})$ contains with every factor w also its reversal $\overline{w}$.

For any factor $w \in \mathcal{L}(\mathbf{u})$, there exists an index i such that w is a prefix of the infinite word $u_iu_{i+1}u_{i+2}\ldots$. Such an index i is called an **occurrence** of w in $\mathbf{u}$. If each factor of $\mathbf{u}$ occurs infinitely many times in $\mathbf{u}$, the infinite word $\mathbf{u}$ is said to be **recurrent**. If the language of $\mathbf{u}$ is closed under reversal, then $\mathbf{u}$ is recurrent (a proof can be found in [14]). The infinite word $\mathbf{u}$ is said to be **uniformly recurrent** if $\mathbf{u}$ is recurrent and for any factor w of $\mathbf{u}$ the distances between successive occurrences of w form a bounded sequence.

The **(factor) complexity** of an infinite word $\mathbf{u}$ is the map $\mathcal{C}\colon \mathbb{N} \mapsto \mathbb{N}$ defined by $\mathcal{C}(n) = \#\mathcal{L}_n(\mathbf{u})$. To determine the increment of complexity of an infinite word $\mathbf{u}$, one has to count the possible **extensions** of factors of length n. A **left extension** of $w \in \mathcal{L}(\mathbf{u})$ is any letter $a \in \mathcal{A}$ such that $aw \in \mathcal{L}(\mathbf{u})$. The set of all left extensions of a factor w will be denoted by $\mathrm{Lext}(w)$. We will mostly deal with recurrent infinite words $\mathbf{u}$. In this case, any factor of $\mathbf{u}$ has at least one left extension. A factor w is called **left special** (or LS for short) if w has at least two left extensions. Clearly, any prefix of a LS factor is LS as well. Similarly, one can define a **right extension**, a **right special** (or RS) factor, $\mathrm{Rext}(w)$. We say that a factor w of $\mathbf{u}$ is a **bispecial** (or BS) factor if it is both RS and LS. Using the introduced terminology, the increment or the **first difference of complexity** $\Delta\mathcal{C}(n) = \mathcal{C}(n+1) - \mathcal{C}(n)$ is given by

$$\Delta\mathcal{C}(n) = \sum_{w\in\mathcal{L}_n(\mathbf{u})} \big(\#\mathrm{Rext}(w) - 1\big) = \sum_{w\in\mathcal{L}_n(\mathbf{u})} \big(\#\mathrm{Lext}(w) - 1\big)\,. \tag{1.3}$$

A non-zero contribution to $\Delta\mathcal{C}(n)$ in the left-hand sum is given only by RS factors $w \in \mathcal{L}_n(\mathbf{u})$, and for recurrent words, a non-zero contribution to $\Delta\mathcal{C}(n)$ in the right-hand sum is provided only by LS factors $w \in \mathcal{L}_n(\mathbf{u})$. If we denote $\mathrm{Bext}(w) = \{awb \in \mathcal{L}(\mathbf{u}) \mid a, b \in \mathcal{A}\}$, then the **second difference of complexity** $\Delta^2\mathcal{C}(n) = \Delta\mathcal{C}(n+1) - \Delta\mathcal{C}(n) = \mathcal{C}(n+2) - 2\mathcal{C}(n+1) + \mathcal{C}(n)$ is given by

$$\Delta^2\mathcal{C}(n) = \sum_{w \in \mathcal{L}_n(\mathbf{u})} \big(\#\mathrm{Bext}(w) - \#\mathrm{Rext}(w) - \#\mathrm{Lext}(w) + 1\big)\,. \tag{1.4}$$

Denote by $\mathrm{b}(w)$ the quantity

$$\mathrm{b}(w) = \#\mathrm{Bext}(w) - \#\mathrm{Rext}(w) - \#\mathrm{Lext}(w) + 1.$$

The number $\mathrm{b}(w)$ is called the **bilateral order** of the factor w and was introduced in [11].

We will moreover need the notion of **palindromic extension**. The set of palindromic extensions of a palindrome $w \in \mathcal{L}(\mathbf{u})$ is defined by $\mathrm{Pext}(w) = \{awa \in \mathcal{L}(\mathbf{u}) \mid a \in \mathcal{A}\}$. The number of palindromes of a fixed length occurring in an infinite word is measured by the so called **palindromic complexity** $\mathcal{P}$, a map which assigns to any non-negative integer n the number

$$\mathcal{P}(n) = \#\{w \in \mathcal{L}_n(\mathbf{u}) \mid w \ \text{ is a palindrome}\}\,.$$

Let $j, k,\ j < k$, be two successive occurrences of a factor w in $\mathbf{u}$. Then $u_j u_{j+1} \ldots u_{k-1}$ is called a **return word** of w. If v is a return word of w, then the word vw is called a **complete return word** of w. It is obvious that an infinite recurrent word is uniformly recurrent if and only if the set of return words of any of its factors is finite.

We will often work with **morphisms**, i.e., mappings $\varphi : \mathcal{A}^* \to \mathcal{A}^*$ satisfying for any $v, w \in \mathcal{A}^*$ that $\varphi(vw) = \varphi(v)\varphi(w)$. A morphism is thus uniquely given if we define images of letters $\varphi(a)$ for all $a \in \mathcal{A}$. A morphism can be naturally extended to infinite words:

$$\varphi(u_0u_1u_2\ldots) = \varphi(u_0)\varphi(u_1)\varphi(u_2)\ldots$$

If an infinite word $\mathbf{u}$ satisfies $\varphi(\mathbf{u}) = \mathbf{u}$, we call $\mathbf{u}$ a **fixed point** of the morphism φ.

Example 1. *Let us illustrate the introduced notions on the infinite word* $\mathbf{u} = (abb)^\omega$. *Its alphabet is* $\mathcal{A} = \{a, b\}$. *The infinite word* $\mathbf{u}$ *is uniformly recurrent and its language is closed under reversal. The word babb is a factor of length* 4 *of* $\mathbf{u}$. *The word abbabba is a prefix of length* 7 *of* $\mathbf{u}$. *The only LS factors are* ε *and b. Consequently* $\Delta\mathcal{C}(n) = 0$ *for all* $n \geq 2$. *It is readily seen that the set of all factors of length* 2 *of* $\mathbf{u}$ *equals* $\mathcal{L}_2(\mathbf{u}) = \{ab, bb, ba\}$. *Therefore* $\mathcal{C}(n) = 3$ *for all* $n \geq 2$. *The only BS factors are* ε *and* b *with* $\mathrm{b}(\varepsilon) = 0$ *and* $\mathrm{b}(b) = -1$. *It is not difficult to see that there is one palindrome of any even length, i.e.,* $\mathcal{P}(2n) = 1$ *for every* $n \in \mathbb{N}$, *and there are two palindromes of length one – the letters* a, b *– and one palindrome of every odd length larger than one, i.e.,* $\mathcal{P}(2n+3) = 1$ *for every* $n \in \mathbb{N}$. *Since* $u_0u_1u_2u_3u_4u_5u_6u_7u_8u_9\ldots = ab\underline{\underline{babba}bba}\ldots$, *the index* $i = 2$ *is an occurrence of the factor babba and bab is a return word of babba and babbabba is a complete return word of babba. If we define a morphism* φ *on* $\{a, b\}^*$ *by* $\varphi(a) = abb$ *and* $\varphi(b) = abb$, *then* $\mathbf{u}$ *is evidently a fixed point of* φ.

1.3 The Fibonacci Word and Sturmian Words

Even more often one considers for the birth of Combinatorics on Words the famous paper [17] by Hedlund and already mentioned Morse from 1940. When studying differential equations of Sturm–Liouville type, they discovered a certain class of infinite words and named them in honour of the French mathematician J. C. F. Sturm. Hedlund and Morse noticed that not every map $f : \mathbb{N} \to \mathbb{N}$ is the factor complexity of an infinite word. Infinite words are **eventually periodic**, i.e., they are of the form wv^ω (where v, w are finite words over the corresponding alphabet and ω denotes an infinite repetition) if and only if their factor complexity is eventually constant, i.e., there exists a constant K such that $\mathcal{C}(n) = K$ for sufficiently large n. The words that are not eventually periodic are called **aperiodic**. For aperiodic words Hedlund and Morse showed that for all $n \in \mathbb{N}$ their factor complexity satisfies

$$\mathcal{C}(n) \geq n + 1.$$

Sturmian words are aperiodic words with the lowest possible complexity.

Definition 1. *An infinite word* $\mathbf{u}$ *is called* **Sturmian** *if for all* $n \in \mathbb{N}$ *it holds*

$$\mathcal{C}(n) = n + 1.$$

They have been intensively studied from the very beginning. Besides their low factor complexity the reason of their popularity is the fact that the famous Fibonacci word belongs to this class.

Example 2. *The fixed point of the morphism* $\varphi_{\mathrm{F}} : \{0,1\}^* \to \{0,1\}^*$ *defined by* $\varphi_{\mathrm{F}}(0) = 01,\ \varphi_{\mathrm{F}}(1) = 0$ *is called the* **Fibonacci word** $\mathbf{u}_{\mathrm{F}}$. *Let us write down a prefix of the Fibonacci word* $\mathbf{u}_{\mathrm{F}} = \lim_{n \to +\infty} \varphi_{\mathrm{F}}^n(0) = 0100101001001010010100100101001001\ldots$

The Fibonacci word is closely connected to the Fibonacci numbers. Let us recall that the Fibonacci numbers were introduced by Leonardo of Pisa, known as Fibonacci, in a mathematical game dealing with rabbits: An adult couple (denote it 0*) has always after one month a pair of young (denote it* 1*) and that pair of young grows up after one month. Fibonacci was interested in how large the population of rabbits would be after* n *months provided the rabbits are immortal.*

It is easy to see that the response is provided by the Fibonacci word. If we denote the length of the n*-iteration* $F_n = |\varphi_{\mathrm{F}}^n(0)|$*, it follows that the number of rabbits after* n *months is equal to* F_n*. It is not difficult to verify that* $F_0 = 1,\ F_1 = 2$ *and* $F_{n+1} = F_n + F_{n-1}$.

Thanks to a long and fruitful study of Sturmian words, a lot of properties and equivalent definitions of these words are known nowadays. Let us mention only those ones that are related to palindromes.

Theorem 1. *Let* $\mathbf{u}$ *be an infinite word. The properties listed below are equivalent:*

(i) $\mathbf{u}$ *is Sturmian, i.e.,* $\mathcal{C}(n) = n + 1$ *for all* $n \in \mathbb{N}$*;*

(ii) $\mathbf{u}$ *contains one palindrome of every even length and two palindromes of every odd length;*

(iii) $\mathbf{u}$ *is binary and every palindrome has a unique palindromic extension.*

This theorem has been proved by Droubay and Pirillo [13]. Notice that the third property can be equivalently rewritten as

$$\mathcal{P}(n) + \mathcal{P}(n+1) = 3 \quad \text{for all } n \in \mathbb{N},$$

and also as

$$\mathcal{P}(n+2) = \mathcal{P}(n) \quad \text{for all } n \in \mathbb{N}.$$

Let us recall that $\mathcal{P}(0) = 1$ since the empty word is considered to be a palindrome.

2 Palindromes in Nature

We know already that a palindrome is a word that stays the same when read backwards. Nobody can be surprised that in natural languages there are no especially long palindromes. The longest palindromic words in Czech are those ones of the type "nepochopen" (not understood), "nepotopen" (not sunk), "nezasazen" (not planted), "nezařazen" (not filed). In English the longest palindromic word is "tattarrattat". Its victory is however doubtful because it is not a common word but an invention by James Joyce who used in his novel Ulysses [15] this neologism to denote strong knocking on the door:

"I was just beginning to yawn with nerves thinking he was trying to make a fool of me when I knew his tattarrattat at the door."

The sentences might be more interesting from the palindromic point of view. They give rise to palindromes if we release gaps between words and eventually diacritics. In Czech the best known palindromic sentences are the following ones:

"Bažantu padá za záda putna žab." (A bucket of frogs is falling behind the pheasant's back.)
"Jelenovi pivo nelej." (Do not pour beer to deer.)
"Kobyla má malý bok." (Mares have small sides.)

Of course, palindromic numbers might be of interest, too. In particular when there is at least a partial explanation for their reason. For instance there is a palindrome consisting of odd ciphers related to the foundation of the Charles Bridge 135797531. The Museum of the Charles Bridge in Prague uses it as a part of its logo. According to the historian of astronomy Zdeněk Horský the foundation stone might have been laid on July, 9, 1357 at 5:31. There was a favorable constellation of Sun and Saturn at that moment. The palindrome thus consists of the following items: year – day – month – hour – minutes.

3 Palindromes in Infinite Words

A much more interesting situation comes to light in the world of infinite words. Palindromes of any length can occur here. However, any anarchy is not reigning here neither.

The palindromic complexity is bounded by the first difference of factor complexity.

Proposition 1 ([1]). *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Then*

$$\mathcal{P}(n) + \mathcal{P}(n+1) \leq \Delta\mathcal{C}(n) + 2 \quad \textit{for all } n \in \mathbb{N}. \tag{3.1}$$

Moreover, an infinite word can contain in any of its factors w at most $|w| + 1$ distinct palindromes (including the empty word). This upper bound on the number of palindromes occurring in a finite word given by Droubay, Justin, and Pirillo [12] initiated many interesting investigations on palindromes in infinite words. A finite word w containing the utmost number $|w| + 1$ of palindromes is called **rich**. An infinite word is said to be **rich** if all its factors are rich. (We keep here the terminology introduced by Glen et al. [14] in 2007, which seems to us to be prevalent nowadays. However, Brlek et al. [8] baptized such words **full** already in 2004.)

There exist several equivalent characterizations of rich words. The most recent one using bilateral orders of factors was stated in the paper **Sturmian Jungle (or Garden?) on Multiliteral Alphabets** [4].

4 Defect and Richness

Brlek et al. [8] suggested to study the **defect** $D(w)$ of a finite word w defined as the difference between the upper bound $|w| + 1$ and the actual number of palindromes contained in w. The defect of an infinite word is then defined as the maximal defect of a factor of the infinite word. In this convention, rich words are precisely the words with zero defect.

Since we focus on palindromes in infinite words, richness, and defect, we will introduce several notions and known results related to this topic. Let us remark that not only all prefixes of rich words are rich, but also all factors are rich. A result from [12] provides us with a handful tool which helps to evaluate the defect of a factor.

Proposition 2 ([12])**.** *A finite or infinite word* $\mathbf{u}$ *is rich if and only if the longest palindromic suffix of* w *occurs exactly once in* w *for any prefix* w *of* $\mathbf{u}$.

The longest palindromic suffix of a factor w will occur often in our considerations, therefore we will denote it by $lps(w)$. In accordance with the terminology introduced in [12], the factor with a unique occurrence in another factor is called **unioccurrent**. The proof of the above proposition is based on the fact that there exists a bijection between the set of palindromes contained in w and the first prefixes of w ending in the corresponding palindromes. It follows that the other prefixes cause the increment of the defect.

Corollary 1. *The defect* $D(w)$ *of a finite word* w *is equal to the number of prefixes* w' *of* w*, for which the* $lps(w')$ *is not unioccurrent in* w'.

This corollary implies that $D(v) \geq D(w)$ whenever w is a factor of v. It enables to give a reasonable definition of the defect of an infinite word (see [8]).

Definition 2. *The defect of an infinite word* $\mathbf{u}$ *is the number (finite or infinite)*

$$D(\mathbf{u}) = \sup\{D(w) \mid w \textit{ is a prefix of } \mathbf{u}\}\,.$$

Let us point out several facts concerning defects that are easy to prove:

1. If we consider all factors of a finite or an infinite word $\mathbf{u}$, we obtain the same defect, i.e.,

$$D(\mathbf{u}) = \sup\{D(w) \mid w \in \mathcal{L}(\mathbf{u})\}\,.$$

2. Any infinite word with finite defect contains infinitely many palindromes.

3. Infinite words with zero defect correspond exactly to rich words.

The authors of [12] who were the first ones to tackle the problem of richness showed that Sturmian and episturmian words are rich. (Let us recall that **episturmian words** are defined as infinite words with the language closed under reversal and having at most one left special factor of every length.)

Let us provide as well an example of an infinite word that is not rich. You will probably guess which word will serve as such an example: the Thue–Morse word $\mathbf{u}_{\mathrm{TM}}$. Consider its prefix $w = 011010011$. Then w contains nine palindromes: $\varepsilon, 0, 1, 00, 11, 010, 101, 0110, 1001$, which is less than $|w| + 1 = 10$. Hence w is not rich (its defect is one), and it is easy to see that $D(\mathbf{u}_{\mathrm{TM}}) = +\infty$.

In [8], an insight into the richness of periodic words can be found. Further on, let us summarize equivalent characterizations of rich words.

Theorem 2. *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Then the following statements are equivalent:*

1. *The word* $\mathbf{u}$ *is rich.*
2. *For any prefix* w *of* $\mathbf{u}$ *the* $lps(w)$ *is unioccurrent in* w*.*
3. *All complete return words of any palindrome in* $\mathbf{u}$ *are palindromes.*
4. *The equality*
$$\mathcal{P}(n) + \mathcal{P}(n+1) = \Delta\mathcal{C}(n) + 2$$
holds for all $n \in \mathbb{N}$*.*
5. *Any bispecial factor* w *of* $\mathbf{u}$ *satisfies:*
   - *if* w *is non-palindromic, then* $\mathrm{b}(w) = 0$*,*
   - *if* w *is a palindrome, then* $\mathrm{b}(w) = \#\mathrm{Pext}(w) - 1$*.*

Glen et al. [14] have proved the characterization based on the notion of complete return words, Bucci et al. [10] have characterized richness using the palindromic and factor complexity, and most recently, we have found a new characterization of rich words considering bilateral orders of factors [4].

Our aim in the sequel is to introduce in the rough the papers [5, 6, 7] devoted to the study of defect. Periodic words with finite defect have been studied in [8] and in [14]. It holds that the defect of an infinite periodic word with the minimal period w is finite if and only if $w = pq$, where both p and q are palindromes. In [14] words with finite defect have been baptized **almost rich**.

4.1 Almost Rich Words

The following characterization of infinite words with finite defect – called as well almost rich words – follows from observations made in [14].

Theorem 3. *Let* $\mathbf{u}$ *be a uniformly recurrent word containing infinitely many palindromes. Then the following statements are equivalent:*

1. *The word* $\mathbf{u}$ *is almost rich.*
2. *There exists an integer* H *such that for any prefix* w *of* $\mathbf{u}$ *with* $|w| \geq H$*, the* $lps(w)$ *is unioccurrent in* w*.*
3. *There exists an integer* K *such that all complete return words of any palindrome in* $\mathbf{u}$ *of length at least* K *are palindromes.*

It is easy to see that the second statement of Theorem 3 can be equivalently rewritten as: There exists an integer H such that for any factor w of $\mathbf{u}$ with $|w| \geq H$, the $lps(w)$ is unioccurrent in w. Let us stress that if we put in the previous theorem $D(\mathbf{u}) = K = H = 0$, all statements become known results for rich words, see Theorem 2.

Example 3. *Let us provide an example of a uniformly recurrent word* $\mathbf{u}$ *with finite defect and let us find for* $\mathbf{u}$ *the lowest values of constants* K *and* H *from Theorem 3. Take the Fibonacci word* $\mathbf{u}_{\mathrm{F}}$*, i.e., the fixed point of* $\varphi_{\mathrm{F}}: 0 \to 01,\ 1 \to 0$. *Define* $\mathbf{u}$ *as its morphic image* $\sigma(\mathbf{u}_{\mathrm{F}})$*, where* $\sigma\colon 0 \to cabcbac,\ 1 \to d$*, i.e.,*

$$\mathbf{u} = cabcbacdcabcbaccabcbacdcabcbacdcabcbaccabcbacdcabcbaccabcbacdcabcbacdcabcbac\ldots$$

It is easy to show that all palindromes of length greater than one and the palindromes a*,* b*, and* d *have only palindromic complete return words. Hint: long palindromes in* $\mathbf{u}$ *contain in their center images of non-empty palindromes from* $\mathbf{u}_{\mathrm{F}}$ *that have palindromic complete return words by the richness of* $\mathbf{u}_{\mathrm{F}}$*. The only non-palindromic complete return word of* c *is* $cabc$*. In order to show that* $D(\mathbf{u}) = 1$*, it suffices to verify that no prefixes longer than* $cabc$ *have* c *as their longest palindromic suffix. This follows directly from the form of* σ*. The lowest values of the constants* K *and* H *are:* $K = 2$*,* $H = 5$*.*

In the paper **Infinite Words with Finite Defect** [5], we have proved a new characterization of infinite words with finite defect based on a relation between the palindromic and factor complexity.

Theorem 4 ([5]). *Let* $\mathbf{u}$ *be a uniformly recurrent word. Then* $\mathbf{u}$ *is almost rich if and only if there exists an integer* N *such that*

$$\mathcal{P}(n) + \mathcal{P}(n+1) = \Delta\mathcal{C}(n) + 2$$

holds for all $n \geq N$*.*

Notice that if we set $N = 0$ in the previous theorem, then we obtain the known characterization of rich words from Theorem 2 (which holds even under a weaker assumption that $\mathcal{L}(\mathbf{u})$ is closed under reversal).

We will present here only the main ingredient of the proof of Theorem 4. Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Using the proof of Proposition 1, those $n \in \mathbb{N}$ for which the equality

$$\mathcal{P}(n) + \mathcal{P}(n+1) = \Delta\mathcal{C}(n) + 2$$

holds can be characterized in the graph language.

An n**-simple path** e is a factor of $\mathbf{u}$ of length at least $n + 1$ such that the only special (right or left) factors of length n occurring in e are its prefix and suffix of length n. If w is the prefix of e of length n and v is the suffix of e of length n, we say that the n-simple path e starts in w and ends in v. We will denote by $G_n(\mathbf{u})$ an undirected graph whose set of vertices is formed by unordered pairs $\{w, \overline{w}\}$ such that $w \in \mathcal{L}_n(\mathbf{u})$ is right or left special. We connect two vertices $\{w, \overline{w}\}$ and $\{v, \overline{v}\}$ by an unordered pair $\{e, \overline{e}\}$ if e or $\overline{e}$ is an n-simple path starting in w or $\overline{w}$ and ending in v or $\overline{v}$. Note that the graph $G_n(\mathbf{u})$ may have multiple edges and loops.

Lemma 1 ([5]). *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Let* $n \in \mathbb{N}$*. Then* $\mathcal{P}(n) + \mathcal{P}(n+1) = \Delta\mathcal{C}(n) + 2$ *if and only if both of the following conditions are met:*

1. *The graph obtained from* $G_n(\mathbf{u})$ *by removing loops is a tree.*
2. *Any* n*-simple path forming a loop in the graph* $G_n(\mathbf{u})$ *is a palindrome.*

4.2 The Brlek–Reutenauer Conjecture

Despite the fact that numerous researchers study palindromes, only recently Brlek and Reutenauer [9] have noticed that the value of defect is closely tied with the expression in (3.1) – let us denote for an infinite word $\mathbf{u}$ by $T_{\mathbf{u}}(n) = \Delta\mathcal{C}(n) + 2 - \mathcal{P}(n) - \mathcal{P}(n+1)$. They have shown that for a periodic infinite word $\mathbf{u}$ with the language closed under reversal, it holds $2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$. Their conjecture says that the same equality holds for all infinite words with the language closed under reversal.

Conjecture 1 (The Brlek–Reutenauer conjecture)**.** *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Then*

$$2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)\,. \tag{4.1}$$

It is known from the paper [10] that Conjecture 1 holds for rich words (both sides equal zero for them). As we have already mentioned, Brlek and Reutenauer provided a result for periodic words.

Theorem 5 ([9])**.** *Let* $\mathbf{u}$ *be a periodic infinite word. Then* (4.1) *holds.*

4.2.1 Partial Proof of the Brlek–Reutenauer Conjecture

In the paper **On the Brlek–Reutenauer Conjecture** [6], we have proved their conjecture for uniformly recurrent words.

Theorem 6 ([6])**.** *If* $\mathbf{u}$ *is a uniformly recurrent infinite word with the language closed under reversal, then* (4.1) *holds.*

In the proof of Theorem 6 we used our result from [5] (recalled here as Theorem 4) to show that either both sides of (4.1) are finite, or both of them are infinite. Further on, the main idea was to construct for any almost rich uniformly recurrent word $\mathbf{u}$ a periodic word $\mathbf{v}$ satisfying $D(\mathbf{u}) = D(\mathbf{v})$ and $T_{\mathbf{u}}(n) = T_{\mathbf{v}}(n)$ for all $n \in \mathbb{N}$. The proof was then finished because by Theorem 5 the conjecture holds for periodic words. However, in the construction we had to prove one more quite interesting statement on increasing squares in uniformly recurrent words with finite defect.

Lemma 2 ([6])**.** *Let* $\mathbf{u}$ *be an almost rich uniformly recurrent infinite word. Then the set*

$$\{w \in \mathcal{A}^* \mid ww \in \mathcal{L}(\mathbf{u})\}$$

is infinite.

The proof relied essentially on the uniform recurrence of the infinite word in question.

4.2.2 Proof of the Brlek–Reutenauer Conjecture

In the paper **Proof of the Brlek–Reutenauer Conjecture** [7], we managed to find completely different arguments than in the previous paper [6] that enabled us to prove Conjecture 1 in full generality without exploiting the result for periodic words.

Theorem 7 ([7])**.** *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Then*

$$2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)\,. \tag{4.2}$$

Let us provide the main ideas of the proof. We divided the proof into two steps:

1. Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Assume $D(\mathbf{u}) < +\infty$ and $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) < +\infty$. Then $2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$.

   Since the defect is finite, there exists H_1 such that $D(q) = D(\mathbf{u})$ for every prefix q of $\mathbf{u}$ of length greater than or equal to $H_1 - 1$. By the finiteness of $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$, there exists H_2 such that $T_{\mathbf{u}}(n) = 0$ for all $n \geq H_2$. Set $H = \max\{H_1, H_2\}$ and find a prefix p of $\mathbf{u}$ containing all factors of length H. Clearly, $D(p) = D(\mathbf{u})$ and $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) = \sum_{n=0}^{H-1} T_{\mathbf{u}}(n)$.

   In the sequel, we will use a theorem by Brlek and Reutenauer for finite words [9].

   Theorem 8 ([9]). *For every finite word w we have*

   $$2D(w) = \sum_{n=0}^{|w|} T_w(n),$$

   where $T_w(n) = \Delta\mathcal{C}_w(n) + 2 - \mathcal{P}_w(n+1) - \mathcal{P}_w(n)$ and the index w means that we consider only factors of w.

   We deduce the following equalities:

   $$2D(\mathbf{u}) = 2D(p) = \sum_{n=0}^{|p|} T_p(n) = \sum_{n=0}^{H-1} T_p(n) + \sum_{n=H}^{|p|} T_p(n) = \sum_{n=0}^{H-1} T_{\mathbf{u}}(n) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n),$$

   where everything follows from the previous arguments; only the equality $\sum_{n=H}^{|p|} T_p(n) = 0$ remains to be explained. Let us rewrite it in the following form:

   $$\begin{aligned} \textstyle\sum_{n=H}^{|p|} T_p(n) &= \textstyle\sum_{n=H}^{|p|} \Big(\mathcal{C}_p(n+1) - \mathcal{C}_p(n) + 2 - \mathcal{P}_p(n+1) - \mathcal{P}_p(n)\Big) \\ &= -\mathcal{C}_p(H) + 2(|p| - H + 1) - 2\textstyle\sum_{n=H}^{|p|} \mathcal{P}_p(n) + \mathcal{P}_p(H). \end{aligned}$$

   We have to explain why the last expression equals zero: The factors of length H are either palindromes of length H – their number is equal to $\mathcal{P}_p(H)$ – or they are non-palindromic factors whose longest palindromic suffix is of course of length less than H. Such non-palindromic factors are certainly not contained in the prefix q of length $H - 1$, therefore their longest palindromic suffix is unioccurrent. Consequently, to any palindrome of length less than H that does not occur in q, there are exactly two non-palindromic factors w and $\overline{w}$ having it as its longest palindromic suffix. The number of such factors is therefore given by twice the number of palindromes of length less than H that are not contained in q, which is equal to $2(|p| - |q| - \sum_{n=H}^{|p|} \mathcal{P}_p(n))$. This concludes the proof.

2. Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Then $D(\mathbf{u})$ is finite if and only if $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ is finite.

   This second part was proved using the graph theory, more precisely, using mainly Lemma 1.

References

[1] P. Baláži, Z. Masáková, E. Pelantová, *Factor versus palindromic complexity of uniformly recurrent infinite words*, Theoret. Comput. Sci. **380** (2007), 266–275.

[2] Ľ. Balková, M. Bucci, A. De Luca, S. Puzynina, *Infinite words with well distributed occurrences*, In: J. Karhumäki, A. Lepistö, L. Zamboni (Eds.), Combinatorics on Words, LNCS **8079**, Springer (2013), 46–57

[3] Ľ. Balková, M. Bucci, A. De Luca, J. Hladký, S. Puzynina, *Pseudorandom number generators based on infinite words*, submitted to Math. Comput. (2014)

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

[5] Ľ. Balková, E. Pelantová, Š. Starosta, *Infinite words with finite defect*, Adv. Appl. Math. **47** (2011), 562–574.

[6] Ľ. Balková, E. Pelantová, Š. Starosta, *On the Brlek–Reutenauer conjecture*, Theoret. Comput. Sci. **412** (2011), 5649–5655.

[7] Ľ. Balková, E. Pelantová, Š. Starosta, *Proof of the Brlek–Reutenauer conjecture*, Theoret. Comput. Sci. **475** (2013), 120–125.

[8] S. Brlek, S. Hamel, M. Nivat, C. Reutenauer, *On the palindromic complexity of infinite words*, Internat. J. Found. Comput. Sci. **2** (2004), 293–306.

[9] S. Brlek, Ch. Reutenauer, *Complexity and palindromic defect of infinite words*, Theoret. Comput. Sci. **412** (2011), 493–497.

[10] M. Bucci, A. De Luca, A. Glen, L. Q. Zamboni, *A connection between palindromic and factor complexity using return words*, Adv. Appl. Math **42** (2009), 60–74.

[11] J. Cassaigne, *Complexity and special factors*, Bull. Belg. Math. Soc. Simon Stevin **4** (1997), 67–88.

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

[13] X. Droubay, G. Pirillo, *Palindromes and Sturmian words*, Theoret. Comput. Sci. **223** (1999), 73–85.

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

[15] J. Joyce, *Ulysses*, Sylvia Beach's Shakespeare and Company in Paris, 1922.

[16] M. Morse, *Recurrent geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc. **22** (1921), 84–100.

[17] M. Morse, G. A. Hedlund, *Symbolic dynamics II - Sturmian trajectories*, Amer. J. Math. **62** (1940), 1–42.

[18] A. Thue, *Über unendliche Zeichenreihen*, Norske vid. Selsk. Skr. Mat. Nat. Kl. **7** (1906), 1–22.

Ing. Ľubomíra Dvořáková (born Balková), Ph. D.

Date of birth:	September 23, 1980
Affiliation:	Department of Mathematics Faculty of Nuclear Sciences and Physical Engineering (FNSPE) Czech Technical University in Prague (CTU)
Web:	kmlinux.fjfi.cvut.cz/∼balkolub

Education

May 2008	defence of the doctoral thesis (obtained degree: Ph.D.)
2005 – 2008	joint thesis between CTU (in Applied Mathematics) and University Paris Diderot – Paris 7 (in Theoretical Physics)
February 2005	defence of the diploma thesis (obtained degree: Engineer)
1999 – 2005	FNSPE CTU (specialization: Applied Mathematics)

Honours and Scholarships

May 2013	Laureate of L'Oréal UNESCO Award for Women in Science
December 2009	CTU Rector Prize of the 1st grade for excellent doctoral theses
December 2009	Honourable Mention within Votruba Prize for the best doctoral thesis in Theoretical Physics
2005	Scholarship of the French government for Ph.D. students "en co-tutelle" (under double supervision)
November 2004	Foundation Preciosa scholarship of excellence for students of CTU
June 2004	McKinsey&Company scholarship for students of engineering

Research Topics

2011 – 2014	• applications of combinatorics on words (CoW) in cryptology: random number generators, dithered hash functions (collaboration with M. Bucci, A. De Luca, J. Hladký, S. Puzynina)
2008 – 2014	• theoretical informatics: beta-integers, non-standard numeration systems, CoW (collaboration with Z. Masáková, E. Pelantová, Š. Starosta, O. Turek)
2005 – 2008	• mathematical models of quasicrystals, mainly beta-integers, combinatorial and arithmetical properties of beta-integers (collaboration with Z. Masáková) physics of beta-integers (diffraction, Schrödinger operator) (collaboration with J.-P. Gazeau) • results in the doctoral thesis: *Beta-integers and quasicrystals*

Professional Experience at the Department of Mathematics, FNSPE, CTU

- **Assistant professor** since 2008
- **Teaching** since 2002:
  lectures in Linear Algebra, Cryptology, History of Mathematics
  tutorials in Linear Algebra and Calculus
- **Supervision of student projects**: six bachelor projects (3rd year) and one research project (4th year)

Research Stays

April – May 2014	University of Oulu, Finland
December 2011	University Paris 7, Paris, France
July 2011	Joint Institute for Nuclear Research, Dubna, Russia
September 2010	Faculty of Informatics, Debrecen, Hungary

Publications

- author or coauthor of 22 papers in refereed scientific journals
- 37 citations in recognized international databases
- author or coauthor of 10 popularization papers in Czech scientific journals
- author of textbooks: Lineární algebra 1, 2

Participation in Conferences

2006 – 2014	lectures at more than 15 international conferences
September 2012	invited lecture at MELA (Meeting on Languages), Telč
May 2012	invited lecture at AM2012 (Applications of Mathematics), Prague

Organization of Conferences

2011	international conference Words 2011
2007 – 2013	student conference Combinatorics on Words
2006 – 2013	Ph.D. student conference Doctoral Days

Grant Investigation

2013 – 2017	Czech Science Foundation grant GAČR (member of the research team)
2009 – 2012	Czech Science Foundation grant GAČR (member of the research team)
2010 – 2014	principal investigator of the grant of the Student Grant Agency of CTU for organization of student conferences SVK
2009	principal investigator of the Internal Grant IGS of CTU
2012 – 2014	guarantor of the Erasmus agreement CTU – University Paris 7
2013 – 2014	guarantor of the Erasmus agreement CTU – University of Oulu

Membership

since 2014	Executive Committee of Czech Mathematical Society
since 2014	Organizing Committee of Mathematical Olympiad in Prague
since 2011	Union of Czech Mathematicians and Physicists
since 2012	Academic Senate of FNSPE CTU

Popularization of Mathematics

lectures for high school students and teachers
public lectures
popularization papers
supervisor and jury member of student scientific projects (SOČ and Amavet)
main administrator of the web for high school students jaderka.cz