On the number of squares in a word

Srečko Brlek*
LaCIM, Université du Québec à Montréal, Canada

Francesco Dolce
FIT, Czech Technical University in Prague, Czech Republic

Shuo Li
LaCIM, Université du Québec à Montréal, Canada

Elise Vandomme
Department of Mathematics, Université de Liege, Belgium

Abstract

A *square* is a word of the form uu. In this contribution we investigate the square complexity of a (finite or infinite) word w, that is the number of distinct squares appearing as factors in w. We introduce the notion of *square defect* and state an identity, inspired by the well-known Brlek-Reutenauer identity for palindromes, relating the number of factors with the number of distinct squares in a word. This identity is established for finite words and for some classes of infinite words such as periodic words and strict standard episturmian ones.

The study of squares, as well as other patterns in a word, is one among the many fundamental topics in combinatorics on words. It was first conjectured by Fraenkel and Simpson in [7] that the number of distinct square factors of a finite word w, which is denoted by $S(w)$, is bounded by its length $|w|$. In the same article, they proved that $S(w) < 2|w|$. This upper bound was first improved by Ilie in [8] who showed that the number of distinct squares is asymptotically bounded by $2|w| - \Theta(\log|w|)$ and later by Deza, Franek and Thierry in [6] who showed that $S(w) \leq \lfloor 11/6 \rfloor |w|$. In a recent article [9], Thierry showed that $S(w) \leq 1.5|w|$. Here we show that $S(w)$ is actually bounded by length of w plus one minus the number of distinct letters appearing in w (Theorem 1). Such an upper bound is not only an improvement of the one conjectured by Fraenkel and Simpson, but it is also sharp for small words.

Let $|\mathrm{Alph}(w)|$ denote the number of distinct letters appearing in a finite word w.

Theorem 1 ([2] Brlek and Li, 2022) *Let w be a finite word.*

$$S(w) \leq |w| - |\mathrm{Alph}(w)| + 1.$$

The proof of the previous theorem is established by using some fundamental properties of Rauzy graphs. Recall that for any finite word w of length k and for any integer n such

*with the support of NSERC (Canada).

that $1 \leq n \leq k$, the Rauzy graph $\Gamma_n(w)$ is the oriented graph defined as follows: the set of vertices is $L_w(n)$ and the set of edges is $L_w(n+1)$; an edge $e \in L_w(n+1)$ starts at the vertex u and ends at the vertex v, if u is a prefix and v is a suffix of e. Let $\Gamma(w) = \cup_{n=1}^{k} \Gamma_n(w)$. Here we define the notion of *small circuit* as follows: a circuit in the graph $\Gamma_n(w)$ is called small if its size is no larger than n.

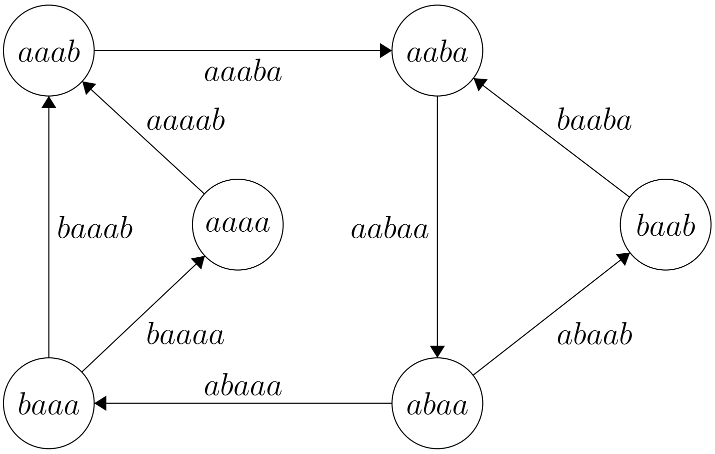
Theorem 1 is proved in two steps: first, the number of small circuits in $\Gamma_n(w)$ is bounded by $|L_w(n+1)| - |L_w(n)| + 1$ for all n satisfying $1 \leq n \leq |w|$, and consequently, the total number of distinct small circuits in $\Gamma(w)$ is bounded by $|w| - |\text{Alph}(w)|$; then an injection is established from the set of non-empty square factors of w to the set of small circuits in $\Gamma(w)$.

Example and notation 2 *Let p be a primitive word and let $[p]$ be the conjugacy class of p. For any positive integer $n \geq |p| - 1$, let us define*

$$[p]_n = \left\{ q^{\frac{n}{|p|}} | q \in [p] \right\},$$

where $q^{\frac{n}{|p|}}$ is a rational power of q. For example, let $p = aba$, then $[p]_4 = \{abaa, baab, aaba\}$ and $[p]_2 = \{ab, ba, aa\}$. Let $C(p, n)$ denote the circuit whose vertex set is $[p]_n$ and the edge set is $[p]_{n+1}$.

Let us consider the word $u = abaaabaaaabaaba$, the Rauzy graph $\Gamma_4(u)$ is as follows:



In this graph, there are three circuits: $C(aaaab, 4)$, $C(aaab, 4)$ and $C(aab, 4)$. Two of them are small, they are $C(aaab, 4)$ and $C(aab, 4)$, while $C(aaaab, 4)$ is not small. We can check that in this graph, $|L_u(5)| = 10$, $|L_u(4)| = 8$ and the total number of small circuits in this graph, which will be denoted by $S_4(u)$, is 2. Thus, $S_4(u) \leq |L_u(5)| - |L_u(4)| + 1$. □

Example 3 *Let us consider the word $w = baababaababbbabbabbbab$. We can check that $|w| = 22$ and there are 14 squares in w:*

ε, aa, bb, abab, baba, abaaba, bbabba, babbab, abbabb, babbbabb,
bbabbbab, baababaaba, aababaabab, babbbabbabbbab.

The nonempty squares can be sent injectively to the small circuits listed as follows:

$C(a, 1), C(b, 1), C(ab, 2), C(ab, 3), C(aba, 3), C(abb, 3), C(abb, 4),$
$C(abb, 5), C(babb, 4), C(babb, 5), C(baaba, 5), C(baaba, 6), C(babbbab, 7).$ □

The *square defect* of a (finite or infinite) word is defined similarly to the palindromic defect [3]:

Definition 4 *The square defect of a finite word w is the number $\mathcal{D}_s(w)$ satisfying*

$$\mathcal{D}_s(w) = |w| + 1 - S(w).$$

Let w be an infinite word, then the square defect of w is defined as

$$\mathcal{D}_s(w) = \sup\{\mathcal{D}_s(u) \mid u \in \mathrm{Fac}(w)\},$$

where $\mathrm{Fac}(w)$ *is the set of factors of w.*

It is clear from Theorem 1 that the square defect of a finite word is always positive; while the behavior of the square defect of an infinite word is less known.

We continue our study on the number of squares by proving the following results.

Proposition 5 *The square defect of any infinite periodic word is infinite.*

Proposition 6 *Any infinite word with finitely many squares has an infinite square defect.*

It is still an open question whether Proposition 6 remains true also when considering a generic infinite word. However, we conjecture that this is the case.

In [5] Brlek and Reutenauer proved the following identity linking the palindromic defect $\mathcal{D}_p(w)$ with the factor and palindromic complexities C_w and P_w:

$$2\mathcal{D}_p(w) = \sum_{n=0}^{|w|} C_w(n+1) - C_w(n) + 2 - P_w(n+1) - P_w(n) \tag{1}$$

It is proved that this identity holds for several examples of infinite words including periodic ones, the Thue-Morse word, all Sturmian ones, the Oldenburger exponent trajectory [5], as well as for languages closed by reversal [1] Later, this identity was extended to σ-palindromes where σ is an involution, also known as anti-palindromes [4].

In this contribution, we consider a similar identity using the square defect of a word. Let us define the identity:

$$2\mathcal{D}_s(w) = \sum_{n=0}^{|w|} C_w(n+1) - C_w(n) + 2 - S_w(n+1) - S_w(n). \tag{3}$$

It is easy to prove that Identity (3) holds for any finite word as well as any infinite periodic word.

We also investigate Identity (3) for the class of strict standard episturmian words. Recall that an infinite word $\mathbf{s} \in A^\omega$ is *standard episturmian* if there exists an infinite word $\Delta(\mathbf{s}) = \Delta_1\Delta_2\cdots$ with $\Delta_i \in A$, called the *directive word* of $\mathbf{s}$, such that the sequence of palindromic prefixes $(u_i)_{i\geq 1}$ of $\mathbf{s}$ is obtained as $u_1 = \varepsilon$ and

$$u_{n+1} = (u_n\Delta_n)^{(+)} \text{ for } n \geq 1,$$

where $w^{(+)}$ of a finite word w is the (unique) shortest palindrome having w as a prefix. A word $\mathbf{s}$ over the k-letter alphabet $A = \{a_1, \ldots, a_k\}$ is *strict standard episturmian* if it has a directive word of the form

$$\Delta(\mathbf{s}) = a_1^{d_1} a_2^{d_2} \cdots a_k^{d_k} a_1^{d_{k+1}} \cdots a_k^{d_{2k}} a_1^{d_{2k+1}} \cdots$$

with $d_i > 0$ for all i.

Theorem 7 *The right hand side of Identity* (3) *is infinite for every strict standard episturmian word.*

We conjecture that Identity (3) is satisfied by every infinite word.

References

[1] L. Balková, E. Pelantová, and Š. Starosta. *Proof of the Brlek-Reutenauer conjecture.* Theor. Comput. Sci. **475** (2013), 120–125.

[2] S. Brlek and S. Li. *On the number of squares in a finite word.* arXiv e-prints (April 2022), arXiv:2204.10204.

[3] S. Brlek, S. Hamel, M. Nivat, and C. Reutenauer. *On the palindromic complexity of infinite words.* International Journal on Foundation of Computer Science **15** (2004), 293–306.

[4] S. Brlek and N. Lafrenière. *Reconstructing words from a σ-palindromic language.* Fund. Inform. **135** (2014), 59–72.

[5] S. Brlek and C. Reutenauer. *Complexity and palindromic defect of infinite words.* Theoretical Computer Science **412** (2011), 493–497.

[6] A. Deza, F. Franek, and A. Thierry. *How many double squares can a string contain?* Discrete Applied Mathematics **180** (2015), 52–69.

[7] A. S. Fraenkel and J. Simpson. *How many squares can a string contain?* J. Comb. Theory, Ser. A **82** (1998), 112–120.

[8] L. Ilie. *A note on the number of squares in a word.* Theor. Comput. Sci. **380** (2007), 373–376.

[9] A. Thierry. A proof that a word of length n has less than $1.5n$ distinct squares, (2020).