# The Formal Language Theory Column

by

**Giovanni Pighizzini**

Dipartimento di Informatica
Università degli Studi di Milano
20135 Milano, Italy
pighizzini@di.unimi.it

# Open and Closed Words

Gabriele Fici
Dipartimento di Matematica e Informatica
Università di Palermo
via Archirafi 34, 90123 Palermo, Italy
gabriele.fici@unipa.it

**Abstract**

Combinatorics on words aims at finding deep connections between properties of sequences. The resulting theoretical findings are often used in the design of efficient combinatorial algorithms for string processing, but may also have independent interest, especially in connection with other areas of discrete mathematics. The property we discuss here is, for a given finite word, that of being closed. A finite word is called closed if it has length $\leq 1$ or it contains a proper factor (substring) that occurs both as a prefix and as a suffix but does not have internal occurrences. Otherwise the word is called open. We illustrate several aspects of open and closed words and factors, and propose some open problems.

## 1 Introduction

In combinatorics on words, one often classifies finite or infinite words according to some combinatorial property. For example, a finite word can be primitive (e.g. *aba*) or a power of another word (e.g. $abaaba = (aba)^2$). Another example is the property of having a *border* (a proper factor that occurs both as a prefix and as a suffix of the word), or being unbordered. Once a property has been chosen, one may look at the factors of a given word by separating those that verify the property from those that do not. For example, the word *abaab* has five distinct bordered factors, namely *aa*, *aba*, *baab*, *abaa* and *abaab* and all the other factors (the empty word $\varepsilon$, *a*, *b*, *ab*, *ba*, *aab* and *baa*) are unbordered. Several papers have been devoted to the study of those words that are extremal with respect to the proportion of factors that verify a given property, for example words with the maximum (or minimum) number of distinct square factors [17, 21, 22], palindromic factors [7, 19, 15], unbordered factors [27, 20], etc.

Despite the simplicity of the definitions, there are several natural questions on these topics that are remaining unanswered for many decades. For example, nobody still knows if the following conjecture, attributed to Fraenkel and Simpson [17], holds true:

**Conjecture 1.** *Any word of length n contains less than n distinct square factors.*

For infinite words, one may check whether there are arbitrarily long factors (or prefixes) verifying a given property. For example, every aperiodic infinite word contains arbitrarily long unbordered factors, while in a purely periodic infinite word the maximum length of an unbordered factor is bounded. [1]

An approach we find particularly promising consists in associating, with a given infinite word, an infinite binary sequence whose $i$-th element is 1 or 0 depending whether the prefix of length $i$ of the word verifies the chosen property or not. This may be seen as the *characteristic sequence* of the property for the given word. For example, take the infinite Fibonacci word [2]

$$F = abaababaabaababaababa\cdots$$

and the property of being a square (concatenation of a word with itself). Then the corresponding characteristic sequence is

$$\chi_{\mathrm{sq}}(F) = 00000100010000010000\cdots$$

where there is a 1 at position $i$ if and only if $i$ is twice a Fibonacci number and $i \geq 6$ (that is, at positions 6, 10, 16, 26, 42, 68, etc.).

In this contribution, we explore the property of being *closed* or *open*.

**Definition 2.** A finite word is *closed* if it has length $\leq 1$ or it contains a proper factor that occurs both as a prefix and as a suffix but does not have internal occurrences. Otherwise the word is *open*.

For example, the word *aba* is closed since the factor *a* appears only as a prefix and as a suffix, while the word *ab* is open since no factor appears only as a prefix and as a suffix.

The first binary closed words are:

$$\varepsilon, a, b, aa, bb, aaa, aba, bab, bbb, aaaa, abab, abba, baab, baba, bbbb.$$

[1] Recall that an infinite word is called *purely periodic* if it has the form $x^\omega$, i.e., it is obtained by concatenating a finite word $x$ with itself infinitely many times; it is called *periodic* if it has the form $yx^\omega$ for two finite words $x$ and $y$; finally, it is called *aperiodic* if it is not periodic.

[2] The Fibonacci word $F$ can be defined as the word over the alphabet $\{a, b\}$ in which the distance between the $n$-th $b$ and the $n$-th $a$ is $n$, for every $n > 0$. The name comes from the fact that this word is intimately related to the well-known sequence of Fibonacci numbers $F_1 = 1$, $F_2 = 1$, $F_n = F_{n-1} + F_{n-2}$ for $n > 2$. For further details, the reader may look at [4] and [14].

In what follows, we call *frontier* the factor of a closed word that occurs in it only as a prefix and as a suffix (the frontier of words of length 1 is the empty word $\varepsilon$). Note that a word cannot have more than one factor that occurs in it only as a prefix and as a suffix, without internal occurrences. Hence the frontier of a closed word is unique. Also note that the frontier of a closed word is its longest border.

The notion of closed word is known in the literature also with the name of *periodic-like* word [11, 8]. An equivalent notion is that of a *complete return* to a factor, as considered in [19]. A complete return to the factor $u$ in a word $w$ is any factor of $w$ having exactly two occurrences of $u$, one as a prefix and one as a suffix. Hence, a word $w$ is closed if and only if it is a complete return to one of its factors; such a factor is clearly both the longest repeated prefix and the longest repeated suffix of $w$ (i.e., the frontier of $w$).

Another related notion is that of *privileged word* [24, 25, 26, 16]. A word $w$ is called privileged if it has length $\leq 1$ or it has a privileged border that appears exactly twice in $w$. Therefore, a privileged word is always closed, but there exist closed words that are not privileged, e.g. *abab*, *ababab*, *ababbabab*, etc.

# 2 General Remarks on Open and Closed Words

The following characterizations of closed words follow easily from the definition:

1. the longest repeated prefix (resp. suffix) of $w$ does not have internal occurrences in $w$, i.e., occurs in $w$ only as a prefix and as a suffix;

2. the longest repeated prefix (resp. suffix) of $w$ does not have two occurrences in $w$ followed (resp. preceded) by different letters;

3. $w$ has a border that does not have internal occurrences in $w$;

4. the longest border of $w$ does not have internal occurrences in $w$.

Obviously, the negations of the previous properties characterize open words.

For any letter $a$ in the alphabet and for any integer $n$, the word $a^n$ is closed, $a^{n-1}$ being a factor occurring only as a prefix and as a suffix in it. This observation can be generalized by considering the *exponent* of a finite word. Recall that the period of a word $w$ is the least positive integer $p$ such that $w_i = w_{i+p}$ for every $i = 1, \ldots, |w| - p$. The exponent of the word $w$ is the ratio between the length and the period of $w$. So for example, the period of the word *abaab* is 3, thus its exponent is $5/3$.

We have the following property:

**Proposition 3.** *Any word whose exponent is at least* 2 *is closed.*

# 3 The Language of Closed Words

There are much more open words than closed words of length $n$ as $n$ grows. Indeed, for any nonempty word $w$, there exists at most one letter $x$ such that $wx$ is closed [12]. Even in the binary case, already at $n = 30$, closed words are less than 3% of the the total [28].

However, the number of closed words of length $n$ grows exponentially in $n$. To see this, it is sufficient to observe that, as a consequence of Proposition 3, the word $ww$ is closed for any choice of the word $w$.

More precise bounds may be derived for the number of closed words of each length. For example, it is known that for every $n$, there are at least $\frac{2^{n-5}}{n^2}$ privileged words [16], and privileged words are closed.

From the point of view of the complexity of the language of closed words in the Chomsky hierarchy, we have that, as soon as the cardinality of the alphabet $\Sigma$ is larger than 1, the language of closed words over $\Sigma$ is not context-free (and its subset formed by the privileged words is also non-context-free) [26].

# 4 Sturmian Words and Rich Words

A deeply studied and particularly interesting class of words is that of Sturmian words. There exist several equivalent definitions of Sturmian words. One is the following: An infinite binary word is Sturmian if it is balanced and aperiodic. Here balanced means that for any two factors $u$ and $v$ of the same length, the difference between the number of occurrences of each letter in $u$ and $v$ is bounded by 1. So, for example, a word containing both *aaa* and *bab* as factors cannot be balanced. Equivalently, an infinite word is Sturmian if it contains exactly $n + 1$ distinct factors of length $n$ for every $n \geq 0$. The Fibonacci word is an example of a Sturmian word.

Among the dozens of equivalent definitions of Sturmian words, there is also one based on closed words: An infinite word $w$ is Sturmian if and only if for every of its factors $v$, there are exactly two closed factors of $w$ whose frontier is $v$ [29]. For example, in the Fibonacci word $F$, take the factor *abaa*. The closed factors of $F$ having *abaa* as their frontier are *abaababaa* and *abaabaa*.

A finite word is called Sturmian if it is a factor of some infinite Sturmian word, i.e., if it is a balanced binary word.

Every finite Sturmian word contains the largest number of distinct palindromic factors a word of the same length can contain, that is equal to the length of the word plus 1 (considering also the empty word). A word (not necessarily binary) with this property is called *rich in palindromes* (or, simply, *rich*). For example, the Sturmian word *abaab* has length 5 and indeed contains 6 distinct palindromic

factors ($\varepsilon$, $a$, $b$, $aa$, $aba$ and $baab$). There exist binary words that are rich but not Sturmian, e.g. $aaabab$ — these examples also show that a word can be rich in palindromes even if itself is not a palindrome.

The following characterization of rich words is related to closed words: A word is rich if and only if every of its closed factors that has a palindromic frontier is itself a palindrome [19].

Sturmian words can be open or closed, but Sturmian palindromes are always closed [9]. Actually, something stronger holds: any rich palindrome is closed. The converse is not true, for example the word $w = aaababbaabbabaaa$ is a closed palindrome of length 16 but contains only 16 palindromic factors, hence it is not rich (there do not exist shorter examples).

# 5 Words with Few Closed Factors

For every length $n$, there exist words with quadratically (in $n$) many distinct closed factors. However, in contrast to the case of palindromic factors, a word of length $n$ must contain *at least* $n+1$ distinct closed factors [2]. A word containing exactly $n+1$ distinct closed factors is called *CR-poor*. As an example, $abca$ is a CR-poor word, since it has length 4 and exactly 5 closed factors, namely $\varepsilon$, $a$, $b$, $c$ and $abca$, whereas the word $ababa$ is not CR-poor since it has length 5 but contains 8 closed factors: $\varepsilon$, $a$, $b$, $aba$, $bab$, $abab$, $baba$ and $ababa$.

There are some relations between rich words and CR-poor words. For example, a palindromic word is rich if and only if all of its palindromic factors are closed [8], while if a word $w$ has the property that all of its closed factors are palindromes, then $w$ is a CR-poor word, and it is also rich [2]. Combining the two results, one has that a word $w$ has the property that its closed factors coincide with its palindromic factors if and only if $w$ is rich and CR-poor.

A characterization of CR-poor words is the following:

**Theorem 4.** *[2] A word is CR-poor if and only if every of its closed factors has a frontier that is a power of some letter.*

The complexity of CR-poor words, however, is rather weak, in fact they form a regular language.

# 6 The Characteristic Sequence of Open/Closed Prefixes

With any finite or infinite word $w$, one can associate the binary sequence $\chi_{\mathrm{cl}}(w)$ whose $i$-th element is 1 if the prefix of length $i$ of $w$ is closed, and 0 if it is open.

So for example, if $w = abaaab$, then $\chi_{\mathrm{cl}}(w) = 101001$.

Recall that a word (or a sequence) is called *recurrent* if each of its factors occurs infinitely often in it. For an infinite word $w$, the sequence $\chi_{\mathrm{cl}}(w)$ is not aperiodic if and only if $w$ is either periodic or not recurrent. In the first case, $\chi_{\mathrm{cl}}(w)$ ends in $1^\omega$, while in the latter case it ends in $0^\omega$. In all the other cases, $\chi_{\mathrm{cl}}(w)$ is an aperiodic sequence [12].

As an example, consider the Fibonacci word $F$. One has

$$\chi_{\mathrm{cl}}(F) = 10101100111000111110\cdots$$

where the lengths of the blocks of consecutive equal symbols form the sequence of Fibonacci numbers $1, 1, 2, 3, 5, \ldots$

The structure of the sequence $\chi_{\mathrm{cl}}$ has been characterized for every Sturmian word [12]. Sturmian words are particularly interesting in this context because of the following remarkable property:

**Theorem 5.** *[12] Every finite or infinite Sturmian word can be uniquely reconstructed (up to renaming letters) from its $\chi_{cl}$ sequence.*

This property does not hold in general, for example the words *aaba* and *aabb* are both associated with the characteristic sequence 1100 (and in fact *aabb* is not Sturmian).

There is an algorithm running in linear time that reconstructs a Sturmian word (up to renaming letters) from its $\chi_{\mathrm{cl}}$ sequence [12]. We now describe a simple linear-time algorithm for computing the $\chi_{\mathrm{cl}}$ sequence of any word.

Recall that the *border array* $B(w)$ of a word $w$ is the integer array whose $i$-th entry is the length of the longest border of the prefix of length $i$ of $w$. For example, if $w = abcaacab$, then $B(w) = [0, 0, 0, 1, 1, 0, 1, 2]$. We also define the array $B'(w)$ by $B'(w)[i] = \max_{j \le i} B(w)[j]$.

**Proposition 6.** *Let $w$ be a nonempty word. Then $\chi_{cl}(w)[1] = 1$ and for every $i > 0$, $\chi_{cl}(w)[i] = B'(w)[i] - B'(w)[i-1]$.*

As an example, for the word $w = abcaacab$, we have that $B'(w) = [0, 0, 0, 1, 1, 1, 1, 2]$, and indeed $\chi_{\mathrm{cl}}(w) = 10010001$.

Since the border array of a word can be computed in linear time with respect to its length [23], Proposition 6 gives a linear-time algorithm to compute the $\chi_{\mathrm{cl}}$ sequence of a word.

# 7 The Longest Closed Factor Array

The Longest Closed Factor Array [3] of a word $w$ of length $n$ is the integer array $LC_w[1 \ldots n]$ such that for every $1 \le i \le n$, $LC_w[i] = \ell$ if and only if $\ell$ is the

length of the longest closed factor of $w$ starting at position $i$. For example, for $w = abcaacab$, one has $LC_w = [8, 7, 5, 2, 3, 1, 1, 1]$.

Note that for every $i$, $LC_w[i]$ is equal to the position of the rightmost 1 in the $\chi_{\mathrm{cl}}$ array of the suffix of $w$ starting at position $i$; that is, by Proposition 6, $LC_w[i]$ is equal to the position of the leftmost occurrence of the maximum in the border array of the suffix of $w$ starting at position $i$.

For example, the border arrays of the suffixes of $w = abcaacab$ are the following (the leftmost occurrence of the maximum is underlined):

$$\begin{aligned} B(abcaacab) &= [0, 0, 0, 1, 1, 0, 1, \underline{2}] \\ B(bcaacab) &= [0, 0, 0, 0, 0, 0, \underline{1}] \\ B(caacab) &= [0, 0, 0, 1, \underline{2}, 0] \\ B(aacab) &= [0, \underline{1}, 0, 1, 0] \\ B(acab) &= [0, 0, \underline{1}, 0] \\ B(cab) &= [\underline{0}, 0, 0] \\ B(ab) &= [\underline{0}, 0] \\ B(b) &= [\underline{0}] \end{aligned}$$

and taking the positions of the underlined elements one gets the Longest Closed Factor Array of $w$.

However, there are more efficient ways to compute the Longest Closed Factor Array of a word. It is known that the Longest Closed Factor Array of a word of length $n$ can be computed in time $O(n\sqrt{\log n})$ [3]. It is an open problem whether it can be computed in linear time (another algorithm exists running in time $O(n\frac{\log n}{\log\log n})$ [1]).

But the Longest Closed Factor Array also has a combinatorial interest, because of the following remarkable result:

**Theorem 7.** *[3]. Every word is uniquely determined (up to renaming letters) by its Longest Closed Factor Array.*

It is known that a word of length $n$ over a fixed-size alphabet can be reconstructed from its Longest Closed Factor Array in time $O(n \log\log n)$ [3]. Here again, it is an open problem whether a word can be reconstructed from its Longest Closed Factor Array in linear time.

# 8 Closed Length

Since single letters are closed, every word can be factored in closed words. Of course, for a word of length $n$ there could be a factorization in less than $n$ closed

words. We are interested in the *minimum* number of closed factors in which a word can be factored. For example, the word *ababc* can be factored in three closed words ($aba \cdot b \cdot c$ or $a \cdot bab \cdot c$) but not in two. We call this minimum number the *closed length* of the word.

The same approach was previously introduced using palindromes in place of closed words. The *palindromic length* of a word is the minimum number of palindromes in which it is possible to factorize that word [18]. Also in this case, the palindromic length of a word of length $n$ is at most $n$ since single letters are palindromes. As an example, the palindromic length of the word *aababb* is 3.

The following nice conjecture is still open in general, even if it has been proved for several classes of infinite words:

**Conjecture 8.** *[18] An infinite word is aperiodic if and only if the palindromic length of its factors is unbounded.*

For example, it is known that the palindromic length of a factor of the Fibonacci word can be arbitrarily large [18]. On the contrary, it can be proved that the closed length of any Sturmian word is at most two [30]. By the way, for any infinite Sturmian word $w$, there exist infinitely many $n$ for which *all* the factors of $w$ of length $n$ have palindromic length two [5].

Very recently, it has been proved that the palindromic length of a word can be computed in linear time with respect to the length of the word [6]. We conclude with the following problem: Is it possible to compute the closed length of a word of length $n$ in time $O(n)$?

# References


[1] Badkobeh, G., Bannai, H., Goto, K., I, T., Iliopoulos, C.S., Inenaga, S., Puglisi, S.J., Sugimoto, S.: Closed factorization. Discr. Appl. Math. 212: 23–29 (2016).

[2] Badkobeh, G., Fici, G., Lipták, Zs.: On the Number of Closed Factors in a Word. In: LATA 2015, 9th International Conference on Language and Automata Theory and Applications. Lecture Notes in Computer Science, vol. 8977, pp. 381–390. Springer (2015).

[3] Bannai, H., Inenaga, S., Kociumaka, T., Lefebvre, A., Radoszewski, J., Rytter, W., Sugimoto, S., Walen, T.: Efficient Algorithms for Longest Closed Factor Array. In: SPIRE 2015, 22nd International Symposium on String Processing and Information Retrieval. Lecture Notes in Computer Science, vol. 9309, pp. 95–102. Springer (2015).

[4] Berstel, J.: Fibonacci Words — A Survey. In G. Rozenberg and A. Salomaa, editors, *The Book of L*, pp. 11–25, Springer-Verlang (1985).

[5] Borchert, A., Rampersad, N.: Words with many Palindrome Pair Factors. Electr. J. Comb. 22(4): P4.23 (2015).

[6] Borozdin, K., Kosolobov, D., Rubinchik, M., Shur, A.M.: Palindromic Length in Linear Time. In: CPM 2017, 28th Annual Symposium on Combinatorial Pattern Matching. LIPIcs vol. 78, 23:1-23:12, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2017).

[7] Brlek, S., Hamel, S., Nivat, M., Reutenauer, C.: On the palindromic complexity of infinite words. Internat. J. Found. Comput. Sci. 15, 293–306 (2004).

[8] Bucci, M., de Luca, A., De Luca, A.: Rich and Periodic-Like Words. In: DLT 2009, 13th International Conference on Developments in Language Theory, Lecture Notes in Comput. Sci., vol. 5583, pp. 145–155. Springer (2009).

[9] Bucci, M., De Luca, A., Fici, G.: Enumeration and Structure of Trapezoidal Words. Theoret. Comput. Sci. 468, 12–22 (2013).

[10] Bucci, M., De Luca, A., Glen, A., Zamboni, L.: A new characteristic property of rich words. Theoret. Comput. Sci. 410(30), 2860–2863 (2009).

[11] Carpi, A., de Luca, A.: Periodic-like words, periodicity and boxes. Acta Inform. 37, 597–618 (2001).

[12] De Luca, A., Fici, G., Zamboni, L.: The Sequence of Open and Closed Prefixes of a Sturmian Word. Adv. Appl. Math. 90, 27–45 (2017).

[13] Fici, G.: A Classification of Trapezoidal Words. In: WORDS 2011, 8th International Conference on Words. pp. 129–137. No. 63 in Electronic Proceedings in Theoretical Computer Science (2011).

[14] Fici, G.: Factorizations of the Fibonacci Infinite Word. J. Integer Seq. 18: Article 15.9.3 (2015).

[15] Fici, G., Zamboni, L.: On the least number of palindromes contained in an infinite word. Theoret. Comput. Sci. 481: 1–8 (2013).

[16] Forsyth, M., Jayakumar, A., Peltomäki, J., Shallit, J.: Remarks on Privileged Words. Int. J. Found. Comput. Sci. 27(4): 431–442 (2016).

[17] Fraenkel, A.S., Simpson, J.: How Many Squares Can a String Contain? J. Comb. Theory, Ser. A 82(1): 112–120 (1998).

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

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

[20] Goc, D., Mousavi, H., Shallit, J.: On the Number of Unbordered Factors. In: LATA 2013, 7th International Conference on Language and Automata Theory and Applications. Lecture Notes in Computer Science, vol. 7810, pp. 299–310. Springer (2013).

[21] Ilie, L.: A simple proof that a word of length n has at most 2n distinct squares. J. Comb. Theory, Ser. A 112(1): 163–164 (2005).

[22] Ilie, L.: A note on the number of squares in a word. Theoret. Comput. Sci. 380(3): 373–376 (2007).

[23] Morris, J.H. , Pratt, V.R.: A Linear Pattern Matching Algorithm, Tech. Rep. 40, Computing Center, University of California, Berkeley, (1970).

[24] Peltomäki, J.: Introducing privileged words: Privileged complexity of Sturmian words. Theoret. Comput. Sci. 500, 57–67 (2013).

[25] Peltomäki, J.: Privileged factors in the Thue-Morse word - A comparison of privileged words and palindromes. Discr. Appl. Math. 193: 187–199 (2015).

[26] Peltomäki, J.: Privileged Words and Sturmian Words. Ph.D. Dissertation, TUCS Dissertations 214, (2016).

[27] Saari, K.: Unbordered words with the smallest number of distinct unbordered factors. Workshop on Challenges in Combinatorics on Words, Fields Institute, Toronto, April 21-26, 2013.

[28] Sloane, N.J.A.: The On-Line Encyclopedia of Integer Sequences. Available electronically at `http://oeis.org`. Sequence A226452.

[29] Vuillon, L.: A Characterization of Sturmian Words by Return Words. Eur. J. Comb. 22(2): 263–275 (2001).

[30] Zamboni, L.: Personal communication (2015).