Subword complexity and decomposition of the set of factors to sets of bounded complexity

Julien Cassaigne*, Anna Frid†, Svetlana Puzynina‡, and Luca Q. Zamboni§

June 23, 2014

Abstract

In this abstract we explore a new hierarchy of classes of languages and infinite words and its connection with complexity classes. Namely, we say that a language belongs to the class $\mathcal{L}_k$ if it is a subset of the catenation of k languages $S_1 \cdots S_k$, where the number of words of length n in each of S_i is bounded by a constant. The class of infinite words whose set of factors is in $\mathcal{L}_k$ is denoted by $\mathcal{W}_k$. In this paper we focus on the relations between the classes $\mathcal{L}_k$, $\mathcal{W}_k$ and the subword complexity of infinite words, which is as usual defined as the number of factors of the word of length n. In particular, we prove that the class $\mathcal{W}_2$ coincides with the class of infinite words of linear complexity. The class $\mathcal{W}_k$ is included to the class of words of complexity $O(n^{k-1})$, but this inclusion is strict for $k \geq 3$. For the class $\mathcal{L}_k$ the inclusions do not hold at all.

The content of the abstract intersects with that of our previous paper [1], but here we try to emphasize results not included to that paper. All the proofs included here are new; the omitted proofs can be found in [1].

1 Classes and basic hierarchy

We consider finite and infinite words over a finite alphabet Σ, i.e., finite or infinite sequences of elements from the set Σ. A *factor* or a *subword* of an infinite word is any sequence of its consecutive letters. The factor $u_i \cdots u_j$ of an infinite word $u = u_1 \cdots u_n \cdots$, with $u_k \in \Sigma$, is denoted by $u[i..j]$. As usual, the set of factors of a finite or infinite word u is denoted by $\mathrm{Fac}(u)$.

The number of factors of length n of a language L is denoted by $p_L(n)$ and is called the *(subword) complexity* function of L. The complexity $p_u(n)$ of an infinite word u is defined as $p_{\mathrm{Fac}(u)}(n)$ [2]. Denote by $\mathcal{P}(\alpha)$ the set of infinite words of complexity $O(n^\alpha)$.

A factor s of a right infinite word u is called *left special* if $as, bs \in \mathrm{Fac}(u)$ for some distinct letters $a, b \in \Sigma$. The length of a finite word s is denoted by $|s|$, and the number of occurrences of a letter a in s is denoted by $|s|_a$. The empty word is denoted ε and we define $|\varepsilon| = 0$. In the paper we mostly follow the terminology and notation from [3].

Let us introduce the classes $\mathcal{L}_k$ of languages and $\mathcal{W}_k$ of infinite words as follows: a language L (infinite word u) belongs to the class $\mathcal{L}_k$ (resp., $\mathcal{W}_k$) if

$$L \subseteq S_1 \cdots S_k$$

*Aix-Marseille Université, France

†Corresponding author, Aix-Marseille Université, France

‡Department of Mathematics and Statistics, University of Turku, Finland, and Sobolev Institute of Mathematics, Novosibirsk, Russia

§Department of Mathematics and Statistics, University of Turku, Finland, and Université de Lyon 1, France

(resp., $\mathrm{Fac}(u) \subseteq S_1 \cdots S_k$) for some languages S_i with $p_{S_i}(n) = O(1)$. In other words, $u \in \mathcal{W}_k$ if and only if $\mathrm{Fac}(u) \in \mathcal{L}_k$, and the condition $p_{S_i}(n) = O(1)$ means exactly that for some constant C we have $p_{S_i}(n) \leq C$ for all n. We also have $\mathcal{P}(0) = \mathcal{W}_1$.

2 The class $\mathcal{W}_2$ and linear complexity

By a simple cardinality argument, we have the following inclusion:

Lemma 1. *For each integer $k > 0$, we have $\mathcal{W}_{k+1} \subseteq \mathcal{P}(k)$.*

Example 1. Let us show that the Thue-Morse word $t = 01101001\cdots$, defined as the fixed point starting with 0 of the morphism $\varphi : 0 \to 01, 1 \to 10$, belongs to $\mathcal{W}_2$. For each n the Thue-Morse word consists of words $t_n = \varphi^n(0)$ and $\overline{t_n} = \varphi^n(1)$, both of them of length 2^n: $t = t_n\overline{t_n}\overline{t_n}t_n\overline{t_n}t_nt_n\overline{t_n}\cdots$. Defining S_1 to be the set of suffixes of all t_n and $\overline{t_n}$, and S_2 to be the set of their prefixes, we see that S_1 and S_2 contain exactly two words of length k each. To cut each factor w of t, we just choose any of its occurrences and a position m in it divided by the maximal power n of 2: $w = t[i..j] = t[i..m]t[m+1..j]$. By the definition of m, $t[i..m]$ is a suffix of t_n or $\overline{t_n}$, and $t[m+1..j]$ is a prefix of one of them, and thus, $w \in S_1S_2$. So, $t \in \mathcal{W}_2$. This construction can be generalized to any fixed point of a primitive morphism but obviously not to fixed points whose complexity is higher than linear (see [4] for examples).

Moreover, in fact, we can prove that all infinite words of linear complexity are in $\mathcal{W}_2$:

Theorem 1 ([1])**.** *An infinite word is of linear complexity if and only if its language of factors is a subset of the catenation of two languages of bounded complexity: $\mathcal{W}_2 = \mathcal{P}(1)$.*

The proof of this theorem is based on the existence of a bounded number of *markers* of each length. A subset M of factors of length n of a word u is said to be a set of *D-markers* if each factor of length Dn of u contains an element of M. In fact, in a word u of linear complexity there exist constants D and R such that we can always choose a set M of D-markers of a given length of cardinality at most R. Indeed, we can choose M to be the set of left special factors of this length [2], as we did in [1]. However, here we give another more direct and more general construction:

Lemma 2. *For each infinite word u which is not ultimately periodic and for each n we can find a set M of 3-markers of length n in $\mathrm{Fac}(u)$ such that $\#M \leq p_u(4n)/n$.*

Proof. Let us construct the set M and the accociated set W of *surrounding words* of length $3n$ inductively as follows: starting from the empty sets M and W, we put the first marker $m_1 = u[2n+1..3n] \in M$ and the surrounding word $w_1 = u[n+1..4n] \in W$. So, $m_1 = w_1[n+1..2n]$.

From now on suppose for each i that we have already added $i-1$ elements to each of M and W and consider the factors of u of length $3n$. If each of them contains a factor from M, we are done. If not, take a word of length $3n$ not containing any element of M as a factor, call it w_i and put to W. Define $m_i = w_i[n+1..2n]$.

Clearly, since there is a finite number of factors of length $3n$ in u, the process is finite. It remains to prove the upper bound $p_u(4n)/n$ to the number of elements of W and thus of M.

To do it, for each element w_i of W, let us consider some its final occurrence which exists since u is not ultimately periodic: $w_i = u[k_i+1..k_i+3n]$. Now for each $j = 0, \ldots, n-1$ consider its *covering factor* $c(i,j) = u[k_i+1+j-n..k_i+3n+j]$. Clearly, the length of $c(i,j)$ is $4n$ and $w_i = c(i,j)[n-j+1..4n-j]$. Let us prove that if $c(i,j) = c(i',j')$, then $i = i'$ and $j = j'$.

Indeed, suppose that $c(i,j) = c(i',j')$ but $i' < i$. Then $w_i = c(i,j)[n-j+1..4n-j]$ and thus $m_i = c(i,j)[2n-j+1..3n-j]$. Analogously, $w_{i'} = c(i,j)[n-j'+1..4n-j']$ and thus $m_{i'} = c(i,j)[2n-j'+1..3n-j']$. But since $j, j' \in \{0,\dots,n-1\}$, we have $2n-j'+1 \geq n-j+1$ and $3n-j' \leq 4n-j$, and thus $m_{i'}$ is a factor of w_i contradicting to the construction of w_i.

So, if $c(i,j) = c(i',j')$, we have $i = i'$. Suppose that $j' < j$; then $w_i = c(i,j)[n-j+1..4n-j] = c(i,j)[n-j'+1..4n-j']$. Consider the word $s = c(i,j)[n-j+1..4n-j']$. It is $(j-j')$-periodic, and in particular, its prefix w_i is $(j-j')$-periodic. So, $p(w_i) \leq j-j' \leq n$, and since each letter of s belongs either to the prefix occurrence of w_i, or to the suffix occurrence of w_i, or to both of them, we see that s is also $p(w_i)$-periodic. It immediately means that the prefix occurrence of w_i to s, $w_i = c(i,j)[n-j+1..4n-j]$, is not a final occurrence of w_i to u, contradicting to its choice. So, $j = j'$.

We have proved that all the words $c(i,j)$ are different. Their total number is $\#W\#\{j = 0,\dots,n-1\} = n\#M$. On the other hand, the length of each of them is $4n$ and thus their number is majorated by $p_u(4n)$. So, $n\#M \leq p_u(4n)$ and $\#M \leq p_u(4n)/n$, which was to be proved. □

The rest of the proof of Theorem 1 is based on the following idea: we build the sets of markers of length 2^k for all $k > 0$ and for each factor of u, find the first occurrence of the longest marker to it. Then we cut this marker in the middle to get $u = st$ with $s \in S$ and $t \in T$. So, we get $\mathrm{Fac}(u) \subseteq ST$; the proof that the complexities of S and T are indeed bounded is omitted here but given in [1].

We continue our attempts to generalize Theorem 1 to words of higher complexity and their decompositions to two sets of lower complexity, but at the moment have no more results to state. At the same time, several other generalizations clearly fail. The rest of the paper is devoted to discussing them.

3 A language of sublinear complexity but not in $\mathcal{L}_k$ for any k

In the previous section, we considered the language of factors of an infinite word u and proved that it is a subset of ST for two languages S and T of bounded complexity. The languages S and T are in general not factorial, but the language $\mathrm{Fac}(u)$ is. In what follows we prove that we cannot abandon the condition that the initial language is factorial: we construct a non-factorial language L whose complexity is less than linear and which is not in $\mathcal{L}_k$ for any k

In order to construct our language L we first introduce a sequence of words x_n on the alphabet $\{0,1,2\}$:

$$x_n = [n]_2 2,$$

where $[n]_2$ is the binary representation of n. For example, $x_5 = 1012$ and $x_{65} = 10000012$; clearly, $|x_n| = \lfloor \log_2 n \rfloor + 2$ Define also $y_n = x_n^{\left\lfloor \frac{n}{|x_n|^2} \right\rfloor}$: for example, $y_5 = \varepsilon$ since $|x_5| = 4$ and $5/4^2 < 1$, and $y_{65} = x_{65} = 10000012$ since $|x_n| = 8$ and $\lfloor 65/8^2 \rfloor = 1$.

Then the desired language L is defined as follows:

$$L = \{y_n : n \in \mathbb{N}\}.$$

Let us prove that $L \notin \mathcal{L}_k$ for any k. Suppose the opposite, i.e., that there exists a k such that $L \subseteq S_1 \cdots S_k$, and the complexity of each of S_i is bounded. Denote $S = \bigcup_{i=1}^{k} S_i$. Clearly, the complexity of S is also bounded.

Take some n_0 such that $|y_n| \geq k+1$ for all $n > n_0$. We can always do it since due to the definition of x_n we have $|x_n| \sim \log n$, and thus $|y_n| \sim n/\log n$.

Now, for any $n > n_0$, there exists $s_n \in S$ such that $2x_n \in \mathrm{Fac}(s_n)$. Indeed, $y_n \in L$ contains at least $k+1$ occurrences of the letter 2, so at least one of the elements of the partition into k elements from S contains two occurrences of 2 and hence has $2x_n$ as a factor. Clearly, $|s_n| \leq |y_n|$, since s_n is a factor of y_n. Besides that, for any $m \neq n$, we have $2x_m \notin \mathrm{Fac}(s_n)$. So, all the words s_n are distinct.

For any $n_1 > n_0$, consider $S_{n_1} = \{s_n : n_0 \leq n \leq n_1\} \subset S$. This set consists of $n_1 - n_0 + 1$ distinct words. All these words are of length at most $|y_{n_2}| = |x_{n_2}| \left\lfloor \frac{n_2}{|x_{n_2}|^2} \right\rfloor \sim \frac{n_2}{\log n_2}$, where $n_0 < n_2 \leq n_1$ and n_2 is chosen to maximize the length. So, the maximal length of a word from S_{n_1} is $o(n_1 - n_0)$ as $n_1 \to \infty$. A contradiction with the fact that the complexity of S is bounded for each length.

It remains to prove that the complexity of L is bounded by a linear function. Indeed, consider the numbers n such that $2^k \leq n < 2^{k+1}$; for all these n we have $|x_n| = k+2$ and $|y_n| = (k+2)\left\lfloor \frac{n}{(k+2)^2} \right\rfloor$. In particular, within this interval, $|y_n|$ takes each value at most $(k+2)^2$ times, since the value of $\lfloor n/p \rfloor$ changes each p units for a fixed p. Also, it is not difficult to check that for sufficiently large n we cannot have $|y_n| = |y_{n'}|$ if $|\lfloor \log_2 n \rfloor - \lfloor \log_2 n' \rfloor| > 1$. Indeed, consider n and n' such that $\lfloor \log_2 n \rfloor = k$, $\lfloor \log_2 n' \rfloor = k+2$ and $|y_{n'}| \leq |y_n|$. Since within each interval of the powers of 2 the length of y_n is non-decreasing, we can assume that $n' = 2^{k+2}$ and $n = 2^{k+1} - 1$. So, $|y_{n'}| \leq |y_n|$ implies

$$(k+4)\left(\frac{2^{k+2}}{(k+4)^2} - 1\right) \leq (k+2)\frac{2^{k+1}}{(k+2)^2},$$

that is,

$$2^{k+1}\frac{k}{(k+2)(k+4)} \leq k+4-\frac{1}{k+2}.$$

Clearly, it is only possible for small k, and if $k = \lfloor \log_2 n \rfloor$ is sufficiently large, then $|y_n| < |y_{n'}|$ for all n' such that $\lfloor \log_2 n' \rfloor = k+2$. Since the length of y_{2^k} is a non-decreasing function, as well as the length of y_n for a given $k = \lfloor \log_2 n \rfloor$, the same is true for all n' such that $\lfloor \log_2 n' \rfloor \geq k+2$.

So, each given value m of the length $|y_n|$ can be reached in intervals corresponding to most two consecutive values of $\lfloor \log_2 n \rfloor$, namely, k and $k+1$. It means that it can be reached at most $(k+2)^2 + (k+3)^2 \sim 2k^2$ times; here $m \sim \frac{2^k}{k}$. The complexity of L is indeed sublinear. □

4 Examples of low-complexity words not in $\mathcal{W}_k$, $k \geq 3$

Lemma 1 and Theorem 1 imply that $\mathcal{W}_2 = \mathcal{P}(1)$, and in general $\mathcal{W}_{k+1} \subseteq \mathcal{P}(k)$ for all k. So, the following natural question arises: is it true that $\mathcal{W}_{k+1} = \mathcal{P}(k)$ for all k?

The answer is negative.

Consider the word $u = ababbabbb\cdots = \prod_{i=1}^{\infty} ab^k$. Its complexity $p_u(n) = \Theta(n^2)$: this can be either proved directly or derived from the famous paper by Pansiot [4], since u is obtained by erasing the first letter c from the fixed point starting with c of the morphism $c \mapsto cab, a \mapsto ab, b \mapsto b$.

Lemma 3 ([1]). *The word u does not belong to $\mathcal{W}_3$.*

On the other hand, it can be proved that the word u belongs to $\mathcal{W}_6$, so, it is natural to reformulate our question as follows: Is it true that $\mathcal{P}(k) \subseteq \mathcal{W}_{f(k)}$ for some function f? In other terms, does there exist a function $f(k)$ such that the language of factors of any word whose complexity is $O(n^k)$ is a subset of a concatenation of $f(k)$ sets of bounded complexity?

The answer is also negative and the counterexample is given by the following construction.

Define the infinite word w as follows: fix a growing integer function $f(n)$ such that $f(1) \geq 1$, $f(n) \leq n$ and $f(n) \to \infty$, and consider the word

$$w = \prod_{p=1}^{\infty} \prod_{q=1}^{f(p)} (a^p b^q)^{k(p,q)},$$

where $k(p,q)$ is a growing function: $k(p,q) \leq k(p,q+1)$ and $k(p,f(p)) \leq k(p+1,1)$ for all p and q.

The complexity of w is $O(n^2 f(n))$, and it does not belong to any class $\mathcal{W}_k$, see [1] for the proof.

5 Open problems

Both counterexamples from the previous section are not recurrent: most factors occur in them only once. We believe that recurrent and even uniformly recurrent examples can also be constructed, and we would appreciate nice constructions.

Also, it is not clear if the complexity of our examples is minimal possible. Can anything be said about the words of complexity $o(n^2)$? Is there an example of complexity $O(n^2)$ not belonging to any $\mathcal{W}_k$?

A generalisation of Theorem 1 to words of higher complexity and the concatenation of two sets would also be appreciated. Of course if $p_u(n)$ grows faster than linearly, the cardinality of the sets cannot be bounded, but how low can it be in general?

References

[1] J. Cassaigne, A. Frid, S. Puzynina, L. Zamboni, Subword complexity and decomposition of the set of factors, accepted to MFCS 2014, `http://arxiv.org/abs/1406.3974`.

[2] J. Cassaigne, F. Nicolas, Factor complexity. Combinatorics, automata and number theory, 163–247, Encyclopedia Math. Appl., 135, Cambridge Univ. Press, 2010.

[3] M. Lothaire, Algebraic combinatorics on words. Cambridge University Press, 2002.

[4] J.-J. Pansiot, Complexité des facteurs des mots infinis engendrés par morphismes itérés. In: Paredaens, J. (ed.) ICALP 1984. LNCS, vol. 172, pp. 380–389. Springer, Heidelberg (1984).