Proof of a conjecture on word complexity

Florence Levé Patrice Séébold

Abstract

An integer n is k-reachable if there exists a word of length k which contains exactly n non-empty different factors. Given k, all the k-reachable integers are between k and $\frac{k(k+1)}{2}$ but, between these two values, not all the integers are k-reachable. We give a general construction which associates to each k a family of words containing for each k-reachable integer n, exactly one word having n different factors. This also proves the conjecture of Kása about the smallest number m_k such that all the integers between m_k and $\frac{k(k+1)}{2}$ are k-reachable.

Résumé

Un entier n est k-atteignable s'il existe un mot de longueur k contenant exactement n facteurs non vides différents. Étant donné k, tous les entiers k-atteignables sont compris entre k et $\frac{k(k+1)}{2}$; mais entre ces deux valeurs, tous les entiers ne sont pas k-atteignables. Nous donnons une construction générale pour associer à chaque k une famille de mots qui, pour tout entier k-atteignable n, contient exactement un mot ayant n facteurs différents. Cette construction nous permet également de prouver la conjecture de Kása concernant le plus petit nombre m_k tel que tous les entiers compris entre m_k et $\frac{k(k+1)}{2}$ sont k-atteignables.

1 Introduction

The combinatorial properties of words have been studied intensively since the 60-70's, although the first papers in this area are those of Thue at the beginning of the century [6, 7] (see [4, 5] for a general overview and a large bibliography). However, most of the works deal with infinite words. Less attention has been given to finite sequences (for more informations see, e.g., [1]).

Bull. Belg. Math. Soc. -1993 (0), 277–291

The present paper addresses the counting of the number of different factors of finite words, depending on their length and the cardinality of the alphabet. It is organized as follows. Section 2 contains definitions and notations. The main results are also introduced in this section. In Section 3, we study the complexity function and obtain a result (Proposition 3.3) which improves one of de Luca. Section 4 is dedicated to the presentation of the family F_k, and the last section contains the statements and proofs of the two main results (Theorems 5.1 and 5.3).

2 Preliminaries

The terminology and notations are mainly those of Lothaire [4, 5].

A *word* is a finite string of elements called *letters*. The *empty word* ε is the neutral element for the concatenation of words (the *concatenation* of two words u and v is the word uv).

The *length* of a word u, denoted by $|u|$, is the number of occurrences of letters in u. In particular $|\varepsilon| = 0$. If n is a nonnegative integer, u^n is the word obtained by concatenating n occurrences of the word u. Of course, $|u^n| = n \times |u|$.

A word w is called a *factor* (resp. a *suffix*) of u if there exist words x, y such that $u = xwy$ (resp. $u = xw$).

Let w be a word and m a positive integer. We denote by $C_m(w)$ the number of different factors of length m in w. For instance, if $w = abcab$ then $C_1(w) = 3$ and $C_2(w) = 3$. By convention, $C_0(w) = 0$.

We denote by $C(w)$ the *complexity* of w, that is, the total number of non-empty different factors of w: $C(w) = \sum_{m=0}^{|w|} C_m(w)$. For example, if $w = abcab$ then $C(w) = 12$. Of course, $C(w) = 0$ if and only if $w = \varepsilon$.

Let k be a nonnegative integer. An integer p is *k-reachable* if there exists a word w of length k such that $C(w) = p$. Remark here that 0 is k-reachable if and only if $k = 0$.

For a given value of k, k-reachable integers are all between k and $\frac{k(k+1)}{2}$. Indeed, the word of length k containing the minimum number of factors is the word a^k which has k different factors (one of each length) and the word having the maximum is $a_1 \cdots a_k$ (where all the letters a_i are different) which contains $\frac{k(k+1)}{2}$ different factors.

On the other hand, if $k \geq 3$ then integers between k and $\frac{k(k+1)}{2}$ are not all k-reachable. For example, the integer 4 is between 3 and 6, but it is not 3-reachable; in the same way, the only 5-reachable integers (thus, between 5 and 15) are 5, 9, 11, 12, 13, 14, 15 (see [3]).

For a given k, we denote by N_k the set of k-reachable integers. From above, $N_3 = \{3, 5, 6\}$ and $N_5 = \{5, 9, 11, 12, 13, 14, 15\}$.

A natural question is to characterize, for all $k \in \mathbb{N} - \{0\}$, the set N_k. We answer this question, associating to each integer k a family F_k of words which, for each k-reachable integer p, contains one and only one word of complexity p (see Theorem 5.3).

Now, for all $k \in \mathbb{N}$, we denote by m_k the smallest number such that all the integers between m_k and $\frac{k(k+1)}{2}$ are k-reachable:

$$m_k = min\{n \in \mathbb{N} \mid \forall p \in \mathbb{N}, n \leq p \leq \frac{k(k+1)}{2} \Rightarrow p \in N_k\}.$$

In the previous example, $m_3 = 5$ and $m_5 = 11$.

Such an integer m_k always exists for all $k \in \mathbb{N}$ (because the integer $\frac{k(k+1)}{2}$ is always k-reachable), and if $k = 0$ or $k = 1$ then $m_k = k = \frac{k(k+1)}{2}$.

Notice that if $k \geq 2$ there exists one and only one nonnegative integer l and one and only one integer i such that

$$k = \frac{l(l+1)}{2} + 2 + i \text{ and } 0 \leq i \leq l. \tag{1}$$

(If $k = 2$, $l = 0$; if $k = 3$ or $k = 4$, $l = 1$; etc.)

Remark that, in every case, $l \leq k - 2$.

Then, with any integer $k \geq 2$ can be associated a unique integer

$$b_k = \frac{l(l^2-1)}{2} + 3l + 2 + i(l+1).$$

Z. Kása conjectured [2, 3] that, for all $k \geq 2$, $m_k = b_k$. The construction of the family F_k also allows us to prove this conjecture (see Theorem 5.1).

3 Some properties of the complexity function

We first translate with our notations a result of de Luca [1, Proposition 4.2].

Proposition 3.1. *For each non empty word w there exist two positive integers x and y such that the function $m \mapsto C_m(w)$ is*

- *strictly increasing in the interval $[1, x]$*
- *non decreasing in the interval $[x, y]$*
- *decreasing by one ($C_{m+1}(w) = C_m(w) - 1$) in the interval $[y, |w|]$.*

In what follows, we shall give a new property of this complexity function. A direct consequence will be a strengthening of de Luca's result: in the interval [x,y] this function is first strictly increasing and then has a constant value (in particular it cannot be constant in different intervals).

The following property clarifies the evolution of $C_m(w)$ as a function of m.

Property 3.2. *For any word w and any integer m, if $C_m(w) = C_{m+1}(w)$ then, for all integers $i \in \mathbb{N}$, $C_{m+i}(w) \leq C_m(w)$.*

Proof. Suppose there exists a shortest word w such that the property is false for an integer m. Let $k = |w|$. One has $k \geq 1$ and $m < k$. Thus there exist a word w' and a letter x such that $w = w'x$, and w has a suffix u of length m. Let $t = C_m(w) = C_{m+1}(w)$. Two cases are possible.

- u is a factor of w'.

  Since the property is false for the integer m, there exists a least integer $i \geq 2$ such that $C_{m+i}(w) \geq t+1$. From Proposition 3.1, $C_{m+i-1}(w) = t$.

  Since u is a factor of w', each factor of w of length m has, in w, an occurrence that is not a suffix of w. In this case, each of these factors extends in exactly one way to the right to give a factor of length $m+1$. Now, since $C_{m+i-1}(w) = t$ and $C_{m+i}(w) \geq t+1$, there exists at least one factor of length $m+i-1$ extending to the right in at least two different ways. But then it would also be the case for its suffix of length m. This leads to a contradiction.

- u is not a factor of w'.

  For any integer i, the suffix of w of length $m+i$ is not a factor of w' (because if such a suffix exists then it ends with u). Thus $C_{m+i}(w') = C_{m+i}(w) - 1$. In particular, $C_m(w') = C_{m+1}(w') = t-1$.

  But, since $|w'| < |w|$, the property is true for w'. Consequently, for any integer i, $C_{m+i}(w') \leq t-1$. This implies $C_{m+i}(w) \leq t$. ∎

A direct consequence of Property 3.2 is that the two integers of Proposition 3.1, x and y, can be chosen such that the function $q \mapsto C_q(w)$ is constant for $x \leq q \leq y$. This property also implies that x is the least integer such that $C_x(w) = max\{C_q(w) | 1 \leq q \leq k\}$. Moreover, since this function is decreasing by one in the interval $[y, |w|]$, one has $y = |w| - C_x(w) + 1$.

Consequently, we obtain the following result which strengthens Proposition 3.1.

Proposition 3.3. *Let w be a word of length k, and m the least integer such that $C_{m+1}(w) \leq C_m(w)$. Then the function $q \mapsto C_q(w)$ is*

- *strictly increasing for q between 1 and m (this means $C_q(w) - C_{q-1}(w) > 0$, $2 \leq q \leq m$);*
- *constant (equal to $C_m(w)$) for q between m and $k - C_m(w) + 1$ (this means $C_m(w) = \ldots = C_{k-C_m(w)+1}(w)$);*
- *decreasing by one for q between $k - C_m(w) + 1$ and k (this means $C_q(w) = C_{q-1}(w) - 1 = k - q + 1$, $k - C_m(w) + 2 \leq q \leq k$).*

With this, the complexity of the word w is given by

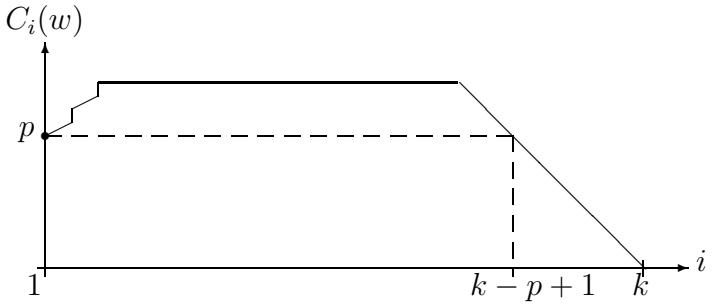
$$\begin{aligned} C(w) &= \sum_{i=0}^{m-1} C_i(w) + (k - C_m(w) + 1 - m + 1)C_m(w) + \sum_{i=1}^{C_m(w)-1} i \\ &= \sum_{i=0}^{m-1} C_i(w) + (k-m)s - \frac{s(s-3)}{2} \end{aligned} \tag{2}$$

where $s = max\{C_q(w) | 1 \leq q \leq k\}$, $m = min\{q \in \mathbb{N} | C_q(w) = s\}$ (and $C_{q+1}(w) - C_q(w) > 0$, $1 \leq q \leq m-1$).

From this we deduce the following property (given without proof in [3]).

Property 3.4. *Let w be a word of length k containing p different letters ($p \leq k$). Then $C(w) \geq p(k-p) + \frac{p(p+1)}{2}$.*

Proof. From Proposition 3.3, $C_i(w) \geq p$ for all $1 \leq i \leq k-p+1$, and $C_{k-j}(w) = j+1$ for $0 \leq j \leq p-2$.



Thus $C(w) \geq p(k-p+1) + \sum_{j=1}^{p-1} j$, i.e., $C(w) \geq p(k-p) + \frac{p(p+1)}{2}$. ∎

Remark that for each value of s, $C(w)$ will be maximal if $m = 1$. In this case

$$C(w) = ks - \frac{s(s-1)}{2} \tag{3}$$

Before continuing, we need to establish a few formulas linking k, l and b_k. We have $k = \frac{l(l+1)}{2} + 2 + i$ and $b_k = \frac{l(l^2-1)}{2} + 3l + 2 + i(l+1)$.
But

$$\begin{aligned} \frac{l(l^2-1)}{2} + 3l + 2 + i(l+1) &= \frac{l(l+1)(l-1)}{2} + (3+i)l + 2 + i \\ &= \frac{l^2(l+1)}{2} + (2+i)l - \frac{l(l+1)}{2} + l + 2 + i \\ &= l[\frac{l(l+1)}{2} + 2 + i] - \frac{l(l-1)}{2} + 2 + i \end{aligned}$$

So

$$\begin{aligned} b_k &= kl - \frac{l(l-1)}{2} + 2 + i & (4) \\ &= k(l+1) - l^2 & (5) \\ &= l(k-l) + k & (6) \end{aligned}$$

We are now able to prove a proposition that will be useful in Section 5.

Proposition 3.5. *Let w be a word of length k and $s = max\{C_q(w) | 1 \leq q \leq k\}$.*
1) If $s \leq l$ then $C(w) \leq b_k - 2$.
2) If $s \geq l+1$ then $C(w) \geq b_k$.

Proof. Let w be a word of length k and $s = max\{C_q(w) | 1 \leq q \leq k\}$. To guarantee the existence of l, we suppose $k \geq 2$.

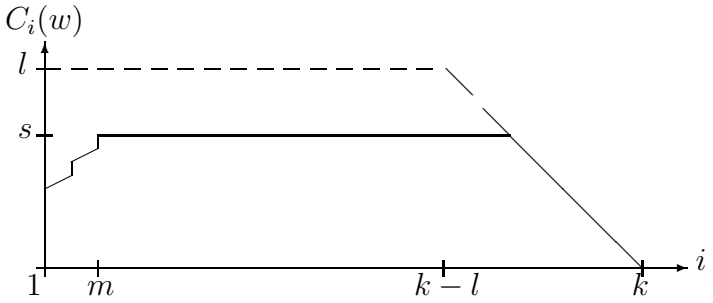
1) If $r \leq l$ then $l = s + p$, $p \in \mathbb{N}$.

Since $l \leq k-2$ (see Relation (1)), we have $2k-2s-p+1 \geq p+5 > 0$. Consequently, $ks - \frac{s(s-1)}{2} \leq kl - \frac{l(l-1)}{2}$ (because $kl - \frac{l(l-1)}{2} - (ks - \frac{s(s-1)}{2}) = \frac{p(2k-2s-p+1)}{2}$).

But, from (4), $b_k - 2 = kl - \frac{l(l-1)}{2} + i$, $0 \leq i \leq l$ thus $kl - \frac{l(l-1)}{2} \leq b_k - 2$. Moreover, $C(w) \leq ks - \frac{s(s-1)}{2}$ (see (3)).

Consequently, $C(w) \leq b_k - 2$.

This case is illustrated by the following picture where the dashed line corresponds to $s = l$ and $m = 1$ (the value of $C(w)$ is then maximal, equal to $b_k - 2 - i = kl - \frac{l(l-1)}{2}$).



2) Suppose now that $s \geq l + 1$.

To make the proof easier to understand, we first consider the case of a word containing at least $l + 1$ letters.

Lemma 3.6. *Let w be a word of length k. If w contains at least $l + 1$ different letters then $C(w) \geq b_k$.*

Proof of Lemma 3.6. Let w be a word of length k containing at least $l + 1$ different letters. According to Property 3.4 and since, when $k \geq 2$, $p(k - p) + \frac{p(p-1)}{2}$ is an increasing function of p for $1 \leq p \leq k$, we have $C(w) \geq (l+1)(k-l-1) + \frac{(l+1)(l+2)}{2}$, that is $C(w) \geq l(k - l) + k + \frac{l(l-1)}{2}$.

But from (6), we have $l(k - l) + k = b_k$. Since for all $l \geq 0$, $\frac{l(l-1)}{2} \geq 0$, we have then $C(w) \geq b_k$. ∎

Assume now that w contains $l - p$ different letters, $0 \leq p \leq l - 1$.

We can suppose that $p \leq l - 2$. Indeed, if $p = l - 1$ then the word w is written over only one letter and it contains only one factor of each length, thus not $l + 1$ for any length.

Let t be the smallest length for which w contains at least $l + 1$ factors ($t \geq 2$). Then w contains at most l factors of length $t - 1$ and so, according to Property 3.3, it contains at most $l - 1$ factors of length $t - 2$, $l - 2$ of length $t - 3$, …, $l - i + 1$ of length $t - i$, the number decreasing at least by 1 at each step. In particular, it contains at most $l - p + 1$ factors of length $t - p$. But w contains $l - p$ factors of length 1 and thus it must contain at least $l - p + 1$ factors of length 2.

This means that $t - p \leq 2$ (because if $t - p > 2$, w would contain a number of factors of length 2 strictly smaller than $l - p + 1$). So $t \leq p + 2$.

Consequently, w contains at least $l + 1$ factors of length $p + 2$, which means that the minimum number of factors corresponds to the following behaviour:

$$\left.\begin{array}{lll} l-p & \text{factors of length} & 1 \\ l-p+1 & & 2 \\ \cdots & \cdots & \cdots \\ l-p+p & & p+1 \\ l-p+p+1=l+1 & & p+2 \end{array}\right\} \quad (l-p)(p+2)+\sum_{i=1}^{p+1} i$$

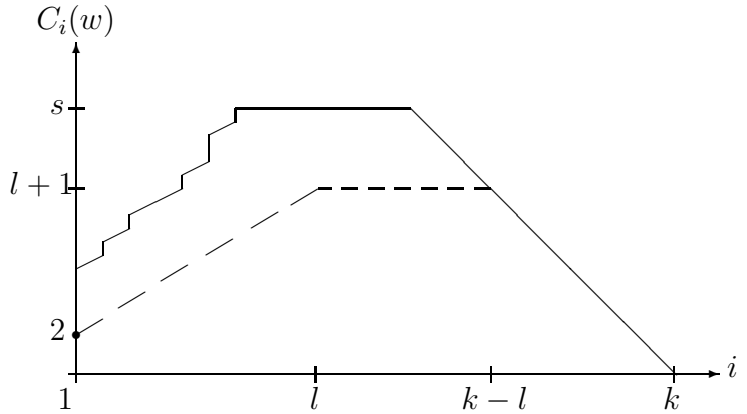
$$\left.\begin{array}{lll} l+1 & & p+3 \\ \cdots & \cdots & \cdots \\ l+1 & & k-l \end{array}\right\} \quad (l+1)(k-l-p-2)$$

$$\left.\begin{array}{lll} l & & k-l+1 \\ \cdots & \cdots & \cdots \\ 1 & & k \end{array}\right\} \quad \sum_{i=1}^{l} i$$

(Remark that this makes sense because $k-l-p-2 \geq 0$.
Indeed $k-l-p-2 = \frac{l(l-1)}{2}+i-p$.
But, since $0 \leq p \leq l-2$ and $0 \leq i \leq l$, $\frac{l(l-1)}{2}+i-p \geq \frac{l(l-1)}{2}-(l-2)$.
And $\frac{l(l-1)}{2}-(l-2) = \frac{(l-2)^2+l}{2} \geq 0$ if $l \geq 0$.)

$$\begin{aligned} \text{Hence } C(w) &\geq (l-p)(p+2)+\sum_{i=1}^{p+1} i+(l+1)(k-l-p-2)+\sum_{i=1}^{l} i \\ &\geq \frac{-p^2-3p-2}{2}-\frac{l(l+1)}{2}+kl+k \\ &\geq -\frac{(p+1)(p+2)}{2}+b_k+\frac{l(l-1)}{2} \\ &\geq b_k \text{ (because, since } 0 \leq p \leq l-2, \tfrac{l(l-1)}{2} \geq \tfrac{(p+1)(p+2)}{2}). \end{aligned}$$

This case is illustrated by the following picture where the dashed line corresponds to $s=l+1$ and $p=l-2$ (so $m=l<k-l$ because $k-2l-1=\frac{(l-1)(l-2)}{2}+i \geq 0$ (see (1)) and the value of $C(w)$ is then $b_k=k(l+1)-l^2$).



■

4 The family F_k

Remark. We wanted to be able to associate with each integer k a set of words which, for each k-reachable integer p, contains one and only one word of complexity p. Although it is not very hard to associate one word with each p, the difficulty is to obtain the unicity. We realize this with the rather complicated construction of a family of words named F_k.

Before describing the family F_k, let us recall that if u is a non-empty word of length q and p an integer ($p \geq q$) then the word $u^{\frac{p}{q}}$ is the *fractional power* of u of length p, which means that $u^{\frac{p}{q}}$ is the prefix of length p of u^p.

For example, if $u = baaaa$ ($q = 5$) and $p = 8$ then $u^{\frac{p}{q}} = u^{\frac{8}{5}} = baaaabaa$.

Let $k \in \mathbb{N}$, $k \geq 3$ (so $l \geq 1$ in (1)).

We consider the following words (all of length k)

$$w_{t,|u|,r} = ba^{k-t}(ba^{t-2-r})^{\frac{t-1-|u|}{t-1-r}}u \text{ with } 1 \leq t \leq k-2 \text{ and}$$

- if $t \leq l$ then $0 \leq |u| \leq t-1$

$$\text{and } \begin{cases} |u| \leq r \leq t-2 & \text{if } |u| \leq t-2 \\ r = t-1 & \text{if } |u| = t-1 \end{cases}$$

- if $t \geq l+1$, then $t-q \leq |u| \leq t-1$ where $q \geq 2$ is such that $1 + \sum_{j=0}^{q-2} j < k - t \leq 1 + \sum_{j=0}^{q-1} j$

$$\text{and } \begin{cases} 2t-k+\sum_{j=0}^{q-2} j \leq r \leq t-2 & \text{if } |u| = t-q \\ |u| \leq r \leq t-2 & \text{if } t-q+1 \leq |u| \leq t-2 \\ r = t-1 & \text{if } |u| = t-1 \end{cases}$$

- each letter of u is different from the others, and differs from a and b

- finally, by convention, if $r = |u| = t-1$ then $(ba^{t-2-r})^{\frac{t-1-|u|}{t-1-r}} = \varepsilon$ (that is $w_{t,t-1,t-1} = ba^{k-t}u$ with $|u| = t-1$).

Remark. Notice that the distinction $t \leq l$ or $t \geq l+1$ has a meaning since from Relation (1), $l \leq k-2$ in all cases.

To each value of k ($k \geq 3$), we associate the family F_k defined as follows: the first word of F_k is the word a^k; the last word of F_k is $a_1a_2 \dots a_k$ where all the letters a_i are different. Between these words is the ordered sequence of all the words $w_{t,|u|,r}$, where the order is given by the following rules:

we increase t starting from 1;
for each value of t, we increase $|u|$;
and for each value of $|u|$, we increase r.

Example. Let us construct the family F_6.

Here $k = 6$ thus, from Relation (1), $l = 2$ and $i = 1$.

We have to compute the words $w_{t,|u|,r}$ for each of the four values of t between 1 and $k-2 = 4$.

- $t = 1$

  In this case $t \leq l$ thus $|u|$ ranges between 0 and $t - 1 = 0$. Consequently, $|u| = t - 1 = 0$ and $r = |u| = t - 1 = 0$.

  Here, the only word is $w_{t,t-1,t-1} = w_{1,0,0} = ba^{k-1} = ba^5$.

- $t = 2$

  In this case $t \leq l$ thus $|u|$ ranges between 0 and $t - 1 = 1$.

  - $|u| = 0$: here $|u| \leq t - 2$ thus r ranges between 0 and $t - 2 = 0$.
    The only word is $w_{2,0,0} = ba^{k-2}b = ba^4b$.
  - $|u| = 1$: here $|u| = t - 1$ thus $r = |u| = t - 1 = 1$.
    Since all the letters of u differ from a and b, the only word is $w_{2,1,1} = ba^{k-2}u = ba^4c$.

- $t = 3$

  In this case $t \geq l + 1$. Let us compute $q \geq 2$ such that $1 + \sum_{j=0}^{q-2} j < k - t \leq 1 + \sum_{j=0}^{q-1} j$. Since $k - t = 3$, $q = 3$.

  Thus $|u|$ ranges between $t - q = 0$ and $t - 1 = 2$.

  - $|u| = 0$: here $|u| = t - q$ thus r ranges between $2t - k + \sum_{j=0}^{q-2} j = 1$ and $t - 2 = 1$.
    The only word is $w_{3,0,1} = ba^{k-3}b^2 = ba^3b^2$.
  - $|u| = 1$: here $t - q + 1 \leq |u| \leq t - 2$ thus r ranges between $|u| = 1$ and $t - 2 = 1$.
    The only word is $w_{3,1,1} = ba^{k-3}bu = ba^3bc$.
  - $|u| = 2$: here $r = |u| = t - 1 = 2$.
    The only word is $w_{3,2,2} = ba^{k-3}u = ba^3cd$.

- $t = 4$

  In this case $(t \geq l + 1)$, $k - t = 2$ thus $q = 2$. Consequently $|u|$ ranges between $t - q = 2$ and $t - 1 = 3$.

  - $|u| = 2$: here $|u| = t - q$ thus r ranges between $2t - k + \sum_{j=0}^{q-2} j = 2$ and $t - 2 = 2$.
    The only word is $w_{4,2,2} = ba^{k-4}bu = ba^2bcd$.
  - $|u| = 3$: here $r = |u| = t - 1 = 3$.
    The only word is $w_{4,3,3} = ba^{k-t}u = ba^2cde$.

Consequently, the (ordered) family F_6 is equal to

$$F_6 = \{a^6, ba^5, ba^4b, ba^4c, ba^3b^2, ba^3bc, ba^3cd, ba^2bcd, ba^2cde, abcdef\}.$$

Remark that $N_6 = \{6, 11, 14, 15, 16, 17, 18, 19, 20, 21\}$ (see [3]).
Moreover, $C(a^6) = 6$, $C(ba^5) = 11$, $C(ba^4b) = 14$, $C(ba^4c) = 15$, $C(ba^3b^2) = 16$, $C(ba^3bc) = 17$, $C(ba^3cd) = 18$, $C(ba^2bcd) = 19$, $C(ba^2cde) = 20$, $C(abcdef) = 21$.

Thus, for each 6-reachable integer n, F_6 contains one and only one word having complexity n.

The important result of this section is Proposition 4.3 which gives the value of $C(w')-C(w)$ when w and w' are two consecutive words of F_k, for all possible values of t, $|u|$ and r.

To establish this result, we must be able to compute the complexity of each word in F_k and so we give a first property of these words.

Remark. In the rest of this section, $k \geq 3$ is a given integer, $l \geq 1$ is defined by Relation (1), and the family F_k as well as the words $w_{t,|u|,r}$ are those specified above.

Property 4.1. *For each word $w_{t,|u|,r}$ in F_k, we have $t-2-r < k-t$.*

Proof. Let $w_{t,|u|,r}$ be a word in F_k. Two cases have to be considered, depending on $t \leq l$ or $t \geq l+1$.

- If $t \leq l$ then $k-l \leq k-t$ and, since $r \geq 0$, $t-2-r \leq l-2$.
  But $k \geq \frac{l(l+1)}{2}+2$, so $\frac{l(l-1)}{2}+2 \leq k-l$.
  Moreover $l-2 < \frac{l(l-1)}{2}+2$.
  Consequently in this case $t-2-r < k-t$.

- If $t \geq l+1$, then, since the value of $t-2-r$ is maximal when the one of r is minimal, we must prove the result in only two cases:
  a) $r = 2t-k+\sum_{j=0}^{q-2} j$
  In this case $t-2-r = t-2-2t+k-\sum_{j=0}^{q-2} j = k-t-2-\sum_{j=0}^{q-2} j$, so $t-2-r < k-t$ because $q \geq 2$.
  b) $r = |u| \geq t-q+1$
  In this case $t-2-r \leq t-2-(t-q+1)$, so $t-2-r \leq q-3$.
  But for any integer q, $q-3 < 1+\frac{(q-1)(q-2)}{2} = 1+\sum_{j=0}^{q-2} j < k-t$. So again $t-2-r < k-t$. ∎

We are now able to compute the complexity of each of the words $w_{t,|u|,r} = ba^{k-t}(ba^{t-2-r})^{\frac{t-1-|u|}{t-1-r}}u$.

- Suppose first that $|u| \leq t-2$ (so $r \leq t-2$).

  a) k factors start with the first b and $(k-1)$ start with the first a (which exists because $t \leq k-2$);

  b) for each of the $k-t-1$ other a in a^{k-t}, there are $t-1$ factors (because the powers of a have already been counted, and $|(ba^{t-2-r})^{\frac{t-1-|u|}{t-1-r}}u| = t-1$). This gives in all $(k-t-1)(t-1)$ factors;

  c) since r and $|u|$ are different from $t-1$, the word $(ba^{t-2-r})^{\frac{t-1-|u|}{t-1-r}}$ is not empty. We just have to count $t-1-(t-1-r) = r$ new factors starting with the first b of this word: indeed, the factors of ba^{t-2-r} have already been counted because $t-2-r < k-t$ (see Property 4.1);

  d) for each letter between this b and the word u, we must count $|u|$ new factors, so that makes $|u|(t-2-|u|)$;

e) finally, u provides $\sum_{i=0}^{|u|} i = \frac{|u|(|u|+1)}{2}$ factors.

In this case, the word $w_{t,|u|,r}$ contains $k + (k-1) + (k-t-1)(t-1) + r + |u|(t-2-|u|) + \frac{|u|(|u|+1)}{2}$ factors.

- Let us now consider the word $w_{t,t-1,t-1} = ba^{k-t}u$.

  In the previous calculation, only the points c) and d) have to be deleted (because here the word $(ba^{t-2-r})^{\frac{t-1-|u|}{t-1-r}}$ is empty), the points a), b) and e) being the same.

  So $C(w_{t,t-1,t-1}) = k + (k-1) + (k-t-1)(t-1) + \frac{|u|(|u|+1)}{2}$.

  But, since $r = |u| = t-1$, here $r + |u|(t-2-|u|) = 0$.

So in all cases the word $w_{t,|u|,r}$ contains $k + (k-1) + (k-t-1)(t-1) + r + |u|(t-2-|u|) + \frac{|u|(|u|+1)}{2}$ factors, which is summarized in

Lemma 4.2. *For any integer $k \geq 3$, $C(w_{t,|u|,r}) = k(t+1) - t^2 + t|u| - \frac{|u|(|u|+3)}{2} + r$.*

We are now able to prove

Proposition 4.3. *Let $k \geq 3$ be an integer and $w_{t,|u|,r}$, $w_{t',|u'|,r'}$ two consecutive words of F_k.*

- *If $t' = t$ then $C(w_{t',|u'|,r'}) = C(w_{t,|u|,r}) + 1$*
- *If $t' = t+1$ and $t \geq l$ then $C(w_{t',|u'|,r'}) = C(w_{t,|u|,r}) + 1$*
- *If $t' = t+1$ and $t < l$ then $C(w_{t',|u'|,r'}) \geq C(w_{t,|u|,r}) + 2$*

Proof. The result has to be proved in all the possible cases of two consecutive words $w_{t,|u|,r}$ and $w_{t',|u'|,r'}$ in F_k, that is:

1) for two consecutive values of r when t and $|u|$ are given;

2) for the greatest value of r corresponding to a given value of $|u|$ and the smallest value of r corresponding to $|u|+1$, with t given;

3) for the greatest values of $|u|$ and r corresponding to a given value of t, and their smallest values corresponding to $t+1$.

So let $w_{t,|u|,r}$ and $w_{t',|u'|,r'}$ be two consecutive words of F_k. We examine the three cases described above in the same order.

1) $t' = t$, $|u'| = |u|$, $r' = r+1$.
In this case, $C(w_{t,|u|,r}) = k(t+1) - t^2 + t|u| - \frac{|u|(|u|+3)}{2} + r$, and $C(w_{t',|u'|,r'}) = k(t+1) - t^2 + t|u| - \frac{|u|(|u|+3)}{2} + r + 1$.
So $C(w_{t',|u'|,r'}) = C(w_{t,|u|,r}) + 1$.

2) $t' = t$, $|u'| = |u| + 1$.
In this case, we necessarily have $|u| \leq t-2$, and so $r = t-2$ and $r' = |u'| = |u| + 1$.
Consequently, $C(w_{t,|u|,r}) = k(t+1) - t^2 + t|u| - \frac{|u|(|u|+3)}{2} + t - 2$, and $C(w_{t',|u'|,r'}) = k(t+1) - t^2 + t(|u|+1) - \frac{(|u|+1)(|u|+4)}{2} + |u| + 1$.

But $t - \frac{(|u|+1)(|u|+4)}{2} + |u| + 1 = -\frac{|u|(|u|+3)}{2} + t - 1$.
So again $C(w_{t',|u'|,r'}) = C(w_{t,|u|,r}) + 1$.
Thus in the two cases where $t' = t$, we have $C(w_{t',|u'|,r'}) = C(w_{t,|u|,r}) + 1$.

3) $t' = t + 1$.
Since $|u|$ and r are the greatest possible values corresponding to t, we have $r = |u| = t - 1$. So the word $w_{t,|u|,r}$ is $w_{t,t-1,t-1}$ and, consequently, $C(w_{t,|u|,r}) = k(t+1) - t^2 + t(t-1) - \frac{(t-1)(t+2)}{2} + t - 1 = k(t+1) - t^2 + \frac{t(t-1)}{2}$.
For the word $w_{t',|u'|,r'}$, since $|u'|$ and r' must be the lowest possible values corresponding to $t' = t + 1$, two cases have to be considered depending on whether $t \geq l$ or $t < l$.

a) $t \geq l$
Then $t' \geq l + 1$, so $|u'| = t' - q$ and $r' = 2t' - k + \sum_{j=0}^{q-2} j$, that is, $|u'| = (t+1) - q$ and $r' = 2(t+1) - k + \sum_{j=0}^{q-2} j$.
Consequently, $C(w_{t',|u'|,r'}) = k(t+2) - (t+1)^2 + (t+1)(t+1-q) - \frac{(t+1-q)(t+4-q)}{2} + 2(t+1) - k + \sum_{j=0}^{q-2} j = k(t+1) - t^2 + \frac{t(t-1)}{2} + 1$.
In this case ($t' = t + 1$ and $t \geq l$) we have again $C(w_{t',|u'|,r'}) = C(w_{t,|u|,r}) + 1$.

b) $t < l$
In this case, $t' \leq l$, so $r' = |u'| = 0$ and $w_{t',|u'|,r'} = w_{t+1,0,0}$.
But $C(w_{t+1,0,0}) = k(t+2) - (t+1)^2 = k(t+1) - t^2 + k - 2t - 1$.
And $t \leq l - 1$, so $k - 2t - 1 \geq k - 2l + 1$.
Since $k = \frac{l(l+1)}{2} + 2 + i$, $0 \leq i \leq l$, we have $C(w_{t+1,0,0}) \geq k(t+1) - t^2 + \frac{l^2-3l}{2} + 3$.
But, since $t \leq l - 1$, we also have $\frac{t(t-1)}{2} \leq \frac{l^2-3l}{2} + 1$, and thus $C(w_{t,|u|,r}) \leq k(t+1) - t^2 + \frac{l^2-3l}{2} + 1$.

Consequently, if $t' = t + 1$ and $t < l$, then $C(w_{t',|u'|,r'}) \geq C(w_{t,|u|,r}) + 2$, and the proposition is proved. ∎

5 Main results

The first direct consequence of Proposition 4.3 is the proof of the conjecture of Z. Kása.

Let us recall that, for any integer $k \geq 2$, $b_k = k(l+1) - l^2$ (see Relation (5)) where l is the unique positive integer such that $k = \frac{l(l+1)}{2} + 2 + i$, $0 \leq i \leq l$.

On the other hand, for all $k \in \mathbb{N}$, m_k is the smallest number such that all the integers between m_k and $\frac{k(k+1)}{2}$ are k-reachable.

Theorem 5.1. *For all integers $k \geq 2$, $m_k = b_k$.*

Proof. If $k = 2$ then $b_k = 2$ and $\frac{k(k+1)}{2} = 3$.

But on the one hand 1 is not 2-reachable, and on the other hand the words aa and ab are respectively of complexity 2 and 3. So $m_2 = 2 = b_2$.

Suppose now $k \geq 3$.

In this case, $l \geq 1$ and $k-2 \geq l$ (see (1)), and the two words $w_{l,0,0}$ and $w_{k-2,k-3,k-3}$ are in F_k. Indeed, $w_{k-2,k-3,k-3} = ba^2u$ is the penultimate word of F_k, and $w_{l,0,0} = \begin{cases} ba^{k-1} & \text{if } l = 1 \\ ba^{k-l}ba^{l-2} & \text{otherwise} \end{cases}$ is the word $w_{t,|u|,r}$ where $t = l$ and $r = |u| = 0$, which is always in F_k.

Since $k - 2 \geq l$, the gap of complexity for all the words in F_k between $w_{l,0,0}$ and $w_{k-2,k-3,k-3}$ is, from Proposition 4.3, always equal to 1. Consequently, all the integers between $C(w_{l,0,0})$ and $C(w_{k-2,k-3,k-3})$ are k-reachable.

But $C(w_{l,0,0}) = k(l+1) - l^2 = b_k$ and $C(w_{k-2,k-3,k-3}) = k(k-1) - (k-2)^2 + (k-2)(k-3) - \frac{(k-3)k}{2} + k - 3 = \frac{k(k+1)}{2} - 1$.

Moreover the last word of F_k is $a_1a_2 \dots a_k$ (where all the letters a_i are different) whose complexity is $\frac{k(k+1)}{2}$.

Consequently $m_k \leq b_k$ and since, from Proposition 3.5, $b_k - 1$ is never k-reachable, the theorem is proved. ∎

We will now show that for any k-reachable integer p, there exists a (unique) word w in F_k such that $C(w) = p$.

We start by proving

Lemma 5.2. *Let w be a word of length k and $t + 1 = max\{C_q(w) | 1 \leq q \leq k\}$. If $t + 1 \leq k - t$ then $k(t+1) - t^2 \leq C(w) \leq k(t+1) - \frac{t(t+1)}{2}$*

Proof. If $w = a^k$ then $t = 0$ and the result is obvious.

If $w \neq a^k$ then the number of letters of w is at least 2, that is, $C_1(w) \geq 2$.

Let m be the least integer such that $C_m(w) = t + 1$.

Relation (2) gives $C(w) = \sum_{i=0}^{m-1} C_i(w) + (k-m)(t+1) - \frac{(t+1)(t-2)}{2}$.

To compute the extremal values of $\sum_{i=0}^{m-1} C_i(w) + (k-m)(t+1)$, notice that since $C_1(w) \geq 2$ and for $1 \leq q \leq m-1$, $C_{q+1}(w) - C_q(w) > 0$ (see Proposition 3.3), one has $C_m(w) \geq m + 1$ so $m \leq t$.

The value of $\sum_{i=0}^{m-1} C_i(w) + (k-m)(t+1)$ will be minimal if $m = t$ (2 factors of length 1, 3 of length 2, ..., t factors of length $t-1$, $t+1$ of length t, which is possible since $t + 1 \leq k - t$), and it will be maximal if $m = 1$ ($t + 1$ factors of each length between 1 and t).

In the first case, $C(w) = \sum_{i=1}^{t-1}(1+i) + (k-t)(t+1) - \frac{(t+1)(t-2)}{2} = k(t+1) - t^2$.

In the second case, from relation (3), $C(w) = k(t+1) - \frac{t(t+1)}{2}$. ∎

We are now able to prove the main result, saying that for each k-reachable integer, F_k contains exactly one word having this integer for complexity.

Theorem 5.3. *Let $k \geq 3$ be an integer. For any k-reachable number p, there exists one and only one word w in F_k such that $C(w) = p$.*

Proof. Let $k \geq 3$ be an integer.

First remark that if a word belongs to F_k then, by definition, its complexity is a k-reachable number. Moreover, from Proposition 4.3, all the words in F_k have different complexities. So it is enough, for any k-reachable integer, to prove the existence in F_k of a word whose complexity is this integer.

So let p be a k-reachable integer.

If $p \geq b_k$, then we saw in the proof of Theorem 5.1 that there exists a word $w \in F_k$ such that $C(w) = p$.

Suppose now $p < b_k$, and let w be a word such that $C(w) = p$ and $t + 1 = max\{C_q(w) | 1 \leq q \leq k\}$.

From Proposition 3.5, $t + 1 \leq l$.

But $k - t = \frac{l(l+1)}{2} + 2 + i - t$, $0 \leq i \leq l$ and if $t < l$ then $k - t > \frac{l(l-1)}{2} + 2$. Since $\frac{l(l-1)}{2} + 2 > l$ for each l, this implies $t + 1 \leq k - t$.

Then, from Lemma 5.2, $k(t + 1) - t^2 \leq C(w) \leq k(t + 1) - \frac{t(t+1)}{2}$.

But because $t < l$, the word $w_{t,0,0}$ is in F_k. On the other hand, the word $w_{t,t-1,t-1}$ also belongs to F_k.

But $C(w_{t,0,0}) = k(t+1)-t^2$ and $C(w_{t,t-1,t-1}) = k(t+1)-t^2+t(t-1)-\frac{(t-1)(t+2)}{2}+t-1 = k(t+1) - \frac{t(t+1)}{2}$.

Moreover, from Proposition 4.3, every integer between $C(w_{t,0,0})$ and $C(w_{t,t-1,t-1})$ is the complexity of some word in F_k.

Thus, since $C(w)$ is between these two values, there is a word $v \in F_k$ such that $C(v) = C(w)$. ∎

Acknowledgements

The authors are very indebted to Julien Cassaigne, Gwenaël Richomme and the three referees whose careful reading and suggestions have greatly improved a first version of this paper. The second author is also indebted to Zoltán Blázsik who introduced him to Kasa's conjecture during his passage in Szeged.

References

[1] A. de Luca. On the combinatorics of finite words. *Theoretical Computer Science*, 218:13–39, 1999.

[2] Z. Kása. On two conjectures on word complexity. In *Second Joint Conference on Modern Mathematics, Ilieni, Rumania*, page 44, June 3-7 1997.

[3] Z. Kása. On the d-complexity of strings. *Pure Mathematics and Applications*, 9(1-2):119–128, 1998.

[4] M. Lothaire. *Combinatorics on Words*, volume 17 of Encyclopedia of Mathematics and its Applications. Addison-Wesley, Reading, Mass., 1983. Reprinted in the Cambridge Mathematical Library, Cambridge University Press, Cambridge UK, 1997.

[5] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, Cambridge UK, to appear in the Cambridge Mathematical Library.

[6] A. Thue. Über unendliche Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter, I. Mat. Nat. Kl.*, 7:1–22, 1906.

[7] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter, I. Mat. Nat. Kl.*, 1:1–67, 1912.

LaRIA, Université de Picardie Jules Verne
5, rue du Moulin Neuf
80000 Amiens
France
e-mail: {leve,seebold}@laria.u-picardie.fr