Words and Repeated Factors

Arturo Carpi   Aldo de Luca

Abstract

In this paper we consider sets of factors of a finite word which permit us to reconstruct the entire word. This analysis is based on the notion of *box*. The *initial* (resp. *terminal*) *box* of w is the shortest prefix (resp. suffix) of w which is an unrepeated factor. A factor u of w is a *proper box* if there are letters a, a', b, b' with $a' \neq a$, $b' \neq b$ such that $u = asb$ and $a's$, sb' are factors of w. A box is called *maximal* if it is not a proper factor of another box. The main result of the paper is the following theorem (*maximal box theorem*): *Any finite word w is uniquely determined by the initial box, the terminal box and the set of maximal boxes.* Another important combinatorial notion is that of *superbox*. A superbox is any factor of w of the kind asb, with a, b letters, such that s is a repeated factor, whereas as and sb are unrepeated factors. A theorem for superboxes similar to the maximal box theorem is proved. Some algorithms allowing us to construct boxes and superboxes and, conversely, to reconstruct the word are given. An extension of these results to languages is also presented.

1 Introduction

The study of the factors of finite as well as infinite words represents a topic of great interest in combinatorics on words. The number of factors of any length of a word (subword complexity) has been viewed as an evaluation of the richness of the structure of the considered word. However, for a finite word, it is sufficient to know its factors up to a certain length to reconstruct the entire word. For instance, it is easily seen that the only word whose factors of length 2 are ab, bc and cd is $abcd$.

The problem of reconstructing a word by knowing its 'short' factors appears in several fields. For instance, in analyzing DNA molecules, it is not possible to read the entire sequence of bases but only segments of limited length.

The aim of this paper is to study some sets of 'short' factors of a word which completely characterize the word itself, as well as the methods to reconstruct the word from such factors.

There are two combinatorial properties of factors of a word which turn out to be crucial for our purpose, namely *repetition* and *extendability*. Indeed, we shall show that the *special factors* (see e.g. [1, 5, 7, 8, 9, 10]), the shortest unrepeated initial and terminal factors and some related concepts are of fundamental importance in determining the 'structure' of the word itself.

We recall that a factor u of a finite word w is called *right* (resp. *left*) *special* if there exist two distinct extensions on the right (resp. on the left) in factors of w. A factor is *bispecial* if it is right and left special.

We shall introduce two classes of factors of a word, both of which uniquely determine the word. The first one is constituted by the *boxes*: the *initial* and *terminal box* of a word w are, respectively, the shortest prefix and suffix of w which are unrepeated factors while a *proper box* is any factor of w of the kind asb with a and b letters and s a bispecial factor of w. The second one is the set of the *superboxes*, which are factors of w of the kind asb with a and b letters, s a repeated factor of w and as, sb unrepeated factors of w.

In Sec. 4 we prove the surprising result, called the *maximal box theorem* (cf. Theorem 1), that any word is completely determined by the set of its (maximal) boxes and the initial and terminal boxes. A simple procedure to construct the word from the boxes is also given. A remarkable corollary is the following. Let K_f be the length of the terminal box of a word f and R_f be the minimal integer such that there are no right special factors of f of length R_f. If g is a word having the same set of factors as f up to the length $n = \max\{R_f, K_f\} + 1$, then $f = g$. The value of n is *optimal* since one can prove that there exists a word g' having the same set of factors of f up to the length $n-1$, and, moreover, all the factors of f of length n are factors of g' and all the factors of g' of length n with the exception of one, are factors of f.

In Sec. 5 we study some interesting structural properties of boxes. In particular it is shown that the set of the maximal boxes in Theorem 1, is 'nearly' optimal because for any given word w one can construct another word v with the property that all boxes of w are factors of v, whereas there exist at most two boxes of v which are not factors of w. Moreover, we show that, starting from the set of maximal boxes of a word, one can construct the so-called 'reduced sets'. An analog of the maximal box theorem for reduced sets holds. These sets are convenient, since they provide a simpler representation of the word.

In Sec. 6 the notion of *superbox* is studied. We prove a theorem showing that the initial and terminal boxes and the set of superboxes determine uniquely the word w. Moreover, we give two simple algorithms: the first allows us for any word w to determine the set of all superboxes and the second permits us to reconstruct the word starting from the set of all superboxes and the initial and terminal boxes.

In Sec. 7 we deal with languages and a box theorem for languages is presented. As a consequence of this result, we give a new proof of the Fine and Wilf 'uniqueness theorem' for periodic functions in the discrete case.

2 Preliminaries

Let A be a finite non-empty set or *alphabet* and A^* the free monoid generated by A. The elements of A are usually called *letters* and those of A^* *words*. The identity element of A^* is called *empty word* and denoted by ϵ. We set $A^+ = A^* \setminus \{\epsilon\}$.

A word $w \in A^+$ can be written uniquely as $w = w_1 w_2 \cdots w_n$, with $w_i \in A$, $1 \leq i \leq n$, $n > 0$. The integer n is called the *length* of w and denoted $|w|$. The length of ϵ is 0. For any $n \geq 0$, we denote by A^n the set of all the words of A^* of length n and set $A^{[n]} = \bigcup_{i=0}^{n} A^n$.

Let $w \in A^*$. The word $u \in A^*$ is a *factor* (or *subword*) of w if there exist $p, q \in A^*$ such that $w = puq$. A factor u of w is called *proper* if $u \neq w$.

If $w = uq$, for some $q \in A^*$ (resp. $w = pu$, for some $p \in A^*$), then u is called a *prefix* (resp. a *suffix*) of w.

For any $w \in A^*$, we denote respectively by $F(w)$, $\mathrm{Pref}(w)$ and $\mathrm{Suff}(w)$ the sets of its factors, prefixes and suffixes.

We shall denote $F(w) \cap A$ by $\mathrm{alph}(w)$. This set represents the subset of the letters of A occurring in the word w.

For any $X \subseteq A^*$, we set

$$F(X) = \bigcup_{u \in X} F(u).$$

An element of $F(X)$ will be also called a *factor* of X.

Let $u \in F(w)$. Any pair $(\lambda, \mu) \in A^* \times A^*$ such that $w = \lambda u \mu$ is called an *occurrence* of u in w. If $\lambda = \epsilon$ (resp. $\mu = \epsilon$), then the occurrence of u is called *initial* (resp. *terminal*). An occurrence is called *internal* if it is neither initial nor terminal. An occurrence (λ, μ) of u in w is called *leftmost* (resp. *rightmost*) if the length of λ (resp. μ) is minimal. A factor u of w is called *internal* if there exists an internal occurrence of u in w.

Let $w = w_1 w_2 \cdots w_n$ ($w_i \in A$, $1 \leq i \leq n$) be a word on the alphabet A. A set $\mathcal{C} \subseteq F(w)$ is a *covering* of w if, for any $k = 1, \ldots, n$, there exist i, j, $1 \leq i \leq k \leq j \leq n$, such that $w_i w_{i+1} \cdots w_j \in \mathcal{C}$.

Let us suppose $\mathrm{Card}(A) = d$ and consider in A^* the prefix ordering of the words. It is well known that the graph associated with this order is a d-ary tree T_d, whose nodes represent the words of A^* and the root represents the empty word.

With each finite word w, one can associate a finite subtree T_w of T_d obtained by taking all the nodes which represent factors of the word w. We call T_w the *factor tree* of w, since any factor of w will be represented by a node n of T_w or, equivalently, by a unique path going from the root to the node n.

If one replaces the prefix ordering of words by the suffix ordering one can construct in a similar way another tree T'_w, whose nodes represent the factors of w.

3 Repeated factors

A factor u of a word w is called *repeated* if there are at least two distinct occurrences of u in w. In the opposite case, the factor u is called *unrepeated*.

A factor u of w is *extendable* on the right (resp. left) in w if there exists a letter $x \in A$ such that $ux \in F(w)$ (resp. $xu \in F(w)$). The factor ux (resp. xu) of w is called a *right* (resp. *left*) *extension* of u in w.

If u is a factor of w, then the *right* (resp. *left*) *valence* of u is the integer $\mathrm{Card}(\{x \in A \mid ux \in F(w)\})$ (resp. $\mathrm{Card}(\{x \in A \mid xu \in F(w)\})$). The right (resp. left) valence of u is then the number, possibly 0, of all the distinct right (resp. left) extensions of u in w. In terms of the factor tree T_w (resp. T'_w) the right (resp. left) valence of u is the degree (that is the number of sons) of the node representing the factor u.

With each word $w \in A^*$ one can associate the factor k_w defined as the shortest suffix of w which is an unrepeated factor of w. This is also equivalent to say that k_w is the shortest factor of w which cannot be extended on the right in w, i.e. it has right valence equal to 0. The set of the factors of w which are not extendable on the right is given by $A^*k_w \cap \mathrm{Suff}(w)$. In the factor tree T_w, these factors are represented by the leaves. In a symmetric way, one can define h_w as the shortest prefix of w which cannot be extended on the left in w. Also in this case, the factors of w which are not extendable on the left are represented by the leaves of the tree T'_w.

In the following, we shall set $K_w = |k_w|$ and $H_w = |h_w|$. If $w \neq \epsilon$, then one has $1 \leq K_w \leq |w|$ and $1 \leq H_w \leq |w|$. If $w = \epsilon$, then $H_\epsilon = K_\epsilon = 0$.

One can remark that all the proper prefixes of h_w and all the proper suffixes of k_w are repeated factors, while h_w and k_w are unrepeated.

In the following, for $w \neq \epsilon$, we shall denote by h'_w (resp. k'_w) the prefix (resp. suffix) of w of length $H_w - 1$ (resp. $K_w - 1$).

A word s is called a *right* (resp. *left*) *special factor* of w if there exist two letters $x, y \in A$, $x \neq y$, such that $sx, sy \in F(w)$ (resp. $xs, ys \in F(w)$).

A right (resp. left) special factor of w is then a repeated factor having at least two distinct extensions on the right (resp. left) in w, i.e. it has a right (resp. left) valence ≥ 2. This implies that a right (resp. left) special factor of w has at least an internal occurrence in w.

We remark that the set of right (resp. left) special factors is closed by suffixes (resp. prefixes).

A factor of w which is right and left special is called *bispecial.*

Let us remark that the empty word ϵ is always a bispecial factor of w, except when w is a power of a single letter. In such a case, $w = k_w = h_w$.

We denote by R_w (resp. L_w) the minimal non-negative integer such that there are no right (resp. left) special factors of length R_w (resp. L_w). One has $0 \leq R_w, L_w \leq |w| - 1$. Since the set of right special factors is closed by suffixes, there are no right special factors of length larger than R_w. Symmetrically, there are no left special factors of length larger than L_w.

A repeated factor of a word w is called *maximal* (with respect to the factorial order) if it is not a proper factor of another repeated factor of w.

Proposition 1 *If a repeated factor u of a word w is maximal, then one of the following conditions is satisfied:*

(i) *u is a bispecial factor,*

(ii) $u = h'_w$,

(iii) $u = k'_w$.

Proof. Let us first suppose that u is a prefix of w. Since u is a repeated factor, then u has to be a prefix of h'_w. From the maximality, it follows that $u = h'_w$. In a symmetric way, one proves that if u is a suffix of w, then $u = k'_w$.

Let us then suppose that u is neither a prefix nor a suffix of w. Then, since u is repeated, there will exist two internal occurrences of u in w, and then letters $x, x', y, y' \in A$ such that xuy, $x'uy'$ are two different occurrences of factors of w. Since u is a maximal repeated factor, it follows $x \neq x'$ and $y \neq y'$. Hence, u is a bispecial factor of w. □

Let $w \in A^*$ be a non-empty word. We denote by G_w the maximal length of a repeated factor of w. The following proposition, whose proof we omit for the sake of brevity, holds [3, 6].

Proposition 2 *Let $w \in A^+$. One has*

$$G_w = \max\{R_w, K_w\} - 1 = \max\{L_w, H_w\} - 1.$$

Let us now for any word $w \in A^*$ introduce an important set of factors that we call boxes.

Let $w \in A^*$ be a word. A factor f of w is called a *proper box* of w if $f = asb$ with $a, b \in A$ and s is a bispecial factor. The factor h_w (resp. k_w) is called the *initial* (resp. *terminal*) *box* of w.

By *box*, without specification, we mean indifferently the initial, the terminal or a proper box.

A box is called *maximal* (with respect to the factorial order) if it is not a proper factor of another box.

The set of maximal boxes of w will be denoted by $\mathcal{B}_w$. We note that, for any word, the initial, the terminal and the maximal boxes can be constructed by a simple algorithm, whose description is omitted.

4 Maximal box theorem

In this section we prove a theorem, called the 'maximal box theorem', which shows that if of a given word one knows the initial box, the terminal box and the set of the maximal boxes, then the word is uniquely determined. A simple procedure in order to reconstruct the word from the boxes is also given. A remarkable consequence of this theorem is that if two words f and g have the same set of factors up to the length $n = \max\{R_f, K_f\} + 1$, then $g = f$. We prove also that this value of n is optimal.

Lemma 1 *Let α be a box of w. Then any internal factor of α is repeated in w.*

Proof. Let us first suppose that $\alpha = k_w$ and f is an internal factor of α. One has that f is a factor of k'_w, so that it is repeated in w. In a symmetrical way, one reaches the same result if f is an internal factor of $\alpha = h_w$. Let us now suppose that $\alpha = asb$ is a proper box of w. An internal factor f of α is a factor of s, which is a bispecial factor of w. Since s is a repeated factor of w, so will be f. □

Example 1 Let w be the word $w = abccbabcab$. One has $h_w = abcc$, $k_w = cab$. The proper boxes of w are ab, ba, bc, ca, cb, cc, abc, bca, bcc, cba, ccb. The maximal boxes are

$$abcc, \quad bca, \quad cab, \quad cba, \quad ccb.$$

Let $w = abaababaaba$. In this case $h_w = abaabab$, $k_w = babaaba$ and these are the only maximal boxes of w.

For all $n \geq 0$, we introduce the binary relation $\preceq_n$ in A^* defined as: for $f, g \in A^*$,

$$f \preceq_n g \quad \text{if and only if} \quad F(f) \cap A^{[n]} \subseteq F(g) \cap A^{[n]}.$$

One easily verifies that the relation $\preceq_n$ is a well-founded quasi-order and, for all $f, u, v \in A^*$, $n \geq 0$, one has $f \preceq_n ufv$. We note that the intersection of the $\preceq_n$ for all $n \geq 0$ is the factorial order, that we denote by $\preceq$.

For all $n \geq 0$, we consider the equivalence relation $\sim_n = \preceq_n \ \cap \preceq_n{}^{-1}$. Thus

$$f \sim_n g \quad \text{if and only if} \quad F(f) \cap A^{[n]} = F(g) \cap A^{[n]}.$$

Theorem 1 (Maximal box theorem) *Let $f, g \in A^*$ be two words such that*

(i) $h_f = h_g, \quad k_f = k_g$,

(ii) $\mathcal{B}_f \subseteq F(g)$,

(iii) $\mathcal{B}_g \subseteq F(f)$.

Then $f = g$.

Proof. We prove, by induction, that for all $n \geq 0$, $f \sim_n g$. Clearly, this implies that $f = g$.

Let us prove first the base of the induction, i.e. $f \sim_1 g$. If $\mathrm{Card}(\mathrm{alph}(f)) \leq 1$, then $f = k_f = k_g$. Thus, $f \in F(g)$. Let us then suppose $\mathrm{Card}(\mathrm{alph}(f)) > 1$. In such a case, ϵ is a bispecial factor and any $w \in A^2 \cap F(f)$ is a box. Since w is included in a maximal box, one has $w \in F(g)$. Thus we have $f \preceq_1 g$. In a symmetrical way, one proves that $g \preceq_1 f$.

Now, let us prove the induction step. We suppose that $n \geq 1$, $f \sim_n g$ and prove that $f \sim_{n+1} g$.

Let w be a factor of f of length $n+1$. If w is a box of f, then w is a factor of a maximal box and, therefore, by condition (ii), w is a factor of g.

Let us then suppose that w is not a box. We factorize w as $w = atb$, with $a, b \in A$, $t \in A^*$, and t is not a bispecial factor of f. Since $|at| = |tb| = n$, by the inductive hypothesis, one has that $at, tb \in F(g)$.

Let us first suppose that t is not a right special factor of f. Since at is extendable on the right in f, $k_g = k_f$ cannot be a suffix of at. This implies that at can be extended on the right in g. Thus there exists a letter c such that $atc \in F(g)$. By the inductive hypothesis, $tc \in F(f)$. Since t

is not right special in f, one obtains $b = c$ and then $atb \in F(g)$. With a symmetrical argument, if t is not a left special factor of f, one proves again that $atb \in F(g)$. Thus we have obtained that $f \preceq_n g$. In a symmetrical way, one derives that $g \preceq_n f$. □

Proposition 3 *Let w be a word. If v is a factor of w, then there exists $u \in \mathrm{Suff}(v)$ such that for any $a \in A$*

$$va \in F(w) \quad \textit{if and only if} \quad (ua \in F(\mathcal{B}_w) \textit{ and } va \notin A^+h_w). \tag{1}$$

Proof. If v is not right extendable, then the statement is trivially verified by $u = v$. If $\mathrm{Card}(\mathrm{alph}(w)) = 1$, then the statement is verified by $u = \epsilon$, since in this case $h_w = w$. If $v = \epsilon$, then $u = \epsilon$ satisfies the condition.

Let us then suppose that $v \neq \epsilon$ is right extendable and that $\mathrm{Card}(\mathrm{alph}(w)) > 1$. We can write

$$v = \lambda bs, \quad b \in A, \ \lambda \in A^*,$$

where s is the longest proper suffix of v which is a bispecial factor of w. We set $u = bs$. Let a be a letter of A and suppose $va \in F(w)$. This trivially implies that $va \notin A^+h_w$. Moreover, $ua = bsa$ is a box so that $ua \in F(\mathcal{B}_w)$.

Conversely, let $ua \in F(\mathcal{B}_w)$, $va \notin A^+h_w$ and suppose, by contradiction, that $va \notin F(w)$. Let t be the longest suffix of v such that $ta \in F(w)$. Since $t \neq v$, one can write

$$v = \mu ct, \text{ with } c \in A, \ \mu \in A^* \text{ and } cta \notin F(w).$$

Since $va \notin A^+h_w$, it follows that $ta \neq h_w$. Thus, since t is left extendable in w, one derives that ta is left extendable in w, so that there exists a letter $x \in A$ such that $xta \in F(w)$. Moreover, since v is right extendable in w, there exists a letter $y \in A$ such that $cty \in F(w)$. One has $x \neq c$ and $y \neq a$, since $cta \notin F(w)$. Hence, t is bispecial. This contradicts the fact that $|t| \geq |u| > |s|$. □

The previous proposition shows that if v is a right extendable factor of w, then to find a right extension of v it is sufficient to determine the longest suffix u of v in w such that there is at least one letter $a \in A$ satisfying the right hand side condition of Eq. (1). For such a letter a one has $va \in F(w)$.

Let us now give a simple procedure, based on Proposition 3, which allows us to construct the word w knowing the initial box h_w, the terminal box k_w and the set $\mathcal{B}_w$ of maximal boxes.

Let us write $h_w = h'_w z$, with $z \in A$.

Initially, we set $p = h_w$. Now suppose that we have already constructed a prefix p of w of length $|p| \geq H_w$.

If $p \in A^*k_w$, then the procedure ends and $w = p$. Otherwise, the right valence of p is 1. In order to extend p in w, we have to distinguish the following cases:

(i) $p \notin A^*h'_w$.

In this case, we search for the shortest suffix u of p which can be extended on the right in $F(\mathcal{B}_w)$ by a unique letter x and replace p by px.

Indeed, by Proposition 3, there exists a suffix of p which can be extended on the right in $F(\mathcal{B}_w)$ by a unique letter which is exactly the letter extending p on the right in w.

(ii) $p \in A^*h'_w$.

In this case, we search for the shortest suffix u of p which can be extended on the right in $F(\mathcal{B}_w)$ by a unique letter $x \in A \setminus \{z\}$ and replace p by px.

Indeed, by Proposition 3, there exists a suffix of p which can be extended on the right in $F(\mathcal{B}_w)$ by a unique letter in the set $A \setminus \{z\}$ which is exactly the letter extending p on the right in w.

Proposition 4 *Let $f \in A^*$ and $n = R_f + 1$. If $g \in A^*$ is such that $g \sim_n f$ and $k_f = k_g$, then $f = g$.*

Proof. Let us first prove that $R_f = R_g$. By the hypothesis, f and g have the same set of factors up to the length $n = R_f + 1$, so that all right special factors of f are also right special factors of g. Thus $R_f \leq R_g$.

Let us suppose that $R_f < R_g$. Since the right special factors are closed by suffixes, there exists a right special factor of g of length R_f. This is also a right special factor of f, since $f \sim_{R_f+1} g$, and this is a contradiction. Thus $R_f = R_g = R$.

Remark that the length of any proper box of f or of g is at most n. Indeed, any proper box of f (resp. of g) can be written as asb with $a, b \in A$ and s a bispecial factor of f (resp. of g). This implies that $|asb| \leq R + 1$.

Hence, by the hypothesis that $g \sim_n f$, one derives that any proper maximal box of f is a factor of g and, conversely, any proper maximal box of g is a factor of f.

Since $k_f = k_g$, in view of Theorem 1, in order to prove $f = g$, it is sufficient to show that $h_f = h_g$.

Let us first suppose that $|h_f| \leq R$. Then h_f is a factor of g which cannot be extended on the left in g because $f \sim_n g$. Thus h_g has to be a prefix of h_f. This implies $|h_g| \leq R$, so that, by using the same argument, it follows that h_f has to be a prefix of h_g, and then $h_f = h_g$.

By a symmetrical argument, one arrives to the same conclusion if one supposes that $|h_g| \leq R$.

Hence, we may suppose that $|h_f|, |h_g| > R$. This implies that h'_f (resp. h'_g) is not a right special factor of f (resp. g). Let us show that h'_f cannot be an internal factor of f. Indeed, otherwise there would exist letters $x, y \in A$ such that $h_f = h'_f x$ and $h'_f y \in F(f)$. Since h_f is unrepeated then $x \neq y$ so that h'_f would be a right special factor of f. Thus, h'_f has to be a suffix of k'_f. Since k'_f is repeated, and h'_f is not an internal factor of f, the only possibility is that $h'_f = k'_f$. In a similar way, one has that $h'_g = k'_g$. By the hypothesis that $k_f = k_g$, one derives $h'_f = k'_f = k'_g = h'_g = u$.

Let $h_f = ux$ and $h_g = uy$, $x, y \in A$. Let v be the suffix of u of length R. Since $f \sim_n g$, vx and vy are factors of both f and g. Since v is not right special, it follows $x = y$ and $h_f = h_g$. □

By a symmetric argument, the following proposition can be proved.

Proposition 5 *Let $f \in A^*$ and $n = L_f + 1$. If $g \in A^*$ is such that $g \sim_n f$, and $h_f = h_g$, then $f = g$.*

Theorem 2 *Let $f \in A^*$ and $n = \max\{R_f, K_f\} + 1$. For any $g \in A^*$, if $g \sim_n f$, then $g = f$.*

Proof. Let $g \sim_n f$. By Proposition 4, it is sufficient to prove that $k_f = k_g$.

Since $f \sim_n g$, then the factor k_f of f cannot be extended on the right in g, since otherwise k_f could be also extended on the right in f. Hence, k_g is a suffix of k_f. If k_g is a proper suffix of k_f, then k_g would be extendable on the right in f and, by hypothesis, also in g, which is a contradiction. Thus, $k_f = k_g$. □

Proposition 6 *Let $f \in A^*$ and $n = \max\{R_f, K_f\}$. There exists $g \in A^*$ such that $g \neq f$ and $g \sim_n f$. Moreover, $f \preceq_{n+1} g$ and all the factors of g of length $n + 1$, with the exception of one, are factors of f.*

Proof. If $f = \epsilon$, then $n = 0$ and the statement is trivially satisfied by $g = a$, $a \in A$. Let us then suppose $f \neq \epsilon$. By Proposition 2 the word f has a repetition of length $n - 1$, i.e. we can write

$$f = prq = p'rq',$$

with $|r| = n - 1$ and $|p| < |p'|$.

Set $g = p'rq$. Let us prove that $f \preceq_{n+1} g$. Since $|p'| > |p|$, one has $p' = p\xi$ with $\xi \in A^+$ and then $f = p\xi rq'$. As $|\xi r| \geq n$, then a factor u of f of length $n + 1$ is either a factor of $p\xi r = p'r$ or a factor of $\xi rq' = rq$, so that u is a factor of g.

Conversely, a factor of g of length n is either a factor of $p'r$ or of rq and, therefore, it is a factor of f. This proves that $f \sim_n g$.

All factors of length $n+1$ of g which occur in $p'r$ or in rq are also factors of f. The only exception is given by xry, where x is the last letter of p' and y is the first letter of q. □

Example 2 Let $A = \{a, b, c\}$ and $w = abccbabcab$. The factor abc is the right special factor of w of maximal length and $k_w = cab$. In this case, $R_w = 4$ and $K_w = 3$. We can write $w = prq = p'rq'$, where $r = abc$, $p = \epsilon$, $q = cbabcab$, $p' = abccb$ and $q' = ab$. Let us then consider the word $g = p'rq = abccbabccbabcab$. One easily verifies that $g \sim_4 w$, $w \preceq_5 g$ and that $babcc$ is the only factor of g of length 5 which is not a factor of w.

5 Boxes and reduced sets

In this section we prove some structural properties of boxes. In particular it is shown that any maximal box is an unrepeated factor. Moreover, one can prove that the set of the maximal boxes in the maximal box theorem is 'nearly' optimal in the sense that for any given word w one can construct another word v with the property that all boxes of w are factors of v, whereas there exist at most two boxes of v which are not factors of w. However, a simpler representation of a word is given by the so-called 'reduced sets' for which a result similar to the maximal box theorem holds.

Proposition 7 *Let α be a maximal box of w. Then α is an unrepeated factor of w.*

Proof. Suppose, by contradiction, that the maximal box α is a repeated factor of w. The box α will be a factor of a maximal repeated factor u of w. By Proposition 1, there are three possibilities:

1. $u = h'_w$. In such a case it follows that α is a proper factor of the initial box h_w, which contradicts the maximality of α as a box.

2. $u = k'_w$. In such a case it follows that α is a proper factor of the terminal box k_w, which contradicts the maximality of α as a box.

3. u is a bispecial factor of w. Since u has always an internal occurrence in w, it can be extended in a box. The same will occur for α, which contradicts again the maximality of α. □

Proposition 8 *Let $w = \lambda asb\mu$, where $asb \neq k_w$ (resp. $asb \neq h_w$) is a proper maximal box of w with $a, b \in A$, $s \in A^*$. If csb (resp. asc) is a box of w with $c \in A$ and $c \neq a$ (resp. $c \neq b$) having the leftmost (resp. rightmost)* *occurrence in w given by (λ', μ'), then $k_w = csb\mu'$ and $\mu' \in \mathrm{Pref}(\mu)$ (resp. $h_w = \lambda' asc$ and $\lambda' \in \mathrm{Suff}(\lambda)$).*

Proof. We can write the word w as

$$w = \lambda asb\mu = \lambda' csb\mu'. \tag{2}$$

First, we show that one of the words μ and μ' is a prefix of the other one. Indeed, if it is not the case, we can write

$$\mu = ux\delta, \quad \mu' = uy\delta'$$

with $x, y \in A$, $x \neq y$, $u, \delta, \delta' \in A^*$. Hence $asbux, csbuy \in F(w)$. This implies that sbu is a bispecial factor, so that $asbux$ is a box properly containing asb. This contradicts the maximality of asb.

Now, let us suppose that μ is a prefix of μ'. By Eq. (2), one has that $asb\mu$ is a suffix of w and $sb\mu$ is repeated in w. This implies that $|asb\mu| \leq K_w$. Since $asb \neq k_w$, it follows that asb is a proper factor of k_w, which contradicts the maximality of asb as a box.

Thus μ' is a prefix of μ. By Eq. (2), one has that $csb\mu'$ is a suffix of w and $sb\mu'$ is repeated in w. Consequently, $csb\mu'$ is a suffix of k_w. If $csb\mu'$ is a proper suffix of k_w, then it will be repeated and this contradicts the fact that (λ', μ') is the leftmost occurrence of csb in w. We conclude that $k_w = csb\mu'$.

The remaining part of the proof is obtained by a symmetrical argument.
□

The following theorem [3], whose proof we omit, shows that for any given word w one can construct another word v with the property that all boxes of w are factors of v, whereas there exist at most two boxes of v which are not factors of w.

Theorem 3 *Let $w = \lambda asb\mu$, where asb is a proper maximal box of w. Then, there exists a further occurrence of s in w, i.e. $w = \xi s\eta$, such that the word v defined by*

$$v = \begin{cases} \lambda as\eta & \text{if } |\lambda a| > |\xi| \\ \xi sb\mu & \text{if } |\lambda a| < |\xi|, \end{cases} \tag{3}$$

satisfies the following conditions:

(i) $\mathcal{B}_w \subseteq F(v)$,

(ii) $1 \leq \mathrm{Card}(\mathcal{B}_v \setminus F(w)) \leq 2$.

Example 3 Let w be the word $w = abccbabcab$ considered in Example 1. One has $h_w = abcc$, $k_w = cab$ and the maximal boxes are

$$abcc, \quad bca, \quad cab, \quad cba, \quad ccb.$$

Let us underline in w the occurrences of the two maximal boxes bca and ccb:

$$w = ab\underline{ccb}abcab = abccba\underline{bca}b.$$

According to Theorem 3, we construct the word

$$v = abccba\underline{bcb}abcab.$$

One easily verifies that the maximal boxes of v are $h_v = abcc$, $k_v = cab$, $ccbabcb$ and $bcbabca$. Moreover, $\mathcal{B}_w \subseteq F(v)$ and $ccbabcb$ and $bcbabca$ are the only two maximal boxes of v which are not factors of v.

Starting from the set of maximal boxes of a given word w, we introduce now the so-called *reduced sets* which also provide a representation of the word. These sets can be convenient since, as we shall see, the representation is simpler in the sense that any element of a reduced set is a factor of a maximal box.

Let f be a word and $\mathcal{B}_f$ be the set of its maximal boxes. We construct a new set $\mathcal{D}_f$ as follows. Let s be a bispecial factor of f and consider the set

$$\Omega_s = \{asb \in \mathcal{B}_f \mid a, b \in A,\ as \notin \mathrm{Suff}(h'_f),\ sb \notin \mathrm{Pref}(k'_f)\}.$$

If Ω_s is not empty, we take arbitrarily one maximal box $\alpha = asb \in \Omega_s$ and replace it by all boxes which are proper factors of it. Iterating these operations over all bispecial factors of f, we obtain a set of boxes. By deleting all elements which are not maximal, with respect to the factor order, we construct a set $\mathcal{D}_f$ that we call a *reduced set* of f.

We observe that, in the construction of $\mathcal{D}_f$, it is sufficient to replace the maximal box $\alpha \in \Omega_s$ by the longest box which is a proper prefix of α and the longest box which is a proper suffix of α. Indeed, it is easily seen that any other box in α will not appear in a reduced set. Moreover, by the construction, any box of f which is not maximal, is a factor of an element of $\mathcal{D}_f$.

We remark that, by the arbitrary choice in each Ω_s of the maximal box to be replaced, one can obtain, in general, several distinct reduced sets.

Example 4 Let w be the word $w = babacbcabaccbb$. One has $h_w = bab$, $k_w = bb$. The bispecial factors of w are

$$\epsilon, \quad a, \quad b, \quad c, \quad cb, \quad abac.$$

The set of maximal boxes of w is

$$\mathcal{B}_w = \{bca,\ acbc,\ ccbb,\ babacb,\ cabacc\}.$$

We can construct a reduced set $\mathcal{D}_w$ in the following way. We make the following replacements:

$$\begin{aligned} bca &\to bc,\ ca, \\ acbc &\to acb,\ cbc, \\ babacb &\to bab,\ acb. \end{aligned}$$

By deleting all non-maximal elements, we obtain the reduced set

$$\mathcal{D}_w = \{bab,\ acb,\ cbc,\ ccbb,\ cabacc\}.$$

The following theorem [3], which is an analog of the maximal box theorem, holds. The proof is omitted for the sake of brevity.

Theorem 4 *Let $f, g \in A^*$ be two words and $\mathcal{D}_f$, $\mathcal{D}_g$ be reduced sets of f and g, respectively. Suppose that*

(i) $h_f = h_g, \quad k_f = k_g,$

(ii) $\mathcal{D}_f \subseteq F(g),$

(iii) $\mathcal{D}_g \subseteq F(f).$

Then $f = g$.

6 Superboxes

In this section we introduce the important notion of *superbox* which is strongly related to that of maximal box. An analog of the maximal box theorem is proved in the case of superboxes. Two simple algorithms are given, the first allows us to determine the superboxes of any word w, the second permits to reconstruct the word starting from h_w, k_w and the set of all superboxes.

Proposition 9 *Let $f = asb$ be a box of the word w with $a, b \in A$, $s \in A^*$ such that as and sb are unrepeated. Then f is a maximal box.*

Proof. Suppose that asb is not maximal. Then, there exists a box α such that $\alpha = \lambda asb\mu$, where $\lambda, \mu \in A^*$ and $\lambda\mu \neq \epsilon$. It follows that either as or sb is an internal factor of α and then, by Lemma 1, a repeated factor of w, which is a contradiction. $\square$

We shall set

$$\mathcal{M}_w = \{asb \in F(w) \mid a, b \in A,\ s \text{ repeated and } as, sb \text{ unrepeated.}\}$$

The elements of $\mathcal{M}_w$ will be called *superboxes*. The reason of this name is due to the fact that, as we shall prove by Proposition 13, any element of $\mathcal{B}_w \setminus \{h_w, k_w\}$ is a factor of an element of $\mathcal{M}_w$.

Let us recall (cf. [2]) that a subset X of A^* is called a *factor code* if no word of X is a proper factor of another word of X.

Proposition 10 *Let w be a word. The set $\mathcal{M}_w$ is a factor code. Moreover, no element of $\mathcal{M}_w$ can be a factor of h_w or of k_w.*

Proof. Suppose by contradiction that $asb, ctd \in \mathcal{M}_w$, with $a, b, c, d \in A$, $s, t \in A^*$, and that asb is a proper factor of ctd. This implies that either as or sb is a factor of t, which is absurd since t is a repeated factor of w. Suppose now that asb is a factor of h_w (resp. k_w). Then as (resp. sb) is a factor of h'_w (resp. k'_w) and then it is a repeated factor, which is a contradiction. $\square$

Proposition 11 *Let $f = asb \in \mathcal{M}_w$ with $a, b \in A$ and $s \in A^*$. If $s \neq k'_w$ and $s \neq h'_w$, then f is a proper maximal box.*

Proof. Since s is a repeated factor of w, there will be at least another occurrence of s in w. Let us prove that we can always reduce ourselves to consider only the case when this occurrence is internal.

Let us suppose that s is a suffix of w. Then there exists a letter c such that cs is a suffix of w, and $c \neq a$, since as is unrepeated. If $|cs| < K_w$, then cs is repeated and, therefore, s has a further internal occurrence in w. If $|cs| > K_w$, then $|s| \geq K_w$. This implies that s is unrepeated, which is a contradiction. Finally, if $|cs| = K_w$ one has $s = k'_w$, which contradicts the hypothesis made. The case when s is a prefix is dealt with a symmetric argument.

Let us consider then the case that the further occurrence of s is internal. In such a case, there exist two letters $c, d \in A$, for which $csd \in F(w)$. One has $c \neq a$ and $d \neq b$, because as and sb are unrepeated. This implies that s is a bispecial factor and that $f = asb$ is a proper box. By Proposition 9, the result follows. $\square$

Proposition 12 *Let asb be a proper maximal box of the word w, with $a, b \in A$ and $s \in A^*$. If sb (resp. as) is neither a prefix (resp. suffix) of k'_w nor of h'_w, then sb (resp. as) is unrepeated.*

Proof. Suppose that sb is a repeated factor of w. If sb is a prefix of w, then sb has to be a prefix of h'_w, which has been excluded. Thus there is a letter c such that csb is a factor of w and $c \neq a$, since asb is unrepeated by Proposition 7. By Proposition 8, one derives that csb is a prefix of k_w and then sb is a prefix of k'_w, which has been excluded. Thus, sb is unrepeated.

The remaining part of the proof is carried out in a symmetrical way. □

Proposition 13 *Any maximal box α of a word w, such that $\alpha \neq h_w$ and $\alpha \neq k_w$ is a factor of a superbox.*

Proof. Let $\alpha = asb$ be a proper maximal box of w such that $\alpha \neq h_w$ and $\alpha \neq k_w$. We consider the set of all factors of w of the kind $f = xry$ where $x, y \in A$, r is a repeated factor of w and α is a factor of f.

Let us take in this set a maximal element, with respect to factor ordering, say $\beta = ctd$, with $c, d \in A$, $t \in A^*$. Let us prove that $\beta \in \mathcal{M}_w$. Let us first suppose that td is repeated. If β is right extendable, then one contradicts the maximality of β. If β is not right extendable, since td is repeated, then $\alpha \preceq \beta = k_w$. Since $\alpha \neq k_w$, one contradicts the maximality of α as a box. This proves that td is unrepeated. In a symmetric way, one shows that ct is unrepeated. □

Proposition 14 *The maximal boxes $\alpha \in \mathcal{B}_w \setminus \{h_w, k_w\}$ are either superboxes or prefixes or suffixes of elements of $\mathcal{M}_w \cap A\{h'_w, k'_w\}A$.*

Proof. Let $\alpha \in \mathcal{B}_w \setminus \{h_w, k_w\}$. By Proposition 13, one has that α is a factor of a superbox β. Hence either $\alpha = \beta$ or α is a proper factor of β. In this latter case, β is not a maximal box. By Proposition 11, it follows that $\beta \in A\{h'_w, k'_w\}A$. Since α is a maximal box, α cannot be a factor of h'_w or of k'_w and then, necessarily, it is a prefix or a suffix of β. □

Proposition 15 *Let $f, g \in A^*$ be two words such that*

(i) $h_f \in \mathrm{Pref}(g)$, $k_f \in \mathrm{Suff}(g)$, $h_g \in \mathrm{Pref}(f)$, $k_g \in \mathrm{Suff}(f)$,

(ii) $\mathcal{M}_f \subseteq F(g)$,

(iii) $\mathcal{M}_g \subseteq F(f)$.

Then $f = g$.

Proof. By Proposition 13, any element of $\mathcal{B}_f$ is a factor of an element of $\mathcal{M}_f \cup \{h_f, k_f\}$ and any element of $\mathcal{B}_g$ is a factor of an element of $\mathcal{M}_g \cup \{h_g, k_g\}$. Hence, $\mathcal{B}_f \subseteq F(g)$ and $\mathcal{B}_g \subseteq F(f)$. Thus, by the maximal box theorem, it is sufficient to prove that $h_f = h_g$ and $k_f = k_g$.

Let us prove that $h_f = h_g$. By (i), we can write f as $f = h_f \xi = h_g \xi'$, $\xi, \xi' \in A^*$ and $g = h_g \lambda = h_f \lambda'$, $\lambda, \lambda' \in A^*$. Let us suppose that $|h_f| < |h_g|$. This implies that h_f is a prefix of h'_g and then it is repeated in g.

We consider a further occurrence of h_f in g and a repeated factor α of g of maximal length containing as a factor such an occurrence of h_f. We can write g as $g = \delta\alpha\mu$ with $\delta, \mu \in A^*$. If δ and μ are non-empty, then there exist letters x and y such that $x\alpha y$ is a factor of g and $x\alpha$, αy are unrepeated. Thus, $x\alpha y \in \mathcal{M}_g$ and then $x\alpha y$ is a factor of f. One derives that in f there is a repeated occurrence of h_f, which is a contradiction. Now, suppose $\delta = \epsilon$. Since α is repeated, α is a proper prefix of h_g and, therefore, h_f has two occurrences in h_g. Since $h_g \in F(f)$, this is a contradiction. Finally, if $\mu = \epsilon$, then α is a suffix of k'_g which, by (i), is a proper suffix of f. Thus α has a non-initial occurrence in f. Consequently h_f is repeated, which is a contradiction.

Thus, $|h_f| \geq |h_g|$. Symmetrically, one can prove that $|h_g| \geq |h_f|$ and then $h_f = h_g$. In a symmetric way, one proves that $k_f = k_g$, which concludes the proof. □

Let $\alpha, \beta \in A^*$. We denote by $\alpha \wedge \beta$ the maximal overlap of α with β, i.e. the suffix of maximal length of α which is a prefix of β. Then α and β can be written as $\alpha = \lambda(\alpha \wedge \beta)$, $\beta = (\alpha \wedge \beta)\mu$, $\lambda, \mu \in A^*$. We shall denote by $\alpha \vee \beta$ the word

$$\alpha \vee \beta = \lambda(\alpha \wedge \beta)\mu = \alpha\mu = \lambda\beta.$$

Lemma 2 *Let u and v be two factors of a word $w \in A^*$. If t is an unrepeated factor of w such that $t \in \mathrm{Suff}(u) \cap \mathrm{Pref}(v)$, then $t = u \wedge v$ and $u \vee v \in F(w)$.*

Proof. One has $u = \xi t$ and $v = t\eta$, where $\xi, \eta \in A^*$. Since $|t| \leq |u \wedge v|$, t is both a prefix and a suffix of $u \wedge v$. But t is unrepeated and, therefore, necessarily $t = u \wedge v$. Moreover, the only occurrence of t in w has to be preceded by ξ and followed by η, so that $u \vee v = \xi t \eta \in F(w)$. □

If $t, u, v \in F(w)$ and both $\mathrm{Suff}(t) \cap \mathrm{Pref}(u)$ and $\mathrm{Suff}(u) \cap \mathrm{Pref}(v)$ contain elements which are unrepeated factors of w, then one has $(t \vee u) \vee v = t \vee (u \vee v)$. Indeed, by an iterated application of Lemma 2, one has that $(t \vee u) \vee v$ and $t \vee (u \vee v)$ are both factors of w beginning by t and ending by v. They must coincide, since t and v are unrepeated in w.

More generally, if $\alpha_0, \alpha_1, \ldots, \alpha_n \in F(w)$ and, for any $i = 0, 1, \ldots, n-1$, $\mathrm{Suff}(\alpha_i) \cap \mathrm{Pref}(\alpha_{i+1})$ contains an element which is an unrepeated factor of w, then one can consider the word $\alpha_0 \vee \alpha_1 \vee \cdots \vee \alpha_n$, in which the parentheses are omitted since its value is independent from the order with which the operations are performed.

Let us now introduce a sequence $\Gamma_w = (\alpha_0, \alpha_1, \ldots, \alpha_n)$ of elements of $\mathcal{M}_w \cup \{h_w, k_w\}$ giving a covering of w. We shall call Γ_w also the *covering sequence* of w.

In the first step, we set $\alpha_0 = h_w$.

Now suppose that we have determined the element $\alpha_i \in \mathcal{M}_w \cup \{h_w\}$ $(i \geq 0)$. Let u be the shortest suffix of α_i which is unrepeated in w and write $u = at$ with $a \in A$ and t repeated in w. One can uniquely write $w = \lambda a t \mu$, with $\lambda, \mu \in A^*$.

If $t\mu$ is a repeated factor of w, then, since $at\mu$ is unrepeated in w, one has $at\mu = k_w$. In this case, we set $\alpha_{i+1} = k_w$ and this is the last element of the sequence. Otherwise, we set $\alpha_{i+1} = ar$, where r is the shortest prefix of $t\mu$ which is unrepeated in w. In this latter case, one has $r = sb$, with $b \in A$ and s is a repeated factor of w, while as is unrepeated in w, since it contains the prefix $u = at$. Thus, $\alpha_{i+1} \in \mathcal{M}_w$.

In other words, each element of the sequence Γ_w, excepted the first one, is obtained by taking the shortest suffix of the previous one which is unrepeated in w and extending it in w until one finds either a superbox or k_w.

Example 5 Let w be the word *abccbabcab*. The previous procedure generates the covering sequence of w

$$abcc, \quad ccb, \quad cba, \quad babca, \quad cab.$$

One has $h_w = abcc$ and $k_w = cab$.

Proposition 16 *Let w be a word and $\Gamma_w = (\alpha_0, \alpha_1, \ldots, \alpha_n)$ be the covering sequence of w. Then one has*

$$w = \alpha_0 \vee \alpha_1 \vee \cdots \vee \alpha_n \tag{4}$$

and

$$\mathcal{M}_w = \{\alpha_1, \alpha_2, \ldots, \alpha_{n-1}\}. \tag{5}$$

Proof. By the definition of Γ_w, for any $i = 0, 1, \ldots, n-1$, there is an unrepeated factor t of w such that $t \in \mathrm{Suff}(\alpha_i) \cap \mathrm{Pref}(\alpha_{i+1})$. Thus, by an iterated application of Lemma 2, one gets that the right hand side of Eq. (4) is a factor of w. Since it contains both h_w and k_w it must be necessarily equal to the entire w, so that Eq. (4) is proved.

Consequently, any factor of w either is a factor of an element of Γ_w or contains a word $\alpha_i \wedge \alpha_{i+1}$ as an internal factor, for some $i = 0, 1, \ldots, n-1$. Since the words $\alpha_i \wedge \alpha_{i+1}$ are unrepeated, they cannot occur as internal factors of a superbox and, therefore, any superbox is necessarily a factor of an element of Γ_w. In view of Proposition 10, one easily derives Eq. (5). $\square$

In the sequel, we shall denote $\mathcal{M}_w \cup \{h_w,\ k_w\}$ by $\mathcal{M}'_w$. Proposition 16 ensures that the elements of $\mathcal{M}'_w$ are exactly the elements of Γ_w and, therefore, $\mathcal{M}'_w$ is a covering of w. Let us observe that, in general, the set of maximal boxes $\mathcal{B}_w$ of the word w is not a covering. For instance, in the case of the word $w = abcdebcd$, one has $\mathcal{B}_w = \{ab, de, ebcd\}$ which is not a covering of w.

Next proposition shows that, once we know the initial box h_w of a word w and the set $\mathcal{M}'_w$, we can effectively order the elements of $\mathcal{M}'_w$ to obtain Γ_w.

Proposition 17 *Let w be a word and $\Gamma_w = (\alpha_0, \alpha_1, \ldots, \alpha_n)$ be the covering sequence of w. Set $\delta_i = |\alpha_i \wedge \alpha_{i+1}|$, $0 \leq i \leq n-1$. Then α_{i+1} is the only element $\beta \in \mathcal{M}'_w \setminus \{\alpha_i\}$ such that $|\alpha_i \wedge \beta| \geq \delta_i$.*

Proof. Suppose $\beta \in \mathcal{M}'_w \setminus \{\alpha_i\}$ and $|\alpha_i \wedge \beta| \geq \delta_i$. Let u be the shortest suffix of α_i which is unrepeated in w. By the definition of Γ_w, u is a prefix of α_{i+1}, so that, by Lemma 2, $|u| = |\alpha_i \wedge \alpha_{i+1}| = \delta_i \leq |\alpha_i \wedge \beta|$. One derives that u occurs in β. But, by the way the procedure for the construction of the covering sequence is carried out, the only elements of Γ_w in which u occurs are α_i and α_{i+1}, so that $\beta = \alpha_{i+1}$. $\square$

To reconstruct a word w, knowing h_w and $\mathcal{M}'_w$, first one has to arrange the elements of $\mathcal{M}'_w$ to obtain Γ_w and then use Eq. (4). The first operation is realized by observing that the first element of Γ_w is h_w and that each non-terminal element α_i of Γ_w is followed by the element $\beta \in \mathcal{M}'_w \setminus \{\alpha_i\}$ such that the overlap of α_i with β has maximal length.

An unrepeated factor of w is called *minimal* if any of its proper factors is repeated. Let us denote by $\mathcal{U}_w$ the set of the minimal unrepeated factors of w.

Proposition 18 *Let w be a word and $\Gamma_w = (\alpha_0, \alpha_1, \ldots, \alpha_n)$ be the covering sequence of w. Then one has*

$$\mathcal{U}_w = \{\alpha_0 \wedge \alpha_1, \alpha_1 \wedge \alpha_2, \ldots, \alpha_{n-1} \wedge \alpha_n\}.$$

Proof. By construction, the maximal overlap of an element α_i with the consecutive α_{i+1}, $0 \leq i \leq n-1$, is a minimal unrepeated factor of w.

Conversely, let $u \in \mathcal{U}_w$. We can write uniquely $w = \lambda u \mu$, with $\lambda, \mu \in A^*$. Since u is minimal unrepeated, one has $u = rb$, with $b \in A$ and r is a repeated factor of w. Thus $w = \lambda r b \mu$. If λr is repeated, since $\lambda r b$ is unrepeated, then $\lambda r b = h_w$. Otherwise, consider the shortest unrepeated suffix of λr; we can write it as as, with $a \in A$ and s is a repeated factor. Thus, in this case, since sb is unrepeated, $asb \in \mathcal{M}_w$. Hence, in any case, u can be extended on the left in w in an element of Γ_w, say α_i, with $0 \leq i \leq n-1$. Moreover, u is the shortest suffix of α_i unrepeated in w, so that it is also a prefix of α_{i+1}. By Lemma 2, one has $u = \alpha_i \wedge \alpha_{i+1}$. □

The following 'uniqueness' result was proved in [3]:

Proposition 19 *Let $w \in A^*$. Then w is the unique word of minimal length which begins by h_w, ends by k_w and contains the elements of $\mathcal{M}_w$ as factors.*

In conclusion of this section, we mention that some general theorems relating Nerode's equivalence of the language of the factors of a word and the theory presented in the previous sections are proved in [3].

7 A box theorem for languages

By language on the alphabet A we mean any non-empty subset L of A^*. A language is called *factorial* if $L = F(L)$. The notions of extendable factor, special factor and box can be naturally extended to languages, as follows.

A factor u of a language L is said to be *right* (resp. *left*) *extendable* in L if there exists a letter $a \in A$ such that $ua \in F(L)$ (resp. $au \in F(L)$). We shall denote by U_L the set of all factors of L which cannot be extended on the left in L, i.e.

$$U_L = \{u \in F(L) \mid Au \cap F(L) = \emptyset\}.$$

In a symmetrical way, V_L will denote the set of all factors of L which cannot be extended on the right in L, i.e.

$$V_L = \{u \in F(L) \mid uA \cap F(L) = \emptyset\}.$$

We shall denote by U_L^0 (resp. V_L^0) the set of the elements of U_L (resp. V_L) which are minimal with respect to the prefix (resp. suffix) order. One has:

$$U_L^0 = U_L \setminus U_L A^+, \quad V_L^0 = V_L \setminus A^+ V_L.$$

Let L be a language over the alphabet A. A word s of A^* is called a *right* (resp. *left*) *special factor* of L if there exist $x, y \in A$, $x \neq y$, such that $sx, sy \in F(L)$ (resp. $xs, ys \in F(L)$). A word of A^* which is a right and left special factor of L is called a *bispecial factor* of L.

We introduce the set B_L of the proper boxes of L. A *proper box* α of L is a factor of L of the kind $\alpha = asb$ with $a, b \in A$ and s a bispecial factor of L. Any element of U_L^0 (resp. V_L^0) is called an *initial* (resp. *terminal*) *box*. By *box*, without specification, we mean indifferently an initial, a terminal or a proper box.

Let L, M be two languages over the alphabet A such that $F(L) \neq F(M)$. A *separating factor* of L and M is any word in the symmetric difference $(F(L) \setminus F(M)) \cup (F(M) \setminus F(L))$. A *minimal separating factor* of L and M is a separating factor of minimal length.

The following lemma, whose proof is in [4], is the key result which allows us to prove a box theorem for languages.

Lemma 3 *Let $L, M \subseteq A^*$ be two languages such that $F(L) \neq F(M)$ and let u be a minimal separating factor of L and M. Set $p = \max\{2, |u|\}$. If the following conditions are satisfied*

1) $U_L^0 \cap A^{p-1} = U_M^0 \cap A^{p-1}, \quad V_L^0 \cap A^{p-1} = V_M^0 \cap A^{p-1}$

2) $B_L \cap A^p \subseteq F(M), \quad B_M \cap A^p \subseteq F(L),$

then $\mathrm{Card}(\mathrm{alph}(L)), \mathrm{Card}(\mathrm{alph}(M)) \leq 1$.

Theorem 5 *Let L and M be languages on the alphabet A. If the following conditions are satisfied:*

1) $U_L^0 = U_M^0, \quad V_L^0 = V_M^0$

2) $B_L \subseteq F(M), \quad B_M \subseteq F(L),$

then either $F(L) = a^*$, $F(M) = b^*$, with $a, b \in A$ and $a \neq b$, or $F(L) = F(M)$.

Proof. We shall prove that if $F(L) \neq F(M)$, then $F(L) = a^*$, $F(M) = b^*$, with $a, b \in A$ and $a \neq b$.

By Lemma 3, one has $\mathrm{Card}(\mathrm{alph}(L)), \mathrm{Card}(\mathrm{alph}(M)) \leq 1$. In such a case, either $V_L^0 \neq \emptyset$ and $F(L) = F(V_L^0)$ or $V_L^0 = \emptyset$ and L is an infinite language; similarly, one has that either $V_M^0 \neq \emptyset$ and $F(M) = F(V_M^0)$ or $V_M^0 = \emptyset$ and M is an infinite language. Thus, if $V_L^0 = V_M^0 \neq \emptyset$, one would have

$$F(L) = F(V_L^0) = F(V_M^0) = F(M),$$

which is a contradiction. The only remaining possibility is that $V_L^0 = V_M^0 = \emptyset$ and $L \subseteq a^*$ and $M \subseteq b^*$ are infinite languages, with $a \neq b$. $\square$

Let us observe that the preceeding theorem is, in fact, a box theorem for factorial languages (with the only exception of a trivial case). However, by using boardmarkers for the words of the language L, a box theorem for languages can be easily derived. More precisely, let us consider a new alphabet $A_0 = A \cup \{\#\}$, where $\# \notin A$. For any language $L \subseteq A^*$, we introduce the language

$$\hat{L} = \#L\#.$$

The following theorem holds [4].

Theorem 6 *Let L and M be languages on the alphabet A such that*

$$\mathcal{B}_{\hat{L}} \subseteq F(\hat{M}), \quad \mathcal{B}_{\hat{M}} \subseteq F(\hat{L}).$$

Then $L = M$.

An infinite word f (from left to right) on the alphabet A is any map $f : \mathbb{N}_+ \to A$. We shall set for any $i \geq 1$, $f_i = f(i)$ and write

$$f = f_1 f_2 \cdots f_n \cdots.$$

A word $u \in A^*$ is a *factor* of f if $u = \epsilon$ or there exist integers i, j such that $1 \leq i \leq j$ and $u = f_i \cdots f_j$. The set of all factors of f is denoted by $F(f)$. A factor u of an infinite word f is called respectively *right special*, *left special*, *bispecial* if it is a right special, left special, bispecial factor of $F(f)$. For an infinite word f we shall write B_f, U_f^0 and V_f^0 instead of $B_{F(f)}$, $U_{F(f)}^0$ and $V_{F(f)}^0$.

An infinite word f is *periodic* of period p if for any $i \geq 1$ one has $f_i = f_{i+p}$.

As an application of Theorem 5, we shall give a new proof of the 'uniqueness theorem' for periodic functions of Fine and Wilf [11], in the discrete case. We recall that this theorem can be stated as follows.

Theorem 7 *Let f and g be two infinite periodic words of periods p and q, respectively. Let $d = \gcd(p, q)$. If*

$$f_1 f_2 \cdots f_{p+q-d} = g_1 g_2 \cdots g_{p+q-d},$$

then $f = g$.

Proof. If $p = q$ the result is trivial. Thus, we can always suppose $p < q$.

Let us first consider the case that

$$\gcd(p, q) = 1.$$

Set $w = f_1 f_2 \cdots f_{p+q-1} = g_1 g_2 \cdots g_{p+q-1}$. By the periodicity of f and g and the fact that $|w| = p + q - 1 \geq 2p - 1$, one has

$$f \sim_p w \sim_p g. \tag{6}$$

Since the finite word w has period p, then as proved in [6], one has

$$R_w + 1 \leq p. \tag{7}$$

By (6), f, g and w have the same right special factors up to the length $p-1$. In particular, they have no right special factor of length R_w. Consequently, the length of a proper box of f or g is at most $R_w + 1$. By Eqs. (6) and (7) one derives

$$B_f \subseteq F(g), \quad B_g \subseteq F(f).$$

Moreover, one has $U_f^0 = U_g^0 = \emptyset$ by the periodicity of f and g, and $V_f^0 = V_g^0 = \emptyset$ because f and g are infinite words (from left to right). Thus by Theorem 5 it follows either $F(f) = a^*$, $F(g) = b^*$, with $a, b \in A$ and $a \neq b$, or

$$F(f) = F(g).$$

The first case cannot occur in our case. Thus, from the previous equality, since f has period p, it follows that also g has period p. Since f and g have the same initial segment of length p, one derives $f = g$.

Let us now suppose that $d > 1$. For any $r = 1, 2, \ldots, d$, consider the infinite words

$$f_r f_{r+d} f_{r+2d} \cdots f_{r+nd} \cdots,$$

$$g_r g_{r+d} g_{r+2d} \cdots g_{r+nd} \cdots.$$

They have respectively periods $\frac{p}{d}$ and $\frac{q}{d}$. Moreover they have the same initial segment of length $\frac{p+q}{d} - 1$ so that, by the preceding result, they are equal. It follows that $f = g$. $\square$

In conclusion, we mention that the notion of 'special factor' can be extended in a natural way to the case of sets of bidimensional words (*pictures*), taking into account *right*, *left*, *up*, and *down extensions* of subpictures. In this way, one can generalize in a suitable way the notion of box. A box theorem for picture languages, extending Theorem 6, is proved in [4].

References

[1] M.-P. Béal, F. Mignosi and A. Restivo, Minimal forbidden words and symbolic dynamics, Proc. STACS '96, Lecture Notes in Computer Science, Springer-Verlag, vol. 1046, 1996, pp. 555–566.

[2] J. Berstel and D. Perrin, *Theory of codes*, Academic Press, New York, 1985.

[3] A. Carpi and A. de Luca, Words and Special Factors, Preprint 98/33, Dipartimento di Matematica dell'Università di Roma 'La Sapienza', 1998.

[4] A. Carpi and A. de Luca, Repetitions and Boxes in Words and Pictures, Preprint 98/44, Dipartimento di Matematica dell'Università di Roma 'La Sapienza', 1998.

[5] J. Cassaigne, Complexité et facteurs speciaux, *Bull. Belg. Math. Soc.* **4** (1997) 67–88.

[6] A. de Luca, On the Combinatorics of Finite Words, *Theoretical Computer Science*, Special Issue for the workshop 'Words', to appear.

[7] A. de Luca and F. Mignosi, Some Combinatorial properties of Sturmian words, *Theoretical Computer Science* **136** (1994) 361–385.

[8] A. de Luca and L. Mione, On bispecial Factors of the Thue-Morse word, *Information Processing Letters* **49** (1994) 361–385.

[9] A. de Luca and S. Varricchio, On the factors of the Thue-Morse word on three symbols, *Information Processing Letters* **27** (1988) 281–285.

[10] A. de Luca and S. Varricchio, Some combinatorial problems of the Thue-Morse sequence and a problem in semigroups, *Theoretical Computer Science* **63** (1989) 333–348.

[11] N. J. Fine and H. S. Wilf, Uniqueness Theorem for Periodic Functions, *Proc. Amer. Math. Soc.*, **16** (1965) 109–114.

Arturo Carpi
Istituto di Cibernetica
CNR
via Toiano, 6
80072 Arco Felice (NA), Italy
arturo@arturo.cib.na.cnr.it

Aldo de Luca
Dipartimento di Matematica
Università di Roma 'La Sapienza'
piazzale Aldo Moro, 2
00185 Roma, Italy
deluca@mercurio.mat.uniroma1.it