RAIRO-Theor. Inf. Appl. 46 (2012) 33–50
DOI: 10.1051/ita/2011131

Available online at:
www.rairo-ita.org

AN APERIODICITY PROBLEM FOR MULTIWORDS

Véronique Bruyère[1], Olivier Carton[2],
Alexandre Decan[1], Olivier Gauwin[1] and Jef Wijsen[1]

Abstract. Multiwords are words in which a single symbol can be replaced by a nonempty set of symbols. They extend the notion of partial words. A word w is *certain* in a multiword M if it occurs in *every* word that can be obtained by selecting one single symbol among the symbols provided in each position of M. Motivated by a problem on incomplete databases, we investigate a variant of the pattern matching problem which is to decide whether a word w is certain in a multiword M. We study the language $\mathsf{CERTAIN}(w)$ of multiwords in which w is certain. We show that this regular language is aperiodic for three large families of words. We also show its aperiodicity in the case of partial words over an alphabet with at least three symbols.

Mathematics Subject Classification. 68R15, 68Q45.

1. Introduction

Given a pattern w and a text t, the *pattern matching problem* is to find all the occurrences of the word w in t. There exist efficient algorithms that solve this problem, like the well-known Knuth-Morris-Pratt algorithm [13] and Boyer-Moore algorithm [5] (see also Chaps. 3 and 4 in [7]).

Several extensions of this problem have been studied. Instead of a single pattern w, the Aho-Corasick algorithm efficiently finds in a text t all the occurrences of words w taken from a finite set of words [1]. A more general problem is the regular expression matching problem where the pattern is a set of words specified by a regular expression (see for instance Chap. 7 in [7]).

Keywords and phrases. Pattern matching, aperiodicity, partial words.

[1] Université de Mons – UMONS, Institut d'Informatique, Belgium. `first.last@umons.ac.be`

[2] Université Paris Diderot, LIAFA, France. `olivier.carton@liafa.jussieu.fr`

Article published by EDP Sciences

Other extensions deal with the pattern matching problem by allowing *don't-care* symbols in the pattern w and/or in the text t. In this case, some positions in the pattern or in the text can contain a set of symbols, instead of a single symbol. A word with don't-care symbols represents a finite set of (classical) words obtained by selecting a single symbol among the symbols provided in each don't-care position. Therefore, if w is a pattern with don't-care symbols and t is a text, the problem consists in finding all the occurrences of words represented by w in the text t. When w is a pattern and t is a text with don't-care symbols, we are interested in finding the occurrences of w in t such that in each don't-care position i, the symbol at the corresponding position of w belongs to the set of symbols of t at position i.

When don't-care symbols are allowed, most of the existing exact methods for pattern matching are useless or have to be adapted. One among the first works in this framework has been presented by Fisher and Paterson in [10]. Without being exhaustive, let us also mention the recent references [12, 14, 18].

The interest in words with don't-care symbols is driven by applications in computational biology, cryptanalysis, musicology, and other areas. The problem studied in this article is motivated by research in incomplete historical databases, as described in [6]. It can be seen as a variant of pattern matching: given a pattern w and a text t with don't-care symbols, does w appear as a factor of each word z represented by t? It is important to notice that we want to be sure that w appears in *each* z, and not in *some* z.

Given a pattern w, the authors of [6] provide a deterministic finite automaton $\mathcal{A}(w)$ recognizing the set $\mathsf{CERTAIN}(w)$ of all words t with don't-care symbols such that w is a factor of each word z represented by t. This automaton is a kind of Knuth-Morris-Pratt automaton (see Chap. 9 of [2]), with a more sophisticated use of the prefixes of w. They also prove that for a particular class of words w, the regular set $\mathsf{CERTAIN}(w)$ is aperiodic, or equivalently [16,19], first-order expressible.

In this article, we study the set $\mathsf{CERTAIN}(w)$ and we partially solve the conjecture proposed in [6] that $\mathsf{CERTAIN}(w)$ is aperiodic for every word w. We prove the aperiodicity of $\mathsf{CERTAIN}(w)$ for three large families of words w including powers of primitive words (for a power greater than or equal to 3) and powers of unbordered words. We also show that, when restricted to partial words, $\mathsf{CERTAIN}(w)$ is aperiodic for alphabets with at least 3 symbols.

In the literature, different terms have been used for words with don't-care symbols like indeterminate words [12], partial words, words with holes or jokers [3,4,8]. In each case, either the don't-care symbol means any symbol of the alphabet, or it has to be selected among a subset of the alphabet depending on its position in the word. In this article, we follow the second approach and we use the term *multiword* coined in [6]. The notion of partial word has been generalized in [11] by the concept of relational word. In this article, the term partial word refers to words where don't-care positions represent the entire alphabet.

The remainder of this article is organized as follows. The next section introduces terminology and notations and formalizes the problems we are interested in. Section 3 proves the aperiodicity of the restriction of $\mathsf{CERTAIN}(w)$ to partial words, when the alphabet has at least 3 symbols. In Section 4, we introduce our decomposition lemma, the main tool used in the technical treatment.

Section 5 contains our main results. It establishes the aperiodicity of $\mathsf{CERTAIN}(w)$ for three large families of words w: powers of primitive words (with powers greater than or equal to 3), powers of unbordered words, and so-called anchored words. We end with a conclusion and some perspectives.

2. Preliminaries

2.1. Words

Let $\Sigma = \{a, b, c, \ldots\}$ be a finite alphabet of symbols. A *word* of *length* $n \geq 0$ is a total function $w \colon \{1, \ldots, n\} \to \Sigma$. As usual, we write such a word $w_1 \ldots w_n$ where $w_i = w(i)$ is the *symbol* at *position* i. The *empty word*, denoted by ϵ, has length 0. The *concatenation* of words u and v is denoted by $u \cdot v$ or uv. If $w = pq$, then p is called a *prefix* of w and q a *suffix*. A prefix (or suffix) of w that is distinct from w is called *proper*. We say that a word w is a *factor* of v, denoted by $v \Vdash w$, if there exist words p and q such that $v = pwq$. We denote as usual by Σ^* the set of all words over Σ, and $\Sigma^+ = \Sigma^* \setminus \{\epsilon\}$. A word w is called *unbordered* if no nonempty proper suffix of w is a prefix of w. A word $w \in \Sigma^+$ is *primitive* if $w = v^k$ implies $k = 1$.

2.2. Multiwords

We define the *powerset alphabet* as $\widehat{\Sigma} = 2^{\Sigma} \setminus \{\emptyset\}$. A *multiword* $M = A_1 A_2 \ldots A_n$ is a finite word over the powerset alphabet $\widehat{\Sigma}$, *i.e.* $A_i \subseteq \Sigma$ and $A_i \neq \emptyset$ for all i.

Given a multiword $M = A_1 A_2 \ldots A_n$, we define the set of words represented by M:

$$\mathsf{words}(M) := \{a_1 a_2 \ldots a_n \mid \forall i \in \{1, \ldots, n\} : a_i \in A_i\}.$$

Let w be a word. We say that a word w is *certain* in M, denoted $M \Vdash_{\mathsf{certain}} w$, if w is a factor of every word in $\mathsf{words}(M)$. Given a word $w \in \Sigma^+$, we are interested in the language $\mathsf{CERTAIN}(w) \subseteq \widehat{\Sigma}^*$ defined as follows:

$$\mathsf{CERTAIN}(w) := \left\{ M \in \widehat{\Sigma}^* \mid M \Vdash_{\mathsf{certain}} w \right\}.$$

Example 2.1. The following multiword M contains two symbols with values $\{a, b\}$ and $\{c, d\}$. Curly braces are omitted for symbols that are singletons; for example, $\{a\}$ is written as a.

So, for $M = abdabca\{a, b\}bdab\{c, d\}abcab$, we have:

$$\begin{aligned} \mathsf{words}(M) = \{ & abdabca\underline{\mathbf{a}bdab\mathbf{c}ab}cab, \\ & abdabca\mathbf{a}bd\underline{ab\mathbf{d}abcab}, \\ & \underline{abdabca\mathbf{b}}bdab\mathbf{c}abcab, \\ & \underline{abdabca\mathbf{b}}bd\underline{ab\mathbf{d}abcab}\}. \end{aligned}$$

Hence, $M \Vdash_{\mathsf{certain}} abdabcab$ because $abdabcab$ is a factor (underlined for readability) of each word in $\mathsf{words}(M)$. So we have $M \in \mathsf{CERTAIN}(abdabcab)$.

2.3. Partial words

A *partial word* is a multiword $M = A_1 A_2 \ldots A_n$ where, for each i, either $A_i = \Sigma$, or $A_i = \{a\}$ with $a \in \Sigma$. The term “partial” refers to the fact that a partial word of length n over Σ can be considered as a partial function $w : \{1, \ldots, n\} \to \Sigma$, where positions i with undefined $w(i)$ correspond to the previous case $A_i = \Sigma$ [3]. In the context of partial words, it is common to use notation $\diamond$ to identify cases $A_i = \Sigma$.

Example 2.2. Let $\Sigma = \{a, b, c\}$. The partial word $M_\diamond = a\{a, b, c\}b\{a, b, c\}c$ is also denoted $a\diamond b\diamond c$ following notation in [3]. The set of possible words of $M_\diamond$ is:

$$\mathsf{words}(M_\diamond) = \{aabac, aabbc, aabcc, abbac, abbbc, abbcc, acbac, acbbc, acbcc\}$$

Let $\mathsf{CERTAIN}_\diamond(w)$ be the restriction of $\mathsf{CERTAIN}(w)$ to partial words:

$$\mathsf{CERTAIN}_\diamond(w) = \{M \mid M \text{ is a partial word and } M \Vdash_{\mathsf{certain}} w\}.$$

2.4. Aperiodicity

In this article, our main motivation is to prove the first-order definability of $\mathsf{CERTAIN}(w)$, for every word w. An intermediate objective is to prove it for $\mathsf{CERTAIN}_\diamond(w)$.

It has been shown in [6] that $\mathsf{CERTAIN}(w)$ is regular, by exhibiting an automaton recognizing this language. As the language $\mathsf{CERTAIN}_\diamond(w)$ is the intersection of $\mathsf{CERTAIN}(w)$ with the (regular) set of partial words, $\mathsf{CERTAIN}_\diamond(w)$ is also regular. Hence the first-order definability of $\mathsf{CERTAIN}(w)$ and $\mathsf{CERTAIN}_\diamond(w)$ reduces to their aperiodicity [16, 19]. A monoid M is aperiodic if there exists an integer n such that $s^{n+1} = s^n$ for any $s \in M$ [17]. By extension, a language L is *aperiodic* if its syntactic monoid $M(L)$ is aperiodic. The syntactic monoid $M(L)$ is equal to $\Sigma^*/{\sim_L}$ where the syntactic congruence $\sim_L$ is defined by

$$u \sim_L u' \iff \forall p, q \in \Sigma^* \ (puq \in L \iff pu'q \in L).$$

It follows that a language is aperiodic if there exists an integer k such that

$$\forall p, u, q \in \Sigma^* \ (pu^k q \in L \iff pu^{k+1} q \in L). \tag{2.1}$$

The question whether $\mathsf{CERTAIN}(w)$ is aperiodic for all $w \in \Sigma^+$ has already been raised in [6]. With our formalism, $\mathsf{CERTAIN}(w)$ is aperiodic if there exists $k > 0$ such that for all $P, U, Q \in \widehat{\Sigma}^*$,

$$PU^k Q \Vdash_{\mathsf{certain}} w \iff PU^{k+1} Q \Vdash_{\mathsf{certain}} w.$$

In [6], the aperiodicity of $\mathsf{CERTAIN}(w)$ has been proved for a particular family of words w, namely the words au (resp. ua) where $a \in \Sigma$ and $u \in (\Sigma \setminus \{a\})^*$. In Section 5, we extend this result to three (much larger) classes of words w, namely, the class of powers of primitive words (with powers greater than or equal to 3) (Thm. 5.1), the class of powers of unbordered words (Thm. 5.5), and the class of so-called anchored words (Thm. 5.7). In Section 3, we prove aperiodicity of $\mathsf{CERTAIN}_\diamond(w)$ for partial words over an alphabet with at least 3 symbols (Thm. 3.2).

We introduce the following lemma, which allows to restrict the proof of aperiodicity to only one implication instead of an equivalence as in (2.1).

Lemma 2.3. *A regular language L is aperiodic if and only if there exists an integer n such that for any integer $k \geq n$*

$$\forall p, u, q \in \Sigma^* \ (pu^k q \in L \implies pu^{k+1} q \in L).$$

Proof. The condition is obviously necessary. To prove that the condition is sufficient, we use the classical result that for any finite monoid M, there exists an integer n such that for any $s \in M$, s^n is an idempotent, that is $s^{2n} = s^n$ [17].

By this result applied to the syntactic monoid of L, there exists an integer m such that

$$\forall p, u, q \in \Sigma^* \ (pu^m q \in L \iff pu^{2m} q \in L).$$

Let $k \geq \mathsf{max}(n, m)$, and let i, j be integers such that $k = i \cdot m + j$ with $0 \leq j < m$. We have

$$pu^{k+1}q \in L \implies pu^{k+2}q \in L \implies \dots \implies pu^{2k-j}q \in L \implies pu^k q \in L$$

since $pu^{2k-j}q \in L \ \Leftrightarrow \ p(u^i)^{2m}u^j q \in L \ \Leftrightarrow \ p(u^i)^m u^j q \in L \ \Leftrightarrow \ pu^k q \in L.$ □

3. Aperiodicity of $\mathsf{CERTAIN}_\diamond(w)$

We start with the case of partial words. We show in this section that the set $\mathsf{CERTAIN}_\diamond(w)$ is aperiodic for alphabets of size greater than or equal to 3. Note that, for an alphabet of smaller size, the notions of multiwords and partial words coincide.

We begin with an interesting lemma dealing with multiwords M containing a symbol A_i with at least three values.

Lemma 3.1. *Let $M = A_1A_2 \dots A_n \in \widehat{\Sigma}^*$ be a multiword, and $w \in \Sigma^+$ a word. Let $i \in \{1, \dots, n\}$ such that $|A_i| \geq 3$. Then, $M \in \mathsf{CERTAIN}(w)$ if and only if $A_1A_2 \dots A_{i-1} \in \mathsf{CERTAIN}(w)$ or $A_{i+1} \dots A_n \in \mathsf{CERTAIN}(w)$.*

Proof. We only need to prove the necessary condition. As $|A_i| \geq 3$, let a, b, c be three distinct symbols in A_i. Assume $M \in \mathsf{CERTAIN}(w)$, $A_1 \dots A_{i-1} \notin \mathsf{CERTAIN}(w)$ and $A_{i+1} \dots A_n \notin \mathsf{CERTAIN}(w)$.

Let $m \in \mathsf{words}(A_1 \dots A_{i-1})$ and $m' \in \mathsf{words}(A_{i+1} \dots A_n)$ such that neither m nor m' contains w as a factor. By hypothesis, w is a factor of mam', mbm', mcm'. Therefore there exist words u, v, x, y such that $w = uavbxcy$ and:

- u is a suffix of m and $vbxcy$ is a prefix of m';
- uav is a suffix of m and xcy is a prefix of m'; and
- $uavbx$ is a suffix of m and y is a prefix of m'.

Graphically,

$\leftarrow\ m\ \rightarrow$		$\leftarrow\ m'\ \rightarrow$
$\dots u$	a	$vbxcy \dots$
$\dots uav$	b	$xcy \dots$
$\dots uavbx$	c	$y \dots$

Let $k = |u| + 1 + |x| + 1$. If we start reading the first line in the array above, and switch to the second one after m, we observe that the kth symbol of w must be c. Let $j = |y| + 1 + |v| + 1$. Then, if we start reading the third line from the end of w, and switch to the second one after reading m', we note that the jth last symbol of w must be a. Since $k + j = |w| + 1$, the kth symbol equals the jth last symbol, hence $a = c$, a contradiction. This concludes the proof. □

Theorem 3.2. *If $|\Sigma| \geq 3$, then $\mathsf{CERTAIN}_{\diamond}(w)$ is aperiodic for each $w \in \Sigma^+$.*

Proof. We prove aperiodicity by using Lemma 2.3. Let $k > |w|$. Let P, U, Q be partial words. Assume $PU^kQ \Vdash_{\mathsf{certain}} w$. For contradiction, suppose $PU^{k+1}Q \nVdash_{\mathsf{certain}} w$. Therefore $U \neq \epsilon$, and two cases occur:

- U is composed only of singleton symbols. Then $\mathsf{words}(U)$ is exactly $\{u\}$ for some u. As $PU^{k+1}Q \nVdash_{\mathsf{certain}} w$, there exist some $p \in \mathsf{words}(P)$, $q \in \mathsf{words}(Q)$ such that $pu^{k+1}q \nVdash w$.
  Since $pu^kq \in \mathsf{words}(PU^kQ)$ and $PU^kQ \Vdash_{\mathsf{certain}} w$, it follows $pu^kq \Vdash w$. As $|u^k| > |w|$ (because $k > |w|$ and $|u| > 0$), we have $pu^k \Vdash w$ or $u^kq \Vdash w$, a contradiction with $pu^{k+1}q \nVdash w$;
- U contains at least one symbol that is not a singleton. In this case, we have $PU^kQ = PM_1 \diamond M_2 Q$ with M_1, M_2 partial words and symbol $\diamond$ is Σ (with size ≥ 3). By Lemma 3.1, either $PM_1 \Vdash_{\mathsf{certain}} w$ or $M_2Q \Vdash_{\mathsf{certain}} w$. Assume $PM_1 \Vdash_{\mathsf{certain}} w$ (the other case is symmetrical). Clearly, for every partial word M', we have $PM_1M' \Vdash_{\mathsf{certain}} w$. Let M' be the partial word such that $PM_1M' = PU^{k+1}Q$. Then, $PU^{k+1}Q \Vdash_{\mathsf{certain}} w$, a contradiction. □

4. Decomposition lemma

Our main aperiodicity results are stated and proved in Section 5. In each case, aperiodicity is established by using Lemma 2.3, *ad absurdum*, in combination with a decomposition lemma. Section 4 is devoted to this lemma which is our main technical tool.

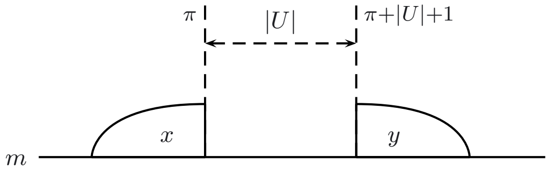

FIGURE 1. ℓ-decomposition (x, y) of w at position π.

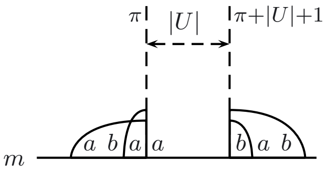

FIGURE 2. Two maximal ℓ-decompositions.

Lemma 4.1 (decomposition lemma). *Let $w \in \Sigma^+$ be a word, and $k \geq 1$. Let $P, U, Q \in \widehat{\Sigma}^*$ be multiwords such that $P\ U^k\ Q \ \Vdash_{\mathsf{certain}}\ w$. Let $p \in \mathsf{words}(P)$, $q \in \mathsf{words}(Q)$ and $u \in \mathsf{words}(U^{k+1})$. If $m = puq$ does not contain w as a factor, then for every position π in m such that $|P| \leq \pi \leq |PU^k|$, there exist $x, y \in \Sigma^+$ such that:*

(1) *$w = xy$;*
(2) *x is a nonempty suffix of $m_1 \dots m_\pi$; and*
(3) *y is a nonempty prefix of $m_{\pi+|U|+1} \dots m_{|m|}$.*

In other words, this lemma states that, for every position π of m (under the hypotheses), there exist a prefix x of w ending at position π and a suffix y of w beginning at position $\pi + |U| + 1$, such that $xy = w$. This situation is depicted in Figure 1.

The pair (x, y) mentioned in Lemma 4.1 is called an *ℓ-decomposition of w at position π* (or simply an ℓ-decomposition at position π, if w is clear from the context), and also an *r-decomposition of w at position $\pi + |U| + 1$*.[3] A position π is *left-maximal* if $m_{\pi+1} \neq m_{\pi+1+|U|}$. Any ℓ-decomposition (x, y) at a left-maximal position is called *maximal*. A position π is *right-maximal* if $m_{\pi-1} \neq m_{\pi-1-|U|}$. Any r-decomposition (x, y) at a right-maximal position is called *maximal*. Finally, a *witness* is a maximal ℓ-decomposition of w at a left-maximal position, or a maximal r-decomposition of w at a right-maximal position. The rationale for calling decompositions maximal comes from the following obvious observations:

- if (x, y) is an ℓ-decomposition of w at a left-maximal position π, then $xm_{\pi+1}$ is not a prefix of w. Intuitively, x cannot be extended to the right;
- if (x, y) is an r-decomposition of w at a right-maximal position π, then $m_{\pi-1}y$ is not a suffix of w. Intuitively, y cannot be extended to the left.

Consider for instance the case $w = abab$. Figure 2 shows two ℓ-decompositions (a, bab) and (aba, b) at a position π of a word m. Both ℓ-decompositions are maximal, since none of them can be extended to position $\pi+1$. Hence π is a left-maximal position, with witnesses (a, bab) and (aba, b).

[3] Notice that in the ℓ-decomposition, x ends at position π, and in the r-decomposition, y begins at position $\pi + |U| + 1$ (see Fig. 1).

Proof of Lemma 4.1. Let π be a position in m such that $|P| \leq \pi \leq |PU^k|$. We can assume $m = pv_1uv_2q$ with $|pv_1| = \pi$ and $|u| = |U|$. Let $m' = pv_1v_2q$. From $PU^kQ \Vdash_{\mathsf{certain}} w$ and $m' \in \mathsf{words}(PU^kQ)$, we have $m' \Vdash w$. The situation is:

$$m = \overbrace{pv_1uv_2q}^{\nVdash w} \;\; m' = \underbrace{\overbrace{pv_1}^{\nVdash w} \;\; \overbrace{v_2q}^{\nVdash w}}_{\Vdash w}.$$

But $m \nVdash w$ implies $pv_1 \nVdash w$ and $v_2q \nVdash w$, so it must be the case that pv_1 ends with some nonempty prefix x of w, that v_2q starts with some nonempty suffix y of w, and that $w = xy$. □

The following lemma shows that every ℓ-decomposition can be extended to the right, until a maximal ℓ-decomposition is reached.

Lemma 4.2. *Let w, k, P, U, Q, m and π be defined as in Lemma 4.1. Let (x, y) be an ℓ-decomposition of w at position π. There exists a left-maximal position $\pi + j$ such that $0 \leq j < |y|$ and with a witness (x', y') where $x' = xm_{\pi+1} \dots m_{\pi+j}$.*

Symmetrically, if (x, y) is an r-decomposition at position π, then there exists a right-maximal position $\pi - j$ such that $0 \leq j < |x|$ and with a witness (x', y') where $y' = m_{\pi-j} \dots m_{\pi-1}y$.

Proof. Suppose for contradiction that for all j satisfying $0 \leq j < |y|$, the position $\pi + j$ is not left-maximal. We show that under these conditions, w is a factor of m, which is impossible. Let a be the first symbol of y, and y' be such that $y = ay'$. As π is not left-maximal, x can be extended to $x' = xa$, and (x', y') is an ℓ-decomposition of w at position $\pi + 1$.

We can repeat this step from $\pi + 1$ to $\pi + 2$, and so on, $|y|$ times. Thus, x can be extended symbol by symbol, until w appears as factor. □

Remark 4.3. In the remainder of the paper, we will apply Lemma 4.1 at several positions π "*far enough from extremities*", without explicitly checking the condition $|P| \leq \pi \leq |PU^k|$. In fact, all our proofs are local, in that they work on a *region* of the word m, which length only depends on $|w|$ and $|U|$. Let us check that we can always find such a region, where Lemma 4.1 can be applied.

We can define such a region as an interval between a leftmost position π_1 and a rightmost position π_2. As mentioned above, the width of the region only depends on $|w|$ and $|U|$, that is $\pi_2 = \pi_1 + i|w| + j|U|$ for some $i, j \geq 0$ depending on the proof we consider. For each proof, i and j are fixed, and for every $w \in \Sigma^+$, we have to find k such that for all P, U, Q, there is a position π_1 for which $|P| \leq \pi_1$ and $\pi_1 + i \cdot |w| + j \cdot |U| \leq |PU^k|$. This is equivalent to $|P| \leq \pi_1 \leq |P| + (k - j) \cdot |U| - i \cdot |w|$. As $|U| \geq 1$ (the case $|U| = 0$ being trivial), it is sufficient that $|P| \leq \pi_1 \leq |P| + (k - j) - i \cdot |w|$. We can choose $k \geq j + i \cdot |w|$, so that for every P, U, Q, we can find a position π (and hence an interval of positions) where Lemma 4.1 can be applied.

5. Aperiodicity of $\mathsf{CERTAIN}(w)$

We now present three large families of words w for which we prove the aperiodicity of $\mathsf{CERTAIN}(w)$. The three proofs are based on two main arguments. The first one is the Decomposition Lemma, and the second one uses a notion of "synchronization", that differs for each family, and restricts the way parts of w can overlap. The aperiodicity proof for the first family is quite long and technical. The two other proofs are much easier.

5.1. Powers of primitive words

The first family contains powers (≥ 3) of a primitive word. This is an interesting family since it is well known [15] that every word is a power (≥ 1) of a primitive word.

Theorem 5.1. *If $w = v^h v'$ where v is primitive, $h \geq 3$, and v' is a proper prefix of v, then $\mathsf{CERTAIN}(w)$ is aperiodic.*

Note that this family includes some primitive words[4], as for instance $(ab)^3 a$. Before proving this theorem, we introduce some terminology [15]. A word w is a *conjugate* of a word w' if $w = uv$ and $w' = vu$ for some nonempty words u, v. It is folklore that all conjugates of a primitive word are primitive. We say that a word x *has period* r if $r \neq \epsilon$ and $x = r^i r'$ with $i \geq 1$ and r' is a proper prefix of r. If v is a primitive word, a *v-factorization* of a word x is a factorization of the form $x = r \cdot v^i \cdot s$ where $i \geq 0$, and r (resp. s) is a proper suffix (resp. prefix) of v. A word x may have several v-factorizations. The following lemma shows that the v-factorization is unique when x is large enough.

Lemma 5.2. *Let v be a primitive word. If x has a v-factorization, and $|x| \geq |v| - 1$, then x has only one v-factorization.*

Proof. We first consider the case where $|x| = |v| - 1$. Assume for contradiction that x has two distinct v-factorizations. We can assume that these two v-factorizations are rs and ϵx (where ϵ is the empty word): this can be obtained w.l.o.g. by considering the suitable conjugate of v.

Hence the situation looks like in Figure 3, with r a proper nonempty suffix of v, s a proper prefix of v, and $xa = v$ where a is a symbol. In particular, s is prefix of x. Let b be the symbol such that sb is a proper prefix of v. We will show in the following that x has period r and also period sb. Then, as $|x| = |r| + |sb| - 1$, we can apply Fine and Wilf's theorem [9, 15], to obtain that r and sb are powers of the same word. Hence $v = (sb)r$ is not primitive, which is a contradiction.

Let us prove that x has period r. Figure 4 illustrates the situation. As $x = rs$, it is sufficient to prove that either s is a proper prefix of r, or s has period r. If $|s| < |r|$, then because of the two v-factorizations, s is a proper prefix of r. If $|s| \geq |r|$, r is now a prefix of s. Let r^j be the prefix of s, with $j \geq 1$ being maximal.

[4] It can also be shown that this family does *not* contain squares of primitive words.

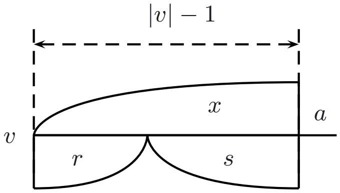

FIGURE 3. Two v-factorizations.

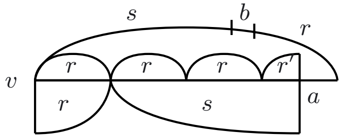

FIGURE 4. x has period r.

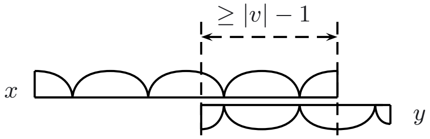

FIGURE 5. Synchronization of words x and y.

The word r^{j+1} is a prefix of rs. But s is also a prefix of rs (consider the two v-factorizations). Hence, either s has period r, or r^{j+1} is a prefix of s. The latter is impossible by definition of j, so s has period r.

Now we prove that sb is a period of x. The proof follows the same line. As $v = (sb)r$, we only have to show that either r is a proper prefix of sb, or r has period sb. Then we just have to consider two cases $|r| < |sb|$ and $|r| \geq |sb|$ as we did before.

We now consider the case where $|x| \geq |v| + 1$. Note that if x has two distinct v-factorizations, then x has a factor x' of length $|x| - 1$ with two distinct v-factorizations. This is impossible according to the preceding arguments. This concludes the proof. □

Let v be a primitive word. If two words x and y have a v-factorization, and have a common factor of length $|v| - 1$, then the preceding lemma implies that their v-factorizations are identical on this common factor, as depicted in Figure 5. In this case, we say that x and y *are synchronized* (according to v).

We now proceed to the proof of Theorem 5.1. This proof relies on Lemma 2.3. It is established by contradiction in a way to use the Decomposition Lemma. Other lemmas are also necessary to complete the proof.

Proof of Theorem 5.1. Let $w \in \Sigma^+$ be a word such that $w = v^h v'$ where v is primitive, $h \geq 3$, and v' is a proper prefix of v. We prove Theorem 5.1 by contradiction, using Lemma 2.3.

Let k be sufficiently large (see Rem. 4.3), and let $P, U, Q \in \widehat{\Sigma}^*$ be multiwords such that $PU^kQ \Vdash_{\mathsf{certain}} w$. Let $p \in \mathsf{words}(P)$, $q \in \mathsf{words}(Q)$ and $u \in \mathsf{words}(U^{k+1})$. Assume for contradiction that $m = puq$ does not contain w as a factor. Therefore $|U| > 0$ and Lemma 4.1 can be applied. We start the proof with two lemmas based on these hypotheses. The following lemma can be paraphrased as follows. Consider

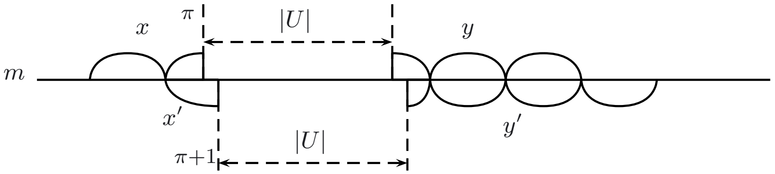

FIGURE 6. Decompositions in the proof of Lemma 5.3, case $|x| < |v|$.

a left-maximal position π with witness (x, y). Then, consider any ℓ-decomposition (x', y') at the next position $\pi+1$. The lemma implies that x and x' cannot overlap much; in particular, the length of each common factor must be less than $|v|-1$.

Lemma 5.3. *Let π be a left-maximal position with witness (x, y). Let (x', y') be an ℓ-decomposition at position $\pi+1$. Then for all common factors f of x and x' (resp. of y and y'), $|f| < |v| - 1$.*

Proof. The words x and x' are proper prefixes of w, while y and y' are proper suffixes of w, so these four words have a v-factorization. Assume that x and x' share a common factor f with $|f| \geq |v|-1$. Then, by Lemma 5.2, they are synchronized. Hence x can be extended to a larger prefix of w (using x'), which contradicts the premise that π is left-maximal. Assume now, that y and y' have a common factor f with $|f| \geq |v|-1$. By Lemma 5.2, they are synchronized, as illustrated in Figure 6. The shift of $|U|$ is the same in both decompositions, so x can be extended to a longer prefix of w by one symbol, which is impossible. □

The second lemma shows that a maximal ℓ-decomposition (x, y) is such that x is large compared to y. In particular, the length of x is always more than twice the length of y. Symmetrically for a maximal r-decomposition.

Lemma 5.4. *If π is a left-maximal position with witness (x, y), then $|x| > |w|-|v|$ and $|y| < |v|$. Symmetrically, if π is right-maximal with witness (x, y), then $|x| < |v|$ and $|y| > |w| - |v|$.*

Proof. We only prove the first part of the lemma (the proof of the second part is symmetrical). Suppose for contradiction that $|x| \leq |w| - |v|$. We distinguish two cases: $|v| \leq |x|$ and $|x| < |v|$.

- Case $|v| \leq |x| \leq |w| - |v|$. We have $|y| \geq |v|$, because $w = xy$. Let (x', y') be an ℓ-decomposition at position $\pi + 1$. As $w = x'y'$ and $|w| \geq 3|v|$, x, x' or y, y' have a common factor f such that $|f| \geq |v| - 1$. This is impossible according to Lemma 5.3;
- case $|x| < |v|$. Now we have $|y| > |w| - |v|$. As $w = v^h v'$ with $h \geq 3$, we get $|y| > 2|v|$. Let us consider an ℓ-decomposition (x', y') at position $\pi + 1$. Lemma 5.3 tells us that $|y'| < |v|$, and thus $|x'| > |w| - |v|$. According to

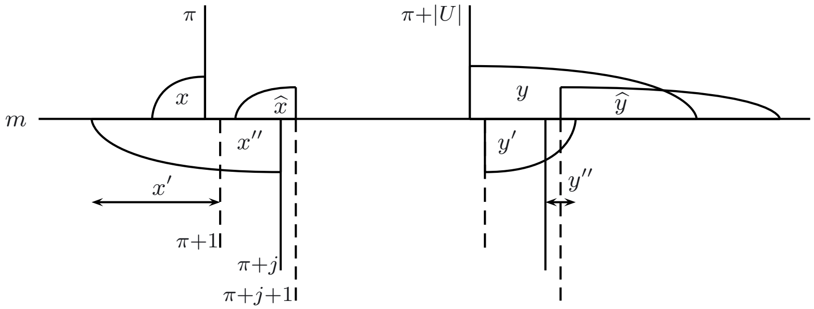

FIGURE 7. Decompositions in the proof of Lemma 5.4.

Lemma 4.2 applied to x' at position $\pi+1$, there exists a left-maximal position $\pi+j$ such that $1 \leq j \leq |y'|$ with witness (x'', y''), x' being a prefix of x''. Now consider an ℓ-decomposition $(\widehat{x}, \widehat{y})$ at position $\pi+j+1$. The situation is illustrated in Figure 7. Recall that $|x'| > |w| - |v|$ and $|y| > |w| - |v|$. As $|x''| \geq |x'|$, we have $|x''| > 2|v|$. Applying Lemma 5.3 at position $\pi + j$, we get $|\widehat{x}| < |v|$ and thus $|\widehat{y}| > |w| - |v|$. We also have $j+1 \leq |y'|+1 \leq |v|$ by Lemma 4.2. So $\widehat{y}$ and y have a common factor f with $|f| > |w| - 2|v| > |v|$. By Lemma 5.2, they are synchronized. Again, two cases can occur:

– y and $\widehat{y}$ end at the same position. In this case, x and $\widehat{x}$ have x as common prefix. Hence, $\widehat{x}$ extends x to the right, yielding a larger prefix of w. This contradicts the premise that π is left-maximal;
– y and $\widehat{y}$ do not end at the same position. As $|\widehat{x}| < |v|$, we know that $\widehat{y}$ ends after y. Moreover, y and $\widehat{y}$ are synchronized, so $\widehat{y}$ is obtained from y by a shift of $|v|^i$ positions in m, for some $i \geq 1$. Let $\widetilde{v}$ be the conjugate of v ending with v' (recall that $w = v^h v'$). The word $y\widetilde{v}$ appears as factor of m. However, $|x| < |v|$, so $|y\widetilde{v}| > |w|$, and thus w is a factor of m. This is impossible.

This concludes the proof of Lemma 5.4. □

We now use the preceding lemmas to complete the proof of Theorem 5.1. By Lemma 4.2, we can assume a left-maximal position π with witness (x, y). By Lemma 5.4, we have $|x| > |w| - |v|$ and $|y| < |v|$. Let (x', y') be an ℓ-decomposition at position $\pi + 1$. According to Lemma 5.3, $|x'| < |v|$. Using Lemma 4.2, we can extend x' until a left-maximal position π' with witness (x'', y''), where x' is a prefix of x''. We know by Lemma 5.4 that $|x''| > |w| - |v|$ and $|y''| < |v|$. In the sequel, we consider the positions π, $\pi - |v|$, α, β and γ, as illustrated in Figure 8:

- α is the position where x' starts, *i.e.* $x' = m_\alpha \ldots m_{\pi+1}$;
- $\beta = \alpha + |v|$;
- $\gamma = \pi - |v| + |y| + 1$ is the position of m corresponding to the $(|w| - |v| + 1)$th position in x.

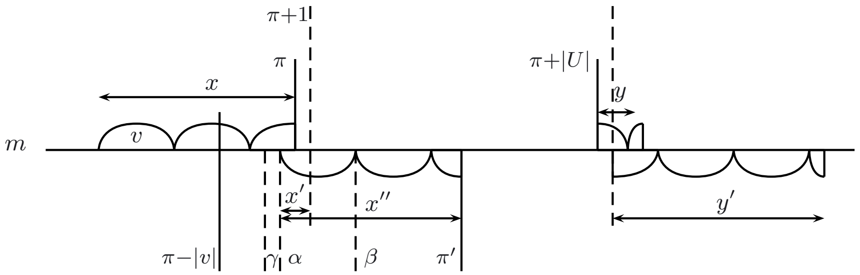

FIGURE 8. Decompositions in the proof of Theorem 5.1, represented in terms of v.

As $1 \leq |y| < |v|$, we know that

$$\pi - |v| + 2 \leq \gamma < \pi + 1.$$

From the definition of α and the fact that $|x'| < |v|$, we have

$$\pi - |v| + 2 < \alpha \leq \pi + 1.$$

Let $(\widehat{x}, \widehat{y})$ be an r-decomposition at position γ. We distinguish two cases, and show that both of them lead to a contradiction.

- Case $|\widehat{y}| < \beta - \gamma$ (*i.e.* $\widehat{y}$ ends before position $\beta - 1$). Then γ is not a right-maximal position. Indeed, $|\widehat{y}| < \beta - \gamma < 2|v|$ because $\beta = \alpha + |v|$ and $\alpha - \gamma < |v|$. However, according to Lemma 5.4, if γ was a right-maximal position, we would have $|\widehat{y}| > |w| - |v| \geq 2|v|$. Hence, by Lemma 4.2, there exists a right-maximal position $\delta < \gamma$ with witness $(\overline{x}, \overline{y})$ such that $\widehat{y}$ is a suffix of $\overline{y}$. This configuration is illustrated in Figure 9. According to Lemma 5.4, $|\overline{y}| > |w| - |v|$. Let us analyze the length of the common factor f of x and $\overline{y}$. We cannot have $|f| < |v| - 1$, because in that case $\overline{y}$ would start after position $\pi - |v| + 1$. As $\overline{y}$ ends before position $\beta - 1$, this would imply that $|\overline{y}| < 2|v|$, a contradiction with $|\overline{y}| > |w| - |v|$. So Lemma 5.2 can be applied, showing that x and $\overline{y}$ are synchronized. As $\overline{y}$ is a suffix of w, and considering the definition of γ, $\overline{y}$ ends after π and allows to extend x, in contradiction with the definition of π;
- case $|\widehat{y}| \geq \beta - \gamma$ (*i.e.* $\widehat{y}$ ends at or after position $\beta - 1$). We will show that the word $m_{\pi-|v|+2} \ldots m_{\pi}$ has two distinct v-factorizations, which constitute a contradiction with Lemma 5.2 (as its length is $|v| - 1$). The first v-factorization comes from x. The second one will be built by extending x' to the left. These two v-factorizations are distinct because π is a left-maximal position with witness (x, y), so x cannot be extended. In the remainder of the proof, we show how to build the second v-factorization from x'. We proceed in two steps, as described in Figure 10.
  - First step. Let z be the common factor between y and y' (we have $|z| = |y| - 1$). From the definition of γ, z appears between positions $\pi - |v| + 2$

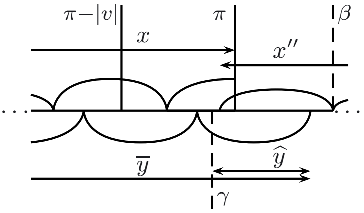

FIGURE 9. Case $|\widehat{y}| < \beta - \gamma$.

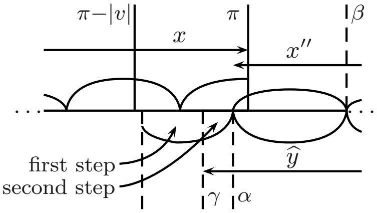

FIGURE 10. Case $|\widehat{y}| \geq \beta - \gamma$.

and $\gamma - 1$ (with its two v-factorizations as suffix of y and prefix of y'): $m_{\pi-|v|+2} \dots m_{\gamma-1} = z$;

- second step. In order to complete the second v-factorization (from position γ to position π), we have to get the suffix of v between positions γ and $\alpha - 1$. If $\alpha \leq \gamma$, the second v-factorization has been completed during the first step. So let us consider that $\gamma < \alpha$. As $|\widehat{y}| \geq \beta - \gamma$ (this corresponds to the second case), x'' and $\widehat{y}$ have a common factor of length greater than $|v| - 1$, so by Lemma 5.2 they are synchronized. Hence $\widehat{y}$ enables to extend x'' to the left until γ, and we obtain the second v-factorization.

This concludes the proof of Theorem 5.1. □

5.2. Powers of unbordered words

The second family of words w for which we prove the aperiodicity of $\mathsf{CERTAIN}(w)$ is composed of every power of an unbordered word. Notice that it contains the words $w = au$ (resp. $w = ua$) with $a \in \Sigma$ and $u \in (\Sigma \setminus \{a\})^*$, for which the aperiodicity of $\mathsf{CERTAIN}(w)$ was proved in [6].

Since every unbordered word is primitive, Theorem 5.1 applies to unbordered words. However, while Theorem 5.1 requires powers greater than or equal to 3, the following theorem admits any power. This second family of words w is incomparable (under set inclusion) with the first one: $(ab)^3a$ belongs to the first family but not to the second one, while ab is unbordered and does not belong to the first family. Of course, there exist words outside both families, like aba.

Theorem 5.5. *If $w = v^h$ with v an unbordered word and $h \geq 1$, then $\mathsf{CERTAIN}(w)$ is aperiodic.*

Proof. The proof is by contradiction, using Lemma 2.3. Let k be sufficiently large (see Rem. 4.3). Let P, U, Q be multiwords such that $PU^kQ \Vdash_{\mathsf{certain}} w$. Assume towards a contradiction that $m = puq$ with $p \in \mathsf{words}(P)$, $u \in \mathsf{words}(U^{k+1})$, and $q \in \mathsf{words}(Q)$ such that $m \nVdash w$. Hence the Decomposition Lemma can be applied.

Lemma 5.6. *There exist a position π in m and an ℓ-decomposition (x, y) at position π such that $|x| \geq |v|$.*

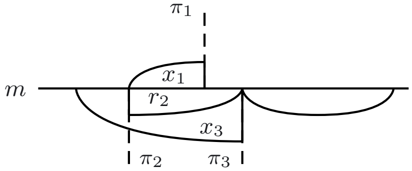

Figure 11. Proof of Lemma 5.6.

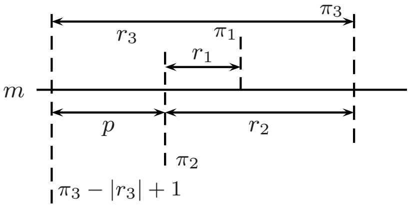

Figure 12. Proof of Theorem 5.5.

Proof. Assume for contradiction that for every position π in m and every ℓ-decomposition (x, y) at π, $|x| < |v|$. For each position π_1, far enough from the borders of m (see Rem. 4.3), there exists an ℓ-decomposition (x_1, y_1) at position π_1 by Lemma 4.1. We choose such position π_1 and ℓ-decomposition (x_1, y_1) with x_1 of maximal length. By our contradiction hypothesis, $|x_1| < |v|$. Let (x_2, y_2) be an r-decomposition at position $\pi_2 = \pi_1 - |x_1| + 1$, and consider r_2 such that $y_2 \in r_2 v^*$ with $0 < |r_2| \leq |v|$. Thus, r_2 is a (not necessarily proper) nonempty suffix of v. As v is unbordered, it must be the case that $|r_2| > |x_1|$. Let (x_3, y_3) be an ℓ-decomposition at position $\pi_3 = \pi_2 + |r_2| - 1$. By our (contradiction) hypothesis, $|x_3| < |v|$. The situation is depicted in Figure 11. As v is unbordered, it must be the case that $|x_3| > |r_2|$. As $|r_2| > |x_1|$, we have a contradiction with the choice of ℓ-decomposition (x_1, y_1) at position π_1 with x_1 of maximal length. □

By Lemma 5.6, we can assume a position π_1 in m and a prefix x_1 of w that ends at position π_1, and such that $|x_1| \geq |v|$. If $w = v^h$ with $h = 1$, it follows that $x_1 = w$ and thus $m \Vdash w$ which is impossible. When $h \geq 2$, we show in the next paragraph that $x_1 m_{\pi_1+1}$ is a prefix of w. Then, by repeated application of the same reasoning, we obtain $m \Vdash w$, again a contradiction.

We can assume $j \geq 1$ such that $x_1 = v^j r_1$ with $0 \leq |r_1| < |v|$. Let (x_2, y_2) be an r-decomposition at position $\pi_2 = \pi_1 - |r_1| + 1$ where $y_2 \in r_2 v^*$ for some r_2 satisfying $0 < |r_2| \leq |v|$. Since v is unbordered it must be the case that $|r_2| > |r_1|$. We distinguish two cases:

- Case $|r_2| = |v|$. Obviously, $x_1 m_{\pi_1+1}$ is a prefix of w;
- case $|r_2| < |v|$. Let (x_3, y_3) be an ℓ-decomposition at position $\pi_3 = \pi_2 + |r_2| - 1$. Let $x_3 \in v^* r_3$ for some r_3 satisfying $0 < |r_3| \leq |v|$. As v is unbordered, $|r_3| > |r_2|$. It follows that the word $p = m_{\pi_3 - |r_3| + 1} \dots m_{\pi_2 - 1}$ must be a nonempty proper prefix of v (see Fig. 12). Since v is unbordered, p cannot be a suffix of v. Since $x_1 = v^j r_1$ is a suffix of $m_1 \dots m_{\pi_1}$ with $j \geq 1$, we have that v^j is a suffix of $m_1 \dots m_{\pi_2 - 1}$. Then, p is a suffix of v, a contradiction. We conclude that this case cannot occur. □

5.3. Anchored words

Given a symbol $a \in \Sigma$, the last family contains words w in which two a-labelled positions are used as anchors in the following way.

Theorem 5.7. *Let $w = savat$ with $a \in \Sigma, v, s, t \in \Sigma^*$ such that:*

(1) *v does not contain a;*
(2) *ava occurs only once in w;*
(3) *If $s \neq \epsilon$, no nonempty prefix of w is a suffix of ava;*
(4) *If $t \neq \epsilon$, no nonempty suffix of w is a prefix of ava.*

Then $\mathsf{CERTAIN}(w)$ *is aperiodic.*

Such a word w is called *anchored* with ava as an anchor. This will be highlighted in the proof. For example, $w = b^2abab^2$ is an anchored word with aba as an anchor; $w = aba$ is also an anchored word.

Let us compare anchored words with our two first families. First, anchored words are primitive words[5] and then cannot be a power of another word, moreover they do not belong to the first family[6]. Second, anchored words and unbordered words are incomparable under set inclusion: b^2abab^2 is anchored and bordered, while ab is unbordered and not anchored. The intersection of anchored and unbordered words is not empty since $abab^2$ is both unbordered and anchored. Third, the family of anchored words covers an important fraction of words. For instance, over the alphabet $\{a, b, c\}$, up to length 14, over a total of 7 174 452 words, the sizes of the three families are 450 for powers of primitive words, 3 999 906 for powers of unbordered words and 6 445 509 for anchored words. Finally, some words do not belong to any of the three families, as for instance $(aba)^2$.

Proof of Theorem 5.7. The proof is again by contradiction, using Lemma 2.3. Let k be large enough (see Rem. 4.3) and let P, U, Q be multiwords such that $PU^kQ \Vdash_{\mathsf{certain}} w$. Assume that there exist $p \in \mathsf{words}(P), u \in \mathsf{words}(U^{k+1}), q \in \mathsf{words}(Q)$ such that $m = puq$ does not contain w as a factor. Therefore Lemma 4.1 can be applied.

Lemma 5.8. *There exist a position π in m and an ℓ-decomposition (x, y) at position π such that either $x \Vdash ava$ or $y \Vdash ava$.*

Proof. Assume the contrary, *i.e.* for all π (far enough from the borders of w, see Rem. 4.3) and all ℓ-decompositions (x, y) at position π, we have $x \nVdash ava$ and $y \nVdash ava$. By Lemma 4.2, we can assume position π is left-maximal and its witness (x, y) is such that $x = sav_1 \dots v_i$ and $y = v_{i+1} \dots v_nat$ where $n = |v|$. Again, by our assumption, there is an ℓ-decomposition (x', y') at position $\pi + 1$ such that $x' = sav_1 \dots v_j$ and $y' = v_{j+1} \dots v_nat$ for some j. If $j \geq 1$, one of $sav_1 \dots v_i$ or $sav_1 \dots v_{j-1}$ must be a suffix of the other (see x, x' in Fig. 13).

We recall that v does not contain the symbol a (by condition (1) of Thm. 5.7). It follows that these two words are equal, and therefore (x, y) is not a maximal

[5] Indeed, if w is anchored but not primitive, $w = u^h$ for some $h > 1$. The anchor ava cannot appear in u by condition (2), so there exists i, $0 \leq i \leq |v|$, such that $av_1 \dots v_i$ is a suffix of u and $v_{i+1} \dots v_{|v|}a$ a prefix of u: this contradicts conditions (3) and (4).

[6] If $w = u^hu'$ with u primitive, $h \geq 3$ and u' a proper prefix of u, then an anchor ava cannot appear in u nor u^2 by condition (2), and can neither appear in uu' (it would appear in u^2).

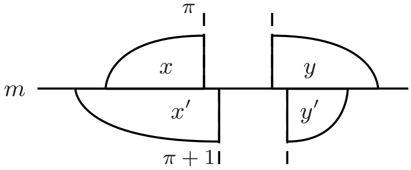

FIGURE 13. Case $j \geq 1$ in the proof of Lemma 5.8

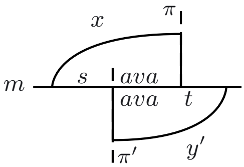

FIGURE 14. Case $x' \Vdash ava$ in the proof of Theorem 5.7.

ℓ-decomposition, a contradiction. So $j = 0$. Assume $i < |v|$. Considering y, the $(|v| - i)$th symbol of y' must be the symbol a, a contradiction as $y' = vat$.

Thus, $j = 0$, $i = |v|$ and $x = sav$, $y = at$, $x' = sa$ and $y' = vat$. Since $m_{\pi+1} = a = m_{\pi+|U|+1}$, it follows that the ℓ-decomposition (x, y) is not maximal, a contradiction. □

By Lemma 5.8, there exist a position π in m and an ℓ-decomposition (x, y) at position π satisfying $x \Vdash ava$ or $y \Vdash ava$. Suppose $x \Vdash ava$ (the other case is symmetrical). By condition (2) of Theorem 5.7, $sava$ is prefix of x. It follows that $t \neq \epsilon$, otherwise $m \Vdash w$ which is impossible. We define $\pi' = \pi - |x| + |s| + 1$ (see Fig. 14). By Lemma 4.1, there is an r-decomposition (x', y') at position π'. By construction, either ava is a proper prefix of y' or y' is prefix of ava. Condition (4) implies that only the first case can happen. By condition (2), we must have $y' = avat$. Recall that $sava$ is prefix of x. It follows that w is factor of m (see Fig. 14), a contradiction. □

6. Conclusions and perspectives

Motivated by a problem in incomplete historical databases, we studied the first-order definability (or aperiodicity) of $\mathsf{CERTAIN}(w)$. Aperiodicity was easy to show for partial words defined relative to an alphabet with at least three symbols. Somewhat surprisingly, aperiodicity proofs turn out to be much harder for multiwords, where uncertain positions can contain exactly two symbols. Using different techniques, we obtained first-order definability for three large classes of words w:

- words of the form $v^h \cdot v'$ with v primitive, $h \geq 3$, and v' a proper prefix of v;
- words of the from v^h with v unbordered and $h \geq 1$; and
- anchored words as defined by Theorem 5.7.

Our proofs are based on synchronization properties of such words, and these techniques do not extend to arbitrary words. It is an open conjecture that $\mathsf{CERTAIN}(w)$ is first-order definable for any word w. This conjecture has been checked experimentally on a large set of words w.

References

[1] A.V. Aho and M.J. Corasick, Efficient string matching: An aid to bibliographic search. *Commun. ACM* **18** (1975) 333–340.

[2] A.V. Aho, J.E. Hopcroft and J.D. Ullman, *The Design and Analysis of Computer Algorithms.* Addison-Wesley (1974).

[3] J. Berstel and L. Boasson, Partial words and a theorem of Fine and Wilf. *Theoret. Comput. Sci.* **218** (1999) 135–141.

[4] F. Blanchet-Sadri, *Algorithmic Combinatorics on Partial Words (Discrete Mathematics and Its Applications).* Chapman & Hall/CRC (2007).

[5] R.S. Boyer and J.S. Mooren, A fast string searching algorithm. *Commun. ACM* **20** (1977) 762–772.

[6] V. Bruyère, A. Decan and J. Wijsen, On first-order query rewriting for incomplete database histories, in *Proc. of the 16th International Symposium on Temporal Representation and Reasoning (TIME)* (2009) 54–61.

[7] M. Crochemore and W. Rytter, *Text Algorithms.* Oxford University Press (1994).

[8] M. Crochemore, C. Hancart and T. Lecroq, *Algorithms on Strings.* Cambridge University Press (2007) 392.

[9] N.J. Fine and H.S. Wilf, Uniqueness theorems for periodic functions. *Proc. of Amer. Math. Soc.* **16** (1965) 109–114.

[10] M.J. Fischer and M.S. Paterson, String matching and other products. *SIAM-AMS Proceedings, Complexity of Computation* **7** (1974) 113–125.

[11] V. Halava, T. Harju and T. Kärki, Relational codes of words. *Theoret. Comput. Sci.* **389** (2007) 237–249.

[12] J. Holub, W.F. Smyth and S. Wang, Fast pattern-matching on indeterminate strings. *J. Discrete Algorithms* **6** (2008) 37–50.

[13] D.E. Knuth, J.H. Morris and V.R. Pratt, Fast pattern matching in strings. *SIAM J. Comput.* **6** (1977) 323–350.

[14] G. Kucherov, L. Noé and M.A. Roytberg, Subset seed automaton, in *Proc. of the 12th International Conference on Implementation and Application of Automata (CIAA).* Springer (2007) 180–191.

[15] M. Lothaire, *Combinatorics on words.* Cambridge University Press (1997).

[16] R. McNaughton and S. Papert, *Counter-free Automata.* MIT Press, Cambridge, MA (1971).

[17] J.-É. Pin, *Varieties of Formal Languages.* North Oxford, London and Plenum, New-York (1986).

[18] M.S. Rahman, C.S. Iliopoulos and L. Mouchard, Pattern matching in degenerate DNA/RNA sequences, in *Workshop on Algorithms and Computation (WALCOM)*, edited by M. Kaykobad and M.S. Rahman. Bangladesh Academy of Sciences (BAS) (2007) 109–120.

[19] M.P. Schützenberger, On finite monoids having only trivial subgroups. *Inform. Control* **8** (1965) 190–194.

Communicated by G. Richomme.
Received November 4, 2010. Accepted October 17, 2011.