Factor automata and special factors

Gabriele Fici

Laboratoire I3S
UMR6070 - CNRS & Université de Nice-Sophia Antipolis, France
fici@i3s.unice.fr

Abstract. The factor automaton of a finite word $w$ is the minimal deterministic automaton recognizing the set of factors of $w$. It is a data structure allowing the search of a pattern in a text in time and space proportional to the length of the pattern. In this paper we link classical combinatorial parameters on words to the size and the structure of the factor automaton. As a main result, we give a characterization of the words having factor automaton with minimal number of states.

1 Introduction

Finite words are sequences of symbols from a finite fixed alphabet. They are used for representing data in several fields of computer science. Indeed, the natural operation on words, concatenation, allows to store an entire data set in a single word, often called the text. For example, the human genome, or an encyclopedia, can be represented as a word over a suitable alphabet.

When a text has been stored, one has often to retrieve local information by looking for special kinds of patterns, for example blocks of consecutive symbols inside the text, called factors.

In this context, an important problem is to construct a data structure for storing a word in order to efficiently find occurrences of factors. For this, several data structures have been introduced over the time. One of these is the factor automaton (or factor DAWG$^1$). The factor automaton of a word $w$ is the minimal deterministic automaton recognizing the language $\text{Fact}(w)$ of the factors of $w$. It allows the search of a factor in $w$ in time and space linear in the length of the searched factor. Moreover, the factor automaton of a word $w$ can be constructed in time and space linear in the length of $w$.

The factor automaton is similar to another well known data structure, the suffix automaton (or suffix DAWG), that is the minimal deterministic automaton recognizing the language $\text{Suff}(w)$ of the suffixes of $w$. In fact, the two data structures are very similar. One can obtain the factor automaton of the word $w$ from the suffix

---


$^1$ DAWG is an acronym for Directed Acyclic Word Graph.
automaton of \( w \) by setting in this latter all the states terminal and then minimizing the resulting automaton. So, in general, the factor automaton of a word has less states than the suffix automaton of the same word. Nevertheless, this difference is negligible in most practical applications.

The structure of the factor and suffix automata of a word \( w \) depends on the combinatorial structure of \( w \), and in particular on its special factors, that are factors appearing in \( w \) within different contexts (the context of a factor is what precedes or follows its occurrence in the word).

We continue the investigation started in a previous paper [6] on the relationships between the structure of suffix and factor automata and the combinatorics of finite words. For the suffix automaton, we gave a formula for the number of states [6]. The situation becomes more difficult for the factor automaton. In Section 3 we discuss upper and lower bounds for the size of the factor automaton of a word over an arbitrary alphabet of size larger than one. However, the existence of a formula for the number of states of the factor automaton remains an open question.

We then focus on the characterization of words \( w \) having the property that their factor automata have exactly \(|w| + 1\) states, that is the minimal possible number of states the factor automaton of a word can have. We denote by \( L_{FA} \) (resp. \( L_{SA} \)) the language of words \( w \) having factor automaton (resp. suffix automaton) with exactly \(|w| + 1\) states. Since the size of the factor automaton of a word is smaller than or equal to the size of the suffix automaton of the same word, we have that the \( L_{SA} \) is contained in \( L_{FA} \), and the inclusion is proper.

The language \( L_{SA} \) coincides with the language of words \( w \) such that every left special factor of \( w \) is a prefix of \( w \) [6, 7]. For binary alphabets, it therefore coincides with the set of finite prefixes of standard sturmian words [7]. In the case of an alphabet larger than two, \( L_{SA} \) contains several known classes of words, e.g. the set of finite prefixes of standard episturmian words [6].

The language \( L_{FA} \) does not seem to have analogous relationships with known classes of words. Actually, words in \( L_{FA} \) are particular extensions of words in \( L_{SA} \). If \( v \) denotes the longest repeated suffix of \( w \), we can write \( v = uv' \), where \( u \) is the longest prefix of \( v \) that is also a prefix of \( w \). We then call \( v' \) the characteristic suffix of \( w \). A word \( w = w'v' \) belongs then to \( L_{FA} \) if and only if its prefix \( w' \) belongs to \( L_{SA} \) (Section 4, Theorem 2).

The paper is organized as follows. In Section 2 we fix notation and recall background on finite words and special factors. In Section 3 we introduce suffix and factor automata and discuss the relationships between their size and classical combinatorial parameters on words. In Section 4 we characterize the language \( L_{FA} \) of words having factor automaton with minimal number of states.

2 Notation and background

An alphabet, denoted by \( \Sigma \), is a finite set of symbols. The size of \( \Sigma \) is denoted by \(|\Sigma|\). A word over \( \Sigma \) is a sequence of symbols from \( \Sigma \). The length of a word \( w \) is
denoted by $|w|$. The set of all words over $\Sigma$ is denoted by $\Sigma^*$. The empty word has length zero and is denoted by $\varepsilon$. The set of all words over $\Sigma$ having length $n \geq 0$ is denoted by $\Sigma^n$. A language over $\Sigma$ is a subset of $\Sigma^*$. For a finite language $L$ we denote by $|L|$ the number of its elements.

Let $w = a_1a_2\ldots a_n$, $n > 0$, be a nonempty word over the alphabet $\Sigma$. Any $i$ such that $1 \leq i \leq n$ is called a position of $w$, and the letter $a_i \in \Sigma$ is called the letter in position $i$.

A prefix of $w$ is any word $v$ such that $v = \varepsilon$ or $v$ is of the form $v = a_1a_2\ldots a_i$, with $1 \leq i \leq n$. A suffix of $w$ is any word $v$ such that $v = \varepsilon$ or $v$ is of the form $v = a_i a_{i+1} \ldots a_n$, with $1 \leq i \leq n$. A factor of $w$ is a prefix of a suffix of $w$ (or, equivalently, a suffix of a prefix of $w$). Therefore, a factor of $w$ is any word $v$ such that $v = \varepsilon$ or $v$ is of the form $v = a_i a_{i+1} \ldots a_j$, with $1 \leq i \leq j \leq n$.

We denote by $\text{Pref}(w)$, $\text{Suff}(w)$ and $\text{Fact}(w)$, respectively, the set of prefixes, suffixes and factors of the word $w$.

A factor $u$ of $w$ is left special in $w$ if there exist two letters $a, b \in \Sigma$, $a \neq b$, such that $au, bu \in \text{Fact}(w)$. A factor $u$ of $w$ is right special in $w$ if there exist $a, b \in \Sigma$, $a \neq b$, such that $ua, ub \in \text{Fact}(w)$. A factor $u$ of $w$ is bispecial in $w$ if it is both left and right special. We denote by $\text{LS}(w)$ (resp. $\text{RS}(w)$, $\text{BS}(w)$) the set of left special (resp. right special, bispecial) factors of the word $w$. We note $S^l(w)$ (resp. $S^r(w)$) the total number of left (resp. right) special factors of $w$.

Remark 1. In the sequel, unless otherwise specified, we assume that the alphabet $\Sigma$ has cardinality greater than one.

We will also use the following parameters:

Definition 1. [5] Let $w$ be a word over $\Sigma$. We note $H_w$ the minimal length of a prefix of $w$ which occurs only once in $w$. We note $K_w$ the minimal length of a suffix of $w$ which occurs only once in $w$.

Definition 2. [5] Let $w$ be a word over $\Sigma$. We note $L_w$ the minimal length for which there are not left special factors of that length in $w$. That is, $L_w = 1 + \max\{|v| : v \in \text{LS}(w)\}$.

Analogously, we note $R_w$ the minimal length for which there are not right special factors of that length in $w$. That is, $R_w = 1 + \max\{|v| : v \in \text{RS}(w)\}$.

Definition 3. [2] Let $w$ be a word over $\Sigma$. We note $P_w$ the minimal length of a prefix of $w$ which is not left special in $w$.

Example 1. Let $w = aababab$. Then $P_w = 2$, $L_w = K_w = 5$ and $H_w = R_w = 2$.

3 Suffix and factor automata

Let $w = a_1a_2\ldots a_n$, $n > 0$, be a nonempty word over the alphabet $\Sigma$. For any $v \in \text{Fact}(w)$ we can define the set of ending positions of $v$ in $w$. It is the set
Endset<sub>w</sub>(v) of the positions of w in which an occurrence of v ends. We assume that Endset<sub>w</sub>(ε) = {0, 1, ..., n}.

Example 2. Let w = aabaab. Then one has Endset<sub>w</sub>(ba) = {4}, Endset<sub>w</sub>(aab) = Endset<sub>w</sub>(ab) = {3, 6}.

In the next proposition we recall some properties of the sets of ending positions (see [4]):

**Proposition 1.** Let u, v ∈ Fact(w). Then one of the three following conditions holds:

1. Endset<sub>w</sub>(v) ⊆ Endset<sub>w</sub>(u);
2. Endset<sub>w</sub>(u) ⊆ Endset<sub>w</sub>(v);
3. Endset<sub>w</sub>(v) ∩ Endset<sub>w</sub>(u) = ∅.

Moreover, if u ∈ Suff(v) then Endset<sub>w</sub>(v) ⊆ Endset<sub>w</sub>(u). If Endset<sub>w</sub>(v) = Endset<sub>w</sub>(u), then v ∈ Suff(u) or u ∈ Suff(v).

One can thus define the following equivalence relation on the set Fact(w):

u ≡<sub>SA</sub> v ⇐⇒ Endset<sub>w</sub>(u) = Endset<sub>w</sub>(v).

The set Fact(w) is then partitioned into a finite number of classes with respect to this equivalence. These classes are called right-equivalence classes.

We note [u]<sub>SA</sub> the right-equivalence class of u. So:

[u]<sub>SA</sub> = {v ∈ Fact(w) : Endset<sub>w</sub>(v) = Endset<sub>w</sub>(u)}.

In the following proposition we gather some useful facts about the right-equivalence classes, that we will use in the sequel.

**Proposition 2.** Let [u]<sub>SA</sub> be a right-equivalence class of factors of the word w. Then:

1. Two distinct elements in [u]<sub>SA</sub> cannot have the same length. If v is the longest element in [u]<sub>SA</sub>, then any other element in [u]<sub>SA</sub> is a proper suffix of v.
2. The class [u]<sub>SA</sub> contains at most one prefix of w; this prefix is the longest element in [u]<sub>SA</sub> and we call [u]<sub>SA</sub> a prefix class.
3. If v ∈ [u]<sub>SA</sub> is a suffix of w, then all the elements in [u]<sub>SA</sub> are suffixes of w. In this case we call [u]<sub>SA</sub> a suffix class.

We now recall the definition and the basic properties of the suffix automaton (for more details see, for instance, [4]).

**Definition 4 ([1, 3]).** The suffix automaton (or suffix DAWG) of a finite word w is the minimal deterministic automaton accepting the language Suff(w). It is denoted by SA(w).
The states of the suffix automaton of $w$ are in fact the right-equivalence classes of factors of the word $w$. For each state $q$ of $SA(w)$, the elements of the class $[u_q]_{SA}$ associated to $q$ are the labeled paths starting at the initial state and ending in $q$. There is an edge from the state $q$ to the state $q'$ of $SA(w)$ labeled by the letter $a \in \Sigma$ if $q'$ is the state associated to the right-equivalence class of $ua$ for any $u$ in the right-equivalence class associated to the state $q$.

An example of suffix automaton is displayed in Figure 1.

![Suffix Automaton](image)

**Fig. 1.** The suffix automaton of the word $w = aabbabb$. Terminal states are double circled.

The size of $SA(w)$, denoted by $|SA(w)|$, is the number of its states. Therefore, $|SA(w)|$ is the number of right-equivalence classes of factors of $w$.

The following bounds on the size of the suffix automaton are well known [4].

**Proposition 3.** Let $w$ be a word over $\Sigma$. If $|w| = 0$ then $|SA(w)| = 1$; if $|w| = 1$ then $|SA(w)| = 2$. If $|w| \geq 2$ then $|w| + 1 \leq |SA(w)| \leq 2|w| - 1$.

**Definition 5.** We denote by $L_{SA}$ the set of words $w$ such that $|SA(w)| = |w| + 1$.

**Example 3.** The word $abc$ is in $L_{SA}$, whereas the word $abcc$ is not.

**Definition 6.** Let $u \in Fact(w)$. The future of $u$ in $w$ is the set

$$\text{Fut}_w(u) = \{ z \in \Sigma^* : uz \in Fact(w) \}.$$ 

The following equivalence, defined on the set $Fact(w)$, is called the Nerode equivalence on $Fact(w)$:

$$u \equiv_{FA} v \iff \text{Fut}_w(u) = \text{Fut}_w(v).$$

It is worth noticing that for any $u, v \in Fact(w)$, if $u \equiv_{SA} v$ then $u \equiv_{FA} v$, but the converse is not always true. As an example, consider the word $w = ababba$. One can check that $\text{Fut}_w(b) = \text{Fut}_w(ab)$. Nevertheless, $5 \in \text{Endset}_w(b) \setminus \text{Endset}_w(ab)$. 


We note \([u]_{FA}\) the Nerode equivalence class of \(u\). So

\[
[u]_{FA} = \{v \in \text{Fact}(w) : \forall z \in \Sigma^*, vz \in \text{Fact}(w) \iff uz \in \text{Fact}(w)\}.
\]

**Lemma 1.** Let \(w\) be a word over \(\Sigma\), and let \(u, v \in \text{Fact}(w)\). If \(u \equiv_{FA} v\), then \(u\) is suffix of \(v\) or \(v\) is suffix of \(u\). In particular, then, two distinct factors of \(w\) of the same length cannot have the same future in \(w\).

**Proof.** Let \(i, j\) be, respectively, the ending positions of the first occurrences of \(u\) and \(v\) in \(w\). Suppose \(i \neq j\). Without loss of generality, we can suppose \(i < j\). But then \(w_{i+1} \cdots w_{|w|}\) cannot belong to the future of \(v\). So \(i = j\), and thus \(u\) is suffix of \(v\) or \(v\) is suffix of \(u\). \(\square\)

We now recall the definition and the basic properties of the factor automaton (for more details see, for instance, [4]).

**Definition 7** ([1, 3]). The factor automaton (or factor DAWG) of a finite word \(w\) is the minimal deterministic automaton that recognizes the languages \(\text{Fact}(w)\) of the factors of \(w\). It is denoted by \(FA(w)\).

One way to construct the factor automaton of the word \(w\) is the following: Build first the suffix automaton \(SA(w)\) of \(w\). Then set all states of \(SA(w)\) terminal and then minimize the resulting automaton.

The states of the factor automaton of \(w\) are in fact the Nerode equivalence classes of factors of the word \(w\). For each state \(q\) of \(FA(w)\), the elements of the class \([u]_{FA}\) associated to \(q\) are the labeled paths starting at the initial state and ending in \(q\). There is an edge from the state \(q\) to the state \(q'\) of \(FA(w)\) labeled by the letter \(a \in \Sigma\) if \(q'\) is the state associated to the Nerode equivalence class of \(ua\) for any \(u\) in the Nerode equivalence class associated to the state \(q\).

![Factor automaton diagram](image)

**Fig. 2.** The factor automaton of the word \(w = aabbabb\). All states are terminal.

An example of factor automaton is displayed in Figure 2.
We denote by \(|FA(w)|\) the number of states of the factor automaton of the word \(w\). Since the factor automaton is obtained from the suffix automaton by minimization, the number of states of the factor automaton of \(w\) is smaller than or equal to the number of states of the suffix automaton of \(w\). The following bounds on the size of the factor automaton are well known [4].

**Proposition 4.** Let \(w\) be a word over \(\Sigma\). If \(|w| \leq 2\), \(|FA(w)| = |w| + 1\). Otherwise \(|w| \geq 3\) and \(|w| + 1 \leq |FA(w)| \leq 2|w| - 2\).

**Definition 8.** We denote by \(L_{FA}\) the set of words \(w\) such that \(|FA(w)| = |w| + 1\).

**Example 4.** The words \(abc\) and \(abcc\) are both in \(L_{FA}\).

Clearly, one has \(L_{SA} \subset L_{FA}\), and the inclusion is proper (as an example, consider the word \(w = abcc\) in Examples 3 and 4).

We now give a formula for the number of states of the suffix automaton of a word \(w\).

**Definition 9.** Let \(w\) be a word over \(\Sigma\). We denote by \(D(w)\) the set of factors \(u\) of \(w\) such that \(u\) is not a prefix of \(w\) and \(u\) is left special in \(w\).

**Proposition 5.** [6] Let \(w\) be a word over \(\Sigma\). Any \(u \in D(w)\) is the longest element in its right-equivalence class \([u]_{SA}\). In particular, then, the elements of \(D(w)\) are each in a distinct right equivalence class.

**Proposition 6.** [6] Let \(w\) be a word over \(\Sigma\) such that \(|w| > 2\) and let \([u]_{SA}\) be a right-equivalence class of factors of \(w\). If \([u]_{SA}\) is not a prefix class, then \([u]_{SA}\) is the class of an element of \(D(w)\). Therefore, the suffix automaton of \(w\) has size:

\[
|SA(w)| = |w| + 1 + |D(w)|
= |w| + 1 + S^l(w) - P_w.
\]

An interesting consequence of Proposition 6 is the following.

**Theorem 1.** [6, 7] Let \(w\) be a word over \(\Sigma\). Then \(w \in L_{SA}\) if and only if every left special factor of \(w\) is a prefix of \(w\).

We now look at the size of the factor automaton. By definition, the number of Nerode equivalence classes of factors of a word \(w\) is smaller than or equal to the number of right-equivalence classes of factors of \(w\), i.e., \(|FA(w)| \leq |SA(w)|\). Therefore, from Proposition 6 we directly obtain

\[
|FA(w)| \leq |w| + 1 + S^l(w) - P_w. \tag{1}
\]
The latter upper bound can be found in a paper of Carpi and de Luca [2], where
the authors also exhibited the following lower bound:

\[
|FA(w)| \geq |w| + 2 + S^l(w) - P_w - K_w
= |SA(w)| - (K_w - 1).
\]

That is, at most \(K_w - 1\) distinct right-equivalence classes of factors are identified
in the Nerode equivalence.

Remark 2. It is possible to prove that the lower bound in (2) can be improved by
replacing the value \(K_w\) by the length of the longest suffix of \(w\) that is left special in
\(w\). However, the existence of a formula for the size of the factor automaton remains
an open question.

4  The language \(L_{FA}\)

In this section we give a characterization of the words \(w\) having factor automata
with exactly \(|w| + 1\) states. Recall that the set of such these words is noted \(L_{FA}\).

Definition 10. [7] Let \(v \in \text{Fact}(w)\). We denote by \(p_w(v)\) the shortest prefix of \(w\)
containing an occurrence of \(v\).

Remark 3. By definition of \(p_w\), it is easy to see that a word \(w\) belongs to \(L_{FA}\) if
and only if every factor \(v\) of \(w\) has the same future in \(w\) as \(p_w(v)\), that is, \(w \in L_{FA}\)
if and only if for every \(v \in \text{Fact}(w)\), one has \(v \equiv_{FA} p_w(v)\).

Lemma 2. Let \(w\) be a word over the alphabet \(\Sigma\). Then \(w \in L_{FA}\) if and only if for
every \(v \in D(w)\), one has \(v \equiv_{FA} p_w(v)\).

Proof. Suppose \(w \in L_{FA}\). Hence, by the previous remark, for every \(v \in \text{Fact}(w)\), one
has \(v \equiv_{FA} p_w(v)\). In particular, then, the same holds for the factors of \(w\) belonging
to \(D(w)\).

Conversely, it is easy to prove (see for instance the proof of Proposition 6 in
[6]) that for every factor \(v \in \text{Fact}(w) \setminus D(w)\), one has \(v \equiv_{SA} p_w(v)\) and hence
\(v \equiv_{FA} p_w(v)\). The statement then follows from the previous remark. \(\Box\)

We now describe some properties of the words in \(L_{FA} \setminus L_{SA}\).

Lemma 3. Let \(w\) be a word over the alphabet \(\Sigma\) such that \(w \in L_{FA} \setminus L_{SA}\). Then,
for every \(n \geq 0\), \(w\) has at most one non-prefix left special factor of length \(n\).

Proof. Suppose by contradiction that there exist two distinct non-prefix left special
factors \(u, v\) such that \(|u| = |v|\). Since \(u\) and \(v\) are left special, \(au, bu, a'v, b'v \in
\text{Fact}(w)\), for some letters \(a, b, a', b' \in \Sigma\) such that \(a \neq b\) and \(a' \neq b'\).
Let \( i_1, i_2, j_1, j_2 \) be, respectively, the ending positions of the first occurrence of \( au, bu, a'v, b'v \) in \( w \).

Without loss of generality, we can suppose that the first occurrence of \( au \) appears in \( w \) before the first occurrence of \( bu \) and that the first occurrence of \( a'v \) appears in \( w \) before the first occurrence of \( b'v \). We can also suppose that the first occurrence of \( au \) appears in \( w \) before the first occurrence of \( a'v \). In other terms, we can suppose that \( i_1 < i_2, j_1 < j_2 \) and that \( i_1 < j_1 \).

By Lemma 2, \( u \) is in the same factor-equivalence class as \( p_w(u) \).

Suppose that \( i_1 < i_2 < j_1 < j_2 \). We would have \( w_{i_2+1} \ldots w_{j_1} \) in the future of \( u \) (because in the future of \( bu \)) but not of \( p_w(u) \), since \( j_1 \) is the ending position of the first occurrence of \( a'v \) in \( w \). Contradiction.

Suppose that \( i_1 < j_1 < i_2 < j_2 \). We would have \( w_{i_2+1} \ldots w_{j_2} \) in the future of \( u \) but not of \( p_w(u) \) for the same argument as above, again a contradiction.

So we must have \( i_1 < j_1 < j_2 < i_2 \). But \( v \) is in the same factor-equivalence class of \( p_w(v) \), by Lemma 2. And we would have \( w_{j_2+1} \ldots w_{i_2} \) in the future of \( v \) but not of \( p_w(v) \), that is, reasoning as above, once again a contradiction. \( \square \)

**Lemma 4.** Let \( w \) be a word over the alphabet \( \Sigma \) such that \( w \in L_{FA} \setminus L_{SA} \). Let \( u \) be a non-prefix left special factor of \( w \) of maximal length, and let \( a \) be the letter preceding the first occurrence of \( u \) in \( w \). Then there exists a letter \( b \neq a \in \Sigma \) such that \( bu \) is suffix of \( w \) and every non-suffix occurrence of \( u \) in \( w \) is preceded by the letter \( a \).

**Proof.** Let \( a \) be the letter preceding the first occurrence of \( u \) in \( w \) and let \( x \) be the letter following the first occurrence of \( u \) in \( w \), i.e., the letter following \( p_w(u) \). Since \( u \) is left special there exists a letter \( b \neq a \) such that \( bu \) appears in \( w \). Suppose by contradiction that \( bu \) appears in \( w \) followed by a letter \( y \). Clearly, \( y \neq x \) by hypothesis on the maximality of \( u \). So \( y \) is in the future of \( u \). By Lemma 2 this implies that \( y \) must be in the future of \( p_w(u) \) too, i.e., \( p_w(u)y \) appears as factor of \( w \), and hence \( auy \) appears as factor of \( w \). Therefore \( uy \) is a non-prefix left special factor of \( w \) longer than \( u \), in contradiction with the maximality of \( u \). \( \square \)

From Lemma 4 and Lemma 3 we obtain:

**Corollary 1.** Let \( w \in L_{FA} \setminus L_{SA} \). Then there exist \( u_1, u_2 \) distinct factors of \( w \) such that \( w = u_1vu_2 \) and \( LS(w) = \text{Pref}(u_1) \cup \text{Pref}(u_2) \). Moreover, one has \( 0 < |u_2| = K_w - 1 \).

**Remark 4.** The previous Corollary also implies that for a word \( w \in L_{FA} \setminus L_{SA} \) one has \( L_w = \max(P_w, K_w) \), and therefore \( K_w \leq L_w \).

**Proposition 7.** Let \( w \in L_{FA} \setminus L_{SA} \). Then the longest repeated prefix of \( w \) is also the longest right special factor of \( w \). In particular, then, \( H_w = R_w \).
Proof. Let $v$ be the longest right special factor of $w$. We prove that $v$ is a prefix of $w$. By contradiction suppose that $v$ is not a prefix of $w$. Let $x$ and $a$ be, respectively, the letters preceding and following the first occurrence of $v$ in $w$. Since $v$ is right special in $w$, there is an occurrence of $vb$, $b \neq a$, in $w$, and any occurrence of $vb$ cannot be preceded by $x$, since otherwise $xv$ would be a right special factor of $w$ longer than $v$. So $xvb$ cannot be a factor of $w$. Thus, $b$ is in the future of $v$ but it cannot be in the future of $p_w(v)$, and, by Lemma 2, this is a contradiction with the fact that $w \in L_{FA}$.

Conversely, let $u$ be the longest repeated prefix of $w$. Then, either $u$ is right special in $w$ or $u$ has exactly two occurrences in $w$, the first as prefix and the second as suffix. Let us prove that the latter situation is impossible. Since $u$ is a suffix of $w$ and $u$ repeats in $w$, then $u$ is a suffix of $u'$, the longest repeated suffix of $w$. By Corollary 1, we know that $u'$ is the longest non-prefix left special factor of $w$. Therefore, $u'$ has at least a second non-prefix occurrence in $w$ and hence $u$ has at least a second non-prefix occurrence in $w$. □

We now characterize the words in $L_{FA}$.

Definition 11. Let $w$ be a word over $\Sigma$. Let $v$ be the longest repeated suffix of $w$, and write $v = uv'$, where $u$ is the longest prefix of $v$ that is also a prefix of $w$. We call $v'$ the characteristic suffix of $w$.

Lemma 5. Let $w$ be a word over $\Sigma$, and let $w = w'v'$, where $v'$ is the characteristic suffix of $w$. Then the longest prefix of $w$ that belongs to $L_{SA}$ is a prefix of $w'$.

Proof. If $v' = \varepsilon$, then $w' = w$ and the claim is trivial. So suppose $v' \neq \varepsilon$. Let $y$ be the first letter of $v'$. By contradiction, suppose that $w'y \in L_{SA}$. Let $v = uv'$ be the longest repeated suffix of $w$ and let $a$ be the letter preceding the occurrence of $v$ as suffix of $w$. Observe that $v$ cannot be a prefix of $w$ since, by definition, $u$ is the longest prefix of $v$ that is also a prefix of $w$, and since $v' \neq \varepsilon$, one has $u \neq v$. Therefore there exists a letter $b \neq a$ such that $bv$ appears in $w$. This implies that $buy$ appears in $w'$. Since $auy$ appears (as suffix) in $w'y$, we have that $uy$ is a left special factor of $w'y$. Since we supposed that $w' \in L_{SA}$, by Theorem 1, we have that $uy$ is prefix of $w'$, and therefore $uy$ is a prefix of $w$, against the hypothesis that $u$ is the longest prefix of $v$ that is also a prefix of $w$. □

Theorem 2. Let $w$ be a word over $\Sigma$, and let $w = w'v'$, where $v'$ is the characteristic suffix of $w$. Then $w \in L_{FA}$ if and only if $w' \in L_{SA}$.

Proof. Suppose first that $w \in L_{SA}$. We have to prove that $w' = w$, i.e., that $v' = \varepsilon$. Let $v = uv'$ be the longest repeated suffix of $w$. If $v = \varepsilon$ the claim trivially holds. So suppose $v$ non-empty and let $a$ be the letter preceding the occurrence of $v$ as suffix of $w$. If, by contradiction, $v' \neq \varepsilon$, then $v$ is not prefix of $w$, and hence $v$ has at least a second non-prefix occurrence in $w$. This second occurrence of $v$ in $w$ must
be preceded by a letter $b \neq a$, since $v$ is the longest repeated suffix of $w$. Thus $v$
would be a non-prefix left special factor of $w$, in contradiction with Theorem 1.

Suppose now $w \in L_{FA} \setminus L_{SA}$. By Corollary 1, the non-prefix left special factors
of $w$ are exactly the prefixes of $v = uv'$ that are longer than $u$. This implies that
the word $w'$ does not contain non-prefix left special factors. Hence, by Theorem 1,
$w' \in L_{SA}$.

So we proved that if $w \in L_{FA}$ then $w' \in L_{SA}$.

Conversely, suppose that $w' \in L_{SA}$. If $v' = \varepsilon$, than $w' = w$ and so $w \in L_{SA} \subset
L_{FA}$. So suppose $v' \neq \varepsilon$ and let $y$ be the first letter of $v'$.

By definition, $v$ is the longest repeated suffix of $w$ and $u$ is the longest (proper)
prefix of $v$ that is also a prefix of $w$. So there exists a letter $x \neq y$ such that $ux$ is
prefix of $w$. Let $a$ be the letter preceding the first occurrence of $v$ in $w$. Then any
non-suffix occurrence of $v$ in $w$ is preceded by $a$, otherwise $uy$ would be left special
in $w'$, a contradiction since $uy$ would be a prefix of $w'$ (by Theorem 1) and so a
prefix of $w$. On the other hand, there exists a letter $b$ such that $bv$ is suffix of $w$,
with $b \neq a$, otherwise $av$ would be a repeated suffix in $w$ longer than $v$.

Let now $t$ be a non-prefix left special factor of $w$. We claim that $t$ is a prefix
of $v$ longer than $u$. Let $a'$ be the letter preceding the first occurrence of $t$ in $w$.
Consider the first occurrence in $w$ of a factor of kind $b't$, for a letter $b' \neq a'$. The
factor $b't$ cannot be a factor of $w'$ since we supposed $w' \in L_{SA}$ and so, by Theorem
1, $t$ would be a prefix of $w'$ and hence of $w$, a contradiction. Let $i$ be the starting
position of the first occurrence of $b't$ in $w$. If $i \geq |w| - |v|$ then $b't$ is a factor of
$v$ and hence $b't$ would appear in $w$ in a position $i' < i$ (since $v$ is by hypothesis a
repeated suffix of $w$), against the definition of $i$. So $i < |w| - |v|$. Let $j$ be the ending
position of the first occurrence of $b't$ in $w$. If $j \leq |w| - |v| + |u|$, then $b't$ would
appear as factor in $w'$ and hence it would be a left special factor of $w'$, that is, as
we saw, a contradiction. So $j > |w| - |v| + |u|$. Now, if $i < |w| - |v| - 1$, then $buy$
is a factor of $t$, and this implies that $buy$ appears as factor in $w'$, a contradiction
since $av$ appears as factor in $w'$ and hence $uy$ would be a left special factor of $w'$
and so, By Theorem 1, a prefix of $w'$.

So we proved that $t$ is a prefix of $v$ longer than $u$.

In order to prove that $w \in L_{FA}$, it is sufficient to show, by Lemma 2, that for
any non-prefix left special factor $t$ of $w$, one has $t \equiv_{FA} p_w(t)$.
Let us write $p_w(t) = p_1p_2\cdots p_n t$. We have two cases:

**Case 1.** Suppose that $t \in \text{Fact}(w')$. We first prove that any occurrence of $t$ as factor of $w'$ is preceded by $p_1p_2\cdots p_n$. By contradiction, suppose that there exists $1 \leq j \leq n$ and a letter $x \neq p_j$ such that $x p_{j+1}\cdots p_n t$ is a factor of $w'$. Then $p_{j+1}\cdots p_n t$ is a left special factor of $w'$. By Theorem 1, this implies that $p_{j+1}\cdots p_n t$ is a prefix of $w'$, and hence of $w$, against the definition of $p_w(t)$.

We now prove that $t$ cannot occur in $w$ in a position starting before $|w| - |v|$ and ending after $|w| - |v'|$. Indeed, this would imply that $vuy$ is a factor of $t$. Since we supposed that $t \in \text{Fact}(w')$, we would have $buy \in \text{Fact}(w')$, that is, as we saw, a contradiction.

This proves that for any $z \in \Sigma^*$ such that $tz \in \text{Fact}(w)$, there is an occurrence of $tz$ in $w$ preceded by $p_1p_2\cdots p_n$. That is, $t$ and $p_w(t)$ have the same future in $w$.

**Case 2.** Suppose that $t \notin \text{Fact}(w')$. In this case the first occurrence of $t$ in $w$ starts in a position smaller than $|w| - |v|$ and ends in a position greater than $|w| - |v'|$. This implies that $buy$ is a factor of $t$. Now observe that the first occurrence of $t$ in $w$ coincides with the first occurrence of $v$ in $w$, since if $t$ had an occurrence in $w$ starting before the first occurrence of $v$, one would have $buy \in \text{Fact}(w')$, that is, as we saw, a contradiction.

Thus, $t$ has exactly two occurrences in $w$: the first preceded by $p_1p_2\cdots p_n$ (by definition of $p_w(t)$), which is also an occurrence of $v$, and the second starting in position $|w| - |v|$. This proves that for any $z \in \Sigma^*$ such that $tz \in \text{Fact}(w)$, there is an occurrence of $tz$ in $w$ preceded by $p_1p_2\cdots p_n$. That is, $t$ and $p_w(t)$ have the same future in $w$. \hfill $\Box$

**Example 5.** Consider the word $w = abcababa \in L_{FA}$. The longest repeated suffix of $w$ is $aba$, and the longest prefix of $aba$ which is also a prefix of $w$ is $ab$. Thus $v' = a$ and $w' = abcabab$. We have $w' \in L_{SA}$, since $LS(w') = Pref(ab) \subseteq Pref(w')$.

**Example 6.** Consider the word $w = abaacaaaa$. The longest repeated suffix of $w$ is $aa$, and the longest prefix of $aa$ which is also a prefix of $w$ is $a$. Then $v' = a$ and $w' = abaacaa$. We have $w \notin L_{FA}$, since $a \in Fut_w(aa) \setminus Fut_w(abaa)$ and $abaa = p_w(aa)$. On the other hand we have $w' \notin L_{SA}$ since $aa$ is a non-prefix left special factor of $w'$.

**Acknowledgements**

The author would like to thank an anonymous referee for her/his careful reading and thoughtful comments.

**References**


