Separating the Words of a Language by Counting Factors

Aleksi Saarela
Department of Mathematics and Statistics
University of Turku, 20014 Turku, Finland
amsaar@utu.fi

Abstract. For a given language L, we study the languages X such that for all distinct words $u, v \in L$, there exists a word $x \in X$ that appears a different number of times as a factor in u and in v. In particular, we are interested in the following question: For which languages L does there exist a finite language X satisfying the above condition? We answer this question for all regular languages and for all sets of factors of infinite words.

Keywords: combinatorics on words, k-abelian equivalence, regular language, infinite word

1. Introduction

The motivation for this article comes from two sources. First, a famous question about finite automata is the *separating words problem*. If $\mathrm{sep}(u, v)$ is the size of the smallest DFA that accepts one of the words u, v and rejects the other, then what is the maximum of the numbers $\mathrm{sep}(u, v)$ when u and v run over all words of length at most n? This question was first studied by Goralčík and Koubek [1], and they proved an upper bound $o(n)$ and a lower bound $\Omega(\log n)$. The upper bound was improved to $O(n^{2/5}(\log n)^{3/5})$ by Robson [2], and this remains the best known result. A survey and some additional results can be found in the article by Demaine, Eisentat, Shallit and Wilson [3]. Several variations of the problem exist. For example, NFAs [3] or context-free grammars [4] could be used instead of DFAs. More generally, we could try to separate two disjoint languages A and B by providing a language X from some specified family of languages such that $A \subseteq X$ and $B \cap X = \varnothing$. As an example related to logic, see [5]. Alternatively, we could try to separate many words $w_1, \ldots, w_k$ by providing languages $X_1, \ldots, X_k$ with some specific properties such that $w_i \in X_j$ if and only if $i = j$. As an example, see [6].

Let $|w|_x$ denote the number of occurrences of a factor x in a word w. A simple observation that can be made about the separating words problem is that if $|u|_x \neq |v|_x$, then $|u|_x \not\equiv |v|_x \pmod{p}$ for some relatively small prime p (more specifically, $p = O(\log(|uv|))$), and the number of occurrences of x modulo p can be easily counted by a DFA with $|x|p$ states. So if u and v have a different number of occurrences of some short factor x, then $\mathrm{sep}(u, v)$ is small, see [3] for more details. Unfortunately, this approach does not provide any general bounds, and more complicated ideas are required to prove the results mentioned in the previous paragraph.

In this article, we are interested in the question of how well words can be separated if we forget about automata and only consider the simple idea of counting occurrences of factors. For any two distinct words u and v of length n, we can find a factor x of length $\lfloor n/2 \rfloor + 1$ or less such that $|u|_x \neq |v|_x$. A proof of this simple fact can be found in an article by Manuch [7]. See [8] for a variation where also the positions of the occurrences modulo a certain number are taken into account. Trying to separate more than two words (possibly infinitely many) at once by counting the numbers of occurrences of more than one factor leads to some interesting questions such as the following one.

Question 1. Given a language L, does there exist a finite language X such that for all distinct words $u, v \in L$, there exists $x \in X$ such that $|u|_x \neq |v|_x$?

The second source of motivation is k-abelian complexity. For a positive integer k, words u and v are said to be k-abelian equivalent if $|u|_x = |v|_x$ for all factors x of length at most k. The factor complexity of an infinite word w is a function that maps a number n to the number of factors of w of length n. The k-abelian complexity of w similarly maps a number n to the number of k-abelian equivalence classes of factors of w of length n. k-abelian equivalence was first studied by Karhumäki [9]. Many basic properties were proved by Karhumäki, Saarela and Zamboni in the article [10], where also k-abelian complexity was introduced. Several articles have been published about k-abelian complexity [11, 12, 13], and about abelian complexity (that is, the case $k = 1$) already earlier [14]. Perhaps the most interesting one from the point of view of this paper is [11], where the relationships between the k-abelian complexities of an infinite word for different values of k were studied. However, the following simple question was not considered in that article.

Question 2. Given an infinite word, does there exist an integer $k \geq 1$ such that the k-abelian complexity of the word is the same as the usual factor complexity of the word?

For a given language, we can define its growth function and k-abelian growth function as concepts analogous to the factor complexity and k-abelian complexity of an infinite word. Then the above question can be generalized. We are specifically interested in the case of regular languages. Some connections between k-abelian equivalence and regular languages have been studied by Cassaigne, Karhumäki, Puzynina and Whiteland [15].

Question 3. Given a language, does there exist an integer $k \geq 1$ such that the growth function of the language is the same as the k-abelian growth function of the language?

In this article, we first define some concepts related to Question 1 and prove basic properties about them. Question 1 and Question 3 are equivalent, but this requires a short proof. We answer

these questions for two families of languages: sets of factors of infinite words (this corresponds to Question 2) and regular languages. In the first case, the answer is positive if and only if the word is ultimately periodic. This result is not very surprising, but it leads to some further questions and results related to aperiodic words. Our main result is a characterization in the case of regular languages: The answer is positive if and only if the language does not have a subset of the form xw^*yw^*z for any words w, x, y, z such that $wy \neq yw$.

This article is an extended journal version of the conference article [16]. The main differences are the following: Theorems 4.3, 4.4, 4.5 and 5.5 are entirely new. The proof of Lemma 5.3 and Lemmas 2.2 and 2.3 that are needed in it were omitted from the conference version, but are now included.

2. Preliminaries

Throughout the article, we use the symbol Σ to denote an alphabet. All words are over Σ unless otherwise specified. The empty word is denoted by ε.

Primitive words and Lyndon words. A nonempty word is *primitive* if it is not a power of any shorter word. The *primitive root* of a nonempty word w is the unique primitive word p such that $w \in p^+$. The primitive root of w is denoted by $\rho(w)$. It is well known that nonempty words u, v have the same primitive root if and only if they commute, that is, $uv = vu$.

Words u and v are *conjugates* if there exist words p, q such that $u = pq$ and $v = qp$. All conjugates of a primitive word are primitive. If two nonempty words are conjugates, then their primitive roots are conjugates. It is well known that if $uw = wv$ and $u \neq \varepsilon$, then there exist words p, q such that $\rho(u) = pq$, $\rho(v) = qp$ and $w \in (pq)^*p$.

We can assume that the alphabet Σ is ordered. This order can be extended to a lexicographic order of Σ^*. A *Lyndon word* is a primitive word that is lexicographically smaller than all of its other conjugates. We use Lyndon words when we need to pick a canonical representative from the conjugacy class of a primitive word. The fact that this representative happens to be lexicographically minimal is not actually important in this article.

The *Lyndon root* of a nonempty word w is the unique Lyndon word that is conjugate to $\rho(w)$. The Lyndon root of w is denoted by $\lambda(w)$. We state here the well known periodicity theorem of Fine and Wilf [17], and we use it to prove a simple result about Lyndon roots.

Theorem 2.1. (Fine and Wilf)
Let u, v be nonempty words. If the infinite words u^ω and v^ω have a common prefix of length $|uv| - \gcd(|u|, |v|)$, then u and v are powers of a common word of length $\gcd(|u|, |v|)$.

Lemma 2.2. Let u, v be nonempty words. If u^m and v^n have a common factor of length $|uv|$, then $\lambda(u) = \lambda(v)$.

Proof:
A factor of u^m of length $|uv|$ is of the form $(u_1)^i u_2$, where u_1 is a conjugate of u, u_2 is a prefix of

u_1, and $i \geq 1$. Similarly, a factor of v^n of length $|uv|$ is of the form $(v_1)^j v_2$, where v_1 is a conjugate of v, v_2 is a prefix of v_1, and $j \geq 1$. If these factors are the same, then $(u_1)^i u_2 = (v_1)^j v_2$, so u_1^ω and v_1^ω have a common prefix of length $|uv|$. It follows from Theorem 2.1 that u_1 and v_1 are powers of a common word and therefore have the same primitive root. This primitive root is conjugate to $\rho(u)$ and to $\rho(v)$, so $\lambda(u) = \lambda(v)$. □

Occurrences. Let u and w be words. An *occurrence of u in w* is a triple (x, u, y) such that $w = xuy$. The number of occurrences of u in w is denoted by $|w|_u$. We allow the case $u = \varepsilon$, and then $|w|_\varepsilon = |w| + 1$.

Let (x, u, y) and (x', u', y') be occurrences in w. If

$$\max(|x|, |x'|) < \min(|xu|, |x'u'|),$$

then we say that these occurrences have an *overlap* of length

$$\min(|xu|, |x'u'|) - \max(|x|, |x'|).$$

If $|x| \geq |x'|$ and $|y| \geq |y'|$, then we say that (x, u, y) is *contained* in (x', u', y').

If (x, u, y) is an occurrence in w and $u \in L$, then (x, u, y) is an *L-occurrence in w*. It is a *maximal L-occurrence in w* if it is not contained in any other L-occurrence in w.

It is well known that if p is a primitive word, then p cannot be a factor of p^2 in a nontrivial way, or more formally, p^2 does not have any other p-occurrences than the trivial ones (ε, p, p) and (p, p, ε). Thus the only p-occurrences in p^n are (p^i, p, p^{n-1-i}) for $i \in \{0, \ldots, n-1\}$, and if pwp is a factor of p^n, then w is a power of p. We can prove the following lemma.

Lemma 2.3. Let w be a word and p be a primitive word. If two p^+-occurrences in w have an overlap of length at least $|p|$, then they are contained in the same maximal p^+-occurrence. Moreover, every p^+-occurrence in w is contained in exactly one maximal p^+-occurrence.

Proof:
To prove the first claim, let (x, p^m, y) and (x', p^n, y') be two p^+-occurrences in w and let $|x| \leq |x'|$. If these occurrences have an overlap of length at least $|p|$, then the occurrence $(x', p, p^{n-1}y')$ is contained in (x, p^m, y). The number $|x'| - |x|$ must be divisible by $|p|$, because otherwise p would be a factor of p^2 in a nontrivial way. Let $|x'| - |x| = kp$. Then (x, p^{k+n}, y') is an occurrence in w. If $|y| \geq |y'|$, then both (x, p^m, y) and (x', p^n, y') are contained in (x, p^{k+n}, y'), which is contained in some maximal occurrence. On the other hand, if $|y| < |y'|$, then (x', p^n, y') is contained in (x, p^m, y), which is contained in some maximal occurrence. This proves the first claim.

If a p^+-occurrence in w is contained in two maximal p^+-occurrences, then those two maximal occurrences are contained in the same maximal occurrence by the first part of the proof. By the definition of maximality, these maximal occurrences are actually the same. This proves the second claim. □

k-abelian equivalence. Let k be a positive integer. Words $u, v \in \Sigma^*$ are *k-abelian equivalent*, denoted by $u \equiv_k v$, if $|u|_x = |v|_x$ for all $x \in \Sigma^{\leq k}$. Clearly, k-abelian equivalence is an equivalence relation. By Theorem 2.4, it is also a congruence, that is, if $u \equiv_k u'$ and $v \equiv_k v'$, then $uv \equiv_k u'v'$. A proof of the next theorem can be found in [10].

Theorem 2.4. Let u, v be words and let $k \geq 1$. If $|u|, |v| \leq 2k-1$, then $u \equiv_k v$ if and only if $u = v$. If $|u|, |v| \geq k-1$, then the following are equivalent:

1. $u \equiv_k v$.
2. $|u|_x = |v|_x$ for all $x \in \Sigma^k$ and u, v have a common prefix of length $k-1$.
3. $|u|_x = |v|_x$ for all $x \in \Sigma^k$ and u, v have a common suffix of length $k-1$.
4. $|u|_x = |v|_x$ for all $x \in \Sigma^k \cup \Sigma^{k-1}$.

We are going to use the following simple lemma a couple of times when showing that two words are k-abelian equivalent.

Lemma 2.5. If $u, v, w, x \in \Sigma^*$, $|v| = k-1$, and $|x| = k$, then $|uvw|_x = |uv|_x + |vw|_x$.

Example 2.6. The words $aabab$ and $abaab$ are 2-abelian equivalent: They have the same prefix of length one, one occurrence of aa, two occurrences of ab, one occurrence of ba, and no occurrences of bb.

The words aba and bab have the same number of occurrences of every factor of length two, but they are not 2-abelian equivalent, because they have a different number of occurrences of a.

Let $k \geq 1$. The words $u = a^k b a^{k-1}$ and $v = a^{k-1} b a^k$ are k-abelian equivalent: They have the same prefix of length $k-1$, and $|u|_x = 1 = |v|_x$ if $x = a^k$ or $x = a^i b a^{k-i-1}$ for some $i \in \{0, \ldots, k-1\}$, and $|u|_x = 0 = |v|_x$ for all other factors x of length k. On the other hand, u and v are not $(k+1)$-abelian equivalent, because they have a different prefix of length k.

Growth functions and factor complexity. The *growth function* of a language L is the function

$$\mathcal{P}_L : \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0},\ \mathcal{P}_L(n) = |L \cap \Sigma^n|$$

mapping a number n to the number of words of length n in L. The *cumulative growth function* of L is the function

$$\overline{\mathcal{P}}_L : \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0},\ \overline{\mathcal{P}}_L(n) = |L \cap \Sigma^{\leq n}| = \sum_{i=0}^{n} \mathcal{P}_L(i)$$

For an infinite word w, let $\mathrm{Fact}(w)$ be the set of factors of w and let $\mathrm{Fact}_n(w) = \mathrm{Fact}(w) \cap \Sigma^n$ be the set of length-n factors of w. The *factor complexity* of an infinite word w, denoted by $\mathcal{P}_w$, is the growth function of $\mathrm{Fact}(w)$. In other words, $\mathcal{P}_w(n) = |\mathrm{Fact}_n(w)|$ for all n.

We can also define k-abelian versions of these functions. The *k-abelian growth function* of a language L is the function

$$\mathcal{P}_L^k : \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0},\ \mathcal{P}_L(n) = |(L \cap \Sigma^n)/\equiv_k|,$$

where $(L \cap \Sigma^n)/\equiv_k$ denotes the set of equivalence classes of elements of $L \cap \Sigma^n$. The *k-abelian complexity* of an infinite word w, denoted by $\mathcal{P}_w^k$, is the k-abelian growth function of the set of factors of w.

An infinite word w is *ultimately periodic* if there exist finite words u, v such that $w = uv^\omega$. An infinite word is *aperiodic* if it is not ultimately periodic. It was proved by Morse and Hedlund [18] that if w is ultimately periodic, then $\mathcal{P}_w(n) = O(1)$, and if w is aperiodic, then $\mathcal{P}_w(n) \geq n+1$ for all n. An infinite word w is called *Sturmian* if $\mathcal{P}_w(n) = n+1$ for all n.

3. Separating sets of factors

A language X is a *separating set of factors* (SSF) of a language L if for all distinct words $u, v \in L$, there exists $x \in X$ such that $|u|_x \neq |v|_x$. The set X is *size-minimal* if no set of smaller cardinality is an SSF of L, and it is *inclusion-minimal* if X does not have a proper subset that is an SSF of L.

Example 3.1. Let $\Sigma = \{a, b\}$. The language a^* has two inclusion-minimal SSFs: $\{\varepsilon\}$ and $\{a\}$. Both of them are also size-minimal. The language $\Sigma^2 = \{aa, ab, ba, bb\}$ has eight inclusion-minimal SSFs:

$$\{a, ab\}, \{a, ba\}, \{b, ab\}, \{b, ba\}, \{aa, ab, ba\}, \{aa, ab, bb\}, \{aa, ba, bb\}, \{ab, ba, bb\}.$$

The first four are size-minimal.

The following lemma contains some very basic results related to the above defitions. In particular, it proves that every language has an inclusion-minimal SSF, and all SSFs are completely characterized by the inclusion-minimal ones.

Lemma 3.2. Let L and X be languages.

1. If $L \neq \varnothing$, then L has a proper subset that is an SSF of L.
2. If X is an SSF of L and $K \subseteq L$, then X is an SSF of K.
3. If X is an SSF of L and $X \subseteq Y$, then Y is an SSF of L.
4. If X is an SSF of L, then X has a subset that is an inclusion-minimal SSF of L.

Proof:
To prove the first claim, let $w \in L$ be of minimal length and let $X = L \setminus \{w\}$. Let $u, v \in L$ and $u \neq v$. By symmetry, we can assume that $|u| \leq |v|$ and $v \neq w$. Then $v \in X$ and $|u|_v = 0 \neq 1 = |v|_v$. This shows that X is an SSF of L.

The second and third claims follow directly from the definition of an SSF.

The fourth claim is easy to prove if X is finite. In the general case, it can be proved by Zorn's lemma as follows. Consider the partially ordered (by inclusion) family of all subsets of X that are SSFs of L. The family contains at least X, so it is nonempty. By Zorn's lemma, if every nonempty chain (that is, a totally ordered subset of the family) C has a lower bound in this family, then the family has a minimal element, which is then an inclusion-minimal SSF of L. We show that the intersection

I of the sets in C is an SSF of L, and therefore it is the required lower bound. For any $u, v \in L$ such that $u \neq v$ and for any $Y \in C$, there exists $y \in Y$ such that $|u|_y \neq |v|_y$. Then y must be a factor of u or v, so if u and v are fixed, then there are only finitely many possibilities for y. Thus at least one of the words y is in all sets Y and therefore also in I. This shows that I is an SSF of L. This completes the proof. □

The next lemma shows a connection between SSFs and k-abelian equivalence.

Lemma 3.3. Let L be a language.

1. Let $k \in \mathbb{Z}_+$. The language $\Sigma^{\leq k}$ is an SSF of L if and only if the words in L are pairwise k-abelian nonequivalent.

2. The language L has a finite SSF if and only if there exists a number k such that the words in L are pairwise k-abelian nonequivalent.

Proof:
The first claim follows directly from the definitions of an SSF and k-abelian equivalence. The "only if" and "if" directions of the second claim can be proved as follows: If a finite set X is an SSF of L, then $X \subseteq \Sigma^{\leq k}$ for some k, and then the words in L are pairwise k-abelian nonequivalent. Conversely, if the words in L are pairwise k-abelian nonequivalent, then $\Sigma^{\leq k}$ is an SSF of L. □

Note that the condition "the words in L are pairwise k-abelian nonequivalent" can be equivalently expressed as "$\mathcal{P}_L = \mathcal{P}_L^k$". This means that Lemma 3.3 implies the equivalence of Questions 1 and 3.

Example 3.4. Let $w, x, y, z \in \{a, b\}^*$ and $L = \{awa, axb, bya, bzb\}$. No two words in L have both a common prefix and a common suffix of length one, so the words are pairwise 2-abelian nonequivalent. By the first claim of Lemma 3.3, $\{a, b\}^{\leq 2}$ is an SSF of L. This SSF is not size-minimal (by the first claim of Lemma 3.2, L has an SSF of size three), but it has the advantage of consisting of very short words and not depending on w, x, y, z. Actually, also $\{\varepsilon, a, aa, ab, ba\}$ is an SSF of L. This follows from the fact that $|u|_b = |u|_\varepsilon - |u|_a - 1$ and $|u|_{bb} = |u|_\varepsilon - |u|_{aa} - |u|_{ab} - |u|_{ba} - 2$ for all $u \in \{a, b\}^*$.

Example 3.5. In a list of about 140 000 English words (found in the SCOWL database[1]), there are no 4-abelian equivalent words. Therefore, by Lemma 3.3, $\Sigma^{\leq 4}$ is an SSF of the language formed by these words (the alphabet Σ here contains the 26 letters from a to z and also many accented letters and other symbols). The only pairs of 3-abelian equivalent words are reregister, registerer and reregisters, registerers. The number of other pairs of 2-abelian equivalent words is also small enough

[1] http://wordlist.aspell.net/

that they can be listed here:

indenter, intender	indenters, intenders
pathophysiologic, physiopathologic	pathophysiological, physiopathological
pathophysiology, physiopathology	pathophysiologies, physiopathologies
tamara, tarama	tamaras, taramas
tantarara, tarantara	tantararas, tarantaras
tantaras, tarantas	

This means that most words of length 4 and 3 are not needed in the SSF. For example, the set $\Sigma^{\leq 2} \cup \{\mathsf{rere}, \mathsf{hop}, \mathsf{ind}, \mathsf{tan}, \mathsf{tar}\}$ is an SSF of the language. We did not try to find a minimal SSF.

In the next lemma, we consider whether the properties of having or not having a finite SSF are preserved under the rational operations union, concatenation and Kleene star.

Lemma 3.6. Let K and L be languages.

1. If L has a finite SSF and F is a finite language, then $L \cup F$ has a finite SSF.
2. If L does not have a finite SSF, then $L \cup K$ does not have a finite SSF.
3. If L has a finite SSF and w is a word, then wL and Lw have finite SSFs.
4. If L does not have a finite SSF and $K \neq \varnothing$, then neither KL nor LK have finite SSFs.
5. L^* has a finite SSF if and only if there exists a word w such that $L \subseteq w^*$.
6. If the symmetric difference of K and L is finite, then either both or neither have a finite SSF.

Proof:

1. Let X be a finite SSF of L. Let $u, v \in L \cup F$ and $u \neq v$. First, if $u, v \in L$, then $|u|_x \neq |v|_x$ for some $x \in X$. Second, if $u \in F$ and $|u| = |v|$, then $|u|_u \neq |v|_u$. Finally, if $|u| \neq |v|$, then $|u|_\varepsilon \neq |v|_\varepsilon$. Thus $X \cup F \cup \{\varepsilon\}$ is an SSF of $L \cup F$.
2. If a finite set is an SSF of $L \cup K$, then it is also an SSF of L.
3. Let wL have no finite SSF. Let $k \in \mathbb{Z}_+$ and $k' = k + |w|$. By Lemma 3.3, there exist two k'-abelian equivalent words $wu, wv \in wL$. Then u and v have a common prefix p of length $k-1$. For all $x \in \Sigma^k$, with the help of Lemma 2.5, we get

$$|u|_x = |wu|_x - |wp|_x = |wv|_x - |wp|_x = |v|_x,$$

so $u \equiv_k v$. We have shown that for all $k \geq 1$, there exist two k-abelian equivalent words in L. By Lemma 3.3, L does not have a finite SSF. The case of Lw is symmetric.

4. Let L have no finite SSF and let $w \in K$. Let $k \in \mathbb{Z}_+$. By Lemma 3.3, there exist two k-abelian equivalent words $u, v \in L$, and then $wu, wv \in KL$ are k-abelian equivalent. We have shown that for all $k \geq 1$, there exist two k-abelian equivalent words in KL. By Lemma 3.3, KL does not have a finite SSF. The case of LK is symmetric.

5. If $L \subseteq w^*$, then $\{w\}$ is an SSF of L. If there does not exist w such that $L \subseteq w^*$, then there exist $u, v \in L$ such that $uv \neq vu$. For all $k \in \mathbb{Z}_+$, the words $u^k v u^{k-1}, u^{k-1} v u^k \in L^*$ are distinct. They have the same prefix of length $k-1$. If u_1 is the prefix and u_2 is the suffix of u^{k-1} of length $k-1$, then, with the help of Lemma 2.5, we get

$$|u^k v u^{k-1}|_x = |u^k|_x + |u_2 v u_1|_x + |u^{k-1}|_x = |u^{k-1}|_x + |u_2 v u_1|_x + |u^k|_x = |u^{k-1} v u^k|_x$$

for all $x \in \Sigma^k$, so $u^k v u^{k-1} \equiv_k u^{k-1} v u^k$. We have shown that for all $k \geq 1$, there exist two k-abelian equivalent words in L^*. By Lemma 3.3, L^* does not have a finite SSF.

6. If K has a finite SSF, then so does $K \cap L$. If $L \setminus K$ is finite, then also L has a finite SSF by the first claim of this lemma. Similarly, if L has a finite SSF and $K \setminus L$ is finite, then also K has a finite SSF.

□

Example 3.7. We give an example showing that the property of having a finite SSF is not always preserved by union and concatenation. Let $L = \{a^k b a^{k-1} \mid k \in \mathbb{Z}_+\}$. Then both L and Laa have the finite SSF $\{\varepsilon\}$. On the other hand, $L\{\varepsilon, aa\} = L \cup Laa$ contains the k-abelian equivalent words $a^k b a^{k-1}$ and $a^{k-1} b a^k$ for all $k \geq 2$, so by Lemma 3.3, $L \cup Laa$ does not have a finite SSF even though both L and Laa do have a finite SSF, and $L\{\varepsilon, aa\}$ does not have a finite SSF even though both L and $\{\varepsilon, aa\}$ do have a finite SSF.

4. Infinite words

In this section, we give an answer to Question 2.

Theorem 4.1. Let w be an infinite word. There exists $k \in \mathbb{Z}_+$ such that $\mathcal{P}_w = \mathcal{P}_w^k$ if and only if w is ultimately periodic.

Proof:
First, let w be ultimately periodic. Then we can write $w = uv^\omega$, where v is primitive and is not a suffix of u. Let $k = |uv| + 1$ and let x, y be k-abelian equivalent factors of w. If x and y are shorter than uv, then $x = y$. Otherwise x and y have a common prefix of length $k - 1 = |uv|$ and we can write $x = u'v'x'$ and $y = u'v'y'$, where $|u'| = |u|$ and $|v'| = |v|$. Here v' is a factor of v^ω, so it must be a conjugate of v, and it is followed by $(v')^\omega$. Thus x' and y' are prefixes of $(v')^\omega$ and they are of the same length, so $x' = y'$ and thus $x = y$. We have proved that no two factors of w are k-abelian equivalent. It follows that $\mathcal{P}_w = \mathcal{P}_w^k$.

Second, let w be aperiodic and let $k \geq 2$ be arbitrary. Let $n = \mathcal{P}_w(k-1) + 1$. There must exist a word u of length $(k-1)n$ that occurs infinitely many times in w as a factor. We can write $u =$

$x_1 \cdots x_n$, where $x_1, \ldots, x_n \in \Sigma^{k-1}$. By the definition of n, there exist two indices $i, j \in \{1, \ldots, n\}$ such that $x_i = x_j$. Let $i < j$, $x = x_i = x_j$ and $y = x_{i+1} \cdots x_{j-1}$. Then xyx is a factor of u and thus occurs infinitely many times in w as a factor. Therefore we can write $w = z_0 xyx z_1 xyx z_2 xyx \cdots$ for some infinite sequence of words $z_0, z_1, z_2, \ldots$. If the words xy and xz_h have the same primitive root p for all $h \in \mathbb{Z}_+$, then $w = z_0 p^\omega$, which contradicts the aperiodicity of w. Thus there exists h such that $\rho(xy) \neq \rho(xz_h)$. Then $xyxz_h \neq xz_hxy$ and thus $xyxz_hx \neq xz_hxyx$. On the other hand, $xyxz_hx$ and xz_hxyx are k-abelian equivalent because they have the same prefix x of length $k-1$ and, by Lemma 2.5,

$$|xyxz_hx|_t = |xyx|_t + |xz_hx|_t = |xz_hx|_t + |xyx|_t = |xz_hxyx|_t$$

for all $t \in \Sigma^k$. Moreover, $xyxz_hx$ and xz_hxyx are factors of w. It follows that $\mathcal{P}_w \neq \mathcal{P}_w^k$. □

Corollary 4.2. The set of factors of an infinite word w has a finite SSF if and only if w is ultimately periodic.

Proof:
Follows from Theorem 4.1 and Lemma 3.3. □

Now we know that if w is aperiodic, then $\mathrm{Fact}(w)$ does not have a finite SSF. We could try to find an SSF that is infinite but small in some sense. A natural way to measure the smallness of an infinite language would be to use the cumulative growth function. Trivially, $\mathrm{Fact}(w)$ has an SSF L such that $\overline{\mathcal{P}}_L(n) \leq n\mathcal{P}_w(n)$ for all n, namely, the set $\mathrm{Fact}(w) \setminus \{\varepsilon\}$. In the next two theorems, we prove that the linear factor n can be replaced by a logarithmic one, and for many families of infinite words, it can be replaced by a constant. The function $\log$ means the binary logarithm.

Theorem 4.3. Let w be an aperiodic infinite word. Then $\mathrm{Fact}(w)$ has an SSF L such that

$$\overline{\mathcal{P}}_L(n) = \sum_{i=1}^{\lfloor \log(n+1) \rfloor} \mathcal{P}_w(2^i - 1) + \sum_{i=1}^{\lfloor \log n \rfloor} \mathcal{P}_w(2^i) \tag{1}$$

$$\leq 2 \log(n+1) \cdot \mathcal{P}_w(n) \tag{2}$$

for all $n \geq 1$.

Proof:
Let

$$L = \bigcup_{i=1}^{\infty} \left(\mathrm{Fact}_{2^i-1}(w) \cup \mathrm{Fact}_{2^i}(w) \right).$$

Then

$$L \cap \Sigma^{\leq n} = \bigcup_{i=1}^{\lfloor \log(n+1) \rfloor} \mathrm{Fact}_{2^i-1}(w) \cup \bigcup_{i=1}^{\lfloor \log n \rfloor} \mathrm{Fact}_{2^i}(w),$$

so (1) is true. The inequality (2) follows from $\mathcal{P}_w$ being strictly increasing. It remains to be shown that L is an SSF of $\mathrm{Fact}(w)$. Let $u, v \in \mathrm{Fact}(w)$ be distinct words. If $|u| \neq |v|$ or $|u| = |v| \leq 1$,

then $|u|_a \neq |v|_a$ for some letter $a \in L$. If $|u| = |v| \geq 2$, let $k = 2^{\lfloor \log(|u|) \rfloor}$, so $k \leq |u| = |v| < 2k$. By Theorem 2.4, u and v are not k-abelian equivalent. Also by Theorem 2.4, there exists a word $x \in \Sigma^k \cup \Sigma^{k-1}$ such that $|u|_x \neq |v|_x$. Then x is a factor of at least one of u and v, and thus $x \in \mathrm{Fact}(w)$. From $|x| \in \{2^{\lfloor \log(|u|) \rfloor}, 2^{\lfloor \log(|u|) \rfloor} - 1\}$ it follows that $x \in L$. This shows that L is an SSF of $\mathrm{Fact}(w)$. □

Theorem 4.4. Let $f : \mathbb{Z}_+ \to \mathbb{R}_+$ be an increasing function and w an infinite aperiodic word such that $\mathcal{P}_w(n) \leq f(n)$ for all $n \geq 1$. If there exists a constant $C > 1$ such that $f(2n) \geq Cf(n)$ for all $n \geq 1$, then $\mathrm{Fact}(w)$ has an SSF L such that

$$\overline{\mathcal{P}}_L(n) \leq \frac{3C-1}{C-1} \cdot f(n)$$

for all $n \geq 1$.

Proof:
Let $m = \lfloor \log n \rfloor$. From $f(2k) \geq Cf(k)$ for all k it follows that $f(2^m) \geq C^{m-i} f(2^i)$ for all $i \leq m$. By Theorem 4.3, $\mathrm{Fact}(w)$ has an SSF L such that (1) holds, and then

$$\begin{aligned}
\overline{\mathcal{P}}_L(n) &\leq \mathcal{P}_w(n) + \sum_{i=1}^{m} \mathcal{P}_w(2^i - 1) + \sum_{i=1}^{m} \mathcal{P}_w(2^i) \\
&\leq f(n) + 2\sum_{i=1}^{m} f(2^i) \\
&\leq f(n) + 2\sum_{i=1}^{m} C^{i-m} f(2^m) \\
&\leq f(n) + 2f(n) \sum_{i=0}^{\infty} C^{-i} = \frac{3C-1}{C-1} \cdot f(n).
\end{aligned}$$

The claim follows. □

For Sturmian words, we get the following result.

Theorem 4.5. Let w be a Sturmian word. Then $\mathrm{Fact}(w)$ has an SSF L such that

$$\overline{\mathcal{P}}_L(n) \leq 4n + \log n - 2$$

for all $n \geq 1$.

Proof:
By Theorem 4.3, $\mathrm{Fact}(w)$ has an SSF L such that (1) holds, and for a Sturmian word w, $\mathcal{P}_w(m) =$

$m+1$ for all m, so

$$\begin{aligned}\overline{\mathcal{P}}_L(n) &= \sum_{i=1}^{\lfloor \log(n+1) \rfloor} \mathcal{P}_w(2^i-1) + \sum_{i=1}^{\lfloor \log n \rfloor} \mathcal{P}_w(2^i) \\ &= \sum_{i=1}^{\lfloor \log(n+1) \rfloor} 2^i + \sum_{i=1}^{\lfloor \log n \rfloor} (2^i+1) \\ &= 2^{\lfloor \log(n+1) \rfloor + 1} - 2 + 2^{\lfloor \log n \rfloor + 1} - 2 + \lfloor \log n \rfloor \\ &\leq 2(n+1) - 2 + 2n - 2 + \log n = 4n + \log n - 2.\end{aligned}$$

The claim follows. □

In the previous theorems, we have proved upper bounds for $\overline{\mathcal{P}}_L(n)$. This leads to several questions: Can we significantly improve these bounds? Can we prove good lower bounds? Can we prove good bounds for $\mathcal{P}_L(n)$? For which infinite words can we find an SSF L such that $\mathcal{P}_L(n) = O(1)$?

5. Regular languages

In this section, we give an answer to Question 1 for regular languages.

Lemma 5.1. If a language L has a subset of the form xw^*yw^*z for some words w, x, y, z such that $wy \neq yw$, then L does not have a finite SSF.

Proof:
For all $k \in \mathbb{Z}_+$, the words $xw^kyw^{k-1}z$ and $xw^{k-1}yw^kz$ are distinct. They have the same prefix of length $k-1$. If w_1 is the prefix and w_2 is the suffix of w^{k-1} of length $k-1$, then, by Lemma 2.5,

$$|xw^kyw^{k-1}z|_t = |xw_1|_t + |w^k|_t + |w_2yw_1|_t + |w^{k-1}|_t + |w_2z|_t = |xw^{k-1}yw^kz|_t$$

for all $t \in \Sigma^k$, so $xw^kyw^{k-1}z \equiv_k xw^{k-1}yw^kz$. We have shown that for all $k \geq 1$, there exist two k-abelian equivalent words in L. By Lemma 3.3, L does not have a finite SSF. □

A language L is *bounded* if it is a subset of a language of the form

$$v_1^* \cdots v_n^*,$$

where $v_1, \ldots, v_n$ are words. It was proved by Ginsburg and Spanier [19] that a regular language is bounded if and only if it is a finite union of languages of the form

$$u_0v_1^*u_1 \cdots v_n^*u_n, \tag{3}$$

where $u_0, \ldots, u_n$ are words and $v_1, \ldots, v_n$ are nonempty words.

Lemma 5.2. Every regular language is bounded or has a subset of the form xw^*yw^*z for some words w, x, y, z such that $wy \neq yw$.

Proof:
The proof is by induction. Every finite language is bounded. We assume that A and B are regular languages that have the claimed property and prove that also $A \cup B$, AB and A^* have the claimed property.

First, we consider $A \cup B$. If both A and B are bounded, then so is $A \cup B$ by the characterization of Ginsburg and Spanier. If at least one of A and B has a subset of the form xw^*yw^*z for some words w, x, y, z such that $wy \neq yw$, then $A \cup B$ has this same subset.

Next, we consider AB. If both A and B are bounded, or if one of them is not bounded but the other one is empty, then AB is bounded by the definition of bounded languages. If A and B are nonempty and at least one of them has a subset of the form xw^*yw^*z for some words w, x, y, z such that $wy \neq yw$, then AB has a subset of the same form with a different x or z.

Finally, we consider A^*. If $A \subseteq u^*$ for some word u, then $A^* \subseteq u^*$ is bounded. If A is not a subset of u^* for any word u, then there exist $w, y \in A$ such that $wy \neq yw$, and A^* has w^*yw^* as a subset. □

By Lemmas 5.1 and 5.2, if a regular language is not bounded, then it does not have a finite SSF. Thus we can concentrate on bounded regular languages. We continue with a technical lemma.

Lemma 5.3. Let L be a bounded regular language. There exist integers $n, k \geq 0$ and a finite set of Lyndon words P such that the following are satisfied:

1. If $p, q \in P$, $p \neq q$, and $l, m \geq 0$, then p^l and q^m do not have a common factor of length n.
2. If $u \in L$ and $p \in P$, then either there is at most one maximal $p^{\geq n}$-occurrence in u or L has a subset of the form $x(p^m)^*y(p^m)^*z$, where $py \neq yp$ and $m \geq 1$.
3. If $u \in L$ and x is a factor of u of length at least k, then x has a factor p^{n+1} for some $p \in P$.

Proof:
If L is finite, then the lemma is basically trivial. For example, we can let $P = \varnothing$ and $n = k = \max\{|w| \mid w \in L\} + 1$. If L is infinite, then, by the characterization of Ginsburg and Spanier, we can write

$$L = \bigcup_{i=1}^{s} u_{i0} \prod_{j=1}^{r_i} v_{ij}^* u_{ij},$$

where $s \geq 1$ and $r_1, \ldots, r_s \geq 0$ are numbers, $r_i \geq 1$ for at least one i, all the u_{ij} are words, and all the v_{ij} are nonempty words. We are going to prove that the three conditions are satisfied for P being the set of Lyndon roots of the words v_{ij} and

$$n = 2 \cdot \max\Big\{ |u_{i0} \prod_{j=1}^{r_i} v_{ij} u_{ij}| \;\Big|\; i \in \{1, \ldots, s\} \Big\},$$

$$k = \max\Big\{ |u_{i0} \prod_{j=1}^{r_i} v_{ij}^{n+2} u_{ij}| \;\Big|\; i \in \{1, \ldots, s\} \Big\}.$$

First, we prove Condition 1. If $p, q \in P$, $l, m \geq 0$, and p^l and q^m have a common factor of length $|pq|$, then $p = q$ by Lemma 2.2. Clearly $n \geq |pq|$, so Condition 1 is satisfied.

Next, we prove Condition 2. This is the most complicated part of the proof. Let $u \in L$ and $p \in P$. There are numbers $i, m_1, \ldots, m_{r_i}$ such that

$$u = u_{i0} \prod_{j=1}^{r_i} v_{ij}^{m_j} u_{ij}.$$

Let (w_1, p^N, w_2) be a maximal $p^{\geq n}$-occurrence in u. If there does not exist an index J such that (w_1, p^N, w_2) and the occurrence

$$\Big(u_{i0} \prod_{j=1}^{J-1} v_{ij}^{m_j} u_{ij},\ v_{iJ}^{m_J},\ u_{iJ} \prod_{j=J+1}^{r_i} v_{ij}^{m_j} u_{ij}\Big) \tag{4}$$

have an overlap of length at least $|pv_{iJ}|$, then

$$|p^N| < \sum_{j=0}^{r_i} |u_{ij}| + \sum_{j=1}^{r_i} |pv_{ij}| \leq \frac{n}{2} + r_i|p| \leq \frac{n}{2} + \frac{n}{2} \cdot |p| \leq n|p|,$$

which is a contradiction. So there exists a number J such that (w_1, p^N, w_2) and (4) have an overlap of length at least $|pv_{iJ}|$, and then $p = \lambda(v_{iJ})$ by Lemma 2.2. We can write $v_{iJ}^{m_J} = p_1 p^M p_2$, where p_1 is a proper suffix of p, p_2 is a proper prefix of p, and $M \geq 1$. Then the occurrences (w_1, p^N, w_2) and

$$\Big(u_{i0}\Big(\prod_{j=1}^{J-1} v_{ij}^{m_j} u_{ij}\Big)p_1,\ p^M,\ p_2 u_{iJ} \prod_{j=J+1}^{r_i} v_{ij}^{m_j} u_{ij}\Big) \tag{5}$$

have an overlap of length at least $|p|$, so (5) is contained in (w_1, p^N, w_2) by Lemma 2.3. If there is another maximal $p^{\geq n}$-occurrence $(w_1', p^{N'}, w_2')$ in u, then similarly there exists a number J' such that $p = \lambda(v_{iJ'})$, $v_{iJ'}^{m_J'} = p_1' p^{M'} p_2'$, where p_1' is a proper suffix of p, p_2' is a proper prefix of p, and $M' \geq 1$, and the occurrence

$$\Big(u_{i0}\Big(\prod_{j=1}^{J'-1} v_{ij}^{m_j} u_{ij}\Big)p_1',\ p^{M'},\ p_2' u_{iJ'} \prod_{j=J'+1}^{r_i} v_{ij}^{m_j} u_{ij}\Big) \tag{6}$$

is contained in the occurrence $(w_1', p^{N'}, w_2')$. It must be $J \neq J'$, because otherwise (5) and (6) would be the same, and then the maximal occurrences (w_1, p^N, w_2) and $(w_1', p^{N'}, w_2')$ would be the same by Lemma 2.3. By symmetry, we can assume $J < J'$. Then L has a subset of the form $x(p^{l_1})^* y (p^{l_2})^* z$, where

$$y = p_2 u_{iJ}\Big(\prod_{j=J+1}^{J'-1} v_{ij}^{m_j} u_{ij}\Big)p_1',$$

and then it also has the subset $x(p^m)^*y(p^m)^*z$, where $m = l_1l_2$. Here $y \notin p^*$ and thus $py \neq yp$, because otherwise (5) and (6) would be contained in the p^+-occurrence

$$\Big(u_{i0}\Big(\prod_{j=1}^{J-1} v_{ij}^{m_j}u_{ij}\Big)p_1,\ p^M y p^{M'},\ p_2'u_{iJ'}\prod_{j=J'+1}^{r_i} v_{ij}^{m_j}u_{ij}\Big),$$

and then the maximal occurrences (w_1, p^N, w_2) and $(w_1', p^{N'}, w_2')$ would be the same by Lemma 2.3.

Finally, we prove Condition 3. Let $u \in L$. There are numbers $i, m_1, \ldots, m_{r_i}$ such that

$$u = u_{i0}\prod_{j=1}^{r_i} v_{ij}^{m_j}u_{ij}.$$

Let x be a factor of u of length at least k. If it does not have a common factor of length at least $|v_{ij}^{n+2}|$ with the factor $v_{ij}^{m_j}$ for any j, then

$$|x| < \sum_{j=0}^{r_i}|u_{ij}| + \sum_{j=1}^{r_i}|v_{ij}^{n+2}| \leq k,$$

which is a contradiction. So there exists a number j such that x and $v_{ij}^{m_j}$ have a common factor of length at least $|v_{ij}^{n+2}|$, and this common factor necessarily has a $\lambda(v_{ij})^{n+1}$-occurrence. □

Now we are ready to prove our main theorem.

Theorem 5.4. A regular language L has a finite SSF if and only if L does not have a subset of the form xw^*yw^*z for any words w, x, y, z such that $wy \neq yw$.

Proof:
The "only if" direction follows from Lemma 5.1. To prove the "if" direction, let n, k, P be as in Lemma 5.3 (L is bounded by Lemma 5.2). Let $u, v \in L$ be k-abelian equivalent. We are going to show that $u = v$. This proves the theorem by Lemma 3.3. If $|u| = |v| < k$, then trivially $u = v$, so we assume that $|u| = |v| \geq k$.

Let $P_j = \{p^i \mid p \in P, i \geq j\}$ for all j. Let the maximal P_n-occurrences in u be

$$(x_1, p_1^{m_1}, x_1'), \ldots, (x_r, p_r^{m_r}, x_r'), \tag{7}$$

where $p_1, \ldots, p_r \in P$. It follows from $|u| \geq k$ and Condition 3 of Lemma 5.3 that $r \geq 1$. We can assume that the occurrences have been ordered so that $|x_1| \leq \cdots \leq |x_r|$. By Condition 2 of Lemma 5.3, the words $p_1, \ldots, p_r$ are pairwise distinct. All P_n-occurrences in u are contained in one of the maximal occurrences (7). By Condition 1 of Lemma 5.3, p^n cannot be a factor of $p_j^{m_j}$ if $p \in P \setminus \{p_j\}$, so if $p \in P \setminus \{p_1, \ldots, p_r\}$, then there are no $p^{\geq n}$-occurrences in u, and all $p_i^{\geq n}$-occurrences are $(x_ip_i^l, p_i^j, p_i^{m_i-j-l}x_i')$ for $j \in \{n, \ldots, m_i\}$ and $l \in \{0, \ldots, m_i - j\}$. In particular, $|u|_{p_i^n} = m_i - n + 1$.

Similarly, let the maximal P_n-occurrences in v be

$$(y_1, q_1^{n_1}, y_1'), \ldots, (y_s, q_s^{n_s}, y_s'),$$

where $s \geq 1$ and $q_1, \ldots, q_s \in P$. As above, we can assume that the occurrences have been ordered so that $|y_1| \leq \cdots \leq |y_s|$, and we can prove that the words $q_1, \ldots, q_s$ are pairwise distinct, p^n cannot be a factor of $q_j^{n_j}$ if $p \in P \setminus \{q_j\}$, and if $p \in P \setminus \{q_1, \ldots, q_s\}$, then there are no $p^{\geq n}$-occurrences in v, all $q_i^{\geq n}$-occurrences are $(y_i q_i^l, q_i^j, q_i^{n_i-j-l} y_i')$ for $j \in \{n, \ldots, n_i\}$ and $l \in \{0, \ldots, n_i - j\}$, and $|v|_{q_i^n} = n_i - n + 1$.

If $p \in P$, then $|p^n| < k$ by Condition 3 of Lemma 5.3, and then $|u|_{p^n} = |v|_{p^n}$ because $u \equiv_k v$. It follows that $r = s$ and $\{p_1, \ldots, p_r\} = \{q_1, \ldots, q_s\}$. We have seen that $|u|_{p_i^n} = m_i - n + 1$ and $|v|_{q_j^n} = n_j - n + 1$, so if $p_i = q_j$, then $m_i = n_j$.

We prove by induction that $(x_i, p_i, m_i) = (y_i, q_i, n_i)$ for all $i \in \{1, \ldots, r\}$. First, we prove the case $i = 1$. The words u and v have prefixes $x_1 p_1^n$ and $y_1 q_1^n$, respectively. There is only one P_n-occurrence and no P_{n+1}-occurrences in $x_1 p_1^n$. Similarly, there is only one P_n-occurrence and no P_{n+1}-occurrences in $y_1 q_1^n$. By Condition 3 of Lemma 5.3, $|x_1 p_1^n| < k$ and $|y_1 q_1^n| < k$. Because u and v are k-abelian equivalent, they have the same prefix of length $k-1$, and thus one of $x_1 p_1^n$ and $y_1 q_1^n$ is a prefix of the other. If, say, $x_1 p_1^n$ is a prefix of $y_1 q_1^n$, then $y_1 q_1^n$ has an occurrence (x_1, p_1^n, z) for some word z, and this must be the unique P_n-occurrence $(y_1, q_1^n, \varepsilon)$. It follows that $x_1 = y_1$ and $p_1 = q_1$, and then also $m_1 = n_1$.

Next, we assume that $(x_i, p_i, m_i) = (y_i, q_i, n_i)$ for some $i \in \{1, \ldots, r-1\}$ and prove that $(x_{i+1}, p_{i+1}, m_{i+1}) = (y_{i+1}, q_{i+1}, n_{i+1})$. Let

$$x_{i+1} = x_i p_i^{m_i - n} x_i'', \qquad y_{i+1} = y_i q_i^{n_i - n} y_i'' = x_i p_i^{m_i - n} y_i''.$$

The unique shortest factor in u beginning with p_i^n and ending with p^n for some $p \in P \setminus \{p_i\}$ is the factor $x_i'' p_{i+1}^n$ with the unique occurrence $(x_i p_i^{m_i - n}, x_i'' p_{i+1}^n, p_{i+1}^{m_{i+1} - n} x_{i+1}')$. Similarly, the unique shortest factor in v beginning with p_i^n and ending with p^n for some $p \in P \setminus \{p_i\}$ is the factor $y_i'' q_{i+1}^n$ with the unique occurrence $(x_i p_i^{m_i - n}, y_i'' q_{i+1}^n, q_{i+1}^{n_{i+1} - n} y_{i+1}')$. There are no P_{n+1}-occurrences in these factors, so they are of length less than k by Condition 3 of Lemma 5.3, and they must be equal because $u \equiv_k v$. It follows that $p_{i+1} = q_{i+1}$, $x_i'' = y_i''$, and $x_{i+1} = y_{i+1}$, and then also $m_{i+1} = n_{i+1}$.

It follows by induction that $x_r p_r^{m_r} = y_r q_r^{n_r}$. Because $|u| = |v|$, it must be $|x_r'| = |y_r'|$. Because x_r' does not have any P_{n+1}-occurrences, $|x_r'| < k$ by Condition 3 of Lemma 5.3. Because u and v are k-abelian equivalent, they have the same suffix of length $k-1$, so $x_r' = y_r'$. Thus $u = v$. This completes the proof. □

Theorem 5.4 is easy to state, but it is not immediately clear what the regular languages without a subset of the specified form look like. If we are given a bounded regular language as a union of languages of the form (3), then it is actually very easy to check whether it has a subset of the specified form, although the formulation of this result, given in the next theorem, is more complicated than Theorem 5.4.

Theorem 5.5. Let

$$L = \bigcup_{i=1}^{s} u_{i0} \prod_{j=1}^{r_i} v_{ij}^* u_{ij},$$

where $s \geq 1$ and $r_1, \ldots, r_s \geq 0$ are numbers, all the u_{ij} are words, and all the v_{ij} are nonempty words. The language L does not have a finite SSF if and only if at least one of the following conditions is true:

1. There exist indices i, j_1, j_2, j_3 such that $j_1 < j_2 < j_3$ and $\lambda(v_{ij_1}) = \lambda(v_{ij_3}) \neq \lambda(v_{ij_2})$.
2. There exist indices i, j such that $\lambda(v_{ij}) = \lambda(v_{i,j+1})$ and $\rho(v_{ij})u_{ij} \neq u_{ij}\rho(v_{i,j+1})$.

Proof:
First, we assume that the first condition is true. Let $w = \lambda(v_{ij_1})$. Then L has a subset

$$x(w^k)^* y_1 v_{ij_2}^* y_2 (w^k)^* z$$

for some words x, y_1, y_2, z and number $k \geq 1$. By Lemma 2.2, a power of v_{ij_2} and a power of w cannot have a common factor of length $|v_{ij_2} w|$. This means that if n is large enough, then $y_1 v_{ij_2}^n y_2$ cannot commute with w^k. Thus L does not have a finite SSF by Theorem 5.4.

Then, we assume that the second condition is true. By $\lambda(v_{ij}) = \lambda(v_{i,j+1})$, there exist words p, q such that $\rho(v_{ij}) = pq$ and $\rho(v_{i,j+1}) = qp$. Then L has a subset

$$x((pq)^k)^* u_{ij} q ((pq)^k)^* z$$

for some words x, z and number $k \geq 1$. By $\rho(v_{ij})u_{ij} \neq u_{ij}\rho(v_{i,j+1})$, $pqu_{ij}q \neq u_{ij}qpq$. Thus L does not have a finite SSF by Theorem 5.4.

Finally, we assume that L does not have a finite SSF. By Theorem 5.4, L has a subset of the form xw^*yw^*z, where $wy \neq yw$. There exists an index i such that for infinitely many n,

$$xw^n yw^n z \in u_{i0} \prod_{j=1}^{r_i} v_{ij}^* u_{ij}.$$

Let $M = \max\{|v_{ij}| \mid j \in \{1, \ldots, r_i\}\}$. If N is so large that

$$|w^N| \geq r_i(M + |w|) + \sum_{j=0}^{r_i} |u_{ij}|,$$

and if

$$xw^N yw^N z = u_{i0} \prod_{j=1}^{r_i} v_{ij}^{m_j} u_{ij},$$

then for some numbers k, l, the occurrences $(x, w^N, yw^N z)$ and $(xw^N y, w^N, z)$ must have overlaps of length at least $M + |w|$ with the occurrences

$$(u_{i0} \prod_{j=1}^{k-1} v_{ij}^{m_j} u_{ij}, v_{ik}^{m_k}, u_{ik} \prod_{j=k+1}^{r_i} v_{ij}^{m_j} u_{ij}), \qquad (u_{i0} \prod_{j=1}^{l-1} v_{ij}^{m_j} u_{ij}, v_{il}^{m_l}, u_{il} \prod_{j=l+1}^{r_i} v_{ij}^{m_j} u_{ij}),$$

respectively. By Lemma 2.2, $\lambda(w) = \lambda(v_{ik}) = \lambda(v_{il})$. If the first condition from the statement of Theorem 5.5 is false, then $\lambda(w) = \lambda(v_{ij})$ for all $j \in \{k, \ldots, l\}$. Let

$$W = \Big(\prod_{j=k}^{l-1} v_{ij}^{m_j} u_{ij} \Big) v_{il}^{m_l}, \qquad U = \prod_{j=k}^{l-1} u_{ij},$$

and let $v_{ij}^{m_j} = \rho(v_{ij})^{n_j}$ for all j. If also the second condition is false, then

$$\rho(v_{ij})^n u_{ij} = u_{ij} \rho(v_{i,j+1})^n$$

for all $j \in \{k, \ldots, l-1\}$ and $n \geq 0$, and therefore

$$\rho(v_{ik})^{n_k + \cdots + n_l} U = W = U \rho(v_{il})^{n_k + \cdots + n_l}.$$

It follows that there exist words p and q such that $\rho(v_{ik}) = pq$, $\rho(v_{il}) = qp$ and $U \in (pq)^* p$, so W is a factor of a power of pq, and therefore a factor of a power of $\rho(w)$. The occurrence $(xw^{N-1}, wyw, w^{N-1}z)$ is contained in the occurrence

$$(u_{i0} \prod_{j=1}^{k-1} v_{ij}^{m_j} u_{ij}, W, u_{il} \prod_{j=l+1}^{r_i} v_{ij}^{m_j} u_{ij}),$$

so $\rho(w) y \rho(w)$ is a factor of W, and therefore a factor of a power of $\rho(w)$. It follows that y is a power of $\rho(w)$, which contradicts $wy \neq yw$. This contradiction shows that both conditions cannot be false. □

Example 5.6. The language $a^*(abab)^*b^*a(ba)^*$ satisfies the first condition of Theorem 5.5, because $\lambda(abab) = \lambda(ba) \neq \lambda(b)$. Thus the language does not have a finite SSF. Alternatively, this can be seen by noticing that the language has a subset $(abab)^*b(abab)^*a$ and using Theorem 5.4.

The language $a^*(abab)^*ba(ba)^*$ satisfies the second condition of Theorem 5.5, because $\lambda(abab) = \lambda(ba)$ and $\rho(abab)ba \neq ba\rho(ba)$. Thus the language does not have a finite SSF. Alternatively, this can be seen by noticing that also this language has the subset $(abab)^*b(abab)^*a$.

The language $a^*(abab)^*aba(ba)^*$ does not satisfy either of the two conditions of Theorem 5.5, so it has a finite SSF.

6. Conclusion

In this article, we have defined and studied separating sets of factors. In particular, we have considered the question of whether a given language has a finite SSF. We have answered this question for sets of factors of infinite words and for regular languages. In the former case, we have also analyzed the cumulative growth functions of infinite SSFs. We can ask the following questions:

- Given a language with a finite SSF, what is the minimal size of an SSF of this language? For example, this question could be considered for Σ^n.
- Given a language with no finite SSF, how "small" can the (cumulative) growth function of an SSF of this language be? For example, this question could be considered for Σ^*, or for sets of factors of infinite words. More specific questions were mentioned at the end of Section 4.

References

[1] Goralčík P, Koubek V. On discerning words by automata. In: Proceedings of the 13th ICALP, volume 226 of *LNCS*. Springer, 1986 pp. 116–122. doi:10.1007/3-540-16761-7_61.

[2] Robson JM. Separating strings with small automata. *Information Processing Letters*, 1989. **30**(4):209–214. doi:10.1016/0020-0190(89)90215-9.

[3] Demaine ED, Eisenstat S, Shallit J, Wilson DA. Remarks on separating words. In: Proceedings of the 13th DCFS, volume 6808 of *LNCS*. Springer, 2011 pp. 147–157. doi:10.1007/978-3-642-22600-7_12.

[4] Currie J, Petersen H, Robson JM, Shallit J. Separating words with small grammars. *Journal of Automata, Languages and Combinatorics*, 1999. **4**(2):101–110.

[5] Place T, Zeitoun M. Separating regular languages with first-order logic. *Logical Methods in Computer Science*, 2016. **12**(1):Paper No. 5, 31. doi:10.2168/LMCS-12(1:5)2016.

[6] Holub v, Kortelainen J. On partitions separating words. *International Journal of Algebra and Computation*, 2011. **21**(8):1305–1316. doi:10.1142/S0218196711006650.

[7] Maňuch J. Characterization of a word by its subwords. In: Proceedings of the 4th DLT. World Sci. Publ., 2000 pp. 210–219. doi:10.1142/9789812792464_0018.

[8] Vyalyĭ MN, Gimadeev RA. On separating words by the occurrences of subwords. *Diskretnyĭ Analiz i Issledovanie Operatsiĭ*, 2014. **21**(1):3–14. doi:10.1134/S1990478914020161.

[9] Karhumäki J. Generalized Parikh mappings and homomorphisms. *Information and Control*, 1980. **47**(3):155–165. doi:10.1016/S0019-9958(80)90493-3.

[10] Karhumäki J, Saarela A, Zamboni LQ. On a generalization of Abelian equivalence and complexity of infinite words. *Journal of Combinatorial Theory. Series A*, 2013. **120**(8):2189–2206. doi:10.1016/j.jcta.2013.08.008.

[11] Cassaigne J, Karhumäki J, Saarela A. On growth and fluctuation of k-abelian complexity. *European Journal of Combinatorics*, 2017. **65**:92–105. doi:10.1016/j.ejc.2017.05.006.

[12] Chen J, Lü X, Wu W. On the k-abelian complexity of the Cantor sequence. *Journal of Combinatorial Theory. Series A*, 2018. **155**:287–303. doi:10.1016/j.jcta.2017.11.010.

[13] Karhumäki J, Saarela A, Zamboni LQ. Variations of the Morse-Hedlund theorem for k-abelian equivalence. *Acta Cybernetica*, 2017. **23**(1):175–189. doi:10.14232/actacyb.23.1.2017.11.

[14] Richomme G, Saari K, Zamboni LQ. Abelian complexity of minimal subshifts. *Journal of the London Mathematical Society*, 2011. **83**(1):79–95. doi:10.1112/jlms/jdq063.

[15] Cassaigne J, Karhumäki J, Puzynina S, Whiteland MA. k-abelian equivalence and rationality. *Fundamenta Informaticae*, 2017. **154**(1–4):65–94. doi:10.3233/FI-2017-1553.

[16] Saarela A. Separating many words by counting occurrences of factors. In: Proceedings of the 23rd DLT, volume 11647 of *LNCS*. Springer, 2019 pp. 251–264. doi:10.1007/978-3-030-24886-4_19.

[17] Fine NJ, Wilf HS. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 1965. **16**:109–114. doi:10.1090/S0002-9939-1965-0174934-9.

[18] Morse M, Hedlund GA. Symbolic dynamics. *American Journal of Mathematics*, 1938. **60**(4):815–866. doi:10.2307/2371264.

[19] Ginsburg S, Spanier EH. Bounded regular sets. *Proceedings of the American Mathematical Society*, 1966. **17**:1043–1049. doi:10.2307/2036087.