Separating Many Words by Counting Occurrences of Factors

Aleksi Saarela[0000−0002−6636−2317]

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 $x \in X$ appearing 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 · Regular language · Infinite word · k-abelian equivalence · Separating words problem

1 Introduction

The motivation for this article comes from three 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 [8], 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 [18], 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 [6]. Several variations of the problem exist. For example, NFAs [6] or context-free grammars [5] 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 [16]. 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 [9].

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 modulo a prime can be easily counted by a DFA. So if u and v have a different number of occurrences of some short factor x, then $\mathrm{sep}(u, v)$ is small, see [6] 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 [13]. See [19] for a variation where also the positions of the occurrences modulo a certain number are taken into account. The question becomes more interesting if we want to separate more than two words (possibly infinitely many) at once, and we can do this by counting the numbers of occurrences of more than one factor. We are particularly interested in the following question.

Question 1.1. 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 an old guessing game for two players, let us call them Alice and Bob: From a given set of options, Alice secretly picks one. Bob is allowed to ask any yes-no questions, and he is trying to figure out what Alice picked. Two famous versions are the game "Twenty Questions" and the children's board game "Guess Who". In their simplest forms, these kinds of games are easy to analyze: The required number of questions is logarithmic with respect to the number of options. However, many more complicated variations have been studied. As examples, see [15] and [1].

In this article, we are interested in a variation where the options are words and, instead of arbitrary yes-no questions, Bob is allowed to ask for the number of occurrences of any factor in the word Alice has chosen. Usually in games like this, Bob can decide every question based on the previous answers, but we can also require that Bob needs to decide all the questions in advance.

Question 1.2. Given a language from which Alice has secretly picked one word w, can Bob find a finite language X such that the answers to the questions "What is $|w|_x$?" for all $x \in X$ are guaranteed to reveal the correct word w?

It is easy to see that Questions 1.1 and 1.2 are equivalent. In this article, we will use the formulation of Question 1.1 instead of talking about games.

The third 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 [10]. Many basic properties were proved by Karhumäki, Saarela and Zamboni in the article [11], where also k-abelian complexity was introduced. Several articles have been published about k-abelian complexity [3,4,12], and about abelian complexity (that is, the case $k = 1$) already earlier [17]. Perhaps the most interesting one from the point of view of this paper is [3], 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 1.3. Given an infinite word, does there exist a number $k \geq 1$ 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 [2].

Question 1.4. Given a language, does there exist a number $k \geq 1$ 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 1.1 and prove basic properties about them. As stated above, Questions 1.1 and 1.2 are equivalent, and so is Question 1.4, 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 1.3) and regular languages. In the first case, the result is not surprising: The answer is positive if and only if the word is ultimately periodic. 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 $wy \neq yw$.

2 Preliminaries

Throughout the article, we use the symbol Σ to denote an alphabet. All words are over Σ unless otherwise specified.

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 $w \in p^+$. 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.

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 the primitive root of w.

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 $|w|_u$.

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.

k-abelian equivalence. Let k be a positive integer. Words $u, v \in \Sigma^*$ are *k-abelian equivalent* if $|u|_x = |v|_x$ for all $x \in \Sigma^{\leq k}$. k-abelian equivalence is an equivalence relation and it is denoted by $\equiv_k$.

Here are some basic facts about k-abelian equivalence (see [11]): $u, v \in \Sigma^{\geq k-1}$ are k-abelian equivalent if and only if they have a common prefix of length $k-1$ and $|u|_x = |v|_x$ for all $x \in \Sigma^k$. The condition about prefixes can be replaced by a symmetric condition about suffixes. Words of length $2k-1$ or less are k-abelian equivalent if and only if they are equal. k-abelian equivalence is a congruence, that is, if $u \equiv_k u'$ and $v \equiv_k v'$, then $uv \equiv_k u'v'$.

We are going to use the following simple fact a couple of times when showing that two words are k-abelian equivalent: If $u, v, w, x \in \Sigma^*$, $|v| = k-1$, and $|x| = k$, then

$$|uvw|_x = |uv|_x + |vw|_x.$$

Example 2.1. 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 $i \in \{0, \dots, k-1\}$, 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 *factor complexity* of an infinite word w, denoted by $\mathcal{P}_w$, is the growth function of the set of factors of w (technically, the domain of $\mathcal{P}_w$ is often defined to be $\mathbb{Z}_+$ instead of $\mathbb{Z}_{\geq 0}$).

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

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

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 $\mathcal{P}^k_w$, 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 [14] 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.

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 $\{a\}$. 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.

Example 3.2. Let $\Sigma = \{a, b, c, d, e, f\}$. The language $L = \{ac, ad, be, bf\}$ has a size-minimal SSF $\{a, c, e\}$. In terms of the guessing game mentioned in the introduction, this means that if Alice has chosen $w \in L$, then Bob can ask for the numbers $|w|_a, |w|_c, |w|_e$, and this will always reveal w. Actually, two questions are enough if Bob can choose the second question after hearing the answer to the first one: He can first ask for $|w|_a$, and then for either $|w|_c$ or $|w|_e$ depending on whether $|w|_a = 1$ or $|w|_a = 0$.

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 3.3. *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 $X = L \smallsetminus \{w\}$. 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 3.4. *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 3.4 proves the equivalence of Questions 1.1 and 1.4.

Example 3.5. 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 3.4, $\{a, b\}^{\leq 2}$

is an SSF of L. This SSF is not size-minimal (by the first claim of Lemma 3.3, 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 3.6. In a list of about 140000 English words (found in the SCOWL database[1]), there are no 4-abelian equivalent words. Therefore, by Lemma 3.4, $\Sigma^{\leq 4}$ 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 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 $\Sigma^{\leq 2} \cup \{rere, hop, ind, tan, tar\}$ 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 3.7. *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.

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

3. Let wL have no finite SSF. Let $k \in \mathbb{Z}_+$ and $k' = k + |w|$. By Lemma 3.4, 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 $x \in \Sigma^k$,
$$|u|_x = |wu|_x - |wp|_x = |wv|_x - |wp|_x = |v|_x,$$
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 3.4, 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 3.4, 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 3.4, 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
$$|u^k v u^{k-1}|_x = |u^k|_x + |u_2 v u_1|_x + |u^{k-1}|_x = |u^{k-1} v u^k|_x$$
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 3.4, L^* does not have a finite SSF.
6. If K has a finite SSF, then so does $K \cap L$. If $L \smallsetminus K$ is finite, then also L has a finite SSF by the first claim of this lemma. Similarly, if L has a finite SSF and $K \smallsetminus L$ is finite, then also K has a finite SSF. □

Example 3.8. 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 3.4, $L \cup Laa$ 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 1.3.

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 v 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 a $(v')^\omega$. 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 $x_1, \dots, x_n \in \Sigma^{k-1}$. By the definition of n, there exist two indices $i, j \in \{1, \dots, n\}$ 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 $w = z_0 xyxz_1 xyxz_2 xyx \cdots$ for some infinite sequence of words $z_0, z_1, z_2, \dots$. If the words xy and xz_i have the same primitive root p for all $i \in \mathbb{Z}_+$, then $w = z_0 p^\omega$, which contradicts the aperiodicity of w. Thus there exists i such that xy and xz_i have a different primitive root. Then $xyxz_i \neq xz_ixy$ and thus $xyxz_ix \neq xz_ixyx$. On the other hand, $xyxz_ix$ and xz_ixyx are k-abelian equivalent because they have the same prefix x of length $k-1$ and

$$|xyxz_ix|_t = |xyx|_t + |xz_ix|_t = |xz_ix|_t + |xyx|_t = |xz_ixyx|_t$$

for all $t \in \Sigma^k$. Moreover, $xyxz_ix$ and xz_ixyx 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 3.4. □

5 Regular languages

In this section, we give an answer to Question 1.1 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

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

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 3.4, 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 $v_1, \dots, v_n$ are words. It was proved by Ginsburg and Spanier [7] that a regular language is bounded if and only if it is a finite union of languages of the form

$$u_0 v_1^* u_1 \cdots v_n^* u_n,$$

where $u_0, \dots, u_n$ are words and $v_1, \dots, v_n$ 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 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 numbers $n, k \geq 0$ 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 claim is basically trivial. If L is infinite, then

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

where $s \geq 1$ and $r_1, \dots, r_s \geq 0$, $r_i \geq 1$ for at least one i, and the words v_{ij} are nonempty. We can let P be the set of Lyndon roots of the words v_{ij} and

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

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

The proof can be found in the arXiv version of this paper.[2] □

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 3.4. 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

$$(x_1, p_1^{m_1}, x_1'), \dots, (x_r, p_r^{m_r}, x_r'), \tag{1}$$

where $p_1, \dots, p_r \in P$. 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 $|x_1| \leq \dots \leq |x_r|$. By Condition 2 of Lemma 5.3, the words $p_1, \dots, p_r$ are pairwise distinct. All P_n-occurrences in u are contained in one of the maximal occurrences (1). By Condition 1 of Lemma 5.3, p^n cannot be a factor of $p_j^{m_j}$ if $p \in P \smallsetminus \{p_j\}$, so if $p \in P \smallsetminus \{p_1, \dots, p_r\}$, then there are no $p^{\geq n}$-occurrences in u, and all $p_i^{\geq n}$-occurrences are $(x_i p_i^l, p_i^j, p_i^{m_i - j - l} x_i')$ for $j \in \{n, \dots, m_i\}$ and $l \in \{0, \dots, m_i - j\}$. In particular, $|u|_{p_i^n} = m_i - n + 1$.

Similarly, let the maximal P_n-occurrences in v be

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

where $s \geq 1$ and $q_1, \dots, q_s \in P$. As above, we can assume that the occurrences have been ordered so that $|y_1| \leq \dots \leq |y_s|$, and we can prove that the words $q_1, \dots, q_s$ are pairwise distinct, p^n cannot be a factor of $q_j^{n_j}$ if $p \in P \smallsetminus \{q_j\}$, and if $p \in P \smallsetminus \{q_1, \dots, q_s\}$, 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 $j \in \{n, \dots, n_i\}$ and $l \in \{0, \dots, n_i - j\}$, 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 $\{p_1, \dots, p_r\} = \{q_1, \dots, q_s\}$. 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 $m_i = n_j$.

[2] `http://arxiv.org/abs/1905.07223`

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 $x_{i+1} = x_i p_i^{m_i - n} x_i''$ and $y_{i+1} = y_i q_i^{n_i - n} y_i'' = x_i p_i^{m_i - n} y_i''$. The unique shortest factor in u beginning with p_i^n and ending with p^n for some $p \in P \smallsetminus \{p_i\}$ is the factor $x_i'' p_{i+1}^n$ starting at position $|x_i p_i^{m_i - n}|$ and ending at position $|x_{i+1} p_{i+1}^n|$. Similarly, the unique shortest factor in v beginning with p_i^n and ending with p^n for some $p \in P \smallsetminus \{p_i\}$ is the factor $y_i'' q_{i+1}^n$ starting at position $|y_i q_i^{n_i - n}| = |x_i p_i^{m_i - n}|$ and ending at position $|y_{i+1} q_{i+1}^n|$. 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. □

Example 5.5. First, consider the language $K = a^*(abab)^*ba(ba)^*$. It has a subset $(abab)^*ba(ba)^* = (abab)^*b(ab)^*a$, which has a subset $(abab)^*b(abab)^*a$. It follows from Theorem 5.4 that K does not have a finite SSF.

Then, consider the language $L = a^*(abab)^*aba(ba)^* = a^*(abab)^*(ab)^*aba = a^*(ab)^*aba$. It can be proved that if L has a subset xw^*yw^*z with $w \neq \varepsilon$, then the Lyndon root of w is a or ab, and $wy = yw$. It follows from Theorem 5.4 that L 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 future, this question could be studied for other families of languages. We can also 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 could be considered for Σ^n.
- Given a language with no finite SSF, how "small" can the growth function of an SSF of this language be? For example, this could be considered for Σ^*.

References

1. Ambainis, A., Bloch, S.A., Schweizer, D.L.: Delayed binary search, or playing twenty questions with a procrastinator. Algorithmica **32**(4), 641–650 (2002). https://doi.org/10.1007/s00453-001-0097-4
2. Cassaigne, J., Karhumäki, J., Puzynina, S., Whiteland, M.A.: k-abelian equivalence and rationality. Fund. Inform. **154**(1–4), 65–94 (2017). https://doi.org/10.3233/FI-2017-1553
3. Cassaigne, J., Karhumäki, J., Saarela, A.: On growth and fluctuation of k-abelian complexity. European J. Combin. **65**, 92–105 (2017). https://doi.org/10.1016/j.ejc.2017.05.006
4. Chen, J., Lü, X., Wu, W.: On the k-abelian complexity of the Cantor sequence. J. Combin. Theory Ser. A **155**, 287–303 (2018). https://doi.org/10.1016/j.jcta.2017.11.010
5. Currie, J., Petersen, H., Robson, J.M., Shallit, J.: Separating words with small grammars. J. Autom. Lang. Comb. **4**(2), 101–110 (1999)
6. Demaine, E.D., Eisenstat, S., Shallit, J., Wilson, D.A.: Remarks on separating words. In: Proceedings of the 13th DCFS. LNCS, vol. 6808, pp. 147–157. Springer (2011). https://doi.org/10.1007/978-3-642-22600-7_12
7. Ginsburg, S., Spanier, E.H.: Bounded regular sets. Proc. Amer. Math. Soc. **17**, 1043–1049 (1966). https://doi.org/10.2307/2036087
8. Goralčík, P., Koubek, V.: On discerning words by automata. In: Proceedings of the 13th ICALP. LNCS, vol. 226, pp. 116–122. Springer (1986). https://doi.org/10.1007/3-540-16761-7_61
9. Holub, v., Kortelainen, J.: On partitions separating words. Internat. J. Algebra Comput. **21**(8), 1305–1316 (2011). https://doi.org/10.1142/S0218196711006650
10. Karhumäki, J.: Generalized Parikh mappings and homomorphisms. Information and Control **47**(3), 155–165 (1980). https://doi.org/10.1016/S0019-9958(80)90493-3
11. Karhumäki, J., Saarela, A., Zamboni, L.Q.: On a generalization of Abelian equivalence and complexity of infinite words. J. Combin. Theory Ser. A **120**(8), 2189–2206 (2013). https://doi.org/10.1016/j.jcta.2013.08.008
12. Karhumäki, J., Saarela, A., Zamboni, L.Q.: Variations of the Morse-Hedlund theorem for k-abelian equivalence. Acta Cybernet. **23**(1), 175–189 (2017). https://doi.org/10.14232/actacyb.23.1.2017.11
13. Maňuch, J.: Characterization of a word by its subwords. In: Proceedings of the 4th DLT. pp. 210–219. World Sci. Publ. (2000). https://doi.org/10.1142/9789812792464_0018
14. Morse, M., Hedlund, G.A.: Symbolic dynamics. Amer. J. Math. **60**(4), 815–866 (1938). https://doi.org/10.2307/2371264
15. Pelc, A.: Solution of Ulam's problem on searching with a lie. J. Combin. Theory Ser. A **44**(1), 129–140 (1987). https://doi.org/10.1016/0097-3165(87)90065-3
16. Place, T., Zeitoun, M.: Separating regular languages with first-order logic. Log. Methods Comput. Sci. **12**(1), Paper No. 5, 31 (2016). https://doi.org/10.2168/LMCS-12(1:5)2016
17. Richomme, G., Saari, K., Zamboni, L.Q.: Abelian complexity of minimal subshifts. J. Lond. Math. Soc. (2) **83**(1), 79–95 (2011). https://doi.org/10.1112/jlms/jdq063
18. Robson, J.M.: Separating strings with small automata. Inform. Process. Lett. **30**(4), 209–214 (1989). https://doi.org/10.1016/0020-0190(89)90215-9

19. Vyalyĭ, M.N., Gimadeev, R.A.: On separating words by the occurrences of subwords. Diskretn. Anal. Issled. Oper. **21**(1), 3–14 (2014). https://doi.org/10.1134/S1990478914020161