On Lengths of Words in Context-free Languages

Lucian ILIE

Turku Centre for Computer Science TUCS
FIN-20520 Turku, Finland

Turku Centre for Computer Science
TUCS Technical Report No 137
October 1997
ISBN 952-12-0089-8
ISSN 1239-1891

Abstract

We consider slender languages, that is, languages for which the number of words of the same length is bounded from above by a constant. We prove that the slenderness problem is decidable for context-free languages and that the maximal number of words of the same length in a given context-free language is computable. Some effective representations of slender context-free languages, as well as other related decidability problems are investigated.

1 Introduction

Considerations concerning lengths of words are an important part of language theory. An infinite sequence $(l_n)_{n\geq 0}$ can be associated in a natural way to a language L: l_n is the number of words of length n in L. The investigation of the case when all terms of the sequence $(l_n)_{n\geq 0}$ are bounded from above by a fixed constant was started by Păun and Salomaa in a series of papers [PS1], [PS2], [PS3]. Such languages are termed *slender*. As shown in [ADPS], these languages are not only important from theoretical point of view, but also they have applications in cryptography. Some theoretical applications can be found also in [CY] and [HI]. The same notion has been investigated for L languages in [DPS], [NS], and the case when l_n is bounded by a polynomial in n has been considered in [Ra], [SYZS]. Honkala has introduced and studied a generalization of the notion of slenderness, called Parikh slenderness, in [Ho1], [Ho2], [Ho3].

One of the most important problems is to find characterizations of such languages. As proved in [PS1], the slender regular languages are exactly the *unions of single loops*, that is, finite unions of sets of the form uv^*w. It was conjectured in [PS1] that a similar characterization holds for slender context-free languages, as *unions of paired loops*, that is, finite unions of sets of the form $\{uv^nwx^ny \mid n \geq 0\}$. This conjecture was shown to be true by the author in [Il].

Another important problem is the decidability of the slenderness problem. This problem was shown to be decidable for unambiguous context-free languages in [PS1]. In the much more difficult case of arbitrary context-free languages, the problem was also proved to be decidable by Raz, [Ra].

In this paper, we present two completely new proofs for the decidability of the slenderness problem for context-free languages. Our approach continues the ideas in [Il] and is based on some upper bounds on the lengths of words in the paired loops composing a slender context-free language, as well as on the effective semilinearity of context-free languages. We are able in this way to find effectively a representation of a given slender context-free language as a union of paired loops and even as a disjoint such union. Also, we show that, for a given slender context-free language, the maximal number of words of the same length is computable. Some undecidability results are also considered.

The paper is organized as follows. The next section contains the definitions and two results from [Il] we need later on. Section 3 contains a useful technical lemma concerning paired loops of a slender language and the upper bounds on the lengths of words in the loops (single or paired) of a slender context-free language. The algorithms for deciding whether or not an arbitrary context-free language is slender are given in section 4. In section 5, we present the other decidability results and effective constructions mentioned above, as well as some related ones. The last section deals with undecidability.

2 Definitions and previous results

For an alphabet Σ, Σ^* denotes the free monoid generated by Σ, λ is its identity, and, for any word $w \in \Sigma^*$, $|w|$ is the length of w.

Let k be a positive integer. A language $L \subseteq \Sigma^*$ is called *k-slender* if

$$card\{w \in L \mid |w| = n\} \leq k,$$

for any $n \geq 0$. L is *slender* if it is k-slender for some $k \geq 1$.

A *single loop* is a set $\{uv^n w \mid n \geq 0\}$, for some $u, v, w \in \Sigma^*$. A *paired loop* is $\{uv^n wx^n y \mid n \geq 0\}$, $u, v, w, x, y \in \Sigma^*$.

We call a *union of single (paired) loops* any finite union of single or paired loops, respectively.

Let $L \subseteq \Sigma^*$ be a context-free language. There is a context-free grammar in Chomsky normal form $G = (N, \Sigma, S, P)$ generating L. Then, by the pumping lemma for context-free languages, there is a constant p such that any word $z \in L$ with $|z| > p$ can be decomposed as $z = uvwxy$, for some $u, v, w, x, y \in \Sigma^*$, such that $vx \neq \lambda$, $|vwx| \leq p$, and $uv^n wx^n y \in L$, for any $n \geq 0$. The constant p is usually taken to be $p = 2^{card(N)}$. In what follows, we shall denote this constant by p_L and it will be understood that it comes from a grammar in Chomsky normal form for L.

We recall the following two results from [Il]. The first one is the characterization of the slender context-free languages as union of paired loops conjectured in [PS1] and the second one is a useful technical lemma we shall need later on.

Theorem 2.1 ([Il]). *If $L \subseteq \Sigma^*$ is a slender context-free language, then L is a union of paired loops. Moreover, we have*

$$L = \bigcup_{i=1}^{m} \{u_i v_i^n w_i x_i^n y_i \mid n \geq 0\}, \tag{1}$$

for some $m \geq 1, u_i, v_i, w_i, x_i, y_i \in \Sigma^*, 1 \leq i \leq m$, such that $|v_i w_i x_i| \leq p_L$, for any $1 \leq i \leq m$.

Lemma 2.2 ([Il]). *If, for some words $u_i, v, w, x, y_i \in \Sigma^*, 1 \leq i \leq 2$, the two sets $A_1 = \{u_1 v^n wx^n y_1 \mid n \geq 0\}$ and $A_2 = \{u_2 v^n wx^n y_2 \mid n \geq 0\}$ have infinitely many common words and $|u_1 y_1| \leq |u_2 y_2|$, then $A_2 \subseteq A_1$.*

Also, for basic results in combinatorics on words and formal language theory, we refer to [Lo] and [HU], [Sa], respectively.

3 Bounds on loops

We intend to show in this section that, if L is a slender context-free language, then it can be decomposed as in (1) such that the lengths of u_i and y_i, $1 \leq i \leq k$, are bounded by some a priori computable constants depending on L only. Notice that in (1) the lengths of v_i, w_i, x_i are already bounded by p_L.

We start with a technical lemma concerning paired loops in a slender language.

Lemma 3.1. *Consider the words $u_i, v, w, x, y \in \Sigma^*, 1 \leq i \leq 5$, such that $vx \neq \lambda$ and the sets*

$$R_m = \{u_1 u_2^m u_3 u_4^m u_5 v^n w x^n y \mid n \geq 0\},$$

for all $m \geq 0$. If the language $R = \bigcup_{m\geq 0} R_m$ is slender, then, for all but finitely many $m \geq 0$, R_m is a subset of $R_{m'}$, for some $0 \leq m' < m$.

Proof. Suppose that there is an infinite sequence $(m_i)_{i\geq 1}$, $0 \leq m_1 < m_2 < \cdots$, such that, for any $i \geq 1$,

$$R_{m_i} \not\subseteq \bigcup_{m=0}^{m_i - 1} R_m.$$

By Lemma 2.2, it follows that, for any $1 \leq i < j$, R_{m_i} and R_{m_j} have finitely many common words. Therefore, if $k \geq 1$ is an arbitrary fixed integer, we can find a positive integer M such that

$$M \geq \max_{1\leq i\leq k|vx|+1} (|u_1 u_3 u_5 w y| + m_i |u_2 u_4|)$$

and all words of length at least M in the set

$$\bigcup_{i=1}^{k|vx|+1} R_{m_i} \tag{2}$$

belong exactly to one set R_{m_i}. (The integer M is chosen such that, for any $1 \leq i \leq k|vx| + 1$, the shortest word in R_{m_i} is shorter than M, as well as all words in the union (2) which belong to at least two sets in the union.)

Consider the interval $[M, M+|vx|-1]$ of integers. For any $1 \leq i \leq k|vx|+1$, the lengths of words in R_{m_i} form an arithmetical progression of ratio $|vx| \neq 0$, having the first element smaller than M. Thus, each set $\{|w| \mid w \in R_{m_i}\}, 1 \leq i \leq k|vx| + 1$, has one element in the interval. Since there are $k|vx| + 1$ sets, we can find $k+1$ words of the same length. As they have the length at least M, they are all different. Consequently, there are $k+1$ different words of the same length in (2). As k has been arbitrarily chosen, this contradicts the slenderness of R. ■

The next lemma settles the bounds on the lengths of words in single loops of slender context-free languages.

Lemma 3.2. *If $L \subseteq \Sigma^*$ is a slender context-free language and $\{uv^n w \mid n \geq 0\} \subseteq L$, for some $u, v, w \in \Sigma^*, v \neq \lambda, |v| \leq p_L$, then there exist $u', w' \in \Sigma^*$ with $|u'| \leq p_L^2, |w'| \leq p_L^2$ such that*

$$\{uv^n w \mid n \geq 0\} \subseteq \{u'v^n w' \mid n \geq 0\} \subseteq L.$$

Proof. We suppose that $|u| > p_L^2$ and prove that there exists $u' \in \Sigma^*$ with $|u'| < |u|$ such that

$$\{uv^n w \mid n \geq 0\} \subseteq \{u'v^n w \mid n \geq 0\} \subseteq L. \tag{3}$$

This is enough for the statement of our lemma because we can continue in this way until $u \in \Sigma^*, |u| \leq p_L^2$, satisfying (3) is found. Also, the same reasoning can be done symmetrically for w.

Consider $G = (N, \Sigma, S, P)$ a context-free grammar in Chomsky normal form for L such that $p_L = 2^{\mathrm{card}(N)}$. Since $|u| > p_L^2 \geq p_L$, it follows that in any derivation tree for uvw in G, in the subtree generating u, there is a path from the root to a leaf on which a nonterminal appears twice. Considering the closest two occurrences of such a nonterminal, we obtain, for some $u_i \in \Sigma^*, 1 \leq i \leq 5$, with $u_2 u_4 \neq \lambda, |u_2 u_3 u_4| \leq p_L$, that $u = u_1 u_2 u_3 u_4 u_5$ and $u_1 u_2^m u_3 u_4^m u_5 v^n w \in L$, for any $m, n \geq 0$. Let us put, for any $m \geq 0$,

$$R_m = \{u_1 u_2^m u_3 u_4^m u_5 v^n w \mid n \geq 0\}.$$

By Lemma 3.1, there is $m_0 \geq 0$ such that

$$\bigcup_{m \geq 0} R_m = \bigcup_{m=0}^{m_0} R_m.$$

We assume in the sequel that $u_2 \neq \lambda$ and $u_4 \neq \lambda$. The cases $u_2 = \lambda$ and $u_4 = \lambda$ are treated similarly.

Consider an integer $M > 0$ such that

$$|u_2^M| \geq |u_2^{m_0} u_3 u_4^{m_0} u_5| + |u_2| + |v|$$

and

$$|u_4^M| \geq |u_4| + |v|.$$

Notice that, obviously, such an M exists and $M > m_0$. The word $u_1 u_2^M u_3 u_4^M u_5 w$ belongs to $\bigcup_{m=0}^{m_0} R_m$ hence, for some $0 \leq m \leq m_0$ and $n \geq 0$, $u_1 u_2^M u_3 u_4^M u_5 w = u_1 u_2^m u_3 u_4^m u_5 v^n w$. Due to the chosen value for M, we have suitable overlappings in order to use Fine and Wilf's theorem. We get that, for a primitive word $z \in \Sigma^*, z \neq \lambda, v$ is a power of z while u_2 and u_4 are powers of conjugates of z. More precisely, we have

$$\begin{aligned}
&v = z^{r_1}, \text{ for some } r_1 > 0,\\
&u_2 = (z_1'' z_1')^{s_1}, \text{ for } z_1', z_1'' \in \Sigma^* \text{ with } z = z_1' z_1'' \text{ and } s_1 \geq 0,\\
&u_4 = (z_2'' z_2')^{s_2}, \text{ for } z_2', z_2'' \in \Sigma^* \text{ with } z = z_2' z_2'' \text{ and } s_2 \geq 0,\\
&u_3 = z_1'' z^{s_3} z_2', \text{ for } s_3 \geq 0,\\
&u_5 = z_2'' z^{s_4}, \text{ for } s_4 \geq 0.
\end{aligned}$$

(Notice that it may also be that $u_3 = {z_1'}^{-1} z_2'$ in the case when z_1' is a prefix of z_2', but the reasoning for this case is completely similar to the one below.)

Because $u_2 u_4 \neq \lambda$, it follows that $s_1 + s_2 > 0$. Also, from $|u_2 u_3 u_4| \leq p_L$ and $z \neq \lambda$, it follows that $s_1 + s_2 \leq p_L$. Denote $r_2 = s_1 + s_2$. Since $|v| \leq p_L$, $r_1, r_2 \in \{1, 2, \ldots, p_L\}$.

For any $m \geq 0, n \geq 0$,

$$u_1 u_2^m u_3 u_4^m u_5 v^n w = u_1 z_1'' z^{s_3 + s_4 + 1} z^{r_2 m} z^{r_1 n} w \in L.$$

If $r_1 = r_2$, then we can replace u by $u' = u_1 z_1'' z^{s_3 + s_4 + 1}$, since

$$\{u z^{r_1 n} w \mid n \geq 0\} \subseteq \{u_1 z_1'' z^{s_3 + s_4 + 1} z^{r_1 n} w \mid n \geq 0\} \subseteq L,$$

and we are done.

Otherwise, we can repeat the procedure above for $y_2 = u_1 z_1'' z^{s_3 + s_4 + 1}$ instead of u since $|y_2| > p_L^2 - p_L \geq p_L$. We obtain, for some $y_3 \in \Sigma^*$ and $1 \leq r_3 \leq p_L$, that $uvw = y_3 z^{r_3} z^{r_2} z^{r_1} w$ and, for any non-negative integers l, m, n, $y_3 z^{r_3 l} z^{r_2 m} z^{r_1 n} w \in L$.

Now, again, if $r_3 \in \{r_1, r_2\}$, then we can take $u' = y_3 z^{r_2}$, otherwise we continue our procedure.

Since $|u| > p_L^2$, we can eventually repeat the procedure p_L times, obtaining

$$uvw = y_{p_L+1} z^{r_{p_L+1}} z^{r_{p_L}} \ldots z^{r_1} w$$

such that, for any $m_i \geq 0, 1 \leq i \leq p_L + 1$,

$$y_{p_L+1} z^{r_{p_L+1} m_{p_L+1}} z^{r_{p_L} m_{p_L}} \ldots z^{r_1 m_1} w \in L.$$

Moreover, $1 \leq r_i \leq p_L$, for any $1 \leq i \leq p_L + 1$, and $\mathrm{card}(\{r_1, r_2, \ldots, r_{p_L}\}) = p_L$. Therefore, there is $1 \leq i \leq p_L$ such that $r_{p_L+1} = r_i$ and we can take $u' = y_{p_L+1} z^{r_{p_L}} z^{r_{p_L-1}} \ldots z^{r_2}$. Obviously $|u'| < |u|$ and u' satisfies (3). ∎

We need one more lemma before proving the main result of this section.

Lemma 3.3. *If $L \subseteq \Sigma^*$ is a slender context-free language and $R = \{uv^n wx^n y \mid n \geq 0\} \subseteq L$, for some $u, v, w, x, y \in \Sigma^*, vx \neq \lambda$, such that $\max(|u|, |y|) > p_L$, then R is a single loop.*

Proof. We consider the case $|u| > p_L$ only. The case $|y| > p_L$ is treated similarly.

As in the proof of the preceding lemma, we have, for some $u_i \in \Sigma^*, 1 \leq i \leq 5$, $u_2 u_4 \neq \lambda$, that $u = u_1 u_2 u_3 u_4 u_5$ and, if denote

$$R_m = \{u_1 u_2^m u_3 u_4^m u_5 v^n w x^n y \mid n \geq 0\},$$

then $R_m \subseteq L$, for any $m \geq 0$. It follows by Lemma 3.1 that there exists $m_0 \geq 0$ such that

$$\bigcup_{m \geq 0} R_m = \bigcup_{m=0}^{m_0} R_m.$$

We suppose that none of the words u_2, u_4, v, and x is empty. (All the other cases are treated similarly.) Again as in the proof of Lemma 3.2, we can arrange suitable overlappings and use Fine and Wilf's theorem. We get, for some $z \in \Sigma^*, z \neq \lambda$, that $u_2 = z^{r_0}, r_0 > 0$, and u_4, v, x are powers of conjugates of z, say $u_4 = (z_1'' z_1')^{r_1}, v = (z_2'' z_2')^{r_2}, x = (z_3'' z_3')^{r_3}$ for $z_i', z_i'' \in \Sigma^*, z = z_i' z_i'', r_i > 0, 1 \leq i \leq 3$. Then $u_3 = z^{r_4} z_1', u_5 = z_1'' z^{r_5} z_2', w = z_2'' z^{r_6} z_3'$, for some $r_i \geq 0, 4 \leq i \leq 6$. (It may happen that some of the words u_3, u_5, and w are empty, but this does not affect our reasoning.) Therefore, we have, for any $n \geq 0$,

$$uv^n wx^n y = u_1 z^{r_0+r_1+r_4+r_5+r_6} \left(z^{r_2+r_3}\right)^n z_3' y,$$

so our paired loop is a single loop, as claimed. ∎

We can prove now the main result of this section, concerning the upper bounds on the lengths of words in the loops of a slender context-free language. The proof will be simple using Lemmas 3.2 and 3.3.

This theorem is of crucial importance in the proof of the decidability of the slenderness problem for context-free languages in the next section.

Theorem 3.4. *If $L \subseteq \Sigma^*$ is a slender context-free language, then L can be written as a union of paired loops such that, for any single loop $\{uv^n w \mid n \geq 0\}$, $|u| \leq p_L^2, |v| \leq p_L, |w| \leq p_L^2$ and, for any paired loop $\{uv^n wx^n y \mid n \geq 0\}$, $|u| \leq p_L, |vwx| \leq p_L, |y| \leq p_L$.*

Proof. We know already from Theorem 2.1 that L can be written as in (1) with $|v_i w_i x_i| \leq p_L$, for any $1 \leq i \leq m$.

Now, for all single loops in the representation (1), we use Lemma 3.2 to reduce the lengths of u_i and y_i such that $|u_i| \leq p_L^2$ and $|y_i| \leq p_L^2$, as required.

In what concerns paired loops, we consider only those paired loops which are not single loops. (All the others are arranged using Lemma 3.2.) For such a paired loop we have $|u_i| \leq p_L$ and $|y_i| \leq p_L$ otherwise, by Lemma 3.3, it is a single loop, a contradiction. ∎

Problem. Are the bounds in Theorem 3.4 optimal?

4 Decidability of slenderness

We give in this section two proofs for the decidability of the slenderness problem for context-free languages.

Theorem 4.1. *It is decidable whether or not an arbitrary context-free language is slender.*

First proof. As it is well-known (see, for instance, [Gi]), for two arbitrary context-free languages L_1 and L_2 one of which is bounded, it is decidable whether or not $L_1 \subseteq L_2$, and so also whether or not $L_1 = L_2$. Using this, we get a very simple proof for the theorem. Indeed, given a context-free language L, we can check through all possible (paired) loops given by Theorem 3.4 for L, find which ones are included in L, and then check whether or not the whole L is obtained.

Second proof. The second proof is more complicated than the first one but it is direct in the sense that it is based on Theorem 3.4 only.

Let $L \subseteq \Sigma^*$ be a context-free language. In what follows, we denote by $\Sigma^{\leq n}$ the set of words of length at most n over Σ.

Algorithm 4.2.
1. Find a context-free grammar in Chomsky normal form $G = (N, \Sigma, S, P)$ generating L and set $p_L = c^{\mathrm{card}(N)}$.

2. Check whether or not

$$L \subseteq \bigcup_{u,vwx,y \in \Sigma^{\le p_L}} uv^*wx^*y \cup \bigcup_{\substack{u,w \in \Sigma^{\le p_L^2} \\ v \in \Sigma^{\le p_L}}} uv^*w$$

and, if not, then answer **no**. Notice that this is decidable and the answer is correct by Theorem 3.4.

3. For each $u, w \in \Sigma^{\le p_L^2}, v \in \Sigma^{\le p_L}$, construct the language $L_{u,v,w} = L \cap uv^*w$.

Claim 1. $L_{u,v,w}$ is effectively regular.

Proof of Claim 1. $L_{u,v,w}$ is effectively context-free and so are the languages $L^{(1)}_{u,v,w} = u^{-1}L_{u,v,w}, L^{(2)}_{u,v,w} = L^{(1)}_{u,v,w}w^{-1}$, and $L^{(3)}_{u,v,w} = g_v(L^{(2)}_{u,v,w})$, where g_v is a gsm which replaces each occurrence of v by the letter $\# \notin \Sigma$. Thus $L^{(3)}_{u,v,w} \subseteq \#^*$ so $L^{(3)}_{u,v,w}$ is regular. Moreover, by Parikh's theorem, it is effectively regular. Performing the above operations in the reverse order, we get that $L_{u,v,w}$ is effectively regular.

4. Construct the language

$$L' = L - \bigcup_{\substack{u,w \in \Sigma^{\le p_L^2} \\ v \in \Sigma^{\le p_L}}} L_{u,v,w}.$$

This construction is effective by Claim 1. Notice that L' does not contain any single loop as in the representation in Theorem 3.4. Also, L is slender if and only if L' is slender.

5. For each $u, v, w, x, y \in \Sigma^*$ with $|u| \le p_L, |vwx| \le p_L, |y| \le p_L$, construct the language

$$L_1 = L' \cap uv^*wx^*y.$$

Clearly, L' is slender if and only if all such languages L_1 are slender. If L_1 is finite or empty, than it is slender, so suppose that L_1 is infinite.

6. Construct the languages $L_2 = u^{-1}L_1, L_3 = L_2y^{-1}$, and $L_4 = g(L_3)$, where g is a gsm which replaces each occurrence of v by the word $\#_1^{|v|}$, erases w, and replaces each occurrence of x by the word $\#_2^{|x|}$, for two new letters $\#_1, \#_2 \notin \Sigma$. It is clear that, since v, w, x are given, such a gsm can be constructed. Also, as proved in Claim 2 below, L_1 and L_4 are simultaneously slender.

Claim 2. L_1 is slender if and only if L_4 is slender.

Proof of Claim 2. Suppose first that L_4 is slender and consider an arbitrary $n \geq 0$. We have that

$$\mathrm{card}(\{w \in L_1 \mid |w| = n\}) \leq \mathrm{card}(\{z \in L_4 \mid |z| = n - |uwy|\}),$$

therefore L_1 is slender.

Conversely, suppose that L_1 is slender and take $n \geq 0$. All words of length n in L_4 come from words of length $n + |uwy|$ in L_1. We show that any word of L_1 cannot produce more than $|wvx|$ different words of L_4. Clearly, this implies the slenderness of L_4.

Suppose that there is $z \in L_1$ which produces al least $|wvx| + 1$ different words of L_4. It follows that there exist two decompositions of z, $z = uv^n wx^m y$ and $z = uv^r wx^s y$ such that $n < r$ and $|v^{r-n}| \geq |wvx|$. Therefore, the words v^n and x^m have a common factor of length $|v| + |x|$. By Fine and Wilf's theorem, v and x are powers of conjugates of the same word $t \in \Sigma^*$, say $x = t^{r_1}, v = (t_2 t_1)^{r_2}$, where $t = t_1 t_2, r_1 > 0, r_2 > 0$. Thus $w = t_2 t^{r_3}$, for some $r_3 \geq 0$. It follows that $uv^n wx^m w = ut_2 t^{r_2 n + r_1 m + r_3} y$, hence $uv^* wx^* y \subseteq ut_2 t^* y$. But, since $|ut_2| \leq 2p_L \leq p_L^2$, $|t| \leq p_L, |y| \leq p_L$, we have, from step 4, $L' \cap ut_2 t^* y = \emptyset$ and so $L_1 = \emptyset$, a contradiction. The claim is proved.

7. By Parikh's theorem, the Parikh set of L_4 is effectively semilinear and let it be $\mathcal{P}(L_4) = \bigcup_{i=1}^{r} P_i$, where P_i is a linear set, for any $1 \leq i \leq r$. Put also $L_4 = \bigcup_{i=1}^{r} R_i$, where R_i corresponds to P_i, $1 \leq i \leq r$. Each R_i is effectively context-free and L_4 is slender if and only if all $R_i, 1 \leq i \leq r$, are slender.

8. For each $1 \leq i \leq r, R_i$ has the form (which is computable)

$$R_i = \{\#_1^{c_0 + \sum_{j=1}^{q} c_j n_j} \#_2^{d_0 + \sum_{j=1}^{q} d_j n_j} \mid n_j \geq 0, 1 \leq j \leq q\}, \tag{4}$$

for some $q \geq 0$ and non-negative constant integers $c_j, d_j, 0 \leq j \leq q$. We may suppose that $c_j + d_j > 0$, for any $1 \leq j \leq q$.

Claim 3. R_i is slender if and only if either $q \leq 1$ or else $\dfrac{c_1}{d_1} = \dfrac{c_2}{d_2} = \cdots = \dfrac{c_q}{d_q}$.

Proof of Claim 3. If $q = 0$, then R_i is finite and hence slender. If $q = 1$, then consider a positive integer k. In order to find the number of words of length k in R_i, we have to solve the system

$$\begin{cases} c_0 + c_1 n_1 = n \\ d_0 + d_1 n_1 = m \\ n + m = k \end{cases}$$

in the unknowns n_1, n, and m. Since the determinant of the matrix of the system is $c_1 + d_1 > 0$, it has a unique solution, so R_i is slender.

Suppose now that $q \geq 2$, $c_j = ld_j$, for $l > 0$, $1 \leq j \leq q$. (Notice that the case when either all c_i's are 0 or all d_i's are 0 is clear.) As above, consider the following system of equations in $n_1, n_2, \ldots, n_q, n, m$:

$$\begin{cases} c_0 + \sum_{j=1}^{q} c_j n_j = n \\ d_0 + \sum_{j=1}^{q} d_j n_j = m \\ n + m = k \end{cases}$$

It implies

$$\begin{cases} n - lm = c_0 - ld_0 \\ n + m = k \end{cases}$$

so n and m are uniquely determined. Consequently, R_i is slender.

In order to prove the converse implication, we argue by contradiction. Suppose that R_i is slender, $q \geq 2$, and, without loss of generality, $\frac{c_1}{d_1} \neq \frac{c_2}{d_2}$. Consider an arbitrary fixed positive integer k and the set

$$S = \{(r(c_2 + d_2), (k - r + 1)(c_1 + d_1)) \mid 1 \leq r \leq k\}.$$

For any $(n_1, n_2) \in S$, we have

$$c_0 + c_1 n_1 + c_2 n_2 + d_0 + d_1 n_1 + d_2 n_2 = c_0 + d_0 + (k + 1)(c_1 + d_1)(c_2 + d_2),$$

which is constant with respect to n_1 and n_2. Also, for any two different pairs $(n_1, n_2), (m_1, m_2) \in S$, we have that $c_0 + c_1 n_1 + c_2 n_2 = c_0 + c_1 m_1 + c_2 m_2$ if and only if $\frac{c_1}{d_1} = \frac{c_2}{d_2}$. Therefore, the words $\#_1^{n_1} \#_2^{n_2}$, $(n_1, n_2) \in S$, are all in R_i, different, and of the same length. Since k has been arbitrarily chosen, it follows that R_i is not slender, a contradiction.

9. Since the conditions equivalent to the slenderness of R_i in Claim 3 are trivially checkable, it follows that it is decidable whether or not R_i is slender.

If all such languages R_i are slender, then answer **yes**, otherwise answer **no**. The proof of the theorem is concluded. ∎

5 Other decidability results

We investigate in this section some connected decision problems. We show that a given slender context-free language L can be effectively written as a union of paired loops, even as a disjoint union of paired loops, and that the smallest k such that L is k-slender is computable. Also the smallest number of loops in a representation of L as a union of paired loops is computable. Some other results are inferred.

Theorem 5.1. *A given slender context-free language can be effectively written as a union of paired loops.*

Proof. Notice first that the theorem follows by the decidability result of [Gi] concerning bounded context-free languages (see the first proof of Theorem 4.1) and Theorem 3.4, but we argue also directly, based on the second proof of Theorem 4.1. We use the same notations. To conclude the statement, it is enough to prove the following two facts:

(i) the language R_i at step 8 in Algorithm 4.2 can be effectively written as a union of paired loops, once we know it is slender, and

(ii) having the language L_4 at step 6 given as a union of paired loops, the language L_3 can be effectively written as a union of paired loops.

Let us prove (i). The case when either all c_i's are 0 or all d_i's are 0 is trivial. Suppose this is not the case and put $\frac{c}{d}$ for $\frac{c_1}{d_1} = \frac{c_2}{d_2} = \cdots = \frac{c_q}{d_q}$ in lowest terms. Then $c_j = ce_j, d_j = de_j, 1 \le j \le q$, and (4) becomes

$$R_i = \{\#_1^{c_0}(\#_1^c)^{\sum_{j=1}^q e_j n_j}(\#_2^d)^{\sum_{j=1}^q e_j n_j}\#_2^{d_0} \mid n_j \ge 0, 1 \le j \le q\}.$$

Now, any integer which is a multiple of

$$e = \gcd(e_1, e_2, \ldots, e_q)$$

and is bigger than or equal to

$$E = \frac{1}{e^q}\prod_{j=1}^q e_j$$

can be expressed as a linear combination of $e_1, e_2, \ldots, e_q$ with positive integer coefficients. Therefore

$$R_i = (R_i \cap \Sigma^{\le c_0+d_0+E(c+d)-1}) \cup \{\#_1^{c_0+cE}(\#_1^{ce})^n(\#_2^{de})^n\#_2^{d_0+dE} \mid n \ge 0\}, \qquad (5)$$

and (i) is proved.

For (ii), we consider a gsm h which restores the modifications of g, that is, it replaces each word $\#_1^{|v|}$ by v, introduces a w between the last $\#_1$ and the first $\#_2$, and replaces each word $\#_2^{|x|}$ by x. Due to the form of R_i, see (5), (ii) is proved if h works as intended, that is, $h(L_4) = L_3$.

We show inclusion in both directions. Notice first that both members of the equality to be proved are included in v^*wx^*.

Take first $v^nwx^m \in L_3$. Then $\#_1^{n|v|}\#_2^{m|x|} \in g(L_3) = L_4$ and it is mapped by h into v^nwx^m. Thus $v^nwx^m \in h(L_4)$.

Conversely, consider $v^nwx^m \in h(L_4)$. Then, for some $n_1, m_1, v^nwx^m = v^{n_1}wx^{m_1}$ and $\#_1^{n_1|v|}\#_2^{m_1|x|} \in L_4 = g(L_3)$. Hence $v^{n_1}wx^{m_1} \in L_3$. The equality is proved. ■

Our next goal is to sharpen the result in Theorem 5.1 showing that a given slender context-free language can be effectively written as a disjoint union of paired loops.

We mention that it was proved in [PS1] that any union of paired loops is a disjoint union of paired loops, but no effective procedure to construct a disjoint union was given.

Our proof relies on the following lemma about differences between paired loops.

Lemma 5.2. *The difference of two given paired loops is effectively a disjoint union of paired loops.*

Proof. Consider two paired loops

$$\begin{aligned} A &= \{uv^nwx^ny \mid n \geq 0\}, u, v, w, x, y \in \Sigma^*, \\ B &= \{u'v'^mw'x'^my' \mid m \geq 0\}, u', v', w', x', y' \in \Sigma^*. \end{aligned}$$

We suppose that all words v, x, v', x' are non-empty. The cases when some of them are empty are treated similarly.

The proof of our lemma will be essentially contained in the following three claims.

Denote, for any $z, t \in \Sigma^*$, the primitive root of z by $\rho(z)$ and by $z \sim t$ the conjugacy relation between z and t. Put also

$$\begin{aligned} n_1 &= \min\{n \geq 2 \mid |uv^n| \geq |u'v'v|, |x^ny| \geq |xx'y'|\}, \\ m_1 &= \min\{m \geq 2 \mid |u'v'^m| \geq |uvv'|, |x'^my'| \geq |x'xy|\}, \text{ and} \\ M_1 &= \max(|uv^{n_1}wx^{n_1}y|, |u'v'^{m_1}w'x'^{m_1}y'|). \end{aligned}$$

Claim 1. If $\rho(v) \not\sim \rho(v')$ or $\rho(x) \not\sim \rho(x')$, then $A \cap B \subseteq \Sigma^{\leq M_1 - 1}$.

Proof of Claim 1. Suppose that there is $z \in A \cap B$ such that $|z| \geq M_1$ and put $z = uv^n wx^n y = u'v'^m w'x'^m y'$. By the choice of M_1, it follows that the words v^n and v'^m have a common factor of length at least $|v| + |v'|$. Therefore, by Fine and Wilf's theorem, $\rho(v) \sim \rho(v')$. The same applies for x^n and x'^m, so also $\rho(x) \sim \rho(x')$, a contradiction.

Denote

$$n_0 = \frac{\mathrm{lcm}(|vx|, |v'x'|)}{|vx|}, m_0 = \frac{\mathrm{lcm}(|vx|, |v'x'|)}{|v'x'|}.$$

Claim 2. If $\rho(v) \sim \rho(v'), \rho(x) \sim \rho(x')$, and $\frac{|v|}{|x|} = \frac{|v'|}{|x'|}$, then

(i) $A \cap B$ is infinite if and only if there is $z \in A \cap B$ such that $M_1 \leq |z| < M_1 + \mathrm{lcm}(|vx|, |v'x'|)$,

(ii) if $A \cap B$ is finite, then $A \cap B \subseteq \Sigma^{\leq M_1 - 1}$,

(iii) if $A \cap B$ is infinite and the shortest word in the set $\{z \in A \cap B \mid |z| \geq M_1\}$ is $uv^p wx^p y$, then

$$A - B = \{z \in A - B \mid |z| < M_1\} \cup \bigcup_{i=1}^{n_0 - 1} \{uv^p (v^i)^n w (x^i)^n x^p y \mid n \geq 0\}, \qquad (6)$$

where the right-hand side of (6) is a disjoint union of paired loops.

Proof of Claim 2. (i) and (ii). Suppose that $A \cap B$ is infinite. Then there is $z \in A \cap B, |z| \geq M_1$. If $|z| < M_1 + \mathrm{lcm}(|vx|, |v'x'|)$, then there is nothing to prove. Otherwise, put $z = uv^n wx^n y = u'v'^m w'x'^m y'$.

Now, since $n_0|v| = m_0|v'|$ and $n_0|x| = m_0|x'|$, it follows that

$$uv^{n-n_0} wx^{n-n_0} y = u'v'^{m-m_0} w'x'^{m-m_0} y'.$$

(We have used also the fact that $|z| \geq M_1 + \mathrm{lcm}(|vx|, |v'x'|)$.)

Consequently, we have obtained a word $z' \in A \cap B$ such that $|z'| = |z| - \mathrm{lcm}(|vx|, |v'x'|)$. The procedure continues until $z \in A \cap B$ as required is found.

Conversely, assume that there is $z \in A \cap B$ such that $|z| \geq M_1$ and put $z = uv^n wx^n y = u'v'^m w'x'^m y'$. As above, if follows that

$$uv^{n+n_0} wx^{n+n_0} y = u'v'^{m+m_0} w'x'^{m+m_0} y',$$

so $A \cap B$ is infinite.

(iii). Consider the shortest word of length at least M_1 in $A \cap B$ as $uv^p wx^p y = u'v'^q w'x'^q y'$. Due to length considerations, we have

$$(A \cap B) - \Sigma^{\leq M_1 - 1} \subseteq \{uv^{p+n_0 n} wx^{p+n_0 n} y \mid n \geq 0\} \cap \{u'v'^{q+m_0 m} w'x'^{q+m_0 m} y' \mid m \geq 0\}.$$

From the proof of (i), if follows that we have, in fact,

$$A \cap B = \{z \in A \cap B \mid |z| < M_1\} \cup \{uv^p(v^{n_0})^n w(x^{n_0})^n x^p y \mid n \geq 0\}.$$

Therefore $A - B$ verifies (6) and, moreover, the right-hand side of (6) is a disjoint union of paired loops. The claim is proved.

Consider now

$$\begin{aligned} &n_2 = \min(\{n \geq 2 \mid |uv^n| \geq |u'v'^m w'x'v|, \text{ for } |uv^n wx^n y| = |u'v'^m w'x'^m y'|\}, \\ &M_2 = |uv^{n_2} wx^{n_2} y|. \end{aligned}$$

(M_2 is intended to assure a long enough overlapping of v^n and x'^n in the conditions of the hypothesis of Claim 3 below.)

The next claim covers one of the two symmetrical cases complementary to the one in Claim 2. (The other one is $\frac{|v|}{|x|} < \frac{|v'|}{|x'|}$ and is similar.)

Claim 3. If $\rho(v) \sim \rho(v'), \rho(x) \sim \rho(x')$, and $\frac{|v|}{|x|} > \frac{|v'|}{|x'|}$, then

(i) $A \cap B$ is infinite if and only if there is $z \in A \cap B$ such that $M_2 \leq |z| < M_2 + \mathrm{lcm}(|vx|, |v'x'|)$,

(ii) if $A \cap B$ is finite, then $A \cap B \subseteq \Sigma^{\leq M_2 - 1}$,

(iii) if $A \cap B$ is infinite and the shortest word in the set $\{z \in A \cap B \mid |z| \geq M_2\}$ is $uv^p wx^p y$, then $A - B$ is given by (6) where M_1 is replaced by M_2.

Proof of Claim 3. The proof is similar to the one of Claim 2. We point out the differences only.

In the present hyphotesis, we do not have any more $n_0|v| = m_0|v'|$ or $n_0|x| = m_0|x'|$, so we cannot insert (or delete) separately at the left or at the right of w or w' respectively. The integer M_2 is chosen big enough to assure that v^n and x'^m have a common factor of length at least $|v| + |x'|$. Therefore, again by Fine and Wilf's theorem, we get that now all primitive roots $\rho(v), \rho(v'), \rho(x)$, and $\rho(x')$ are conjugates. (In this case A and B are, in fact, single loops.) Consequently, we can still perform under these conditions the insertion (deletion) of the words v^{n_0}, x^{n_0} and v'^{m_0}, x'^{m_0}, respectively.

Due to some length considerations, (6) is fulfilled with M_2 instead of M_1. Let us further remark that now (6) is a disjoint union of single loops.

Finally, it should be clear from the exhaustive case analysis above that $A - B$ is always a disjoint union of paired loops which can be effectively found.

■

It is now an easy task to give, based on Theorem 5.1 and the precedind lemma, an algorithm for proving the next result and it is left to the reader.

Theorem 5.3. *A given slender context-free language can be effectively written as a disjoint union of paired loops.*

Using Theorem 5.3, we can show now that, for a given slender context-free language L, the maximal number of words of the same length in L is computable. Clearly, this is the smallest k such that L is k-slender, that is

$$\max_{n \geq 0}(\mathrm{card}(\{w \in L \mid |w| = n\})) = \min(\{k \mid L \text{ is } k\text{-slender}\}).$$

Theorem 5.4. *For a given slender context-free language L, the smallest k such that L is k-slender is computable.*

Proof. By Theorem 5.3, we can write L as a disjoint union of paired loops, say

$$L = \bigcup_{i=1}^{m} \{u_i v_i^n w_i x_i^n y_i \mid n \geq 0\}.$$

In order to determine whether or not two paired loops $\{u_i v_i^n w_i x_i^n y_i \mid n \geq 0\}$ and $\{u_j v_j^m w_j x_j^m y_j \mid m \geq 0\}$ have words of the same length, we have to solve the linear diophantine equation in two variables

$$|u_i w_i y_i| + n|v_i x_i| = |u_j w_j y_j| + m|v_j x_j|, \tag{7}$$

which is possible (see, for instance, [Ro]). Moreover, we can find all solutions of (7) and so all lengths common to the considered paired loops. This extends to any subset of the set of paired loops composing L.

Since the paired loops in the representation of L are pairwise disjoint, any two words of the same length in two different paired loops are different. Therefore, the maximal number of words of the same length in L, hence the minimal k such that L is k-slender, can be computed. ∎

Corollary 5.5. *For a given $k \geq 0$, it is decidable whether or not an arbitrary context-free language is k-slender.*

The last results of this section concern the number of loops in the representation of a context-free language as a union of paired loops.

Theorem 5.6. *For a given slender context-free language L, the smallest m such that L can be written as a union of m paired loops is computable.*

Proof. Suppose that m is the smallest integer such that L can be written as a union of m paired loops, say

$$L = \bigcup_{i=1}^{m} \{u_i v_i^n w_i x_i^n y_i \mid n \geq 0\}.$$

Using techniques similar to the ones in the proofs of Lemmas 3.2 and 3.3, it can be shown that there is a representation of L as a union of paired loops which satisfies the length conditions in Theorem 3.4. The idea is to reduce first the lengths of the pumping words and then, as in Lemmas 3.2 and 3.3, the lengths of the others. The details are left to the reader.

Then, from all representations of L as in Theorem 3.4, we choose one with minimal number of loops. Its number of loops is m. ∎

Corollary 5.7. *For a given $m \geq 0$, it is decidable whether or not an arbitrary (slender) context-free language can be written as a union of m paired loops.*

Remark. Notice that, as soon as the constant m in Theorem 5.6 for a slender context-free language L is at least two, there may be infinitely many representations of L as a union of m paired loops, as seen in the next example.

Consider the regular language

$$L = \{a\} \cup \{a^n \mid n \geq 3\}.$$

L is 1-slender and it cannot be written with less than two paired loops. But

$$L = \{a(a^k)^n \mid n \geq 0\} \cup \{a^n \mid n \geq 3\},$$

for any $k \geq 2$, so there are infinitely many possibilities of decomposing L into two paired loops.

Problem. We did not investigate the number of paired loops in decompositions as disjoint unions of paired loops. Do Theorem 5.6 and Corollary 5.7 hold in this case? Also, what about the previous remark?

6 Undecidability

The result in Theorem 4.1 is no longer true when we consider complements or intersections of context-free languages; in both cases we have undecidable problems. This fact was proved in [PS1]. We give here a much simpler argument which works in both cases.

Theorem 6.1. *It is undecidable whether or not the complement of an arbitrary context-free language is slender.*

Proof. Consider an arbitrary instance of the Post Correspondence Problem,

$$\{(x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n)\}, \tag{8}$$

$n \geq 1, x_i, y_i \in \{a, b\}^+, 1 \leq i \leq n$, and two new letters $c, \# \notin \{a, b\}$. Denote $x_{n+1} = \# = y_{n+1}$ and construct the well-known languages (see, for instance, [Gi])

$$\begin{aligned}
&L_x = \{ba^{i_1}ba^{i_2}\ldots ba^{i_k}cx_{i_k}x_{i_{k-1}}\ldots x_1 \mid k \geq 0, 1 \leq i_j \leq n+1, 1 \leq j \leq k\},\\
&L_y = \{ba^{i_1}ba^{i_2}\ldots ba^{i_k}cy_{i_k}y_{i_{k-1}}\ldots y_1 \mid k \geq 0, 1 \leq i_j \leq n+1, 1 \leq j \leq k\},\\
&L_{x,y} = L_x c L_y^R,\\
&L_R = \{w_1cw_2cw_2^Rcw_1^R \mid w_1, w_2 \in \{a, b, \#\}^*\},\\
&L_{PCP} = L_R \cap L_{x,y},
\end{aligned}$$

where w^R denotes the reversal of w and $L^R = \{w^R \mid w \in L\}$. It is known that the complement $\overline{L_{PCP}}$ of L_{PCP} is context-free.

L_{PCP} is the set of all solutions of the following instance of the Post Correspondence Problem

$$\{(x_1, x_2, \ldots, x_{n+1}), (y_1, y_2, \ldots, y_{n+1})\}. \tag{9}$$

Since $x_{n+1} = y_{n+1} = \#$, all words in $\#^*$ are solutions of (9). Moreover, since $\# \notin \{a, b\}$, all solutions of (9) are catenations of solutions of (8) and words in $\#^*$. It follows that L_{PCP} is slender if and only if the set of solutions of (9) is $\#^*$, that is, if and only if (8) has no solution, which is undecidable. ■

Theorem 6.2. *It is undecidable whether or not the intersection of two arbitrary context-free languages is slender.*

Proof. In the notations of the preceding proof, L_R and $L_{x,y}$ are context-free and $L_R \cap L_{x,y}$ is slender if and only if (8) has no solution. ■

References

[ADPS] M. Andraşiu, J. Dassow, Gh. Păun and A. Salomaa, Language-theoretic problems arising from Richelieu cryptosystems, *Theoret. Comput. Sci.* **116** (1993) 339–357.

[CY] C. Calude and S. Yu, Language-theoretic complexity of disjunctive sequences, submitted.

[DPS] J. Dassow, G. Păun and A. Salomaa, On thinness and slenderness of L languages, *Bull. EATCS* **49** (1993) 152–158.

[Gi] S. Ginsburg, *The Mathematical Theory of Context-Free Languages* (McGraw-Hill, New York, 1966).

[HI] T. Harju and L. Ilie, Languages obtained from infinite words, submitted.

[Ho1] J. Honkala, On Parikh slender languages and power series, *J. Comp. Syst. Sci.* **52** (1996) 185–190.

[Ho2] J. Honkala, On images of algebraic series, *J. Univ. Comp. Sci.* **2**(4) (1996) 217–223.

[Ho3] J. Honkala, A decision method for Parikh slenderness of context-free languages, *Discrete Appl. Math.* **73** (1997) 1–4.

[HU] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation* (Addison-Wesley, Reading, Mass., 1979).

[Il] L. Ilie, On a conjecture about slender context-free languages, *Theoret. Comput. Sci.* **132** (1994) 427–434.

[Lo] M. Lothaire, *Combinatorics on Words* (Addison-Wesley, Reading, Mass., 1983).

[NS] T. Nishida and A. Salomaa, Slender 0L languages, *Theoret. Comput. Sci.* **158** (1996) 161–176.

[PS1] Gh. Păun and A. Salomaa, Thin and slender languages, *Discrete Appl. Math.* **61** (1995) 257–270.

[PS2] Gh. Păun and A. Salomaa, Decision problems concerning the thinness of D0L languages, *Bull. EATCS* **46** (1992) 171–181.

[PS3] Gh. Păun and A. Salomaa, Closure properties of slender languages, *Theoret. Comput. Sci.* **120** (1993) 293–301.

[Ra] D. Raz, Length considerations in context-free languages, *Theoret. Comput. Sci.*, to appear.

[Ro] K. Rosen, *Elementary number theory and its applications* (Addison-Wesley, Reading, Mass., 1988).

[Sa] A. Salomaa, *Formal Languages* (Academic Press, New York, 1973).

[SYZS] A. Szilard, S. Yu, Z. Zhang and J. Shallit, Characterizing regular languages with polynomial densities, *Proc. 17th MFCS*, LNCS 629 (1992) 494–503.

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi

University of Turku
- **Department of Mathematical Sciences**

Åbo Akademi University
- **Department of Computer Science**
- **Institute for Advanced Management Systems Research**

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**