Electronic Notes in Theoretical Computer Science 53 (2002)
URL: http://www.elsevier.nl/locate/entcs/volume53.html 9 pages

Chomsky-Schützenberger Type Characterizations of Poly-Slender and Parikh Slender Context-Free Languages [1]

Masami Ito [2]

Department of Mathematics
Faculty of Science,
Kyoto Sangyo University
Kyoto 603, Japan

Carlos Martín-Vide [3]

Research Group in Mathematical Linguistics
Rovira i Virgili University
Pça. Imperial Tàrraco 1, 43005 Tarragona, Spain

Victor Mitrana [4]

Faculty of Mathematics
University of Bucharest
Str. Academiei 14, 70109, Bucharest, Romania

Abstract

In this paper we propose a Chomsky-Schützenberger type characterization of k-poly-slender context-free languages, as the homomorphical image of an intersection of a Dyck language and a $(2k+1)$-poly-slender regular language. A stronger result is provided, namely the homomorphism and the Dyck language are determined irrespective of the given poly-slender context-free language, when considering the family of all poly-slender context-free languages. A similar characterization is obtained for Parikh slender context-free languages.

[1] Work supported by the Grants-in Aid for Scientific Research No. 10440034, Japan Society for the Promotion of Sciences and the Dirección General de Enseñanza Superior e Investigación Cientifica, SB 97–00110508
[2] Email: ito@ksuvx0.kyoto-su.ac.jp
[3] Email: cmv@astor.urv.es
[4] Email: mitrana@funinf.cs.unibuc.ro

1 Introduction

Considerations concerning the length of words in languages have been in the focus of interest of many researchers for a long time. Thus, in [5] (see also [3,14]) for a given language L, one associates with each nonnegative integer n, the number of words of length at most n in L, denoted by $\#_L(n)$. Then one defines the "generating function" of L by

$$f_L(z) = \sum_{n>0} \#_L(n) z^n$$

and proves that this function is an algebraic function over the field of rationals, provided that L is an unambiguous context-free language. Moreover, the transcendentality of generating function is the main tool in the study of inherent ambiguity of context-free languages in [7].

Also deep results involving words length considerations in Lindenmayer systems have been reported in a series of papers [6,9,15,20].

Motivated by some problems arising in cryptography, [1] introduces *slender languages* as being those languages which contain at most a constant number of words of any length. Roughly speaking, the following formal language approach of the Richelieu cryptosystem is considered in [1]: a plaintext is encripted using an encription key of length n yielding a cryptotext of the same length of the encription key. If the legal receiver knows the key, then it can easily recover the original text. The problem appears when transmitting all possible keys to the legal receiver. In order to decrease the recovering complexity, the keys should come from a slender language such that the receiver may test a limited number of keys for each cryptotext. The reader interestead in further details may consult [1].

Characterizations of slender regular and context-free languages have been reported in [18,21] and [11], respectively. Many aspects of different types of slender languages have been investigated in a series of papers [16,17]

Raz extends in [19] the concept of slenderness to poly-slenderness, a poly-slender language being a language whose number of words of length n is bounded by the value of a polynomial in n for all $n \geq 0$. A characterization of poly-slender regular languages has been provided in [21] while a characterization of poly-slender context-free languages has been proposed in [12].

In this paper we propose a characterization of k-poly-slender context-free languages in the Chomsky-Schützenberger type [5], as the homomorphical image of an intersection of a Dyck language with a $(2k+1)$-poly-slender regular language. When dropping k, a stronger result is provided, namely the homomorphism and the Dyck language are determined irrespective of the given poly-slender context-free language. Moreover, the converse assertion is also true. A similar characterization is obtained for

Parikh slender context-free languages, namely each Parikh slender context-free language is the homomorphical image of an intersection of a Dyck language with a Parikh slender regular language, but the converse does not hold.

2 Basic Definitions and Previous Results

The set of all words over the alphabet V is denoted by V^*, while $V^+ = V^* \setminus \{\varepsilon\}$, where ε is the empty word. For each word $x \in V^+$, whose length is denoted by $|x|$, $|x|_a$ delivers the number of occurrences of the letter a in x. Any subset of V^* is called a language over V. For a nonnegative integer k a language $L \subseteq V^*$ is called *k-bounded* if there are the words $x_1, x_2, \ldots, x_k$ over V^* such that $L \subseteq \{x_1\}^*\{x_2\}^* \ldots \{x_k\}^*$. L is bounded if it is k-bounded for some $k \geq 0$. For basic notions and results of combinatorics on words the reader is refered to [4,13].

Given a language L the mapping $\#_L(n)$ gives the number of all words of length n in L for every n. Formally, $\#_L(n) = \text{card}(\{x \in L \mid |x| = n\})$. For a nonnegative integer k, the language L is said to be *k-poly-slender* if $\#_L(n) \in \mathcal{O}(n^k)$. Furthermore, L is *poly-slender* if it is k-poly-slender for some $k \geq 0$. Any 0-poly-slender language is called also *slender*.

The following result appears in [19].

Theorem 2.1 *The class of poly-slender context-free languages is exactly the class of bounded context-free languages.*

Moreover, a direct consequence of the proof of Proposition 1 in [19] useful in the sequel is

Proposition 2.2 *A k-bounded language is (k-1)-poly-slender for any* $k \geq 1$.

For an alphabet V we denote by $\bar{V} = \{\bar{a} \mid a \in V\}$. The language over $(V \cup \bar{V})^*$ generated by the context-free grammar having the set of rules $\{S \to SS, S \to \varepsilon\} \cup \{S \to aS\bar{a} \mid a \in V\}$ is called the Dyck language over V denoted $\mathcal{D}_V$. We write $\bar{w} = \bar{a}_1\bar{a}_2 \ldots \bar{a}_n$ for $w = a_1a_2 \ldots a_n$, $a_i \in V$, $1 \leq i \leq n$.

Now we recall the definition of Dyck loops following [12]. Let $V_k = \{a_1, a_2, \ldots, a_k\}$ and $z = z_1z_2 \ldots z_{2k} \in \mathcal{D}_{V_k}$ be a word with $z_j \in V_k \cup \bar{V}_k$, $1 \leq j \leq 2k$, $|z|_{a_i} = |z|_{\bar{a}_i} = 1$ for all $1 \leq i \leq k$, and U be an alphabet disjoint from $V_k \cup \bar{V}_k$. For a homomorphism $h : (U \cup V_k \cup \bar{V}_k)^* \longrightarrow U^*$ and some integers $n_i \geq 0$, $1 \leq i \leq k$, we define the homomorphism

$$h_{n_1,n_2,\ldots,n_k} : (U \cup V_k \cup \bar{V}_k)^* \longrightarrow U^*$$

by

$$\begin{aligned} h_{n_1,n_2,\ldots,n_k}(a) &= a \text{ for all } a \in U, \\ h_{n_1,n_2,\ldots,n_k}(a_i) &= u_i^{n_i},\ 1 \leq i \leq k \\ h_{n_1,n_2,\ldots,n_k}(\bar{a}_i) &= v_i^{n_i},\ 1 \leq i \leq k, \end{aligned}$$

for some $u_i, v_i \in U^*$, $1 \leq i \leq k$. Then for some words $w_0, w_1, \ldots, w_{2k} \in U^*$ the language

$$D = \{h_{n_1,n_2,\ldots,n_k}(w_0 z_1 w_1 z_2 w_2 \ldots z_{2k} w_k) \mid n_i \geq 0,\ 1 \leq i \leq k\}$$

is called a *k-Dyck loop*. Clearly, the same set D can be obtained by changing the homomorphism $h_{1,1,\ldots,1}$ or the Dyck word z. We say that z is a *underlying word* of D and $h = h_{1,1,\ldots,1}$ is a *underlying homomorphism* of D.

According to the above notations, the main results of [11] and [12] can be written as

Theorem 2.3 *A context-free language is k-poly-slender if and only if it is a finite union of (k+1)-Dyck loops.*

3 Characterizations of Poly-Slender Context-Free Languages

We say that the class of languages $\mathcal{L}_1$ is weakly characterized in the Chomsky-Schützenberger type with the class of languages $\mathcal{L}_2$ if for any language $L_1 \in \mathcal{L}_1$ over an alphabet V, there exist an alphabet U, a language $L_2 \in \mathcal{L}_2$ over U and a homomorphism h from U into V with $L_1 = h(\mathcal{D}_U \cap L_2)$. Further, if $h(\mathcal{D}_U \cap L_2) \in \mathcal{L}_1$ for any homomorphism h from U into an alphabet V, and any language $L_2 \in \mathcal{L}_2$ over U, then we say that the class of languages $\mathcal{L}_1$ is strongly characterized in the Chomsky-Schützenberger type with the class of languages $\mathcal{L}_2$.

Such a strong characterization is well known for one of the most investigated families of languages in the formal language theory [5].

Theorem 3.1 *The family of context-free languages is strongly characterized in the Chomsky-Schützenberger type with the family of regular languages.*

Now we are ready to prove the main result of this section.

Theorem 3.2 *For each $k \geq 0$, the family of k-poly-slender context-free languages is weakly characterized in the Chomsky-Schützenberger type with the family of (2k+1)-poly-slender regular languages.*

Proof. Let V be a given alphabet and L be a k-poly-slender context-free language over V. By Theorem 2.3, L can be written as a finite union of $(k+1)$-Dyck loops. Let us denote these $(k+1)$-Dyck loops by $D_1, D_2, \ldots, D_m$ for some $m \geq 1$. Further, assume that

$$D_i = \{h^{(D_i)}_{n_1,n_2,\ldots,n_{k+1}}(w_0^{(D_i)} z_1^{(D_i)} w_1^{(D_i)} z_2^{(D_i)} w_2^{(D_i)} \ldots z_{2k+2}^{(D_i)} w_{2k+2}^{(D_i)}) \mid n_j \geq 0,\ 1 \leq j \leq k+1\},$$

with $w_0^{(D_i)}, w_1^{(D_i)}, \ldots, w_{2k+2}^{(D_i)} \in V^*$ for any $1 \leq i \leq m$, and

$$V_{k+1}^{(D_i)} \cup \bar{V}_{k+1}^{(D_i)} = \{z_1^{(D_i)}, z_2^{(D_i)}, \ldots z_{2k+2}^{(D_i)}\},$$

for all $1 \leq i \leq m$. Moreover, all alphabets $V_{k+1}^{(D_i)}$ may be supposed mutually disjoint.

We consider the alphabet

$$U = \bigcup_{i=1}^{m} V_{k+1}^{(D_i)} \cup \bigcup_{i=1}^{m} \{c_0^{(i)}, c_1^{(i)}, \ldots, c_{2k+2}^{(i)}\},$$

where $c_j^{(i)}$, $1 \leq i \leq m, 0 \leq j \leq 2k+2$, are new letters not in $\bigcup_{i=1}^{m}(V_{k+1}^{(D_i)} \cup \bar{V}_{k+1}^{(D_i)})$.

Now we define the regular language

$$R = \bigcup_{i=1}^{m} \{c_0^{(i)}\bar{c}_0^{(i)}\}\{z_1^{(D_i)}\}^*\{c_1^{(i)}\bar{c}_1^{(i)}\}\{z_2^{(D_i)}\}^*\{c_2^{(i)}\bar{c}_2^{(i)}\}\ldots\{z_{2k+2}^{(D_i)}\}^*\{c_{2k+2}^{(i)}\bar{c}_{2k+2}^{(i)}\},$$

and the homomorphism $g : (U \cup \bar{U})^* \longrightarrow V^*$ defined by

$$\begin{aligned} g(c_j^{(i)}) &= w_j^{(D_i)}, \\ g(\bar{c}_j^{(i)}) &= \varepsilon, \\ g(z_j^{(D_i)}) &= h_{1,1,\ldots,1}^{(D_i)}(z_j^{(D_i)}), \end{aligned}$$

for all $1 \leq i \leq m$ and $1 \leq j \leq 2k+2$. Clearly, for each $1 \leq i \leq m$ the following relation holds:

$$\begin{aligned} D_i = g(\mathcal{D}_U \cap \{c_0^{(i)}\bar{c}_0^{(i)}\}\{z_1^{(D_i)}\}^*\{c_1^{(i)}\bar{c}_1^{(i)}\}\{z_2^{(D_i)}\}^*\{c_2^{(i)}\bar{c}_2^{(i)}\}\ldots \\ \{z_{2k+2}^{(D_i)}\}^*\{c_{2k+2}^{(i)}\bar{c}_{2k+2}^{(i)}\}). \end{aligned}$$

Consequently, $L = g(\mathcal{D}_U \cap R)$ holds as well. By Proposition 2.2 it follows that the regular language R from above is $(2k+1)$-poly-slender which concludes the proof. □

Since the homomorphical image of every k-Dyck loop is a k-Dyck loop, it follows that all families of k-poly-slender context-free languages, $k \geq 0$, are closed under homomorphisms. We do not know whether the family of $(2k+1)$-poly-slender regular languages in the above characterization can be replaced by the family of k-poly-slender regular languages. This would imply a strong characterization of the class of k-poly-slender context-free languages for each $k \geq 0$. Another related problem regards the minimal value of $m \geq k$ such that each class of k-poly-slender languages can be weakly characterized in the Chomsky-Schützenberger type with the family of m-poly-slender regular languages. We conjecture that this minimal value is exactly $2k+1$, so that the previous result is optimal in this respect.

Several decidability problems, which remain *open* in this paper, naturally arise if our conjecture is false:

1. Given a k-poly-slender language, is the minimal value of m from above computable?

2. Are there k-poly-slender languages, which are not $(k-1)$-poly-slender, such that the aforementioned minimal value is k? In the affirmative, can we decide whether a given k-poly-slender language has this property, namely does it admit a Chomsky-Schützenberger characterization with a k-poly-slender regular language?

However, it is obvious that

Theorem 3.3 *The class of poly-slender context-free languages is strongly characterized with the class of poly-slender regular languages.*

As one can easily see, in the above Chomsky-Schützenberger type characterizations, the alphabet of the regular and the Dyck language, together with the homomorphism depend on the given poly-slender context-free language. This result can be still improved such that all the aforementioned objects can be determined irrespective of the given poly-slender context-free language. This improvment is based on the same characterization of poly-slender context-free languages, by [8] via Theorem 2.1.

A family of languages $\mathcal{L}$ is closed under paired loop if for any language $L \in \mathcal{L}$ and any two words x, y, the language $\bigcup_{n\geq 0}\{x^n\}L\{y^n\}$ is in $\mathcal{L}$.

Theorem 3.4 *The family of poly-slender context-free languages is the smallest family which contains all singleton languages and is closed under the following operations:*

(i) union,

(ii) concatenation,

(iii) paired loop.

We call the minimal number of the above operations needed for obtaining a poly-slender context-free language L by the *order* of L. Now we state

Theorem 3.5 *Let V be an alphabet; a new alphabet U and a homomorphism h from U^* into V^* can be determined such that for each poly-slender context-free language L over V there exists a poly-slender regular language R over U such that $L = h(\mathcal{D}_U \cap R)$. Moreover, $h(\mathcal{D}_U \cap R)$ is a poly-slender context-free for any homomorphism h from U^* into V^* and any poly-slender regular language R.*

Proof. For the given alphabet V we construct the alphabet $W = V \cup \{\#\} \cup V'$, where $V' = \{a' \mid a \in V\}$, and $\#$ is a new symbol. For $w = a_1a_2\ldots a_n$, $a_i \in V$, $1 \leq i \leq n$, we denote by w' the word $a'_1a'_2\ldots a'_n$. Put $U = W \cup \bar{W}$.

The homomorphism h is defined by

$$h(a) = h(\bar{a}') = a,\ a \in V, \text{ and } h(a) = \varepsilon,\ a \in U \setminus (V \cup \bar{V}').$$

Let L be a poly-slender context-free language over V. We prove the statement by induction on the order of L, say k. If $k = 0$, then L is a singleton

language, $L = \{x\}$. One takes the singleton regular language $R = \{x\bar{x}^R\}$, where x^R delivers the mirror image of x. Clearly, $L = h(\mathcal{D}_U \cap R)$.

Let L_1, L_2 be two poly-slender context-free languages over V of orders smaller than k. By the induction hypothesis,

$$L_i = h(\mathcal{D}_U \cap R_i), i = 1, 2.$$

If $L = L_1 \cup L_2$, then $L = h(\mathcal{D}_U \cap (R_1 \cap R_2))$ trivially holds.
If $L = L_1 L_2$, then $L = h(\mathcal{D}_U \cap (R_1\{\#\}R_2\{\bar{\#}\}))$ holds.
If $L = \bigcup_{i\geq 0}\{x^i\}L_1\{y^i\}$, then $L = h(\mathcal{D}_U \cap (\{xy'^R\}^* R_1\{\bar{y}'\bar{x}^R\}^*))$ holds.
Consequently, the first part of the theorem is proved. The second part is obvious. □

4 Characterizations of Parikh Slender Languages

Instead of words of length n, in [10] one counts the number of words with the same Parikh vector. In particular, a very nice and simple proof of the result of Autebert, Flajolet, and Gabarro [2] concerning the inherent ambiguity of coprefix languages of infinite words is obtained.

By definition, if $V = \{a_1, a_2, \ldots, a_m\}$ is an alphabet and $w \in V^*$ is a word, the Parikh vector $\psi_V(w)$ is defined by

$$\psi_V(w) = (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_m}).$$

A language $L \subseteq V^*$ is termed *Parikh slender* if there is a positive integer k such that for each vector of positive integers $(i_1, i_2, \ldots, i_m)$ there are at most k words in L with the Parikh vector $(i_1, i_2, \ldots, i_m)$. A natural question asks whether or not the family of Parikh slender context-free languages admits a Chomsky-Schützenberger type characterization. As we have seen, this problem has been left open for slender languages in the previous section.

A language $L \subseteq V*$ is said to be a *multiple loop language* if there exist $k \geq 1$ and the strings $u_1, v_1, u_2, v_2, \ldots, u_k, v_k, u_{k+1} \in V^*$ such that the following two conditions are satisfied:

(i) $L = u_1 v_1^* u_2 v_2^* \ldots u_k v_k^* u_{k+1}$
(ii) $\psi_V(v_1), \psi_V(v_2), \ldots, \psi_V(v_k)$ are linearly independent over $\mathbf{Q}$.

The following characterization of Parikh slender regular languages appears in [10].

Theorem 4.1 *A regular language is Parikh slender if and only if it is a finite union of disjoint multiple loop languages.*

It is quite easy to notice that each Parikh slender context-free language is poly-slender and the regular language from the proof of Theorem 3.2 is Parikh slender in accordance with Theorem 4.1. Therefore, we have

just got a Chomsky-Schützenberger type characterization of Parikh slender context-free languages.

Theorem 4.2 *The family of Parikh slender context-free languages is weakly characterized with the family of Parikh slender regular languages.*

However, the family of Parikh slender languages cannot be strongly characterized with the family of Parikh slender regular languages as shown by the following example. We take the Dyck language $\mathcal{D}_V$ over the alphabet $V = \{a, b, c\}$, the Parikh slender regular language $R = a^+\bar{a}^+b^+\bar{b}^+c^+\bar{c}^+$ and the homomorphism h from $(V \cup \bar{V})^*$ into $\{a, b\}^*$ defined by $h(a) = h(c) = a$, $h(b) = b$, and $h(\bar{x}) = \varepsilon$, $x \in V$. We have

$$h(\mathcal{D}_V \cap R) = a^+b^+a^+$$

which is not a Parikh slender language.

References

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

[2] Autebert, J.-M., P. Flajolet, and J. Gabarro, *Prefixes of infinite words and ambiguous context-free languages*, Inform. Process. Letters **25** (1987), 211–216.

[3] Berstel, J. *Sur la densité asymptotique de langages formels*. In M. Nivat, ed.,"Automata, Languages and Programming" North-Holland, 1972, 345–358.

[4] Choffrut, C., and J. Karhumäki, *Combinatorics on words*. In G. Rozenberg and A. Salomaa, eds. "The Handbook of Formal Languages", vol. I, Springer-Verlag, Berlin, 1997, 329–437.

[5] Chomsky, N., and M. P. Schützenberger, *The algebraic theory of context-free languages*. In P. Bradford and D. Hirschberg, eds. "Computer Programming and Formal Systems", North-Holland, Amsterdam, 1963, 118–161.

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

[7] Flajolet, P. *Analytic models and ambiguity of context-free languages*, Theoret. Comput. Sci. **49** (1987), 283–309.

[8] Ginsburg, S. "The Mathematical Theory of Context-Free Languages", McGraw-Hill, New York, 1966.

[9] Honkala, J. *Decision problems concerning thiness and slenderness of formal languages*, Acta Inform. **35** (1998), 625–636.

[10] Honkala, J. *On Parikh slender languages and power series*, J. Comput. System Sci. **52** (1996), 185–190.

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

[12] Ilie, L., G. Rozenberg, and A. Salomaa, *A characterization of poly-slender context-free languages*, Theor. Inform. Appl. **34** (2000), 77–86.

[13] Lothaire, M. "Combinatorics on Words", Addison-Wesley, Reading, MA., 1983.

[14] Maurer, H.A., and M. Nivat, *Rational bijection of rational sets*, Acta Inform. **13** (1980), 365–378.

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

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

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

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

[19] Raz, D. *Length considerations in context-free languages*, Theoret. Comput. Sci. **183** (1997), 21–32.

[20] Rozenberg, G., and A. Salomaa, "The Mathematical Theory of L Systems", Academic Press, New York, 1980.

[21] Shallit, J. *Numeration systems, linear recurrences, and regular sets*, Inform. and Comput. **113** (1994), 331–347.