Acta Cybernetica 15 (2001) 25-32.

A Chomsky-Schützenberger-Stanley Type Characterization of the Class of Slender Context-Free Languages*

Pál Dömösi † Satoshi Okawa ‡

Abstract

Slender context-free languages have a complete algebraic characterization by L. Ilie in [13]. In this paper we give another characterization of this class of languages. In particular, using linear Dyck languages instead of unrestricted ones, we obtain a Chomsky-Schützenberger-Stanley type characterization of slender context-free languages.

1 Introduction

We consider slender languages, that is, languages for which the number of words of the same length is bounded by a constant. As proved in [16], the slender regular languages are exactly the disjoint unions of single loops, that is, disjoint finite unions of sets of the form uv^*w. A similar characterization holds for slender context-free languages as disjoint unions of paired loops, that is, finite unions of sets of the form $\{uv^nwx^ny \mid n \geq 0\}$ [13, 17].

The characterization of language classes by algebraic operations is one of the most important issues in formal language theory. Chomsky-Schützenberger-Stanley's characterization [1, 2, 20, 21] for the class of context-free languages was the first well-known result in this direction and is stated as follows: For any context-free language L, there exists a regular language R such that $L = h(R \cap D)$ where D is a Dyck language and h is a homomorphism. Moreover, it is clear that $h(R \cap D)$

*This work has been supported by the joint project "Automata & Formal Languages" of the Hungarian Academy of Sciences and Japanese Society for Promotion of Science (No. 15) "Automata & Formal Languages"and by the Hungarian National Foundation for Scientific Research (OTKA T030140).

†Institute of Mathematics and Informatics, Debrecen University, Debrecen, Egyetem tér 1, H-4032, Hungary, e-mail: domosi@math.klte.hu

‡Faculty of Computer Science and Engineering, the University of Aizu, Aizu-Wakamatsu, 965-8580, Japan, e-mail: okawa@u-aizu.ac.jp

is a context-free language for each regular language R by the closure properties of the class of context-free languages. A refinement of this classical result is shown in [11].

For recursively enumerable languages, a Chomsky-Schützenberger-Stanley type characterization is given in [10]. It is also proved [15] that there exists no characterization of this type for context-sensitive languages. (See some other types of homomorphic characterizations of recursively enumerable languages in [3, 4, 5, 7, 9].)

In this paper we investigate a characterization of Chomsky-Schützenberger-Stanley type for slender context-free languages.

However, a Chomsky-Schützenberger-Stanley type characterization of the class of slender context-free languages is almost meaningless, because a slender context-free language is linear but a Dyck language is not linear. If we use a Dyck language for characterization, then, in fact, we use complex languages to characterize simpler ones. We consider another characterization which is similar to Chomsky-Schützenberger-Stanley's one.

This paper is organized as follows. In Section 2, we introduce some fundamental notions, notations, definitions of slender languages, and the loop characterization results for slender languages. In Section 3, we give our main theorem, which offers a Chomsky - Stanley type characterization of the class of slender context-free languages. Section 4 gives some concluding remarks.

2 Preliminaries

For all notions and notations not defined here, see [6, 8, 12, 14, 18, 19]. By an *alphabet* Σ we mean a finite nonvoid set. An element of Σ is called a *letter*. A *word* over Σ is a finite sequence of elements in Σ. For a word w, the *length* $|w|$ is the number of letters in w, where each letter is counted as many times as it occurs. The set of all the words over Σ is denoted by Σ^*. In particular, λ is a word in Σ^* and is called the *empty word*. Thus $|\lambda| = 0$. If u and v are words over an alphabet Σ, then their *catenation* uv is also a word over Σ. It is clear that λ acts as identity for this operation, that is, for every word u over Σ, $u\lambda = \lambda u = u$. Therefore, Σ^* becomes a free monoid with catenation as the multiplication and λ as the identity, and $\Sigma^+ = \Sigma^* \setminus \{\lambda\}$ is a semigroup.

Any subset of Σ^* is referred to as a *language* over Σ.

Now we define slender languages. A language $L \subseteq \Sigma^*$ is said to be *k-slender* if $card\{w \in L \mid |w| = n\} \leq k$ for every $n \geq 0$. A language is *slender* if it is k-slender for some positive integer k. In particular, a 1-slender language is called a *thin* language.

For the loop characterization of slenderness, some notation and definitions are introduced. For a word u, setting $u^0 = \lambda$ and $u^n = u^{n-1}u$ for $n > 0$, we define u^* and u^+ as usual, by $u^* = \{u^n \mid n \geq 0\}$ and $u^+ = u^* \setminus \{\lambda\}$.

A language L is said to be a *union of single loops* (or, in short, USL) if for some

positive integer k and some words $u_i, v_i, w_i, 1 \leq i \leq k$,

$$(*) \quad L = \bigcup_{i=1}^{k} \{u_i\}\{v_i\}^*\{w_i\}.$$

A language L is called a *union of paired loops* (or UPL, in short) if for some positive k and some words $u_i, v_i, w_i, x_i, y_i, 1 \leq i \leq k$,

$$(**) \quad L = \bigcup_{i=1}^{k} \{u_i v_i^n w_i x_i^n y_i \mid n \geq 0\}.$$

A USL language L is said to be a *disjoint union of single loops* (DUSL, in short) if the sets in the union $(*)$ are pairwise disjoint. The notion of a *disjoint union of paired loops* (DUPL) is defined analogously considering $(**)$.

A *grammar* is an ordered quadruple $G = (N, \Sigma, S, P)$ where N and Σ are disjoint alphabets of *variables* and *terminals*, respectively, the *start symbol* $S \in N$, and P is a finite set of ordered pairs (α, β) called *productions*, such that β is a word over the alphabet $N \cup \Sigma$ and α is a word over $N \cup \Sigma$ containing at least one letter of N. Usually, a production is written $\alpha \to \beta$ instead of (α, β).

For a word ξ over $N \cup \Sigma$, if ξ is decomposed into

$$\xi = \xi_1 \alpha \xi_2$$

and $\alpha \to \beta$ is a production in P, then $\alpha \to \beta$ is applicable to ξ and the result of the application is a word $\eta = \xi_1 \beta \xi_2$. We say that ξ *derives directly* η, and write $\xi \Rightarrow \eta$.

The *language generated by a grammar* $G = (N, \Sigma, S, P)$ is the set $L(G) = \{w \mid w \in \Sigma^*$ and $S \Rightarrow^* w\}$, where $\Rightarrow^*$ denotes the reflexive and transitive closure of $\Rightarrow$.

If $\alpha \to \beta \in P$ implies $\alpha \in N$ then G is called *context-free*. A context-free grammar is said to be *linear* if for every production $\alpha \to \beta \in P$, $\beta \in \Sigma^* N \Sigma^* \cup \Sigma^*$ A linear grammar is called *right-linear* or *regular* if for every production $\alpha \to \beta \in P$, $\beta \in \Sigma^* N \cup \Sigma^*$. $L \subseteq \Sigma^*$ is a *regular (linear, context-free) language* if we have $L = L(G)$ for some regular (linear, context-free) grammar G.

Let $G = (N, \Sigma, S, P)$ be a context-free grammar with $N = \{S\}$, $\Sigma = \{a_i, a_i' \mid i = 1, \ldots, n\}$, and $P = \{S \to \lambda, S \to SS\} \cup \{S \to a_i S a_i', \mid i = 1, \ldots, n\}$. We say that G and $L(G)$ are a *Dyck grammar* and the *Dyck language over the 2n-letter alphabet* Σ, respectively. Furthermore, if the set of productions of a grammar $G_{\mathcal{L}}$ is $P_{\mathcal{L}} = \{S \to \lambda\} \cup \{S \to a_i S a_i', \mid i = 1, \ldots, n\}$, then $G_{\mathcal{L}}$ is called a *linear Dyck grammar* and its language $L(G_{\mathcal{L}})$ is called a *linear Dyck language*.

We shall use the following well-known results about slender languages.

Theorem 2.1. [16] *The following conditions are equivalent for a language L:*

(i) L is regular and slender.

(ii) L is USL.

(iii) L is DUSL. □

Theorem 2.2. [16] *Every UPL language is DUPL, slender, linear and unambiguous.* □

Theorem 2.3.[13, 17] *Every slender context-free language is UPL.* □

We have the following direct consequence of Theorems 2.2 and 2.3:

Proposition 2.4. *The class of slender linear languages coincides with the class of slender context-fee languages. In addition, the class of slender context-free languages contains only unambiguous languages.* □

3 Results

As stated in the Introduction, a Chomsky-Schützenberger-Stanley type characterization of the class of slender context-free languages is almost meaningless, because a slender context-free language is linear but a Dyck language is not linear. Therefore, we use linear Dyck languages instead of Dyck languages in our Chomsky-Stanley type characterization.

Theorem 3.1. *Let Σ be an alphabet. Then a Δ, a homomorphism $h : \Delta^* \to \Sigma^*$ and a linear Dyck language $D_{\mathcal{L}}$ on Δ can be determined from Σ, such that for every slender context-free language $L \subseteq \Sigma^*$ there can be found a regular language $R \subseteq \Delta^*$ with $L = h(R \cap D_{\mathcal{L}})$.*

Proof. Let Σ be an alphabet. Then we first define an alphabet Δ, a homomorphism h, and the linear Dyck language $D_{\mathcal{L}}$ on Δ as follows:

An alphabet Δ is defined by

$$\Delta = \Sigma \cup \Sigma' \cup \bar{\Sigma} \cup \bar{\Sigma}',$$

where

$$\Sigma' = \{a' \mid a \in \Sigma\}, \bar{\Sigma} = \{\bar{a} \mid a \in \Sigma\}, \text{ and } \bar{\Sigma}' = \{\bar{a}' \mid a \in \Sigma\}.$$

The homomorphism $h : \Delta^* \to \Sigma^*$ is defined by

$$h(a) = h(\bar{a}') = a,\ a \in \Sigma \text{ and } h(x) = \lambda,\ x \in \Delta \setminus (\Sigma \cup \bar{\Sigma}').$$

The linear Dyck language $D_{\mathcal{L}}$ over Δ is the language generated by

$$G_{\mathcal{L}} = (\{S\}, \Delta, S, P_{\mathcal{L}}),$$

where

$$P_{\mathcal{L}} = \{S \to aSa', S \to \lambda \mid a \in \Sigma \cup \bar{\Sigma}\}.$$

In order to simplify the notations, we use the following abbreviations. For a word $w = a_1 \dots a_\ell \in \Sigma^*$, we have $w^R = a_\ell \dots a_2 a_1$, $w' = a'_1 \dots a'_\ell, \bar{w} = \bar{a}_1 \dots \bar{a}_\ell$, and $\bar{w}' = \bar{a}'_1 \dots \bar{a}'_\ell$.

Let L be any slender context-free language over Σ. By Theorem 2.3 we can find a finite index set I and words u_i, v_i, w_i, x_i, y_i, for all $i \in I$ with $L = \bigcup_{i \in I} \{u_i v_i^n w_i x_i^n y_i \mid n \geq 0\}$.

Define a regular grammar $G_L = (N, \Delta, A, P_R)$, where $N = \{A\} \cup \{B_i, C_i \mid i \in I\}$, $P_R = P_1 \cup P_2 \cup P_3 \cup P_4 \cup P_5$ as

$P_1 = \{A \to u_i \bar{y_i}^R B_i \mid i \in I\}$,

$P_2 = \{B_i \to v_i \bar{x_i}^R B_i \mid i \in I\}$,

$P_3 = \{B_i \to w_i w_i'^R C_i \mid i \in I\}$,

$P_4 = \{C_i \to \bar{x_i}' v_i'^R C_i \mid i \in I\}$, and

$P_5 = \{C_i \to \bar{y_i}' u_i'^R \mid i \in I\}$.

Let R be a language generated by G_L, i.e., $R = L(G_L)$. Then $L = h(R \cap D_{\mathcal{L}})$ can be proved by

a.) $L \subset h(R \cap D_{\mathcal{L}})$.

Suppose w is in L and w is of the form $u_i v_i^n w_i x_i^n y_i$ for some i and n.

By the definition of G_L, it is clear that a word

$$\xi = u_i \bar{y_i}^R (v_i \bar{x_i}^R)^n w_i w_i'^R (\bar{x_i}' v_i'^R)^n \bar{y_i}' u_i'^R$$

is generated by G_L as follows:

$A \Rightarrow u_i \bar{y_i}^R B_i \Rightarrow u_i \bar{y_i}^R v_i \bar{x_i}^R B_i \Rightarrow^* u_i \bar{y_i}^R (v_i \bar{x_i}^R)^n B_i \Rightarrow u_i \bar{y_i}^R (v_i \bar{x_i}^R)^n w_i w_i'^R C_i$
$\Rightarrow u_i \bar{y_i}^R (v_i \bar{x_i}^R)^n w_i w_i'^R \bar{x_i}' v_i'^R C_i \Rightarrow^* u_i \bar{y_i}^R (v_i \bar{x_i}^R)^n w_i w_i'^R (\bar{x_i}' v_i'^R)^n C_i$
$\Rightarrow^* u_i \bar{y_i}^R (v_i \bar{x_i}^R)^n w_i w_i'^R (\bar{x_i}' v_i'^R)^n \bar{y_i}' u_i'^R$.

Moreover, it is clear that ξ is in $D_{\mathcal{L}}$, and therefore ξ is in $R \cap D_{\mathcal{L}}$. By the definition of h, $h(\xi)$ is a word $u_i v_i^n w_i x_i^n y_i$, i.e., w. So w is a word in $h(R \cap D_{\mathcal{L}})$.

b.) $h(R \cap D_{\mathcal{L}}) \subset L$.

Let $w \in h(R \cap D_{\mathcal{L}})$. Then, there is a word ξ in $R \cap D_{\mathcal{L}}$ such that $w = h(\xi)$. By the definition of G_L, ξ should be of the form

$$\xi = u_i \bar{y_i}^R (v_i \bar{x_i}^R)^m w_i w_i'^R (\bar{x_i}' v_i'^R)^n \bar{y_i}' u_i'^R$$

for some $i \in I$. By the definition of $D_{\mathcal{L}}$, $n = m$. Hence, $\xi = u_i \bar{y_i}^R (v_i \bar{x_i}^R)^n w_i w_i'^R (\bar{x_i}' v_i'^R)^n \bar{y_i}' u_i'^R$ and $h(\xi) = u_i v_i^n w_i x_i^n y_i$. Therefore, $w = h(\xi)$ is in L.

This completes the proof. □

Remark. There exists a regular language R such that $h(R \cap D_{\mathcal{L}})$ is not slender.

For example, choose a regular language Δ^* as R. Then, by the fact that $R \cap D_{\mathcal{L}}$ is $D_{\mathcal{L}}$ and the fact that $h(D_{\mathcal{L}})$ is Σ^*, the remark follows. Because of the previous observation, it is interesting to find a class $\mathcal{C}$ of regular languages satisfying the following condition: For any slender context-free language L, we can find R in $\mathcal{C}$ such that $L = h(R \cap D_{\mathcal{L}})$, and for any R in $\mathcal{C}$, $h(R \cap D_{\mathcal{L}})$ is a slender context-free language.

We denote by $\mathcal{R}_D$ the class of regular languages that satisfy the condition mentioned above.

A language L is called a *union of double loops* (or UDL, in short) if for words u_i, v_i, w_i, x_i, y_i where $1 \leq i \leq k$,

$$L = \bigcup_{i=1}^{k} \{u_i v_i^* w_i x_i^* y_i\}.$$

It is clear that L is regular by the definition. However, it is clear by Theorem 2.1 that L is not slender.

Now we have the following result, a little stronger than Theorem 3.1:

Theorem 3.2. *Let Σ be an alphabet. Then an alphabet Δ, a homomorphism $h : \Delta^* \to \Sigma^*$ and a linear Dyck language $D_{\mathcal{L}}$ on Δ can be determined from Σ such that for every slender context-free language $L \subseteq \Sigma^*$, there can be found a UDL regular language $R \subseteq \Delta^*$ such that $L = h(R \cap D_{\mathcal{L}})$. Moreover, for any UDL regular language R, $h(R \cap D_{\mathcal{L}})$ is slender context-free.*

Proof. In the proof of Theorem 3.1, one can see the fact that R is a UDL regular language, so the first part of the theorem holds.

Now we consider the second part. Let R be a UDL regular language. Then, since the class of linear context-free languages is closed under the operation of intersection with a regular set, $R \cap D_{\mathcal{L}}$ is linear. Furthermore, by counting the number of words of length n in $R \cap D_{\mathcal{L}}$, we can find that $R \cap D_{\mathcal{L}}$ is slender. Indeed, by the simmetricity of the elements of $D_{\mathcal{L}}$, every $uv^\ell wx^m y \in D_{\mathcal{L}}$ has the form $a_1 \dots a_f (a_{f+1} \dots a_g)^\ell a_{g+1} \dots a_h a'_h \dots a'_{g+1} (a'_g \dots a'_{f+1})^m a'_f \dots a'_1$ with $k = \ell$ and $|v| = |x|$. Hence, by $L = \cup_{i=1}^{k} \{u_i v_i^* w_i x_i^* y_i\}$, the language $R \cap D_{\mathcal{L}}$ has at most k words of length n for every $n \geq 1$. Since the class of slender context-free languages is closed under homomorphisms, $h(R \cap D_{\mathcal{L}})$ is slender context-free.

This completes the proof. □

4 Concluding Remarks

In this paper, we investigated some Chomsky-Schützenberger-Stanley type homomorphic characterizations for slender context-free languages and obtained the first characterization as Theorem 3.1 and the second characterization as Theorem 3.2, by which any slender language can be represented by the homomorphic image of the intersection of a linear Dyck language and a UDL regular language, and for any UDL regular language, the homomorphic image of its intersection with a linear Dyck language is slender. This means that the second result is stronger than the first one.

References

[1] Chomsky, N., Context-free grammars and pushdown storage, *M.I.T. Res. Lab., Electron Quart. Prog. Rept.*, **65** (1962) 187-194.

[2] Chomsky, N., Schützenberger, M., The algebraic theory of context-free languages, *Comput. Programming and Formal Systems, North-Holland, Amsterdam* (1963), 118–161.

[3] Culik, K. II, A purely homomorphic characterization of recursively enumerable sets, *J. ACM* **26** (1979) 345-350.

[4] Engelfriet, J., Rozenberg, G., Fixed point languages, equality languages and representation of recursively enumerable languages, *J. ACM* **27** (1980) 499-518.

[5] Fülöp, Z., Vágvölgyi, S., On ranges of compositions of deterministic root-to-frontier tree transformations, *Acta Cybernet.* **8** (1988), 259-266.

[6] Ginsburg, S., The Mathematical Theory of Context-Free Languages, *McGraw-Hill Book Company, New York, St Louis, San Francisco, Auckland, Bogota, Hamburg, Johannesburg, London, Madrid, Mexico, Montreal, New Delhi, Panama, Paris, Sao Paulo, Singapore, Sydney, Tokyo, Toronto,* 1966.

[7] Ginsburg, S., Greibach, S. A., Harrison, M. A., One-way stack automata, *J. ACM* **14** (1967) 389/418.

[8] Harrison, M. A., Introduction to Formal Language Theory, *Addison-Wesley Publishing Company, Reading , Massachusetts, Menlo Park, California, London, Amsterdam, Don Mils, Ontario, Sidney,* 1978.

[9] Hirose, S., Nasu, M., Left universal context-free grammars and homomorphic characterizations of languages, *Inform. and Control,* **50** (1981) 110-118.

[10] Hirose, S., Okawa, S., Yoneda, M., A homomorphic characterization of recursively enumerable languages, *Theoret. Comput.Sci.* **35** (1985) 261-269.

[11] Hirose, S., Yoneda, M., On the Chomsky's and Stanley's homomorphic characterization of context-free languages, *Theoret. Comput.Sci.* **36** (1985) 109-112.

[12] Hopcroft, J. E. & Ullmann, J. D., Introduction to Automata Theory, Languages, and Computation , *Addison-Wesley, Reading , Massachusetts, Menlo Park, California, London, Amsterdam, Don Mils, Ontario, Sidney,* 1979.

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

[14] Imreh, B., Ito, M., A note on the regular strongly shuffle-closed languages, *Acta Cybernet.* **12** (1995), 11-22.

[15] Okawa, S., Hirose, S., Yoneda, M., On the impossibility of the homomorphic characterization of context-sensitive languages, *Theoret. Comput.Sci.* **44** (1986) 225-228.

[16] Păun, G., Salomaa, A., Thin and slender languages, *Discrete Appl. Math.*, **61** (1995) 257-270.

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

[18] Révész, Gy. E., Introduction to Formal Languages, *McGraw-Hill, New York, St Louis, San Francisco, Auckland, Bogota, Hamburg, Johannesburg, London, Madrid, Mexico, Montreal, New Delhi, Panama, Paris, Sao Paulo, Singapore, Sydney, Tokyo, Toronto,* 1983.

[19] Salomaa, A., Formal Languages, *Academic Press, New York, London,* 1973.

[20] Schützenberger, M., On context-free languages and push-down automata, *Inf. Control* **6** (1963), 246–264.

[21] Stanley, R. J., Finite state representations of context-free languages, *M.I.T. Res. Lab., Electron Quart. Prog. Rept.*, **76** (1965) 276-279.

Received November, 2000