INFORMATION
SCIENCE
TECHNICAL
REPORT

NAIST-IS-TR2010003
ISSN 0919-9527

Chomsky-Schützenberger-Type Characterization of Multiple Context-Free Languages

Ryo Yoshinaka, Yuichi Kaji, and Hiroyuki Seki

April 2010

NAIST
〒 630-0192
奈良県生駒市高山町 8916-5
奈良先端科学技術大学院大学
情報科学研究科

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan

Chomsky-Schützenberger-Type Characterization of Multiple Context-Free Languages

Ryo Yoshinaka[1⋆], Yuichi Kaji[2], and Hiroyuki Seki[2]

[1] Graduate School of Information Science and Technology, Hokkaido University,
ry@ist.hokudai.ac.jp
[2] Graduate School of Information Science, Nara Institute of Science and Technology,
{kaji,seki}@is.naist.jp

Abstract. It is a well-known theorem by Chomsky and Schützenberger (1963) that every context-free language can be represented as a homomorphic image of the intersection of a Dyck language and a regular language. This paper gives a Chomsky-Schützenberger-type characterization for multiple context-free languages, which are a natural extension of context-free languages, with introducing the notion of multiple Dyck languages, which are also a generalization of Dyck languages.

1 Introduction

A multiple context-free grammar (mcfg) is a natural extension of a context-free grammar (cfg). A nonterminal symbol in an mcfg derives tuples of strings by synchronized parallel derivation. The direct derivation relation of an mcfg is defined by a function over tuples of strings (of terminal symbols) such that each component of the function value is defined by a concatenation of some components of arguments and constant strings of terminal symbols with a linearity condition on components of arguments. Let us call such a function *linear regular*. The language generated by an mcfg is called a multiple context-free language (mcfl).

The generative power of mcfgs is properly larger than cfgs and properly smaller than context-sensitive grammars (csgs). There are several computational models that have the same generative power as mcfgs, e.g., string version of linear context-free rewriting systems, finite copying tree-to-string transducers, string generating context-free hypergraph grammars and local unordered scattered context grammars (see [2, 6] for the discussion of these equivalences). Mcfgs share many properties with cfgs such as closure properties. There are other grammatical formalisms of which generative power is between cfgs and csgs such as indexed grammars. In contrast to indexed grammars, the membership problem for an mcfl is solvable in polynomial time in the length of an input string and each mcfl is semilinear. These properties are due to the synchronized parallel derivation realized by linear regular functions. Generally, each component of a

⋆ He is now concurrently working in ERATO MINATO Discrete Structure Manipulation System Project, Japan Science and Technology Agency.

tuple of strings appearing in the derivation is not adjacent with one another in the resultant string of terminal symbols. However, these components always share synchronized structure of derivation. To capture this property of mcfls, we will introduce *multiple Dyck languages* and show a theorem that is an extension of the representation theorem of cfls. It is well-known that any cfl can be represented as a homomorphic image of the intersection of a regular language and a Dyck language (Chomsky-Schützenberger theorem). A Dyck language is the set of well-nested parentheses (brackets). A multiple Dyck language is the set of 'well-nested tuples of parentheses.' The main theorem of this paper is that for a given mcfl L, there exists a multiple Dyck language D, a regular language R and a homomorphism h such that $L = h(D \cap R)$. As is the same with cfls, this representation theorem for mcfls can be easily lifted to the generator theorem.

The main results of this paper were partially published in and are partially based on Chapter 4 of [3].

2 Preliminaries

For an alphabet Σ, Σ^* denotes the set of all strings over Σ and $(\Sigma^*)^m$ denotes the set of all m-tuples of strings over Σ. The empty string is denoted by ε.

2.1 Context-Free Grammars

A *context-free grammar (cfg)* is a tuple $G = \langle \Sigma, N, P, S \rangle$, where Σ is a finite set of *terminal symbols*, N is a finite set of *nonterminal symbols*, $P \subseteq N \times (\Sigma \cup N)^*$ is a finite set of *rules*, which are denoted by $A \to \alpha$ for $A \in N$ and $\alpha \in (\Sigma \cup N)^*$, and $S \in N$ is called *the start symbol*. Elements of $P \cap (N \times \Sigma^*)$ are called *terminating rules*. $\Rightarrow_G$ and $\Rightarrow_G^*$ denote derivations of one step and any steps (including zero-step), respectively. The language generated by a cfg G, which is called a *context-free language (cfl)*, is the set $L(G) = \{ w \in \Sigma^* \mid S \Rightarrow_G^* w \}$. If $P \subseteq N \times (\Sigma^* \cup \Sigma^* N)$, G is called a *right-linear grammar* and $L(G)$ is called a *regular language*.

Let $\overline{\Sigma}$ denote an alphabet disjoint from Σ that admits a bijection $\overline{(\cdot)}$ from Σ to $\overline{\Sigma}$. The *Dyck grammar* over $\Sigma \cup \overline{\Sigma}$ is the cfg that has S as its unique nonterminal symbol and whose rules are $S \to \varepsilon$ and $S \to SaS\bar{a}$ for all $a \in \Sigma$. The language generated by a Dyck grammar is called a *Dyck language*. A string on $\Sigma \cup \overline{\Sigma}$ is *well-bracketed* if it is an element of the Dyck language. An occurrence of $a \in \Sigma$ and an occurrence of $\bar{a} \in \overline{\Sigma}$ in a well-bracketed string are *corresponding* if they are derived at the same derivation step. Note that the Dyck grammar is unambiguous. According to the custom, we call elements of $\Sigma \cup \overline{\Sigma}$ *parentheses*.

Chomsky and Schützenberger [1] gave a characterization of cfls by Dyck languages.

Theorem 1. *A language L over Σ is context-free if and only if there are an alphabet Δ, a homomorphism $h : (\Delta \cup \overline{\Delta})^* \to \Sigma^*$ and a regular language R over $\Delta \cup \overline{\Delta}$ such that $L = h(D \cap R)$ where D is the Dyck language over $\Delta \cup \overline{\Delta}$.*

Theorem 1 can be stated in an even stronger (for the 'only if' direction) form:

Theorem 2. *For a given alphabet Σ, there are an alphabet Δ and a homomorphism $h : (\Delta \cup \overline{\Delta})^* \to \Sigma^*$ such that for any language L over Σ, L is context-free if and only if there is a regular language R such that $L = h(D \cap R)$ where D is the Dyck language over $\Delta \cup \overline{\Delta}$.*

2.2 Multiple Context-Free Grammars

We assume a countably infinite set X of *variables*. A function from $(\Sigma^*)^{m_1} \times \cdots \times (\Sigma^*)^{m_n}$ to $(\Sigma^*)^m$ is said to be *linear regular*, if there are $t_1, \ldots, t_m \in (\Sigma \cup \{ x_{i,j} \in X \mid 1 \leq i \leq n, 1 \leq j \leq m_i \})^*$ such that each variable $x_{i,j}$ occurs at most once in $t_1 \ldots t_m$ and for any $\vec{w}_i = \langle w_{i,1}, \ldots, w_{i,m_i} \rangle \in (\Sigma^*)^{m_i}$ with $1 \leq i \leq n$, it holds that

$$f(\vec{w}_1, \ldots, \vec{w}_n) = \langle v_1, \ldots, v_m \rangle$$

where each v_k for $k = 1, \ldots, m$ is obtained from t_k by substituting $w_{i,j}$ for $x_{i,j}$ for all i and j. We simply write $f(\langle x_{1,1}, \ldots, x_{1,m_1} \rangle, \ldots, \langle x_{n,1}, \ldots, x_{n,m_n} \rangle) = \langle t_1, \ldots, t_m \rangle$ to denote the definition of f. For example, both $f(\langle x_1, x_2 \rangle) = \langle ax_1b, cx_2d \rangle$ and $g(\langle x_1, x_2 \rangle, \langle y_1, y_2 \rangle) = \langle x_1y_1, y_2x_2 \rangle$ are linear regular functions where $x_1, x_2, y_1, y_2 \in X$ and $a, b, c, d \in \Sigma$, while $h(\langle x_1 \rangle) = \langle x_1x_1 \rangle$ is not, since x_1 appears twice in the right-hand side. f is said to be *nonerasing*, if every variable in the left-hand side of the definition of f appears in the right-hand side. f is *terminal-free*, if the right-hand side of its definition contains no symbols from Σ.

An alphabet N is said to be *indexed* when we assume a function dim that assigns positive integers to symbols in N.

A *multiple context-free grammar (mcfg)* is a tuple $G = \langle \Sigma, N, F, P, S \rangle$, where

- Σ is an (unindexed) alphabet whose elements are called *terminal symbols*,
- N is an indexed alphabet whose elements are called *nonterminal symbols*,
- F is a finite set of linear regular functions,
- P is a finite set of *rules* of the form $A \to f(B_1, \ldots, B_n)$ where $A, B_1, \ldots, B_n \in N$ and $f : (\Sigma^*)^{\dim(B_1)} \times \cdots \times (\Sigma^*)^{\dim(B_n)} \to (\Sigma^*)^{\dim(A)} \in F$,
- $S \in N$ is called *the start symbol* whose dimension is 1.

For a rule $\pi = A \to f(B_1, \ldots, B_n)$, the *head* and the *body* of π refer to A and $f(B_1, \ldots, B_n)$, respectively, and the *rank* of π is defined to be $\mathrm{rank}(\pi) = n$. If $\mathrm{rank}(\pi) = 0$ and $f() = \vec{w}$, we simply write $A \to \vec{w}$ for π with suppressing f. If f is terminal-free, π is also said to be *terminal-free*.

For each $A \in N$, $L_G(A)$ is recursively defined as the smallest set of $\dim(A)$-tuples of strings satisfying that if $A \to f(B_1, \ldots, B_n) \in P$ and $\vec{w}_i \in L_G(B_i)$ for $i = 1, \ldots, n$, then $f(\vec{w}_1, \ldots, \vec{w}_n) \in L_G(A)$. The *language $L(G)$ generated by G* is the set $\{ w \in \Sigma^* \mid \langle w \rangle \in L_G(S) \}$. $L(G)$ is called a *multiple context-free language (mcfl)*. Two grammars G and G' are *equivalent* if $L(G) = L(G')$.

Example 1. Let G_1 be the mcfg $\langle \Sigma_1, N_1, F_1, P_1, S \rangle$ such that $\Sigma_1 = \{a, b, c, d\}$, $N_1 = \{S, A, B\}$ with $\dim(S) = 1$, $\dim(A) = \dim(B) = 2$, F_1 consists of e, f, g and the constant functions appearing in the body of rules in P_1 below

where $e(\langle x_1, x_2\rangle, \langle y_1, y_2\rangle) = \langle x_1 y_1 x_2 y_2\rangle$, $f(\langle x_1, x_2\rangle) = \langle a x_1, b x_2\rangle$, $g(\langle x_1, x_2\rangle) = \langle c x_1, d x_2\rangle$, and $P_1 = \{S \to e(A, B),\ A \to f(A),\ A \to \langle a, b\rangle,\ B \to g(B),\ B \to \langle c, d\rangle\}$. Let us call the rules in P_1 $\pi_1, \pi_2, \ldots, \pi_5$ in the order written above. For example, $\langle a, b\rangle \in L_{G_1}(A)$ by π_3, $\langle aa, bb\rangle \in L_{G_1}(A)$ by π_2, $\langle c, d\rangle \in L_{G_1}(B)$ by π_5 and $\langle aacbbd\rangle \in L_{G_1}(S)$ by π_1. We have $L(G_1) = \{\, a^m c^n b^m d^n \mid m, n \geq 1 \,\}$.

For a nonterminal symbol A of an mcfg G, a series of rule application steps to obtain a tuple of strings of terminal symbols $\vec{w} \in L_G(A)$ is called a *derivation* of $\vec{w}$ in G.

By q-MCFG(r) we denote the collection of mcfgs G such that $\dim(A) \leq q$ for all $A \in N$ and $\mathrm{rank}(\pi) \leq r$ for all $\pi \in P$. q-MCFL(r) is the class of mcfls generated by grammars in q-MCFG(r).

G is said to be *nonerasing*, if all $f \in F$ are nonerasing. It is known that every $G \in q$-MCFG(r) has an equivalent nonerasing grammar in q-MCFG(r) [8]. Grammars from 1-MCFG(r) are identified with cfgs.

Proposition 1 (Seki et al. [8] and Rambow and Satta [6]). *For $q \geq 1$, q-MCFL(r) $\subsetneq$ (q+1)-MCFL(r). For $q \geq 2$, $r \geq 1$, q-MCFL(r) $\subsetneq$ q-MCFL(r+1) except for* 2-MCFL(2) = 2-MCFL(3). *For $q \geq 1$, $r \geq 3$ and $1 \leq k \leq r-2$, q-MCFL(r) $\subseteq$ $(k+1)q$-MCFL($r-k$).*

Proposition 2 (Rambow and Satta [6]). *Each family q-MCFL(r) for $r \geq 2$ is a substitution closed full AFL. That is, they are closed under homomorphism, inverse homomorphism, intersection with regular languages, union, concatenation, the Kleene plus and substitution.*

Proposition 3 (Seki et al. [8]). *Let $G \in q$-MCFG(r) be given. It is decidable in $O(|w|^{q(r+1)})$ time whether $w \in L(G)$ for any $w \in \Sigma^*$.*

2.3 Multiple Dyck Languages

Let q and r be fixed. We define the notion of the multiple Dyck language in q-MCFL(r) on an indexed alphabet, where we assume that the maximum dimension of elements of the indexed alphabet does not exceed r. For an indexed alphabet Δ, let

$$\widehat{\Delta} = \{\, a^{[i]}, \bar{a}^{[i]} \mid a \in \Delta,\ 1 \leq i \leq \dim(a) \,\}.$$

Definition 1. The *multiple Dyck grammar* D_Δ on an indexed alphabet Δ is the mcfg that has nonterminal symbols S_m with $\dim(S_m) = m$ for $m \leq q$, among which the start symbol is S_1, and that has rules of the following three types:

1. all the possible terminal-free rules allowed in q-MCFG(r);
2. rules of the form $S_m \to f_a(S_m)$ where $f_a(\langle x_1, \ldots, x_m\rangle) = \langle a^{[1]} x_1 \bar{a}^{[1]}, \ldots, a^{[m]} x_m \bar{a}^{[m]}\rangle$ for $a \in \Delta$ with $\dim(a) = m$;
3. rules of the form $S_m \to f(S_m)$ with $f(\langle x_1, \ldots, x_m\rangle) = \langle t_1, \ldots, t_m\rangle$ where each t_i is either x_i, $x_i a^{[1]} \bar{a}^{[1]}$ or $a^{[1]} \bar{a}^{[1]} x_i$ for some $a \in \Delta$ of dimension 1.

The language $L(D_\Delta)$ is called the *multiple Dyck language* over $\Delta \cup \overline{\Delta}$.

We note that rules of type 3 are redundant if $r > 1$. If $q = 1$ and $r > 1$, $L(D_\Delta)$ is indeed the (context-free) Dyck language over $\Delta \cup \overline{\Delta}$.

Every element of tuples in $L_{D_\Delta}(S_m)$ is well-bracketed. Moreover pairs of corresponding parentheses in a string from $L(D_\Delta)$ are partitioned into groups each of which consists of exactly $\langle a^{[1]}, \bar{a}^{[1]} \rangle, \ldots, \langle a^{[\dim(a)]}, \bar{a}^{[\dim(a)]} \rangle$ for some $a \in \Delta$. If some member of such a group A is inside some member of a group B, then all members from A are inside some member of B. For example, $b^{[1]}a^{[1]}\bar{a}^{[1]}a^{[2]}\bar{a}^{[2]}\bar{b}^{[1]}$ $b^{[2]}a^{[3]}\bar{a}^{[3]}\bar{b}^{[2]}$ is allowed, while $b^{[1]}a^{[1]}\bar{a}^{[1]}\bar{b}^{[1]}a^{[2]}\bar{a}^{[2]}b^{[2]}a^{[3]}\bar{a}^{[3]}\bar{b}^{[2]}$ is not, where $\dim(a) = 3$ and $\dim(b) = 2$. The way of combining parentheses is restricted by available means in q-MCFG(r).

3 Theorem

This section discusses Chomsky-Schützenberger type characterization of mcfls.

3.1 Informal Example of Construction

We first review an idea of the proof of Theorem 1 by using a simple example. Let G_0 be the cfg $\langle \Sigma_0, N_0, P_0, S \rangle$ where $\Sigma_0 = \{a, b, c\}$, $N_0 = \{S, A, B\}$, $P_0 = \{S \to aA,\ A \to bAB,\ A \to a,\ B \to c\}$. We call the four rules π_1, π_2, π_3 and π_4 in the order as written above. Let $\Delta = \{ \mathbb{(}_{\pi_1,1}, \mathbb{(}_{\pi_2,1}, \mathbb{(}_{\pi_2,2}, \mathbb{(}_a, \mathbb{(}_b, \mathbb{(}_c \}$ and let us write $\mathbb{)}_x$ to denote $\overline{\mathbb{(}_x}$ for each $\mathbb{(}_x \in \Delta$. Also let $h : (\Delta \cup \overline{\Delta})^* \to \Sigma_0^*$ be the homomorphism defined by $h(\mathbb{(}_x) = x$ for $x \in \Sigma_0$ and $h(z) = \varepsilon$ for other $z \in \Delta \cup \overline{\Delta}$. Figure 1 shows an example of a derivation tree (called t_0) in G_0. Intuitively, $\mathbb{(}_{\pi,i}$ and $\mathbb{)}_{\pi,i}$ mean the left end and the right end of a derivation

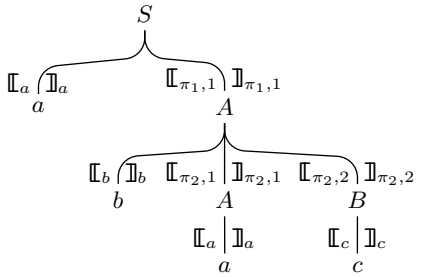

Fig. 1. A derivation tree in G_0

starting from the i-th nonterminal symbol in the body of the rule π. For $x \in \Sigma_0$, a pair $\mathbb{(}_x$ and $\mathbb{)}_x$ denotes x. In the figure, paired symbols in $\Delta \cup \overline{\Delta}$ are placed on the left-side and the right-side of each edge. For a tree t, let $\alpha(t)$ denote

the string over $\Delta \cup \overline{\Delta}$ obtained by concatenating these labels in the depth-first left-to-right order. For example,

$$\alpha(t_0) = \mathbb{[}_a\ \mathbb{]}_a\ \mathbb{[}_{\pi_1,1}\ \mathbb{[}_b\ \mathbb{]}_b\ \mathbb{[}_{\pi_2,1}\ \mathbb{[}_a\ \mathbb{]}_a\ \mathbb{]}_{\pi_2,1}\ \mathbb{[}_{\pi_2,2}\ \mathbb{[}_c\ \mathbb{]}_c\ \mathbb{]}_{\pi_2,2}\ \mathbb{]}_{\pi_1,1}$$

for t_0 in the figure. For a tree t, let $\mathrm{yield}(t)$ denote the string obtained by concatenating the labels of leaf nodes of t from left to right. Then, $\mathrm{yield}(t) = h(\alpha(t))$ for a derivation tree t in G_0 and $L(G_0) = h(\{\, \alpha(t) \mid t \text{ is a derivation tree in } G_0 \,\})$. Therefore, what we should do is to construct a right-linear grammar G_{R_0} such that $L(G_{R_0}) \cap D = \{\, \alpha(t) \mid t \text{ is a derivation tree in } G_0 \,\}$ in this particular example where D is the Dyck language over $\Delta \cup \overline{\Delta}$. G_{R_0} can be defined by considering the finite-state tree traversal that emits $\mathbb{[}_x$ and $\mathbb{]}_x$ when it visits $x \in \Sigma_0$, emits $\mathbb{[}_{\pi,i}$ when it visits the i-th nonterminal symbol in the body of π, and emits $\mathbb{]}_{\pi,i}$ when it returns from that nonterminal symbol. Note that nonterminal symbols in N_0 are used as 'finite states' (nonterminal symbols of G_{R_0}) when the traversal goes down while a new nonterminal symbol T is used when it goes up.

$$\begin{array}{llll}
S \to \mathbb{[}_a\ \mathbb{]}_a\ \mathbb{[}_{\pi_1,1}\ A & T \to\ \mathbb{]}_{\pi_1,1}\ T & & \\
A \to \mathbb{[}_b\ \mathbb{]}_b\ \mathbb{[}_{\pi_2,1}\ A & T \to\ \mathbb{]}_{\pi_2,1}\ \mathbb{[}_{\pi_2,2}\ B & T \to\ \mathbb{]}_{\pi_2,2}\ T & \\
A \to \mathbb{[}_a\ \mathbb{]}_a\ T & B \to \mathbb{[}_c\ \mathbb{]}_c\ T & & \\
T \to \varepsilon & & &
\end{array}$$

A similar idea can be applied to mcfg. Let G_1 be the mcfg from Example 1. Figure 2 shows a tree that illustrates the derivation in the example. (This kind

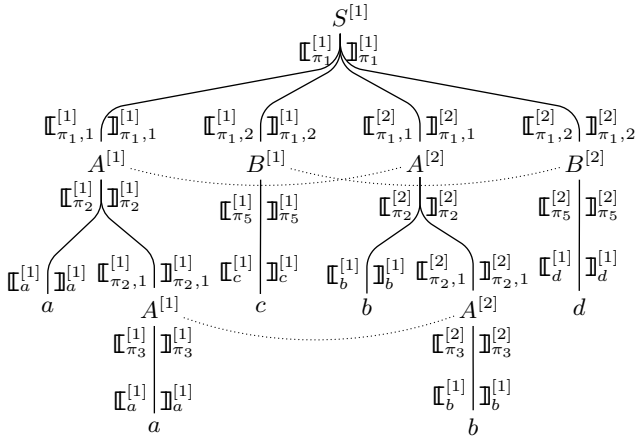

Fig. 2. A derived tree in G_1

of tree is called a *derived tree* in mcfg. Here we use derived tree without formal definition since derived tree is not needed in the formal proofs in the rest of this paper.) In the figure, $A^{[j]}$ $(j = 1, 2)$ denotes a (hypothetical) nonterminal symbol that derives the j-th component w_j of $\langle w_1, w_2 \rangle \in L_{G_1}(A)$. $S^{[1]}$, $B^{[1]}$ and

$B^{[2]}$ are used in the same purpose. A horizontal arc between $A^{[1]}$ and $A^{[2]}$ means that these two nodes together represent an instance of A in the derivation. Let

$$\Gamma = \{\pi_1, \pi_2, \ldots, \pi_5, \langle \pi_1, 1\rangle, \langle \pi_1, 2\rangle, \langle \pi_2, 1\rangle, \langle \pi_4, 1\rangle, a, b, c, d\}.$$

The symbol $\langle \pi, i\rangle$ $(1 \leq i \leq \mathrm{rank}(\pi))$ corresponds to the i-th nonterminal symbol in the body of the rule π. For example, $\langle \pi_1, 1\rangle$ and $\langle \pi_1, 2\rangle$ correspond to A and B, respectively. For each $\pi \in P_1$, define $\dim(\pi)$ to be the dimension of the head of π. For each $\pi \in P_1$ and i $(1 \leq i \leq \mathrm{rank}(\pi))$, define $\dim(\langle \pi, i\rangle)$ to be the dimension of the i-th nonterminal symbol in the body of π. For each $x \in \Sigma_1$, define $\dim(x) = 1$. Thus, $\dim(\pi_1) = \dim(a) = \cdots = \dim(d) = 1$ and $\dim(x) = 2$ for other $x \in \Gamma$. We abbreviate symbols in $\widehat{\Gamma}$ as $\mathbb{L}^{[1]}_{\pi_1}$, $\mathbb{R}^{[1]}_{\pi_1}$, $\mathbb{L}^{[1]}_{\pi_2}$, $\mathbb{R}^{[1]}_{\pi_2}$, $\mathbb{L}^{[2]}_{\pi_2}$, $\mathbb{R}^{[2]}_{\pi_2}$, ..., $\mathbb{L}^{[1]}_{\pi_1,1}$, $\mathbb{R}^{[1]}_{\pi_1,1}$, $\mathbb{L}^{[2]}_{\pi_1,1}$, $\mathbb{R}^{[2]}_{\pi_1,1}$, ... Similarly to cfg's case, $\mathbb{L}^{[j]}_{\pi,i}$ (rsp. $\mathbb{R}^{[j]}_{\pi,i}$) denotes the left (rsp. right) end of a derivation for the j-th component of the i-th nonterminal symbol in the body of π. For the mcfg G_1, we also have $L(G_1) = h(\{\, \alpha(t) \mid t$ is a 'derived tree' in $G_1 \,\})$ where h is defined similarly to cfg's case. Hence, it suffices to give a right-linear grammar G_{R_1} such that $L(G_{R_1}) \cap L(D_\Gamma) = \{\, \alpha(t) \mid t$ is a 'derived tree' in $G_1 \,\}$. The construction of G_{R_1} is a little cumbersome but not difficult:

$$\begin{array}{llll}
S^{[1]} \to \mathbb{L}^{[1]}_{\pi_1}\ \mathbb{L}^{[1]}_{\pi_1,1} A^{[1]} & T \to \mathbb{R}^{[1]}_{\pi_1,1}\ \mathbb{L}^{[1]}_{\pi_1,2}\ B^{[1]} & T \to \mathbb{R}^{[1]}_{\pi_1,2}\ \mathbb{L}^{[2]}_{\pi_1,1}\ A^{[2]} & \\
T \to \mathbb{R}^{[2]}_{\pi_1,1}\ \mathbb{L}^{[2]}_{\pi_1,2}\ B^{[2]} & T \to \mathbb{R}^{[2]}_{\pi_1,2}\ \mathbb{R}^{[1]}_{\pi_1}\ T & & \\
A^{[1]} \to \mathbb{L}^{[1]}_{\pi_2}\ \mathbb{L}^{[1]}_{a}\ \mathbb{R}^{[1]}_{a}\ \mathbb{L}^{[1]}_{\pi_2,1}\ A^{[1]} & T \to \mathbb{R}^{[1]}_{\pi_2,1}\ \mathbb{R}^{[1]}_{\pi_2}\ T & & \\
A^{[2]} \to \mathbb{L}^{[2]}_{\pi_2}\ \mathbb{L}^{[1]}_{b}\ \mathbb{R}^{[1]}_{b}\ \mathbb{L}^{[2]}_{\pi_2,1}\ A^{[2]} & T \to \mathbb{R}^{[2]}_{\pi_2,1}\ \mathbb{R}^{[2]}_{\pi_2}\ T & & \\
A^{[1]} \to \mathbb{L}^{[1]}_{\pi_3}\ \mathbb{L}^{[1]}_{a}\ \mathbb{R}^{[1]}_{a}\ \mathbb{R}^{[1]}_{\pi_3}\ T & A^{[2]} \to \mathbb{L}^{[2]}_{\pi_3}\ \mathbb{L}^{[1]}_{b}\ \mathbb{R}^{[1]}_{b}\ \mathbb{R}^{[2]}_{\pi_3}\ T & & \\
B^{[1]} \to \mathbb{L}^{[1]}_{\pi_4}\ \mathbb{L}^{[1]}_{c}\ \mathbb{R}^{[1]}_{c}\ \mathbb{L}^{[1]}_{\pi_4,1}\ B^{[1]} & T \to \mathbb{R}^{[1]}_{\pi_4,1}\ \mathbb{R}^{[1]}_{\pi_4}\ T & & \\
B^{[2]} \to \mathbb{L}^{[2]}_{\pi_4}\ \mathbb{L}^{[1]}_{d}\ \mathbb{R}^{[1]}_{d}\ \mathbb{L}^{[2]}_{\pi_4,1}\ B^{[2]} & T \to \mathbb{R}^{[2]}_{\pi_4,1}\ \mathbb{R}^{[2]}_{\pi_4}\ T & & \\
B^{[1]} \to \mathbb{L}^{[1]}_{\pi_5}\ \mathbb{L}^{[1]}_{c}\ \mathbb{R}^{[1]}_{c}\ \mathbb{R}^{[1]}_{\pi_5}\ T & B^{[2]} \to \mathbb{L}^{[2]}_{\pi_5}\ \mathbb{L}^{[1]}_{d}\ \mathbb{R}^{[1]}_{d}\ \mathbb{R}^{[2]}_{\pi_5}\ T & & \\
T \to \varepsilon\ . & & &
\end{array}$$

3.2 Formal Construction

Let us arbitrarily fix positive integers q and r. We now give our Chomsky-Schützenberger type characterization for q-MCFL(r). Without loss of generality, we may assume that any $G \in q$-MCFG(r) satisfies the following conditions:

- G is nonerasing;
- if G has a rule $A \to f(B_1, \ldots, B_n)$ and $1 \leq i < j \leq n$, then $B_i \neq B_j$.

Indeed every mcfg in q-MCFG(r) has an equivalent one in q-MCFG(r) with this property.

Let $G = \langle \Sigma, N, F, P, S\rangle \in q$-MCFG($r$) be given. Our goal is to find an indexed alphabet Δ, a right-linear grammar R over $\widehat{\Delta}$, and a homomorphism $h : \widehat{\Delta}^* \to \Sigma^*$ such that $L(G) = h(L(D_\Delta) \cap L(R))$.

Let

$$\Delta = \{\, \mathbb{[}_a \mid a \in \Sigma \,\} \cup \{\, \mathbb{[}_\pi \mid \pi \in P \,\} \cup \{\, \mathbb{[}_{\pi,i} \mid 1 \le i \le \mathrm{rank}(\pi),\ \pi \in P \,\}$$

where $\dim(\mathbb{[}_a) = 1$ for $a \in \Sigma$, $\dim(\mathbb{[}_\pi) = \dim(A)$ and $\dim(\mathbb{[}_{\pi,i}) = \dim(B_i)$ if $\pi \in P$ is of the form $A \to f(B_1, \ldots, B_n)$. Hereafter we write $\mathbb{]}_*$ instead of $\overline{\mathbb{[}_*}$ for each $\overline{\mathbb{[}_*} \in \overline{\Delta}$. By $\widetilde{(\cdot)}$ we denote the homomorphism from Σ^* to $\widehat{\Delta}^*$ such that $\widetilde{a} = \mathbb{[}_a^{[1]}\, \mathbb{]}_a^{[1]}$.

The nonterminal symbols of the right-linear grammar R is

$$\{\, T \,\} \cup \{\, A^{[k]} \mid A \in N,\ 1 \le k \le \dim(A) \,\}$$

and the start symbol is $S^{[1]}$. The rules of R are given as follows. Suppose that G has a rule π of the form $A \to f(B_1, \ldots, B_n)$ and f is represented as

$$f(\langle x_{1,1}, \ldots, x_{1,m_1}\rangle, \ldots, \langle x_{n,1}, \ldots, x_{n,m_n}\rangle) = \langle t_1, \ldots, t_m\rangle$$
$$\text{where } t_k = u_{k,0} x_{i_{k1},j_{k1}} u_{k,1} \ldots x_{i_{kp_k},j_{kp_k}} u_{k,p_k} \text{ with } u_{k,0}, \ldots, u_{k,p_k} \in \Sigma^* \text{ for } k = 1, \ldots, m.$$

For each $k = 1, \ldots, m$, if $p_k = 0$, then R has the rule

$$A^{[k]} \to \mathbb{[}_\pi^{[k]}\, \widetilde{u}_{k,0}\, \mathbb{]}_\pi^{[k]}\, T$$

and otherwise, R has the following $p_k + 1$ rules:

$$\begin{aligned} A^{[k]} &\to \mathbb{[}_\pi^{[k]}\, \widetilde{u}_{k,0}\, \mathbb{[}_{\pi,i_{k1}}^{[j_{k1}]}\, B_{i_{k1}}^{[j_{k1}]}, \\ T &\to \mathbb{]}_{\pi,i_{k(l-1)}}^{[j_{k(l-1)}]}\, \widetilde{u}_{k,l-1}\, \mathbb{[}_{\pi,i_{kl}}^{[j_{kl}]}\, B_{i_{kl}}^{[j_{kl}]} \text{ for } 1 < l \le p_k, \\ T &\to \mathbb{]}_{\pi,i_{kp_k}}^{[j_{kp_k}]}\, \widetilde{u}_{k,p_k}\, \mathbb{]}_\pi^{[k]}\, T. \end{aligned}$$

Moreover R has

$$T \to \varepsilon,$$

which is the unique terminating rule of R.

We define the homomorphism $h : \widehat{\Delta}^* \to \Sigma^*$ so that for $z \in \widehat{\Delta}$,

$$h(z) = \begin{cases} a & \text{if } z = \mathbb{[}_a^{[1]} \text{ for some } a \in \Sigma; \\ \varepsilon & \text{otherwise.} \end{cases}$$

3.3 Correctness

Lemma 1. $L(G) \subseteq h(L(R) \cap L(D_\Delta))$.

Proof. By induction we show that if $\langle w_1, \ldots, w_m\rangle \in L_G(A)$, then there are $v_1, \ldots, v_m \in \widehat{\Delta}^*$ such that $\langle v_1, \ldots, v_m\rangle \in L_{D_\Delta}(S_m)$ and $A^{[k]} \Rightarrow_R^* v_k$ and $h(v_k) = w_k$ for each $k = 1, \ldots, m$.

Suppose that $\langle w_1, \ldots, w_m \rangle \in L_G(A)$ due to $\pi = A \to f(B_1, \ldots, B_n) \in P$ and $\langle w_{i,1}, \ldots, w_{i,m_i} \rangle \in L_G(B_i)$ for $i = 1, \ldots, n$ where $f(\langle w_{1,1}, \ldots, w_{1,m_1} \rangle, \ldots, \langle w_{n,1}, \ldots, w_{n,m_n} \rangle) = \langle w_1, \ldots, w_m \rangle$. Note that the case of $n = 0$ provides the basis of the induction.

The induction hypothesis says that for each $i = 1, \ldots, n$ we have $v_{i,1}, \ldots, v_{i,m_i} \in \widehat{\Delta}^*$ such that $\langle v_{i,1}, \ldots, v_{i,m_i} \rangle \in L_{D_\Delta}(S_{m_i})$, $h(v_{i,j}) = w_{i,j}$ and $B_i^{[j]} \Rightarrow_R^* v_{i,j}$ for $j = 1, \ldots, m_i$, where we have $B_i^{[j]} \Rightarrow_R^* v_{i,j} T \Rightarrow_R v_{i,j}$ because $T \to \varepsilon$ is the unique terminating rule of R. Let us represent f as

$$f(\langle x_{1,1}, \ldots, x_{1,m_1} \rangle, \ldots, \langle x_{n,1}, \ldots, x_{n,m_n} \rangle) = \langle t_1, \ldots, t_m \rangle$$
$$\text{where } t_k = u_{k,0} x_{i_{k1},j_{k1}} u_{k,1} \ldots x_{i_{kp_k},j_{kp_k}} u_{k,p_k} \text{ with } u_{k,0}, \ldots, u_{k,p_k} \in \Sigma^* \text{ for } k = 1, \ldots, m.$$

We define v_k by

$$v_k = \mathbb{L}_\pi^{[k]} \widetilde{u}_{k,0} \mathbb{L}_{\pi,i_{k1}}^{[j_{k1}]} v_{i_{k1},j_{k1}} \mathbb{I}_{\pi,i_{k1}}^{[j_{k1}]} \widetilde{u}_{k,1} \ldots \mathbb{L}_{\pi,i_{kp_k}}^{[j_{kp_k}]} v_{i_{kp_k},j_{kp_k}} \mathbb{I}_{\pi,i_{kp_k}}^{[j_{kp_k}]} \widetilde{u}_{k,p_k} \mathbb{I}_\pi^{[k]}. \quad (1)$$

It is easy to see that for each k, $h(v_k) = w_k$ and $A^{[k]} \Rightarrow_R^* v_k$ by $B_i^{[j]} \Rightarrow_R^* v_{i,j} T$.

Hence it is enough to show that $\langle v_1, \ldots, v_m \rangle \in L_{D_\Delta}(S_m)$. Let

$$v'_{i_{kl},j_{kl}} = \mathbb{L}_{\pi,i_{kl}}^{[j_{kl}]} v_{i_{kl},j_{kl}} \mathbb{I}_{\pi,i_{kl}}^{[j_{kl}]} \widetilde{u}_{k,l}, \quad (2)$$
$$v'_k = v'_{i_{k1},j_{k1}} \ldots v'_{i_{kp_k},j_{kp_k}} \quad (3)$$

for $l = 1, \ldots, p_k$ and $k = 1, \ldots, m$. By (1), (2), (3),

$$v_k = \mathbb{L}_\pi^{[k]} \widetilde{u}_{k,0} v'_k \mathbb{I}_\pi^{[k]}. \quad (4)$$

Applying appropriate rules of type 2 and type 3 of Definition 1 to

$$\langle v_{i,1}, \ldots, v_{i,m_i} \rangle \in L_{D_\Delta}(S_{m_i}),$$

for $i = 1, \ldots, n$, we have

$$\langle v'_{i,1}, \ldots, v'_{i,m_i} \rangle \in L_{D_\Delta}(S_{m_i})$$

by (2). Applying to those the rule $S_m \to f'(S_{m_1}, \ldots, S_{m_n})$ of type 1 where f' is obtained by removing all the occurrences of terminal symbols in the definition of f, we get

$$\langle v'_1, \ldots, v'_m \rangle \in L_{D_\Delta}(S_m)$$

by (3). By (4), appropriate rules of type 3 and type 2 provide

$$\langle v_1, \ldots, v_m \rangle \in L_{D_\Delta}(S_m). \qquad \square$$

Lemma 2. *Suppose that $A^{[k]} \Rightarrow_R^* w$ and w is well-bracketed. Then there is a rule $\pi \in P$ such that the head of π is A and the outermost parentheses of w are just $\mathbb{L}_\pi^{[k]}$, $\mathbb{I}_\pi^{[k]}$.*

Proof. The first rule for deriving w applied to $A^{[k]}$ is either $A^{[k]} \to \mathbb{L}_{\pi}^{[k]} \widetilde{u}_{k,0} \mathbb{J}_{\pi}^{[k]} T$ or $A^{[k]} \to \mathbb{L}_{\pi}^{[k]} \widetilde{u}_{k,0} \mathbb{L}_{\pi,i_{k1}}^{[j_{k1}]} B_{i_{k1}}^{[j_{k1}]}$ for some $\pi \in P$. In the former case, only $T \to \varepsilon$ can be used due to the well-bracketedness of w, because all the other rules of G' whose heads are T start by a closing parenthesis. In the latter case, the open parenthesis $\mathbb{L}_{\pi}^{[k]}$ must be closed by the succeeding derivation process. The only rule of G' for $\mathbb{J}_{\pi}^{[k]}$ is $T \to \mathbb{J}_{\pi,i_{kp_k}}^{[j_{kp_k}]} \widetilde{u}_{k,p_k} \mathbb{J}_{\pi}^{[k]} T$. Thus $A^{[k]} \Rightarrow_{G'}^{*} \mathbb{L}_{\pi}^{[k]} w' \mathbb{J}_{\pi}^{[k]} T \Rightarrow_{G'}^{*} w$, where the occurrences of $\mathbb{L}_{\pi}^{[k]}$ and $\mathbb{J}_{\pi}^{[k]}$ are corresponding. By the same reason for the former case, we must apply the rule $T \to \varepsilon$ and obtain $w = \mathbb{L}_{\pi}^{[k]} w' \mathbb{J}_{\pi}^{[k]}$. □

Lemma 3. $h(L(R) \cap L(D_{\Delta})) \subseteq L(G)$.

Proof. We show by induction that whenever $\langle w_1, \ldots, w_m \rangle \in L_{D_{\Delta}}(S_m)$ and $A^{[k]} \Rightarrow_R^{*} w_k$ for $k = 1, \ldots, m$ where $m = \dim(A)$, we have $\langle h(w_1), \ldots, h(w_m) \rangle \in L_G(A)$.

Let us consider the derivation of w_k in R. By Lemma 2, each w_k has the form $w_k = \mathbb{L}_{\pi_k}^{[k]} w'_k \mathbb{J}_{\pi_k}^{[k]}$ for some rule $\pi_k \in P$ and $w'_k \in \widehat{\Delta}^{*}$. The outermost parentheses of $\langle w_1, \ldots, w_m \rangle$ are exactly $\mathbb{L}_{\pi_1}^{[1]}, \mathbb{J}_{\pi_1}^{[1]}, \ldots, \mathbb{L}_{\pi_m}^{[m]}, \mathbb{J}_{\pi_m}^{[m]}$ and thus $\langle w_1, \ldots, w_m \rangle \in L_{D_{\Delta}}(S_m)$ implies that $\pi_1 = \pi_2 = \cdots = \pi_m$. We may hereafter omit the subscript of π_k as π. Let π be $A \to f(B_1, \ldots, B_n)$ and f represented as

$$f(\langle x_{1,1}, \ldots, x_{1,m_1} \rangle, \ldots, \langle x_{n,1}, \ldots, x_{n,m_n} \rangle) = \langle t_1, \ldots, t_m \rangle$$
$$\text{where } t_k = u_{k,0} x_{i_{k1},j_{k1}} u_{k,1} \ldots x_{i_{kp_k},j_{kp_k}} u_{k,p_k} \text{ with } u_{k,0}, \ldots, u_{k,p_k} \in \Sigma^{*} \text{ for } k = 1, \ldots, m. \quad (5)$$

If $p_k = 0$, the only rule of R that derives $\mathbb{L}_{\pi}^{[k]}$ is $A^{[k]} \to \mathbb{L}_{\pi}^{[k]} \widetilde{u}_{k,0} \mathbb{J}_{\pi}^{[k]} T$ and thus $w_k = \mathbb{L}_{\pi}^{[k]} \widetilde{u}_{k,0} \mathbb{J}_{\pi}^{[k]}$. If $p_k \geq 1$, we have

$$A^{[k]} \underset{R}{\Rightarrow} \mathbb{L}_{\pi}^{[k]} \widetilde{u}_{k,0} \mathbb{L}_{\pi,i_{k1}}^{[j_{k1}]} B_{i_{k1}}^{[j_{k1}]} \underset{R}{\overset{*}{\Rightarrow}} w_k.$$

Corresponding to the occurrence of $\mathbb{L}_{\pi,i_{k1}}^{[j_{k1}]}$, $\mathbb{J}_{\pi,i_{k1}}^{[j_{k1}]}$ must occur in w_k. The only rule that provides $\mathbb{J}_{\pi,i_{k1}}^{[j_{k1}]}$ is $T \to \mathbb{J}_{\pi,i_{k1}}^{[j_{k1}]} \widetilde{u}_{k,1} \mathbb{L}_{\pi,i_{k2}}^{[j_{k2}]} B_{i_{k2}}^{[j_{k2}]}$ unless $p_k = 1$. Thus

$$\begin{aligned} A^{[k]} &\underset{R}{\Rightarrow} \mathbb{L}_{\pi}^{[k]} \widetilde{u}_{k,0} \mathbb{L}_{\pi,i_{k1}}^{[j_{k1}]} B_{i_{k1}}^{[j_{k1}]} \\ &\underset{R}{\overset{*}{\Rightarrow}} \mathbb{L}_{\pi}^{[k]} \widetilde{u}_{k,0} \mathbb{L}_{\pi,i_{k1}}^{[j_{k1}]} v_{k,1} T \\ &\underset{R}{\Rightarrow} \mathbb{L}_{\pi}^{[k]} \widetilde{u}_{k,0} \mathbb{L}_{\pi,i_{k1}}^{[j_{k1}]} v_{k,1} \mathbb{J}_{\pi,i_{k1}}^{[j_{k1}]} \widetilde{u}_{k,1} \mathbb{L}_{\pi,i_{k2}}^{[j_{k2}]} B_{i_{k2}}^{[j_{k2}]} \\ &\underset{R}{\overset{*}{\Rightarrow}} w_k. \end{aligned}$$

for some $v_{k,1}$, which must be well-bracketed. Then we need $\rrbracket_{\pi,i_{k2}}^{[j_{k2}]}$ corresponding to the occurrence of $\llbracket_{\pi,i_{k2}}^{[j_{k2}]}$. Repeatedly applying this discussion, we finally get

$$\begin{aligned}
A^{[k]} &\underset{R}{\Rightarrow} \llbracket_{\pi}^{[k]}\, \widetilde{u}_{k,0}\, \llbracket_{\pi,i_{k1}}^{[j_{k1}]}\, B_{i_{k1}}^{[j_{k1}]} \\
&\underset{R}{\overset{*}{\Rightarrow}} \llbracket_{\pi}^{[k]}\, \widetilde{u}_{k,0}\, \llbracket_{\pi,i_{k1}}^{[j_{k1}]}\, v_{k,1} T \\
&\underset{R}{\overset{*}{\Rightarrow}} \llbracket_{\pi}^{[k]}\, \widetilde{u}_{k,0}\, \llbracket_{\pi,i_{k1}}^{[j_{k1}]}\, v_{k,1}\, \rrbracket_{\pi,i_{k1}}^{[j_{k1}]}\, \widetilde{u}_{k,1} \dots\, \llbracket_{\pi,i_{kp_k}}^{[j_{kp_k}]}\, v_{i_{kp_k},j_{kp_k}} T \\
&\underset{R}{\Rightarrow} \llbracket_{\pi}^{[k]}\, \widetilde{u}_{k,0}\, \llbracket_{\pi,i_{k1}}^{[j_{k1}]}\, v_{k,1}\, \rrbracket_{\pi,i_{k1}}^{[j_{k1}]}\, \widetilde{u}_{k,1} \dots\, \llbracket_{\pi,i_{kp_k}}^{[j_{kp_k}]}\, v_{k,p_k}\, \rrbracket_{\pi,i_{kp_k}}^{[j_{kp_k}]}\, \widetilde{u}_{k,p_k}\, \rrbracket_{\pi}^{[k]}\, T \\
&\underset{R}{\overset{*}{\Rightarrow}} w_k.
\end{aligned}$$

This holds for any $p_k \geq 1$. By Lemma 2

$$w_k = \llbracket_{\pi}^{[k]}\, \widetilde{u}_{k,0}\, \llbracket_{\pi,i_{k1}}^{[j_{k1}]}\, v_{k,1}\, \rrbracket_{\pi,i_{k1}}^{[j_{k1}]}\, \widetilde{u}_{k,1} \dots\, \llbracket_{\pi,i_{kp_k}}^{[j_{kp_k}]}\, v_{k,p_k}\, \rrbracket_{\pi,i_{kp_k}}^{[j_{kp_k}]}\, \widetilde{u}_{k,p_k}\, \rrbracket_{\pi}^{[k]}\,.$$

Let $w_{i,j} = v_{k,l}$ if $x_{i,j}$ occurs as the l-th variable in t_k, i.e., $w_{i_{kl},j_{kl}} = v_{k,l}$. We note that $B_i^{[j]} \Rightarrow_R^* w_{i,j} T \Rightarrow_R w_{i,j}$ and

$$h(w_k) = u_{k,0} h(w_{i_{k1},j_{k1}}) u_{k,1} \dots h(w_{i_{kp_k},j_{kp_k}}) u_{k,p_k}. \tag{6}$$

Applying Lemma 2 to each $w_{i,j}$, which must be well-bracketed, we have $w_{i,j} = \llbracket_{\rho_{i,j}}^{[j]}\, w'_{i,j}\, \rrbracket_{\rho_{i,j}}^{[j]}$ for some rule $\rho_{i,j}$ of G. Here the third outermost parentheses of $\langle w_1, \dots, w_m \rangle$ consist of exactly $\sum_{1 \leq i \leq n} m_i$ pairs

$$\langle\, \llbracket_{\rho_{i,1}}^{[1]}\,,\, \rrbracket_{\rho_{i,1}}^{[1]}\, \rangle, \dots, \langle\, \llbracket_{\rho_{i,m_i}}^{[m_i]}\,,\, \rrbracket_{\rho_{i,m_i}}^{[m_i]}\, \rangle \text{ for } i = 1, \dots, n.$$

By $\langle w_1, \dots, w_m \rangle \in L_{D_\Delta}(S_m)$, for each $i = 1, \dots, n$ and $j = 1, \dots, m_i$, all of

$$\langle\, \llbracket_{\rho_{i,j}}^{[1]}\,,\, \rrbracket_{\rho_{i,j}}^{[1]}\, \rangle, \dots, \langle\, \llbracket_{\rho_{i,j}}^{[m_i]}\,,\, \rrbracket_{\rho_{i,j}}^{[m_i]}\, \rangle$$

must occur as third outermost parentheses in $\langle w_1, \dots, w_m \rangle$. Recall that the head of the rule $\rho_{i,j}$ is B_i and that $B_i = B_{i'}$ implies $i = i'$. Hence $i \neq i'$ implies $\rho_{i,j} \neq \rho_{i',j'}$ for any j and j'. Therefore for any i, j, it holds that

$$\langle\, \llbracket_{\rho_{i,1}}^{[1]}\,,\, \rrbracket_{\rho_{i,1}}^{[1]}\, \rangle = \langle\, \llbracket_{\rho_{i,j}}^{[1]}\,,\, \rrbracket_{\rho_{i,j}}^{[1]}\, \rangle, \dots, \langle\, \llbracket_{\rho_{i,m_i}}^{[m_i]}\,,\, \rrbracket_{\rho_{i,m_i}}^{[m_i]}\, \rangle = \langle\, \llbracket_{\rho_{i,j}}^{[m_i]}\,,\, \rrbracket_{\rho_{i,j}}^{[m_i]}\, \rangle$$

and we have ρ_i such that $\rho_i = \rho_{i,1} = \cdots = \rho_{i,m_i}$. In the derivation of $\langle w_1, \dots, w_m \rangle \in L_{D_\Delta}(S_m)$, at some point the rule $S_{m_i} \to f_{\llbracket_{\rho_i}}(S_{m_i})$ of type 2, where $f_{\llbracket_{\rho_i}}(\langle x_1, \dots, x_m \rangle) = \langle\, \llbracket_{\rho_i}^{[1]}\, x_1\, \rrbracket_{\rho_i}^{[1]}, \dots,\, \llbracket_{\rho_i}^{[m_i]}\, x_m\, \rrbracket_{\rho_i}^{[m_i]}\, \rangle$, must be applied to $\langle w_{i,1}, \dots, w_{i,m_i} \rangle \in L_{D_\Delta}(S_{m_i})$. By the induction hypothesis, we have $\langle h(w_{i,1}), \dots, h(w_{i,m_i}) \rangle \in L_G(B_i)$ for $i = 1, \dots, n$. Applying the rule π to those tuples, we obtain by (5) and (6)

$$\begin{aligned}
&f(\langle h(w_{1,1}), \dots, h(w_{1,m_1}) \rangle, \dots, \langle h(w_{n,1}), \dots, h(w_{n,m_n}) \rangle) \\
&\qquad = \langle h(w_1), \dots, h(w_m) \rangle \in L_G(A). \quad \square
\end{aligned}$$

Theorem 3. *A language L is in q-MCFL(r) if and only if there are a multiple Dyck language $D \in q$-MCFL(r), a regular language R and a homomorphism h such that*

$$L = h(D \cap R).$$

Proof. By Lemmas 1 and 3 and Proposition 2. □

3.4 Generator Theorem

It is easy to get the stronger Chomsky-Schützenberger-type characterization for q-MCFL(r) by the standard technique.

Let

$$\Delta' = \{\, \mathbb{[}_a \mid a \in \Sigma \,\} \cup \{\, \mathbb{[}_m, \mathbb{[}_m \mid 1 \le m \le q \,\}$$

where $\dim(a) = 1$ and $\dim(\mathbb{[}_m) = \dim(\mathbb{[}_m) = m$ and $h' : \widehat{\Delta'}^* \to \Sigma^*$ be the homomorphism mapping each $\mathbb{[}_a$ to a for $a \in \Sigma$ and other symbols to the empty string.

For a given mcfg $G \in q$-MCFG(r), let Δ and R be the indexed alphabet and the right-linear grammar from Section 3.2, respectively. Let us enumerate all the elements of dimension m in $\Delta \setminus \{\, \mathbb{[}_a \mid a \in \Sigma \,\}$ and denote them by $\mathbb{[}_{m,1}, \ldots, \mathbb{[}_{m,k_m}$ for each m. We then define a right-linear grammar R' from R by replacing $\mathbb{[}^{[j]}_{m,i}$ with $\mathbb{[}^{[j]}_m \underbrace{\mathbb{[}^{[j]}_m \ldots \mathbb{[}^{[j]}_m}_{i\text{-times}} \mathbb{[}^{[j]}_m$ and $\mathbb{]}^{[j]}_{m,i}$ with $\mathbb{]}^{[j]}_m \underbrace{\mathbb{]}^{[j]}_m \ldots \mathbb{]}^{[j]}_m}_{i\text{-times}} \mathbb{]}^{[j]}_m$.

We have

$$L(G) = h'(L(D_{\Delta'}) \cap L(R')).$$

Corollary 1. *There are a multiple Dyck language $D \in q$-MCFL(r) and a homomorphism h such that a language L is in q-MCFL(r) if and only if there is a regular language R such that $L = h(D \cap R)$.*

4 Conclusion

This paper introduced multiple Dyck languages and then proved a Chomsky-Schützenberger-type representation theorem for each class q-MCFL(r) as well as the generator theorem. The literature (e.g. [4,7]) has proposed other parameters such as *degree* and *well-nestedness* that give further classifications of mcfls. Theorem 3 and Corollary 1 hold for those subclasses as well by accordingly modifying the definition of rules of type 1 of multiple Dyck grammars in Definition 1.

Logical characterizations for several classes of languages have been obtained in the literature. For example, the class of regular languages coincides with the class of languages that are definable in monadic second-order logic (see [9]). Also, the class of cfls is exactly the class of languages definable in an existential second-order logic where the second-order variable ranges only over matching predicates [5]. A *matching* predicate M is a binary predicate over the set of positions of symbols in a given string such that each position belongs to at most one pair (i, j) satisfying $M(i, j)$ and M is not crossing ($(i, j) \in M$, $(k, l) \in M$ and

$i < k < j$ imply $i < l < j$). Intuitively, $M(i, j)$ means that the symbols occurring at the positions i and j form a pair of a left parenthesis and its corresponding right one. This suggests us to extend a matching predicate to a $2r$-ary predicate M_r to express r pairs of left and right parentheses in $\widehat{\Delta}$ of Section 2.3. It is left as future study to give a logic that characterizes mcfls by using these extended matching predicates.

Acknowledgement

This work was partially supported by the NII joint research project "Open Problems on Multiple Context-Free Grammars" from National Institute of Informatics, Grant-in-Aid for Young Scientists (B-20700124), and a grant from the Global COE Program "Center for Next-Generation Information Technology based on Knowledge Discovery and Knowledge Federation" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Noam Chomsky and Marcel-Paul Schützenberger. The algebraic theory of context-free languages. In P. Braffort and D. Hirschberg, editors, *Computer Programming and Formal Systems*, pages 118–161. North Holland, 1963.
2. Joost Engelfriet. *Context-Free Graph Grammars.* in *Handbook of Formal Languages*, volume 3, chapter 3, section 6, Springer, 1997.
3. Yuichi Kaji. Universal recognition problems and a representation theorem using Dyck-type languages for multiple context-free grammars. Bachelor's Thesis, Osaka University, 1991.
4. Makoto Kanazawa. The convergence of well-nested mildly context- sensitive grammar formalisms. An invited talk given at the 14th Conference on Formal Grammar, Bordeaux, France. Slides available at `http://research.nii.ac.jp/~kanazawa/`.
5. Clemens Lautemann, Thomas Schwentick, and Denis Thérien. Logics for context-free languages. In *Computer Science Logic*, Volume 933 of *Lecture Notes in Computer Science*, pages 205–216. Springer, 1995.
6. Owen Rambow and Giorgio Satta. Independent parallelism in finite copying parallel rewriting systems. *Theoretical Computer Science*, 223(1-2):87–120, 1999.
7. Hiroyuki Seki and Yuki Kato. On the generative power of multiple context-free grammars and macro grammars. *IEICE Transactions*, 91-D(2):209–221, 2008.
8. Hiroyuki Seki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. On multiple context-free grammars. *Theoretical Computer Science*, 88(2):191–229, 1991.
9. Wolfgang Thomas. *Languages, Automata, and Logic*, in *Handbook of Formal Languages*, volume 3, chapter 7, Springer, 1997.

A Appendix

The *degree* of a linear regular function $f : (\Sigma^*)^{m_1} \times \cdots \times (\Sigma^*)^{m_n} \to (\Sigma^*)^m$ is defined to be $\deg(f) = m + m_1 + \cdots + m_n$ and the *degree* of an mcfg G is the maximum of $\deg(f)$ for $f \in F$. The functions of rules of type 2 and type 3

of multiple Dyck grammar D_Σ in Definition 1 have degree $2r$ where r is the maximum dimension of letters of Σ. Indeed we may assume without loss of generality that $\deg(G) \geq 2\dim(G)$, because of the following lemma.

Lemma 4. *If* $2\dim(G) > \deg(G) \geq 2$, *there is* G' *equivalent to* G *such that* $\dim(G') < \dim(G)$ *and* $\deg(G') \leq \deg(G)$.

Proof. Without loss of generality, we assume that G is nonerasing. We present a method for eliminating rules whose right hand side has a nonterminal B such that $2\dim(B) > \deg(G)$. Let

$$A \to f(B_1, \ldots, B_n) \text{ with } f(\vec{x}_1, \ldots, \vec{x}_n) = \vec{t}$$

be a rule such that $2\dim(B_k) > \deg(G)$ for some B_k. Since $2|\vec{x}_k| > \deg(G) \geq |\vec{t}| + |\vec{x}_1| + \cdots + |\vec{x}_n|$, there is t in $\vec{t}$ such that $t = t_1 x_{k,i} u x_{k,j} t_2$ for some $x_{k,i}, x_{k,j}$ from $\vec{x}_k = \langle x_{k,1}, \ldots, x_{k,m} \rangle$, $u \in \Sigma^*$ and $t_1, t_2 \in (\Sigma \cup X)^*$. Then we introduce a fresh nonterminal B'_k with $\dim(B'_k) = \dim(B_k) - 1$ and replace the rule by

$$\begin{gathered} A \to f'(B_1, \ldots, B_{k-1}, B'_k, B_{k+1}, \ldots, B_n) \\ \text{with } f'(\vec{x}_1, \ldots, \vec{x}_{k-1}, \vec{x}'_k, \vec{x}_{k+1}, \ldots, \vec{x}_n) = \vec{t}' \end{gathered}$$

where $\vec{x}'_k$ is obtained from $\vec{x}_k$ by deleting $x_{k,j}$ and $\vec{t}'$ is obtained from $\vec{t}$ by replacing t with $t_1 x_{k,i} t_2$. For each rule whose head is B_k

$$B_k \to g(C_1, \ldots, C_m) \text{ with } g(\vec{x}_1, \ldots, \vec{x}_m) = \vec{s},$$

we add the rule

$$B'_k \to g'(C_1, \ldots, C_m) \text{ with } g'(\vec{x}_1, \ldots, \vec{x}_m) = \vec{s}'$$

where $\vec{s}'$ is obtained from $\vec{s} = \langle s_1, \ldots, s_m \rangle$ by deleting s_j and replacing s_i with $s_i u s_j$. We note that every C_l in the body of the new rule satisfies $2\dim(C_l) < \deg(G)$ by $\deg(G) \geq \dim(B_k) + \dim(C_l)$, and hence the introduced rule will not be a target of repetitive applications of this procedure. It is easy to see that this procedure preserves the language and never increases the degree or the branching factor. □