PAPER

On the Generative Power of Cancel Minimal Linear Grammars with Single Nonterminal Symbol except the Start Symbol

Kaoru FUJIOKA[†a)] *and* **Hirofumi KATSUNO**[††b)], ***Members***

SUMMARY This paper concerns cancel minimal linear grammars ([5]) that was introduced to generalize Geffert normal forms for phrase structure grammars. We consider the generative power of restricted cancel minimal linear grammars: the grammars have only one nonterminal symbol C except the start symbol S, and their productions consist of context-free type productions, the left-hand side of which is S and the right-hand side contains at most one occurrence of S, and a unique cancellation production $C^m \rightarrow \epsilon$ that replaces the string C^m by the empty string ϵ. We show that, for any given positive integer m, the class of languages generated by cancel minimal linear grammars with $C^m \rightarrow \epsilon$, is properly included in the class of linear languages. Conversely, we show that for any linear language L, there exists some positive integer m such that a cancel minimal linear grammar with $C^m \rightarrow \epsilon$ generates L. We also show how the generative power of cancel minimal linear grammars with a unique cancellation production $C^m \rightarrow \epsilon$ vary according to changes of m and restrictions imposed on occurrences of terminal symbols in the right-hand side of productions.
key words: *minimal linear languages, linear languages, Geffert normal forms, generative power*

1. Introduction

Among the variety of normal forms for phrase structure (or type-0) grammars ([1], [3], [7]), Geffert normal forms [1] are unique in the sense that each of them has two different kinds of productions: context-free type productions with only the start symbol S on the left-hand side, and the one or two cancellation productions that replace a sequence of nonterminal symbols except S with the empty string ϵ. The cancellation productions in each Geffert normal form play a vital role to generate any recursively enumerable languages. Furthermore, the cancellation productions are related to cutting operations of DNA strands, and Geffert normal forms are used to examine the generative power of DNA computing models ([6], [8]). Each Geffert normal form also provides an "intermediate grammar" that can bridge a gap between context-free and recursively enumerable languages ([2], [4], [9]).

Onodera [5] gives a framework in which each Geffert normal form is uniformly described as a grammar referred to as a *cancel minimal linear grammar*. A cancel minimal linear grammar has two kinds of productions: context-free type productions, the left-hand side of which is the start symbol S and the right-hand side contains at most one occurrence of S, and cancellation productions similarly defined as the case of Geffert normal forms. Note that, if we regard each nonterminal symbol except S as a terminal symbol, then the context-free type productions above are considered to be minimal linear, and hence we call the context-free type productions minimal linear type productions in this paper. Within the framework of cancel minimal linear grammars, one of the Geffert's results ([1]) means that the cancel minimal linear grammar with only two cancellation productions $AB \rightarrow \epsilon$ and $CC \rightarrow \epsilon$ has the power of generating any recursively enumerable language.

Onodera [5] examines the generative power of the cancel minimal linear grammars with only one of the two above cancellation productions, under the assumption of dealing with only ϵ-free languages. She shows that the language generated by any cancel minimal linear grammar with $AB \rightarrow \epsilon$ is context-free, and that any linear language can be generated by such a grammar. Furthermore, she shows that the class of languages generated by the cancel minimal linear grammars with $CC \rightarrow \epsilon$ is a proper subset of the class of linear languages.

In this paper, we study the generative power of cancel minimal linear grammars with $C^m \rightarrow \epsilon$ for an arbitrarily fixed $m \geq 1$ without assuming that only ϵ-free languages are allowed. We show that for any given $m \geq 1$, cancel minimal linear grammars with $C^m \rightarrow \epsilon$ only generate linear languages. In contrast to this, for $C^m \rightarrow \epsilon$ with m not bounded, we show that the class of languages generated by those grammars is equivalent to the class of linear languages. We also examine the difference in the generative power of the cancel minimal linear grammars between with $C^m \rightarrow \epsilon$ and with $C^n \rightarrow \epsilon$ for $m \neq n$.

We impose some restrictions on occurrences of terminal symbols in the minimal linear type productions, and show how the restrictions affect the generative power of the cancel minimal linear grammars with $C^m \rightarrow \epsilon$.

These results may shed some new light on relations of language classes between minimal linear languages and linear languages.

2. Preliminaries

We assume the reader to be familiar with the rudiments of formal language theory (see, e.g., Rozenberg and Salomaa [7]).

A *phrase structure grammar* (a *grammar* for short) is a construct $G = (N, T, P, S)$, where N is a set of *nonterminal*

Manuscript received November 24, 2010.
Manuscript revised May 21, 2011.
[†]The author is with the Office for Strategic Research Planning, Kyushu University, Fukuoka-shi, 812–8581 Japan.
[††]The author is with the Department of Science and Engineering, Tokyo Denki University, Saitama-ken, 350–0394 Japan.
a) E-mail: kaoru@tcslab.csce.kyushu-u.ac.jp
b) E-mail: katsuno@mail.dendai.ac.jp
DOI: 10.1587/transinf.E94.D.1945

Copyright © 2011 The Institute of Electronics, Information and Communication Engineers

symbols, T is a set of *terminal symbols*, P is a set of *productions*, and S in N is the *start symbol*. A production in P is of the form $\pi_1 \to \pi_2$, where $\pi_1 \in (N \cup T)^* N (N \cup T)^*$ and $\pi_2 \in (N \cup T)^*$. For any α_1 and α_2 in $(N \cup T)^*$, if $\alpha_1 = \alpha_{11}\pi_1\alpha_{12}$, $\alpha_2 = \alpha_{11}\pi_2\alpha_{12}$, and $r : \pi_1 \to \pi_2 \in P$, then we say that α_2 *is derivable from* α_1 *by* r, and write $\alpha_1 \overset{r}{\Longrightarrow}_G \alpha_2$. If G is understood, we write $\alpha_1 \overset{r}{\Longrightarrow} \alpha_2$. Similarly, for a sequence of productions γ, we simply write $\alpha_1 \overset{\gamma}{\Longrightarrow} \alpha_2$. Further, if there is no need to refer to productions, then we simply write $\alpha_1 \Longrightarrow \alpha_2$, and we denote the reflexive and transitive closure of $\Longrightarrow$ by $\Longrightarrow^*$. A string in $(N \cup T)^*$ derivable from the start symbol S is called a *sentential form*.

We define the *language* $L(G)$ generated by a grammar $G = (N, T, P, S)$ as follows: $L(G) = \{z \in T^* \mid S \Longrightarrow^* z\}$. It is well known that the class of languages generated by the phrase structure grammars is equal to the class of *recursively enumerable languages*.

A language L is said to be ϵ*-free*, if it contains no *empty string* ϵ. In this paper, we mainly deal with ϵ-free languages.

A grammar $G = (N, T, P, S)$ is *linear* if each production in P is of the form $N_i \to \alpha$, where $N_i \in N$ and α contains at most one nonterminal symbol. A language generated by any linear grammar is also called *linear*. It is obvious that any linear language can be generated by a linear grammar each of whose productions is of the form $N_1 \to uN_2$, $N_1 \to N_2u$, or $N_1 \to u$, where $N_1, N_2 \in N$ and $u \in T^*$.

A grammar $G = (N, T, P, S)$ is *right* (resp. *left*) *linear* if it is linear and every production in P is of the form $N_1 \to uN_2$ or $N_1 \to u$ (resp. $N_1 \to N_2u$ or $N_1 \to u$), where $N_1, N_2 \in N$ and $u \in T^*$. Any language generated by such a grammar is called *right* (resp. *left*) *linear*. It is well known that the class of right linear languages is equivalent to that of left linear languages, which is also called the class of *regular languages*.

A grammar $G = (N, T, P, S)$ is *minimal linear* if $N = \{S\}$ and every production in P is of the form $S \to uSv$ or $S \to w$, where $u, v, w \in T^*$. Any language generated by such a grammar is called *minimal linear*.

Geffert [1] shows the following theorem.

Theorem 1: Any recursively enumerable language can be generated by a grammar $G = (\{S\} \cup N_C, T, P \cup P_C, S)$ satisfying the following conditions:

- Every production in P is of the form $S \to \alpha_1 S \alpha_2$ or $S \to \alpha$, where $\alpha_1, \alpha_2, \alpha \in (T \cup N_C)^*$,
- $N_C = \{A, B, C\}$ and $P_C = \{AB \to \epsilon,\ CC \to \epsilon\}$.

Note that Geffert examines the other four cases as variations of Theorem 1:

(1) $N_C = \{A, B, C, D\}$, $P_C = \{AB \to \epsilon,\ CD \to \epsilon\}$.
(2) $N_C = \{A, B\}$, $P_C = \{AB \to \epsilon,\ BBB \to \epsilon\}$.
(3) $N_C = \{A, B\}$, $P_C = \{ABBBA \to \epsilon\}$.
(4) $N_C = \{A, B, C\}$, $P_C = \{ABC \to \epsilon\}$.

He shows that in each case any recursively enumerable language can be generated by a grammar $G = (\{S\} \cup N_C, T, P \cup P_C, S)$.

Generalizing these *Geffert normal forms*, Onodera [5] introduces a new class of grammars as follows.

Definition 1: A grammar $G = (\{S\} \cup N_C, T, P, S)$ is an Ω*-cancel minimal linear grammar* (Ω*-cml grammar* for short) if it satisfies the following:
(1) S is the start symbol.
(2) N_C is a finite set of nonterminal symbols except S.
(3) T is a finite set of terminal symbols.
(4) Ω is a finite set of strings in N_C^+.
(5) P is a finite set of productions and is partitioned into two parts P_M and P_C defined as follows:
  (a) $P_M \subseteq \{ S \to \alpha_1 S \alpha_2,\ S \to \alpha \mid \alpha_1, \alpha_2, \alpha \in (T \cup N_C)^* \}$,
  (b) $P_C = \{ \omega \to \epsilon \mid \omega \in \Omega \}$.

We call a production in P_M a *minimal linear type production* (an *ml-production* for short) and call a production in P_C a *cancellation production* (a *c-production* for short).

A language L is an Ω*-cancel minimal linear language* (Ω*-cml language* for short) if there is an Ω-cml grammar G such that $L = L(G)$.

For a string α, α^R represents the reverse of α, and let $|\alpha|_T$ be the number of terminal symbols in α.

Definition 2: If an ml-production has the right side with no terminal symbol, then the production is called a *terminal-free ml-production*, otherwise it is called a *terminal ml-production*. If a terminal ml-production is of one of the forms $S \to \alpha_1 S \alpha_2$, $S \to \alpha S$, $S \to S\alpha$, $S \to \alpha$, where $|\alpha_1|_T, |\alpha_2|_T, |\alpha|_T > 0$, then it is called *strict terminal ml-production* (*s-terminal ml-production* for short).

An Ω-cml grammar G is called a *terminal* (resp. *strict terminal* or *s-terminal* for short) Ω*-cml grammar*, if any ml-production in P is a terminal (resp. an s-terminal) production. A language L is called a *terminal* (resp. an *s-terminal*) Ω*-cml language* if there is a terminal (resp. an s-terminal) Ω-cml grammar that generates L.

When we deal with only ϵ-free languages, the classes of linear, minimal linear, Ω-cancel minimal linear, and regular languages are denoted by LIN, ML, CML_Ω, and REG, respectively. If we deal with Ω-cancel minimal linear languages containing the empty string, then we denote the language class by $\mathrm{CML}_\Omega^\epsilon$.

Note that terminal Ω-cancel minimal linear languages and strict terminal Ω-cancel minimal linear languages do not contain the empty string. Let r_ϵ be the terminal-free ml-production $S \to \epsilon$.

Definition 3: An Ω-cml grammar which has no terminal-free ml-production except r_ϵ is called an *extended terminal* Ω*-cml grammar* and a language L is called an *extended terminal* Ω*-cml language* if there is an extended terminal Ω-*cml grammar* that generates L.

The classes of terminal Ω-cml, strict terminal Ω-cml, and extended terminal Ω-cml languages are denoted by t-CML_Ω, st-CML_Ω, and t-$\mathrm{CML}_\Omega^\epsilon$, respectively.

Onodera [5] examines the generative power of some classes of $\{AB\}$-cml grammars and $\{C^2\}$-cml grammars.

Concerning $\{C^2\}$-cml grammars, she proves the following theorem.

Theorem 2:

1. ML $\subset$ t-CML$_{\{C^2\}}$ $\subset$ LIN
2. REG and t-CML$_{\{C^2\}}$ are incomparable.

In this paper, for any positive integer m, we deal with $\{C^m\}$-cml grammars.

3. $\{C^m\}$-cml Languages

In this section, we consider the generative power of terminal $\{C^m\}$-cml grammars. In the case of $m = 1$, a $\{C\}$-cml grammar is context-free, and we can remove the nonterminal symbol C from the original grammar by the standard technique for eliminating ϵ-rules from a context-free grammar. Therefore, the following lemma is obvious.

Lemma 1: st-CML$_{\{C\}}$ =t-CML$_{\{C\}}$ =CML$_{\{C\}}$ =ML.

In the following, we consider the case $m \geq 2$.

3.1 $\{C^m\}$-cml Languages and Terminal $\{C^m\}$-cml Languages

In this subsection, for any given integer $m \geq 2$, we consider the generative power of $\{C^m\}$-cml grammars and terminal $\{C^m\}$-cml grammars. In particular, we examine influences of terminal-free ml-productions on the generative power.

In every $\{C^m\}$-cml grammar $G = (\{S, C\}, T, P, S)$, we may assume that any ml-production in P is of one of the six forms

(1)	$S \to C^iuC^kSC^lvC^j$,	(2)	$S \to C^iuC^kSC^j$,
(3)	$S \to C^iSC^lvC^j$,	(4)	$S \to C^iuC^j$,
(5)	$S \to C^iSC^j$,	(6)	$S \to C^i$,

where $u, v \in T^+$, $0 \leq i, j, k, l < m$.

This is because any ml-production can be transformed into one of the above forms by using the c-production $r_C : C^m \to \epsilon$, or the ml-production makes no contribution to producing a string in T^*. For example, an ml-production $S \to C^{m+i}uC^kSC^{2m+l}vC^j$ with $u, v \in T^+$ and $0 \leq i, j, k, l < m$, is equivalent to $S \to C^iuC^kSC^lvC^j$, whereas an ml-production $S \to uC^ivS$ with $u, v \in T^+$ and $0 < i < m$ is useless to produce a string in T^*.

Note that when we use a notation $S \to C^iuC^{m-i}S$ with $i = 0$, the ml-production is not of the form (2) above, because C^m is not allowed in the form. Hence, in the following, we regard C^m as ϵ in such cases.

According to the six forms above, we partition the set of ml-productions P_M into six sets $P(1), P(2), \ldots, P(6)$ such that for each n ($1 \leq n \leq 6$), $P(n)$ consists of ml-productions in the n-th form above. For example,

$$P(1) = \{r \mid r : S \to C^iuC^kSC^lvC^j \text{ in } P \text{ and } u, v \in T^+\}.$$

Each terminal ml-production in P is in $P(1) \cup P(2) \cup P(3) \cup P(4)$, while each terminal-free ml-production in P is in $P(5) \cup P(6)$. Hence, we denote $P(1) \cup P(2) \cup P(3) \cup P(4)$ by $P(t)$, and $P(5) \cup P(6) \cup \{r_C\}$ by $P(tf)$. In the following, we call a production in $P(tf)$ a terminal-free production.

If we use only productions in $P(tf)$, we cannot produce strings of terminal symbols except for ϵ, hence there is a possibility that t-CML$_{\{C^m\}}$ may be equal to CML$_{\{C^m\}}$. To prove the claim, we define a set of terminal ml-productions derived from a terminal ml-production by using terminal-free productions.

Definition 4: Let $G = (\{S, C\}, T, P, S)$ be a $\{C^m\}$-cml grammar. For an ml-production $r : S \to C^iuC^kSC^lvC^j$ in $P(1)$, we define the *closure of r under terminal-free productions*, denoted by $cl(r)$, as $cl(r) = cl_1(r) \cup cl_2(r)$, where

- $S \to C^{i'}uC^{k'}SC^{l'}vC^{j'}$ is in $cl_1(r)$
  iff there exists a derivation γ_1 such that
  (1) $S \xRightarrow{\gamma_1}_G C^{i'}uC^{k'}SC^{l'}vC^{j'}$,
  (2) r occurs only once in γ_1, and the rest of γ_1 are productions in $P(tf)$,
  (3) $0 \leq i', j', k', l' < m$,

- $S \to C^{i'}uvC^{j'}$ is in $cl_2(r)$
  iff there exists a derivation γ_2 such that
  (1) $S \xRightarrow{\gamma_2}_G C^{i'}uvC^{j'}$,
  (2) r occurs only once in γ_2, and the rest of γ_2 are productions in $P(tf)$,
  (3) $0 \leq i', j' < m$.

Similarly, for $r : S \to C^iuC^kSC^j$ in $P(2)$ and $r : S \to C^iSC^lvC^j$ in $P(3)$, we define $cl(r)$ as $cl(r) = cl_1(r) \cup cl_2(r)$, and for $r : S \to C^iuC^j$ in $P(4)$, we define $cl(r)$ as $cl(r) = cl_2(r)$.

By using $cl(r)$ with r in $P(t)$, we define the *closure of $P(t)$ under terminal-free productions* as

$$cl(P(t)) = \cup_{r \in P(t)} cl(r).$$

Lemma 2: $cl(P(t))$ is a finite set of terminal ml-productions.

Proof : Since each definition of $cl(r)$ with r in $P(t)$ imposes the condition $0 \leq i', j', k', l' < m$, the number of terminal ml-productions in $cl(r)$ is finite. Therefore, it follows from the definition of $cl(P(t))$ that $cl(P(t))$ is a finite set of terminal ml-productions. □

Next, by using the closure $cl(P(t))$, we construct a terminal $\{C^m\}$-cml grammar $\overline{G}$ from a $\{C^m\}$-cml grammar $G = (\{S, C\}, T, P, S)$ which satisfies that if $L(G)$ is ϵ-free then $L(\overline{G}) = L(G)$.

Definition 5: A terminal $\{C^m\}$-cml grammar $\overline{G} = (\{S, C\}, T, \overline{P}, S)$ is the *transformed terminal $\{C^m\}$-cml grammar of* $G = (\{S, C\}, T, P, S)$ if $\overline{P} = cl(P(t)) \cup P_C$.

Theorem 3: CML$_{\{C^m\}}$ =t-CML$_{\{C^m\}}$.

Proof : It is obvious that t-CML$_{\{C^m\}}$ $\subseteq$CML$_{\{C^m\}}$ holds. We will prove the converse by showing that for any $\{C^m\}$-cml grammar G, if $L(G)$ is ϵ-free then its transformed terminal $\{C^m\}$-cml grammar $\overline{G}$ generates the same language as $L(G)$.

The inclusion $L(\overline{G}) \subseteq L(G)$ is obvious from the definition of the closure of a terminal ml-production under terminal-free productions. In order to prove the converse inclusion, we will show that, for $0 \le i, j, i', j' < m$ and $w \in T^+$, if $C^iSC^j \overset{\gamma}{\Longrightarrow}_G C^{i'}wC^{j'}$ then $C^iSC^j \Longrightarrow^*_{\overline{G}} C^{i'}wC^{j'}$ by using the induction on the number n of terminal ml-productions applied in the derivation γ. We note that when $i = j = i' = j' = 0$, the claim implies $L(G) \subseteq L(\overline{G})$.

Base step, $n = 1$: Consider a derivation γ such that $C^iSC^j \overset{\gamma}{\Longrightarrow}_G C^{i'}wC^{j'}$, where a terminal ml-production r in $P(t)$ occurs only once in γ, and the rest of γ are productions in $P(tf)$. Then, there exists a derivation γ' such that $S \overset{\gamma'}{\Longrightarrow}_G C^pwC^q$, where $p = (m + i' - i) \bmod m$, $q = (m + j' - j) \bmod m$, r occurs only once in γ, and the rest of γ are productions in $P(tf)$. It follows from the definition of $cl(r)$ that $S \to C^pwC^q$ is a member of $cl(r)$. Therefore, $C^iSC^j \Longrightarrow^*_{\overline{G}} C^{i'}wC^{j'}$ holds.

Induction step: Assume that $C^iSC^j \overset{\gamma}{\Longrightarrow}_G C^{i'}wC^{j'}$ and the total number of terminal ml-production occurrences in γ is $n + 1$. Then, there exists one of the following derivations:

(1) $C^iSC^j \overset{\gamma_1}{\Longrightarrow}_G C^{i'}uC^kSC^lvC^{j'} \overset{\gamma_2}{\Longrightarrow}_G C^{i'}wC^{j'}$,
(2) $C^iSC^j \overset{\gamma_1}{\Longrightarrow}_G C^{i'}uC^kSC^l \overset{\gamma_2}{\Longrightarrow}_G C^{i'}wC^{j'}$,
(3) $C^iSC^j \overset{\gamma_1}{\Longrightarrow}_G C^kSC^lvC^{j'} \overset{\gamma_2}{\Longrightarrow}_G C^{i'}wC^{j'}$,

where $0 \le k, l < m$, γ_1 contains only one terminal ml-production occurrence, and the total number of terminal ml-production occurrences in γ_2 is n.

We will show that, in the second case, $C^iSC^j \Longrightarrow^*_{\overline{G}} C^{i'}wC^{j'}$ holds. Since γ_1 is a sequence of productions in $P(tf)$ except one occurrence of a terminal ml-production r, there exists a derivation γ'_1 such that $S \overset{\gamma'_1}{\Longrightarrow}_G C^puC^kSC^q$, where $p = (m + i' - i) \bmod m$, $q = (m + l - j) \bmod m$, r occurs only once in γ'_1, and the rest of γ'_1 are productions in $P(tf)$. From the definitions of $cl(r)$ and $\overline{G}$, $S \to C^puC^kSC^q$ is a member of both $cl(r)$ and $\overline{P}$. Therefore, $C^iSC^j \Longrightarrow^*_{\overline{G}} C^{i'}uC^kSC^l$ holds.

On the other hand, it follows from $C^{i'}uC^kSC^l \overset{\gamma_2}{\Longrightarrow}_G C^{i'}wC^{j'}$ that there exists a string $w' \in T^+$ such that $w = uw'$ and $C^kSC^l \overset{\gamma_2}{\Longrightarrow}_G w'C^{j'}$. Then, by the induction hypothesis, $C^kSC^l \Longrightarrow^*_{\overline{G}} w'C^{j'}$ holds. Therefore, there is a derivation $C^iSC^j \Longrightarrow^*_{\overline{G}} C^{i'}uC^kSC^l \Longrightarrow^*_{\overline{G}} C^{i'}uw'C^{j'} = C^{i'}wC^{j'}$.

The proof of the first and the third cases is analogous to the proof of the second case. □

Corollary 1: $\mathrm{CML}^{\epsilon}_{\{C^m\}}$ =t-$\mathrm{CML}^{\epsilon}_{\{C^m\}}$. Moreover, for any $\{C^m\}$-cml grammar G, if $\epsilon \in L(G)$ then $L(G) = L(\overline{G}) \cup \{\epsilon\} = L(\overline{G}')$, where $\overline{G}' = (\{S, C\}, T, \overline{P} \cup \{r_\epsilon\}, S)$.

3.2 Terminal $\{C^m\}$-cml Grammars and Nondeterministic Finite Automata

For any terminal $\{C^m\}$-cml grammar G, we construct a nondeterministic finite automaton M_G such that a derivation step in G corresponds to a transition in M_G.

In the following, let $S \to C^iuC^kSC^lvC^j$ be an ml-production in $P(1) \cup P(2) \cup P(3)$ with $u, v \in T^*$ and $uv \neq \epsilon$. Then, we assume that if $u = \epsilon$ then $k = 0$, and that if $v = \epsilon$ then $l = 0$.

Definition 6: For a terminal $\{C^m\}$-cml grammar $G = (\{S, C\}, T, P, S)$, $M_G = (Q, \Sigma_G, \delta, q_{0,0}, \{q_f\})$ is a *nondeterministic finite automaton derived from G*, where

$$Q = \{q_{i,j} \mid 0 \le i, j < m\} \cup \{q_f\},$$
$$\Sigma_G = \{[u|v] \mid S \to C^iuC^kSC^lv^RC^j \in P(1) \cup P(2) \cup P(3)\} \cup \{[u] \mid S \to C^iuC^j \in P(4)\},$$

$q_{0,0}$ is the start state, and q_f is the final state. The transition mapping δ is defined as follows:

If $S \to C^{i'}uC^kSC^lv^RC^{j'}$ is in $P(1)$, then $\delta(q_{i,j}, [u|v]) \ni q_{k,l}$ with $i = (m - i') \bmod m$ and $j = (m - j') \bmod m$.
If $S \to C^{i'}uC^kSC^{j'}$ is in $P(2)$, then for each j $(0 \le j < m)$ $\delta(q_{i,j}, [u|\epsilon]) \ni q_{k,l}$ with $i = (m - i') \bmod m$ and $l = (j + j') \bmod m$.
If $S \to C^{i'}SC^lv^RC^{j'}$ is in $P(3)$, then for each i $(0 \le i < m)$ $\delta(q_{i,j}, [\epsilon|v]) \ni q_{k,l}$ with $k = (i + i') \bmod m$ and $j = (m - j') \bmod m$.
If $S \to C^{i'}uC^{j'}$ is in $P(4)$, then $\delta(q_{i,j}, [u]) = \{q_f\}$ with $i = (m - i') \bmod m$ and $j = (m - j') \bmod m$.

We extend δ by induction to a function $\delta^* : Q \times \Sigma_G^+ \to \mathcal{P}(Q)$ according to the rules:

$$\delta^*(q, \sigma) = \delta(q, \sigma),$$
$$\delta^*(q, \alpha\sigma) = \cup_{q' \in \delta^*(q,\alpha)} \delta(q', \sigma),$$

where $\sigma \in \Sigma_G$ and $\alpha \in \Sigma_G^+$.

Moreover, if $\alpha = [u_1|v_1^R] \cdots [u_k|v_k^R]$, then we use the notation $\delta^*(q, [u_1 \cdots u_k|(v_1 \cdots v_k)^R])$ to denote $\delta^*(q, \alpha)$ for simplicity. Note that $[u_1 \cdots u_k|(v_1 \cdots v_k)^R]$ cannot be in Σ_G in general.

We note the following points about M_G in Definition 6.

1. Intuitively, the state $q_{i,j}$ $(0 \le i, j < m)$ in M_G corresponds to the set consisting of sentential forms $\tau_1C^iSC^j\tau_2$ in G such that $\tau_1, \tau_2 \in \{C^m\}^*T^*\{C^m\}^*$.
2. An ml-production in $P(1) \cup P(4)$ produces a unique transition, while an ml-production in $P(2) \cup P(3)$ produces m kinds of transitions.

The following lemmas are obvious from Definition 6.

Lemma 3: If M_G has a transition such that either $j \neq l$ and $\delta(q_{i,j}, [u|\epsilon]) \ni q_{k,l}$ or $i \neq k$ and $\delta(q_{i,j}, [\epsilon|v]) \ni q_{k,l}$, then G is not a strict terminal $\{C^m\}$-cml grammar.

Lemma 4: If a string $\alpha \in \Sigma_G^*$ is in $L(M_G)$, then α is one of the forms: $[u]$ and $[u_1|v_1] \cdots [u_n|v_n][u]$ $(n \ge 1)$.

In the following, for simplicity, we assume that if $n = 0$ then $[u_1|v_1] \cdots [u_n|v_n][u] = [u]$.

Theorem 4: For the nondeterministic finite automaton M_G derived from a terminal $\{C^m\}$-cml grammar G, if a string $[u_1|v_1]\cdots[u_n|v_n][u]$ is in $L(M_G)$, then $u_1\cdots u_n u v_n^R\cdots v_1^R$ is in $L(G)$.

Proof : Consider a terminal $\{C^m\}$-cml grammar $G = (\{S,C\},T,P,S)$ and the nondeterministic finite automaton $M_G = (Q,\Sigma_G,\delta,q_{0,0},\{q_f\})$ derived from G.

We will show that if $\delta(q_{i,j},[u_1|v_1]\cdots[u_n|v_n][u]) \ni q_f$ then there is a derivation $C^iSC^j \Longrightarrow^* u_1\cdots u_n u v_n^R\cdots v_1^R$ by using the induction on n. Note that for the case $i = j = 0$, this implies Theorem 4.

Base step, $n = 0$: Assume that $\delta(q_{i,j},[u]) \ni q_f$. By the construction of δ, there is a production $r : S \to C^{i'}uC^{j'}$ with $i = (m-i') \bmod m$ and $j = (m-j') \bmod m$. Therefore, there is a derivation $C^iSC^j \overset{r}{\Longrightarrow} C^iC^{i'}uC^{j'}C^j \Longrightarrow^* u$.

Induction step: For $n \geq 1$, assume that q_f is an element of $\delta(q_{i,j},[u_1|v_1]\cdots[u_n|v_n][u])$. Then, there is a state $q_{k,l}$ such that $\delta(q_{i,j},[u_1|v_1]) \ni q_{k,l}$ and $\delta(q_{k,l},[u_2|v_2]\cdots[u_n|v_n][u]) \ni q_f$. From the induction hypothesis, there is a derivation $C^kSC^l \Longrightarrow^* u_2\cdots u_n u v_n^R\cdots v_2^R$.

There are three cases for u_1,v_1: (1) $u_1,v_1 \neq \epsilon$; (2) $u_1 = \epsilon$, $v_1 \neq \epsilon$; (3) $u_1 \neq \epsilon$, $v_1 = \epsilon$. We prove only the first case, since the proof of the other cases is quite similar to the proof of the first case.

Assume that $u_1,v_1 \neq \epsilon$. By the construction of δ, there is a production $r : S \to C^{i'}u_1C^kSC^lv_1^RC^{j'}$ in P with $i = (m-i') \bmod m$ and $j = (m-j') \bmod m$. Therefore, there is a derivation

$$\begin{aligned} C^iSC^j \overset{r}{\Longrightarrow} &\quad C^iC^{i'}u_1C^kSC^lv_1^RC^{j'}C^j \Longrightarrow^* u_1C^kSC^lv_1^R \\ \Longrightarrow^* &\quad u_1u_2\cdots u_n u v_n^R\cdots v_2^Rv_1^R. \end{aligned}$$

□

Theorem 5: For a terminal $\{C^m\}$-cml grammar $G = (\{S,C\},T,P,S)$, if a string $w \in T^+$ is in $L(G)$, then there exists a string $[u_1|v_1]\cdots[u_n|v_n][u] \in \Sigma_G^+$ with $n \geq 0$ such that $w = u_1\cdots u_n u v_n^R\cdots v_1^R$ and $[u_1|v_1]\cdots[u_n|v_n][u] \in L(M_G)$.

Proof : We will show that for $0 \leq i,j < m$ and $w \in T^+$, if there is a derivation $C^iSC^j \overset{\gamma}{\Longrightarrow} w$ such that terminal ml-productions occur $n+1$ ($n \geq 0$) times in γ, then there exists a string $[u_1|v_1]\cdots[u_n|v_n][u]$ such that $\delta^*(q_{i,j},[u_1|v_1]\cdots[u_n|v_n][u]) \ni q_f$ and $w = u_1\cdots u_n u v_n^R\cdots v_1^R$. We will prove this by induction on n. We note that for the case $i = j = 0$, this implies Theorem 5.

Base step, $n = 0$: Assume that there is a derivation $C^iSC^j \overset{\gamma}{\Longrightarrow} w$, where $0 \leq i,j < m$, $w \in T^+$, and only one terminal ml-production occurs in γ. Then, the terminal ml-production is $S \to C^{i'}wC^{j'}$ with $i = (m-i') \bmod m$ and $j = (m-j') \bmod m$. By the construction of δ, there is a transition $\delta(q_{i,j},[w]) \ni q_f$.

Induction step: Assume that there is a derivation $C^iSC^j \overset{\gamma}{\Longrightarrow} w$ such that terminal ml-productions occur $n+2$ times in γ. Let r be the first used terminal ml-production in γ. There are three cases: $r \in P(1)$; $r \in P(2)$; $r \in P(3)$. We prove only the case $r \in P(1)$, since the proof of other cases is similar to the proof of the first case.

Suppose that r is $S \to C^{i'}uC^kSC^lv^RC^{j'}$ in $P(1)$. Then, there exists a derivation

$$C^iSC^j \overset{r}{\Longrightarrow} C^{i+i'}uC^kSC^lv^RC^{j'+j} \overset{\gamma_1}{\Longrightarrow} uC^kSC^lv^R \overset{\gamma_2}{\Longrightarrow} uw'v^R,$$

such that $uw'v^R = w$, only the c-production is applied in γ_1, and ml-productions occur $n+1$ times in γ_2.

Since only the c-production is applied in γ_1, it follows from the definition of δ that $\delta(q_{i,j},[u|v]) \ni q_{k,l}$. By the induction hypothesis and $C^kSC^l \overset{\gamma_2}{\Longrightarrow} w'$, there exists a string $\alpha \in \Sigma_G^+$ such that $\alpha = [u_1|v_1]\cdots[u_n|v_n][u']$, $\delta^*(q_{k,l},\alpha) \ni q_f$, and $w' = u_1\cdots u_n u' v_n^R\cdots v_1^R$. Hence, $\delta^*(q_{i,j},[u|v]\alpha) \ni q_f$ and $w = uu_1\cdots u_n u' v_n^R\cdots v_1^Rv^R$ hold. □

3.3 Inclusion Relation of Terminal $\{C^m\}$-cml Language Classes

For two distinct positive integers m and n, we examine inclusion relations between t-CML$_{\{C^m\}}$ and t-CML$_{\{C^n\}}$. First, we show that if n is a multiple of m then t-CML$_{\{C^n\}}$ includes t-CML$_{\{C^m\}}$.

Theorem 6: For given integers $m,h \geq 2$, t-CML$_{\{C^m\}} \subseteq$ t-CML$_{\{C^{hm}\}}$.

Proof : For a terminal $\{C^m\}$-cml language $L(G)$ with $G = (\{S,C\},T,P,S)$, we construct a terminal $\{C^{hm}\}$-cml grammar $G' = (\{S,C\},T,P',S)$, where ml-productions in P' are defined as follows: For $u,v \in T^*$, $0 \leq i,j,k,l < m$,
if $S \to C^iuC^kSC^lvC^j$ is in $P(1) \cup P(2) \cup P(3)$
then $S \to C^{hi}uC^{hk}SC^{hl}vC^{hj}$ is in P',
if $S \to C^iuC^j$ is in $P(4)$ then $S \to C^{hi}uC^{hj}$ is in P'.

To prove the theorem, it is enough to show that for $0 \leq x,y < m$ and $w \in T^+$, there is a derivation $C^xSC^y \overset{\gamma}{\Longrightarrow}_G w$ if and only if $C^{hx}SC^{hy} \overset{\gamma'}{\Longrightarrow}_{G'} w$. Since the only if part is obvious from the construction of G', we will prove the if part by induction on the number n of ml-productions occur in the derivation γ'.

Base step, $n = 1$: Consider a derivation $C^{hx}SC^{hy} \overset{\gamma'}{\Longrightarrow}_{G'} w$, where only one ml-production occurs in γ'. Then, the ml-production is $S \to C^{hi}wC^{hj}$ in P' such that

$$(hx+hi) \bmod hm = (hy+hj) \bmod hm = 0.$$

This implies $(x+i) \bmod m = (y+j) \bmod m = 0$. Therefore, there exist a production $S \to C^iwC^j$ in P and a derivation $C^xSC^y \Longrightarrow_G C^xC^iwC^jC^y \Longrightarrow_G^* w$.

Induction step: Assume that $C^{hx}SC^{hy} \overset{\gamma'}{\Longrightarrow}_{G'} w$ and ml-productions of P' are applied $n+1$ times in γ'. Then, there are an ml-production $r : S \to C^{hi}uC^{hk}SC^{hl}vC^{hj}$ and two derivations, γ_1 and γ_2, such that

$$\begin{aligned} C^{hx}SC^{hy} &\overset{r}{\Longrightarrow}_{G'} C^{hx}C^{hi}uC^{hk}SC^{hl}vC^{hj}C^{hy} \\ &\overset{\gamma_1}{\Longrightarrow}_{G'} uC^{hk}SC^{hl}v \overset{\gamma_2}{\Longrightarrow}_{G'} uw'v, \end{aligned}$$

$uw'v = w$, only the c-production is applied in γ_1, and ml-productions occur n times in γ_2. There are three cases for

u, v: (1) $u, v \neq \epsilon$; (2) $u \neq \epsilon$, $v = \epsilon$; (3) $u = \epsilon$, $v \neq \epsilon$. We show only the first case, since the proof of the other cases is similar to the proof of this case.

Assume that $u, v \neq \epsilon$. Then, we have $(hx + hi)$ mod $hm = (hj + hy)$ mod $hm = 0$, which implies $(x + i)$ mod $m = (j + y)$ mod $m = 0$. From the induction hypothesis, there is a derivation $C^k S C^l \Longrightarrow_G^* w'$. Therefore, there is a derivation $C^x S C^y \Longrightarrow_G C^x C^i u C^k S C^l v C^j C^y \Longrightarrow_G^* u C^k S C^l v \Longrightarrow_G^* w$. □

If n is greater than m then we can show that t-CML$_{\{C^n\}}$ is not included in t-CML$_{\{C^m\}}$.

Theorem 7: If $n > m \geq 2$ then there exists a terminal $\{C^n\}$-cml language that is not a terminal $\{C^m\}$-cml language.

Proof : For each k $(0 \leq k < n^2)$, we define $f(k)$ as a pair of integers (i, j) such that $k = i \cdot n + j$ and $0 \leq j < n$. Consider a terminal $\{C^n\}$-cml grammar $G = (\{S, C\}, T, P, S)$ such that

$$T = \{a_0, b_0\} \cup \bigcup_{1 \leq k < n^2} \{a_k, b_k, d_k\}$$

and

$$\begin{aligned} P = \{ & S \rightarrow C^{n-i} a_k C^i S C^j a_k C^{n-j} \mid \\ & \qquad 0 \leq k < n^2 \text{ and } f(k) = (i, j)\} \\ \cup & \{S \rightarrow C^{n-i} b_k C^{n-j} \mid 0 \leq k < n^2 \text{ and } f(k) = (i, j)\} \\ \cup & \{S \rightarrow d_k C^i S C^j d_k \mid 1 \leq k < n^2 \text{ and } f(k) = (i, j)\} \\ \cup & \{C^n \rightarrow \epsilon\}. \end{aligned}$$

Let $L_0 = \{a_0^l b_0 a_0^l \mid l \geq 0\}$ and $L_k = \{d_k a_k^l b_k a_k^l d_k \mid l \geq 0\}$ for k $(1 \leq k < n^2)$. Then, $L(G) = \cup_{0 \leq k < n^2} L_k$.

Assume that there is a terminal $\{C^m\}$-cml grammar G' such that $L(G') = L(G)$. Let $M_{G'}$ be the nondeterministic finite automaton $(Q', \Sigma', \delta', q'_{0,0}, \{q'_f\})$ derived from G', where

$$\begin{aligned} Q' = \; & \{q'_{i,j} \mid 0 \leq i, j < m\} \cup \{q'_f\}, \\ \Sigma' = \; & \{[u|v] \mid S \rightarrow C^i u C^k S C^l v^R C^j \in P'\} \\ & \cup \{[u] \mid S \rightarrow C^i u C^j \in P'\}. \end{aligned}$$

For each k $(0 \leq k < n^2)$, if a_k occurs in $w \in L(G)$, then $w \in L_k$. Hence, by Theorem 5, there is a state $\widehat{q_k} \in Q'$ such that $(\delta')^*(\widehat{q_k}, [a_k^{p_k} | a_k^{p_k}]) \ni \widehat{q_k}$. The set Q' consists of $m^2 + 1$ states, and there is no transition from the final state q'_f. Hence, it follows from $m < n$ that there are two distinct integers s and t such that $0 \leq s, t < n^2$ and $\widehat{q_s} = \widehat{q_t}$. Therefore, $(\delta')^*(\widehat{q_s}, [a_s^{p_s} | a_s^{p_s}][a_t^{p_t} | a_t^{p_t}]) \ni \widehat{q_s}$ holds. By Theorem 4, this implies that there is a string w in $L(G)$ such that both a_s and a_t occur in w. This contradicts $L(G) = \cup_{0 \leq k < n^2} L_k$.

Corollary 2: For given integers $m, h \geq 2$, CML$_{\{C^m\}} \subset$ CML$_{\{C^{hm}\}}$.

3.4 Terminal $\{C^m\}$-cml Languages and Strict Terminal $\{C^m\}$-cml Languages

We show that the class of terminal $\{C^m\}$-cml languages properly includes the class of s-terminal $\{C^m\}$-cml languages.

Theorem 8: For a given integer $m \geq 2$, st-CML$_{\{C^m\}} \subset$t-CML$_{\{C^m\}}$.

Proof : Since st-CML$_{\{C^m\}} \subseteq$t-CML$_{\{C^m\}}$ immediately follows from the definitions of the language classes, we show the proper inclusion.

We prove only the case $m = 2$. We show the outline of the proof of the case $m > 2$ in Appendix.

Let $m = 2$. Consider a terminal $\{C^2\}$-cml grammar $G = (\{S, C\}, T, P, S)$ such that $T = \{a_0, a_1, a_2, a_3, d_0, d_1, d_2, d_3, b_0, b_1, b_2, b_3, e_0, e_1, g, h\}$, and

$$\begin{aligned} P = \{ \; & S \rightarrow a_0 S d_0, & & S \rightarrow a_1 S C d_1 C, & & S \rightarrow C a_2 C S d_2, \\ & S \rightarrow C a_3 C S C d_3 C, & & S \rightarrow e_0 S, & & S \rightarrow C e_1 C S, \\ & S \rightarrow b_0, & & S \rightarrow b_1 C, & & S \rightarrow C b_2, \\ & S \rightarrow C b_3 C, & & S \rightarrow h S C, & & S \rightarrow g C S, \\ & C^2 \rightarrow \epsilon \; \}. \end{aligned}$$

We will show that $L(G)$ is not in st-CML$_{\{C^2\}}$.

Let $M_G = (Q, \Sigma_G, \delta, q_{0,0}, \{q_f\})$ be the nondeterministic finite automaton derived from G. Figure 1 shows the transition diagram of M_G. By using the transition diagram and Theorem 4, we can easily show that the following eight sets are subsets of $L(G)$.

$$\begin{aligned} L_0 &= \{a_0^n b_0 d_0^n \mid n \geq 0\}, & L_1 &= \{h a_1^n b_1 d_1^n \mid n \geq 0\}, \\ L_2 &= \{g a_2^n b_2 d_2^n \mid n \geq 0\}, & L_3 &= \{h g a_3^n b_3 d_3^n \mid n \geq 0\}, \\ L_4 &= \{e_0^n b_0 \mid n \geq 0\}, & L_5 &= \{h e_0^n b_1 \mid n \geq 0\}, \\ L_6 &= \{g e_1^n b_2 \mid n \geq 0\}, & L_7 &= \{h g e_1^n b_3 \mid n \geq 0\}. \end{aligned}$$

It is also easy to show that $L(G)$ has the nine properties:

- **(P1)** if $w \in L(G)$, then one and only one of b_0, b_1, b_2 and b_3 occurs in w only one time.
- **(P2)** $a_0^n b_0 d_0^k \in L(G)$ if and only if $n = k \geq 0$.
- **(P3)** $h a_1^n b_1 d_1^k \in L(G)$ if and only if $n = k \geq 0$.
- **(P4)** $g a_2^n b_2 d_2^k \in L(G)$ if and only if $n = k \geq 0$.
- **(P5)** $h g a_3^n b_3 d_3^k \in L(G)$ if and only if $n = k \geq 0$.
- **(P6)** if b_0 occurs in $w \in L(G)$ then w has an even number (including zero) of h occurrences, and g does not occur in w.
- **(P7)** if b_1 occurs in $w \in L(G)$ then w has an odd number of h occurrences, and g does not occur in w.

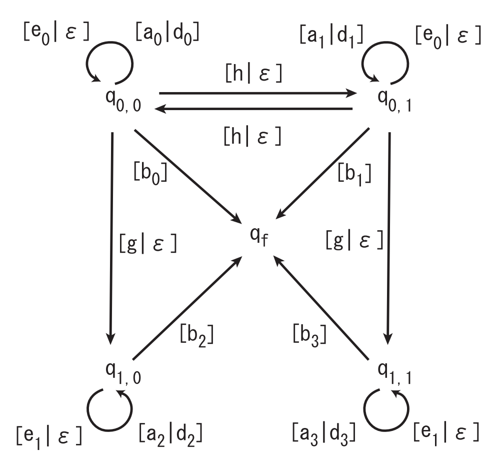

Fig. 1 The transition diagram of M_G.

(P8) if b_2 occurs in $w \in L(G)$ then w has only one occurrence of g and an even number (including zero) of h occurrences.

(P9) if b_3 occurs in $w \in L(G)$ then w has only one occurrence of g and an odd number of h occurrences.

We prove by contradiction that $L(G)$ is not in st-CML$_{\{C^2\}}$. Suppose that there is an s-terminal $\{C^2\}$-cml grammar $G' = (\{S, C\}, T, P', S)$ such that $L(G) = L(G')$. Let $M_{G'} = (Q', \Sigma', \delta', q'_{0,0}, \{q'_f\})$ be the nondeterministic finite automaton derived from G', where

$$\begin{aligned} Q' = &\ \{q'_{i,j} \mid 0 \le i, j < 2\} \cup \{q'_f\}, \\ \Sigma' = &\ \{[u|v] \mid S \to C^i u C^k S C^l v^R C^j \in P'\} \\ &\cup \{[u] \mid S \to C^i u C^j \in P'\}. \end{aligned}$$

For any $n \ge 0$, each L_i $(0 \le i \le 7)$ contains a string w with $|w| > n$. Furthermore, Theorems 4 and 5 show that there is a correspondence between strings in $L(G')$ and strings in $L(M_{G'})$. Hence, by a similar argument used in the proof of the pumping lemma for regular languages ([7]), we can prove that, for each i $(0 \le i \le 7)$, there are states $\widehat{q_i}, q_i^* \in Q'$, and strings $x_i, y_i, z_i, \alpha_i, \beta_i, t_i, u_i \in T^*$ such that the following seven conditions hold:

(1) if $x_i\alpha_i \ne \epsilon$ then $(\delta')^*(q'_{0,0}, [x_i|\alpha_i^R]) \ni \widehat{q_i}$ else $\widehat{q_i} = q'_{0,0}$,
(2) $(\delta')^*(\widehat{q_i}, [y_i|\beta_i^R]) \ni \widehat{q_i}$,
(3) if $z_it_i \ne \epsilon$ then $(\delta')^*(\widehat{q_i}, [z_i|t_i^R]) \ni q_i^*$ else $\widehat{q_i} = q_i^*$,
(4) $\delta'(q_i^*, [u_i]) = \{q'_f\}$,
(5) the string $x_iy_iz_iu_it_i\beta_i\alpha_i$ is in L_i,
(6) for each $k \ge 0$, the string $x_iy_i^kz_iu_it_i\beta_i^k\alpha_i$ is in $L(G)$,
(7) $y_i\beta_i \ne \epsilon$ and $u_i \ne \epsilon$.

We show the following seven claims.

Claim 1: For each i $(0 \le i \le 3)$, the following hold.

(a) For each $k \ge 0$, the string $x_iy_i^kz_iu_it_i\beta_i^k\alpha_i$ is in L_i.
(b) There exists a positive integer p_i such that $y_i = a_i^{p_i}$ and $\beta_i = d_i^{p_i}$.
(c) The string $z_iu_it_i$ has only one occurrence of b_i,
(d) x_1 has only one occurrence of h, x_2 has only one occurrence of g, and x_3 has only one occurrence of both h and g.

Proof : We prove only the case $i = 1$, since the proof of other cases is similar to the proof of this case.

By Condition (5), there exists $p \ge 0$ such that $x_1y_1z_1u_1t_1\beta_1\alpha_1 = ha_1^pb_1d_1^p$. Hence, it follows from Conditions (6), (7) and Property (P1) that b_1 occurs once in $z_1u_1t_1$. Similarly, from Conditions (6), (7) and Property (P7), it holds that h occurs once in x_1. Therefore, it follows from Condition (6) and Property (P3) that, for each $k \ge 0$, the string $x_1y_1^kz_1u_1t_1\beta_1^k\alpha_1$ is in L_1, and that there exists a positive integer p_1 such that $y_1 = a_1^{p_1}$ and $\beta_1 = d_1^{p_1}$.

Claim 2: For each i $(1 \le i \le 3)$, $\widehat{q_i}$ is different from $q'_{0,0}$.

Proof : First, we note that, since the string b_0 is in $L(G) = L(G')$, P' must include the production $S \to b_0$, which implies $\delta'(q'_{0,0}, [b_0]) = \{q'_f\}$.

Next, assume that $\widehat{q_i} = q'_{0,0}$. Then, $(\delta')^*(q'_{0,0}, [y_i|\beta_i^R]) \ni q'_{0,0}$ holds. Hence, it follows from $\delta'(q'_{0,0}, [b_0]) = \{q'_f\}$ that for each $k \ge 0$, $a_i^{kp_i}b_0d_i^{kp_i}$ is in $L(G') = L(G)$. However, $a_i^{kp_i}b_0d_i^{kp_i}$ is not in $L(G)$, which is a contradiction. Therefore, Claim 2 holds.

Claim 3: The states $\widehat{q_0}$, $\widehat{q_1}$, $\widehat{q_2}$ and $\widehat{q_3}$ are all distinct.

Proof : We show that $\widehat{q_0}$ and $\widehat{q_1}$ are distinct. If the two states are the same, then it follows from $(\delta')^*(q'_{0,0}, [x_1|\alpha_1^R]) \ni \widehat{q_1}$, $(\delta')^*(\widehat{q_1}, [y_1|\beta_1^R]) \ni \widehat{q_1}$, $(\delta')^*(\widehat{q_1}, [z_0|t_0^R]) \ni q_0^*$ and $\delta'(q_0^*, [u_0]) = \{q'_f\}$ that for each $k \ge 0$, $x_1y_1^kz_0u_0t_0\beta_1^k\alpha_1$ is in $L(G)$. On the other hand, it follows from Claim 1 that $x_1y_1^kz_0u_0t_0\beta_1^k\alpha_1$ has only one occurrence of both h and b_0. This contradicts Property (P6).

We can prove the other cases by using Properties (P6)–(P9) and Claim 1 in similar ways.

Claim 4: The state $\widehat{q_0}$ is $q'_{0,0}$.

Proof : $M_{G'}$ has four states except for the final state q'_f. Therefore, Claim 4 follows from Claims 2 and 3.

Claim 5: For each i $(4 \le i \le 7)$, let $j = i - 4$. Then, b_j is a suffix of u_i, and $t_i\beta_i\alpha_i = \epsilon$.

Proof : We prove only the cases $i = 4$ and 5, because we can similarly prove the cases $i = 6$ and 7. By a similar argument used in the proof of Claim 1, we can show that there exists a positive integer p_i such that $y_i\beta_i = e_0^{p_i}$. Hence, it follows from Condition (7) that b_j is a suffix of α_i, t_i or u_i.

Assume that b_j is a suffix of α_i or t_i. Then, Conditions (5), (6) and (7) imply that $u_i = e_0^{k_i}$ for some $k_i \ge 1$. Since it follows from Claim 3 that q_i^* is one of $\widehat{q_0}$, $\widehat{q_1}$, $\widehat{q_2}$ and $\widehat{q_3}$, let $q_i^* = \widehat{q_k}$ $(0 \le k \le 3)$. Then, the string $x_iy_iz_iz_ku_kt_kt_i\beta_i\alpha_i$ is in $L(G')$. On the other hand, from the assumption that b_j is a suffix of α_i or t_i, $t_i\beta_i\alpha_i$ has an occurrence of b_j. Furthermore, it follows from Claim 1 that $z_ku_kt_k$ has an occurrence of b_k, which contradicts Property (P1). Therefore, b_j is a suffix of u_i.

The equation $t_i\beta_i\alpha_i = \epsilon$ follows from Condition (5) and the fact that b_j is a suffix of u_i.

Claim 6: For each i $(5 \le i \le 7)$, $x_i \ne \epsilon$. In particular, x_5 has only one occurrence of h, x_6 has only one occurrence of g, and x_7 has only one occurrence of both h and g.

Proof : As shown in the proof of Claim 5, if $i = 5$ then $y_i\beta_i = e_0^{p_i}$ for some $p_i \ge 1$ holds, and if $i = 6$ or 7 then $y_i\beta_i = e_1^{p_i}$ for some $p_i \ge 1$ holds. On the other hand, if $w \in L_i$ $(5 \le i \le 7)$ then neither e_0 nor e_1 is a prefix of w. Therefore, $x_i \ne \epsilon$, and h (resp. g, hg) is a prefix of x_5 (resp. x_6, x_7). Then, Claim 6 holds.

Claim 7: The states $\widehat{q_5}$, $\widehat{q_6}$ and $\widehat{q_7}$ are all distinct, and none of them is $q'_{0,0}$.

Proof : The proof of the fact that $\widehat{q_5}$, $\widehat{q_6}$ and $\widehat{q_7}$ are all distinct is similar to the proof of Claim 3.

We will prove that $\widehat{q_5}$ is not equal to $q'_{0,0}$. Suppose that the two states are the same. Then, x_5b_0 is in $L(G') = L(G)$. On the other hand, it follows from Claim 6 that x_5b_0 has only one occurrence of h. This contradicts Property (P6). Similarly, we can prove that neither $\widehat{q_6}$ nor $\widehat{q_7}$ is equal to $q'_{0,0}$.

We will conclude the proof of Theorem 8. It follows from Claim 7 that one of $\widehat{q_5}$, $\widehat{q_6}$ and $\widehat{q_7}$ is equal to $q'_{0,1}$. Suppose that $\widehat{q_p}$ $(5 \le p \le 7)$ is equal to $q'_{0,1}$. Then, since $\alpha_p = \epsilon$ follows from Claim 5, $(\delta')^*(q'_{0,0}, [x_p|\epsilon]) \ni q'_{0,1}$ holds. This contradicts Lemma 3 and the assumption that G' is an s-terminal $\{C^2\}$-cml grammar. □

3.5 Linear Languages and Regular Languages

We show that the class of ϵ-free linear languages properly includes the class of terminal $\{C^m\}$-cml languages.

Theorem 9: For a given integer $m \ge 2$, every terminal $\{C^m\}$-cml language is linear.

Proof : For a terminal $\{C^m\}$-cml grammar G, consider a nondeterministic finite automaton $M_G = (Q, \Sigma, \delta, q_{0,0}, \{q_f\})$ derived from G. Based on M_G, construct a linear grammar $G_l = (N, T, P_l, N_{0,0})$, where

$$\begin{aligned} N &= \{N_{i,j} \mid q_{i,j} \in Q\}, \\ P_l &= \{N_{i,j} \to uN_{k,l}v^R \mid \delta(q_{i,j}, [u|v]) \ni q_{k,l}\} \ \cup \\ &\qquad \{N_{i,j} \to u \mid \delta(q_{i,j}, [u]) \ni q_f\}. \end{aligned}$$

From Theorems 4 and 5, it is obvious that $L(G) = L(G_l)$. □

We will show that the class of languages generated by terminal $\{C^m\}$-cml (resp. s-terminal $\{C^m\}$-cml) grammars and the class of ϵ-free regular languages are incomparable.

Theorem 10: For a given integer $m \ge 2$, t-CML$_{\{C^m\}}$ (resp. st-CML$_{\{C^m\}}$) and REG are incomparable.

Proof : Since ML and REG are incomparable ([3]) and ML is included in st-CML$_{\{C^m\}}$, it suffices to show that there exists a regular language that is not a terminal $\{C^m\}$-cml language.

Consider a regular language

$$L_r = \{(a_0)^{k_0}(a_1)^{k_1}\cdots(a_{2m^2})^{k_{2m^2}} \mid k_0, k_1, \cdots, k_{2m^2} \ge 0\}.$$

Assume that there is a terminal $\{C^m\}$-cml grammar $G = (\{S, C\}, T, P, S)$ such that $T = \{a_0, a_1, \ldots, a_{2m^2}\}$ and $L_r = L(G)$. Let $M_G = (Q, \Sigma_G, \delta, q_{0,0}, \{q_f\})$ be the nondeterministic finite automaton derived from G.

For each l $(0 \le l \le 2m^2)$, since $\{(a_l)^k \mid k \ge 0\}$ is a subset of L_r, it follows from Theorem 5 and $L_r = L(G)$ that there exist a state $\widehat{q_l} \in Q$, and integers $i_l, j_l \ge 0$ such that $\delta^*(\widehat{q_l}, [a_l^{i_l}|a_l^{j_l}]) \ni \widehat{q_l}$, and at least one of i_l and j_l is greater than 0. Similarly, if there exist strings $u, v \in T^*$ such that $\delta^*(\widehat{q_l}, [u|v^R]) \ni \widehat{q_l}$, then $a_l^{i_l}ua_l^{i_l}$ and $a_l^{j_l}va_l^{j_l}$ are substrings of some $w \in L_r$. Hence, if $i_l > 0$ (resp. $j_l > 0$) then u (resp. v) is a sequence of a_l. Therefore, if $\widehat{q_{l_1}} = \widehat{q_{l_2}}$ and $l_1 < l_2$, then both $j_{l_1} = 0$ and $i_{l_2} = 0$ hold. This implies that there exist no three mutually distinct integers l_1, l_2, l_3 such that $0 \le l_1, l_2, l_3 \le 2m^2$ and $\widehat{q_{l_1}} = \widehat{q_{l_2}} = \widehat{q_{l_3}}$. That is, M_G must have at least $\lceil (2m^2+1)/2 \rceil = m^2 + 1$ states except for the final state, whereas Q consists of m^2 states except for the final state. This is a contradiction. Therefore, L_r is not a terminal $\{C^m\}$-cml language. □

Since REG is included in LIN, the following proper inclusion follows from Theorems 9 and 10.

Theorem 11: For a given integer $m \ge 2$, CML$_{\{C^m\}} \subset$ LIN.

Note that Theorem 11 can be derived also from Theorems 7 and 9.

4. $\{C^*\}$-cml Languages

We consider the union of CML$_{\{C^m\}}$ over all $m \ge 1$ in this section.

Definition 7: A language L is a $\{C^*\}$-*cml language* (resp. *terminal* $\{C^*\}$-*cml language*) if there is some integer $m \ge 1$ such that L is a $\{C^m\}$-cml language (resp. terminal $\{C^m\}$-cml language). Let CML$_{\{C^*\}}$ (resp. t-CML$_{\{C^*\}}$) be the class of $\{C^*\}$-cml languages (resp. terminal $\{C^*\}$-cml languages).

From Definition 7 and Theorems 3 and 9, the following are obvious.

$$\cup_{m\ge1}\text{t-CML}_{\{C^m\}} = \text{t-CML}_{\{C^*\}} = \text{CML}_{\{C^*\}} \subseteq \text{LIN}.$$

Lemma 5: An ϵ-free linear language is a terminal $\{C^*\}$-cml language.

Proof : Consider an ϵ-free linear language $L = L(G)$, where $G = (N, T, P, N_0)$ and $N = \{N_0, \cdots, N_{n-1}\}$. Without loss of generality, we may assume that any production in P is of one of the forms $N_p \to \tau N_q$, $N_p \to N_q\tau$, $N_p \to \tau$, where $\tau \in T^+$ and $N_p, N_q \in N$.

We construct a terminal $\{C^n\}$-cml grammar $G' = (\{S, C\}, T, P', S)$ as follows: $P' = P'_l \cup P'_r \cup P'_f \cup P_C$, where

$$\begin{aligned} P'_l &= \{S \to C^{n-p}\tau C^q S C^y \mid \\ &\qquad N_p \to \tau N_q \in P, \quad y = (n+q-p) \bmod n\} \\ P'_r &= \{S \to C^x S C^q \tau C^{n-p} \mid \\ &\qquad N_p \to N_q\tau \in P, \quad x = (n+q-p) \bmod n\} \\ P'_f &= \{S \to C^{n-p}\tau C^{n-p} \mid N_p \to \tau \in P\} \\ P_C &= \{C^n \to \epsilon\}. \end{aligned}$$

We will show that for any $z \in T^+$ and any $N_p \in N$, there is a derivation $\phi : N_p \overset{\phi}{\Longrightarrow}_G z$ if and only if there is a derivation $\gamma : C^pSC^p \overset{\gamma}{\Longrightarrow}_{G'} z$. Note that for the case $p = 0$, this implies that a string z is in $L(G)$ if and only if z is in $L(G')$.

[Only-if part]: We use induction on the length k of ϕ.

Base step, $k = 1$: Assume that there is a derivation $\delta : N_p \Longrightarrow_G z$, where $N_p \in N$ and $z \in T^+$. For a production $N_p \to z$ in P, from the construction of P'_f, there is a production $r : S \to C^{n-p}zC^{n-p}$ in P'. Therefore, there is a derivation $C^pSC^p \overset{r}{\Longrightarrow}_{G'} C^pC^{n-p}zC^{n-p}C^p \Longrightarrow^*_{G'} z$.

Induction step: Consider a derivation $\phi : N_p \overset{r}{\Longrightarrow}_G \alpha \Longrightarrow^*_G z$, where the length of ϕ is $k+1$, $N_p \in N$, $z \in T^+$, and

$r \in P$. There are two cases for r: (1) r is $N_p \to \tau N_q$, and (2) r is $N_p \to N_q \tau$. We prove only the first case, since the proof of the second case is similar to the proof of the first case.

Then, the derivation ϕ becomes $\phi : N_p \overset{r}{\Longrightarrow}_G \tau N_q \Longrightarrow^*_G \tau z' = z$. For the production r, from the construction of P'_l, a production $r' : S \to C^{n-p}\tau C^q S C^y$ is in P', where $y = (n+q-p) \bmod n$. For a derivation $N_q \Longrightarrow^*_G z'$, from the induction hypothesis, there is a derivation $C^q S C^q \Longrightarrow^*_{G'} z'$. Therefore, there is a derivation $C^p S C^p \overset{r'}{\Longrightarrow}_{G'} C^p C^{n-p}\tau C^q S C^y C^p \overset{\sigma_c}{\Longrightarrow}_{G'} \tau C^q S C^q \Longrightarrow^*_{G'} \tau z'$, where σ_c is a sequence of the c-production.

[If part]: We use induction on the number k of ml-productions that occur in γ.

Base step, $k = 1$: Assume that there is a derivation $\gamma : C^p S C^p \Longrightarrow^*_{G'} z$, where $0 \le p < n$, $z \in T^+$, and only one ml-production occurs in γ. Then, the ml-production is $r : S \to C^{n-p} z C^{n-p}$. Since r is in P'_f, it follows from the construction of P' that $N_p \to z$ is in P. Therefore, there is a derivation $N_p \Longrightarrow_G z$.

Induction step: Consider a derivation $\gamma : C^p S C^p \overset{r}{\Longrightarrow}_{G'} \alpha \overset{\gamma_1}{\Longrightarrow}_{G'} z$, where r is an ml-production, ml-productions occur k times in γ_1, $0 \le p < n$, and $z \in T^+$. There are two cases for r: (1) $r \in P'_l$; (2) $r \in P'_r$. We prove only the first case, since the proof of the second case is similar to the proof of the first case.

Let $r \in P'_l$. Then, it follows from the definition of P'_l that r is $S \to C^{n-p}\tau C^q S C^y$, $y = (n+q-p) \bmod n$, and $N_p \to \tau N_q \in P$. Hence, the derivation γ is $C^p S C^p \overset{r}{\Longrightarrow}_{G'} C^p C^{n-p}\tau C^q S C^y C^p \overset{\gamma_1}{\Longrightarrow}_{G'} \tau z' = z$. Therefore, there is a derivation $\gamma_2 : C^q S C^q \overset{\gamma_2}{\Longrightarrow}_{G'} z'$ such that ml-productions occur k times in γ_2. From the induction hypothesis, there is a derivation $N_q \Longrightarrow^*_G z'$. Therefore, there is a derivation $N_p \Longrightarrow_G \tau N_q \Longrightarrow^*_G \tau z' = z$. □

From Lemma 5, we have the following theorem.

Theorem 12: $\mathrm{CML}_{\{C^*\}}$=t-$\mathrm{CML}_{\{C^*\}}$=LIN.

5. Concluding Remarks

We examined the generative power of $\{C^m\}$-cml grammars. Figure 2 illustrates major results proved in this paper. We also showed the following:

1. if n is a multiple of m and $n > m$ then t-$\mathrm{CML}_{\{C^n\}}$ properly includes t-$\mathrm{CML}_{\{C^m\}}$,
2. if $n > m \ge 2$ then t-$\mathrm{CML}_{\{C^n\}}$ is not included in t-$\mathrm{CML}_{\{C^m\}}$.

The question of whether t-$\mathrm{CML}_{\{C^m\}}$ and t-$\mathrm{CML}_{\{C^n\}}$ are incomparable for $n > m \ge 2$ is open except for the case where n is a multiple of m.

In this paper, we only considered the generative power of cancel minimal linear grammars with a unique nonterminal symbol except S. As noted in Sect. 2, Geffert [1] shows other types of cml grammars, for example,

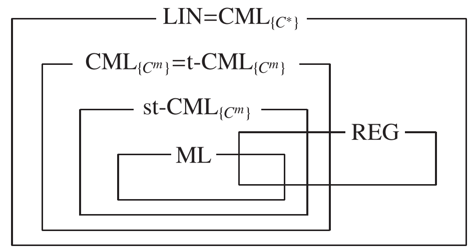

Fig. 2 Language hierarchy.

(1) $P_C = \{AB \to \epsilon, BBB \to \epsilon\}$, $N_C = \{A, B\}$,
(2) $P_C = \{ABBBA \to \epsilon\}$, $N_C = \{A, B\}$.

The question of deciding the generative power of cml grammars with two nonterminal symbols except S is open and of great interest to be studied.

Acknowledgments

The authors would like to express deeply their gratitude to T. Yokomori for useful discussions and valuable suggestions. The authors are also indebted to anonymous reviewers for their valuable comments that improved the contents of this paper. This work was supported in part by Grants-in-Aid for scientific research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (No. 23740081).

References

[1] V. Geffert, "Normal forms for phrase-structure grammars. Theoretical Informatics and Applications," RAIRO, vol.25, no.5, pp.473–496, 1991.
[2] S. Hirose and M. Yoneda, "On the Chomsky and Stanley's homomorphic characterization of context-free languages," Theor. Comput. Sci., vol.36, pp.109–112, 1985.
[3] S. Okawa and S. Hirose, "Homomorphic characterizations of recursively enumerable languages with very small language classes," Theor. Comput. Sci., vol.250, pp.55–69, 2001.
[4] S. Okawa, S. Hirose, and M. Yoneda, "On the impossibility of the homomorphic characterization of context-sensitive languages," Theor. Comput. Sci., vol.44, pp.225–228, 1986.
[5] K. Onodera, "On the generative powers of some extensions of minimal linear grammars," IEICE Trans. Inf. & Syst., vol.E90-D, no.6, pp.895–904, June 2007.
[6] G. Paun, G. Rozenberg, and A. Salomaa, DNA Computing — New Computing Paradigms, Springer, 1998.
[7] G. Rozenberg and A. Salomaa, Eds., Handbook of Formal Languages, Springer, 1997.
[8] Y. Sakakibara and H. Imai, "A DNA-based computational model using a specific type of restriction enzyme," M. Hagiya and A. Obuchi Eds.: DNA8, LNCS 2568, Springer, pp.315–325, 2003.
[9] T. Yokomori, "On purely morphic characterizations of context-free languages," Theor. Comput. Sci., vol.51, pp.301–308, 1987.

Appendix: Proof Outline of Theorem 8: General Case

We show that for $m \ge 2$, there exists a terminal $\{C^m\}$-cml grammar G such that no strict terminal $\{C^m\}$-cml grammar G' generates $L(G)$.

The outline of the proof is similar to the proof of the case $m = 2$. Consider a terminal $\{C^m\}$-cml grammar $G = (\{S, C\}, T, P, S)$ such that

$$\begin{aligned} T = &\{a_{i,j} \mid 0 \le i, j < m\} \cup \{b_{i,j} \mid 0 \le i, j < m\} \\ &\cup \{d_{i,j} \mid 0 \le i, j < m\} \cup \{e_i \mid 0 \le i < m\} \\ &\cup \{g_1, g_2, \ldots, g_{m-1}, h\}, \text{ and} \\ P = &\{S \to C^{m-i} a_{i,k} C^i S C^k d_{i,k} C^{m-k} \mid 0 \le i, k < m\} \\ &\cup \{S \to C^{m-i} b_{i,k} C^{m-k} \mid 0 \le i, k < m\} \\ &\cup \{S \to C^{m-i} e_i C^i S \mid 0 \le i < m\} \\ &\cup \{S \to C^{m-i+1} g_i C^i S \mid 1 \le i < m\} \\ &\cup \{S \to hSC,\ \ C^m \to \epsilon\}. \end{aligned}$$

We can construct from G the nondeterministic finite automaton $M_G = (Q, \Sigma, \delta, q_{0,0}, \{q_f\})$. The transition mapping δ is defined as, for $0 \le i, j < m$ and $1 \le k < m$,

$$\begin{aligned} &\delta(q_{i,j}, [a_{i,j} \mid d^R_{i,j}]) = \{q_{i,j}\}, \quad &&\delta(q_{i,j}, [b_{i,j}]) = \{q_f\}, \\ &\delta(q_{i,j}, [e_i \mid \epsilon]) = \{q_{i,j}\}, &&\delta(q_{k-1,j}, [g_k \mid \epsilon]) = \{q_{k,j}\}, \\ &\delta(q_{0,j}, [h \mid \epsilon]) = \{q_{0,j+1}\}, \end{aligned}$$

where we assume $q_{0,m}$ is equal to $q_{0,0}$.

In the following, for $i = 0$, we assume that $h^0 = g_1 \cdots g_i = \epsilon$. Then, it is easy to show that, for $0 \le i, j < m$, the following sets are subsets of $L(G)$:

$$\begin{aligned} L_{i,j} &= \{h^j g_1 \cdots g_i a^n_{i,j} b_{i,j} d^n_{i,j} \mid n \ge 0\} \\ L_{m+i,m+j} &= \{h^j g_1 \cdots g_i e^n_i b_{i,j} \mid n \ge 0\} \end{aligned}$$

The language $L(G)$ has the properties:

1. if $w \in L(G)$, then w has only one occurrence of $b_{i,j}$ $(0 \le i, j < m)$, and none of them occur in w at the same time.
2. for $0 \le i, j < m$, $h^j g_1 \cdots g_i a^p_{i,j} b_{i,j} d^q_{i,j} \in L(G)$ if and only if $p = q \ge 0$.
3. if $b_{i,j}$ occurs in $w \in L(G)$ then the number of h occurrences in w is congruent to j modulo m, and g_k $(1 \le k \le i)$ occurs in w only once.

Assume that there exists an s-terminal $\{C^m\}$-cml grammar $G' = (\{S, C\}, T, P', S)$ such that $L(G) = L(G')$. Let $M_{G'} = (Q', \Sigma', \delta', q'_{0,0}, \{q'_f\})$ be the nondeterministic finite automaton derived from G'.

We can show several claims similar to the claims showed in the proof of the case $m = 2$. Therefore, we can derive a contradiction. □

Kaoru Fujioka was born in Tokyo, Japan. Her maiden name is Kaoru Onodera. She received B.S., M.S. and Ph.D. degrees from Waseda University, Tokyo, Japan, in 2000, 2002, and 2007, respectively. From 2006 to 2009, she was an instructor in Department of Science and Engineering, Tokyo Denki University. She is currently an Assistant Professor in the Office for Strategic Research Planning, Kyushu University, Fukuoka, Japan. Her current research interests include formal language theory, DNA computing, and computational learning theory.

Hirofumi Katsuno received B.S. and M.S. degrees in mathematics, and Ph.D. degree in mathematical science from the University of Tokyo. From 1976 to 2000, he was with NTT (Nippon Telegraph and Telephone Corporation). He is currently a professor in Department of Science and Engineering, Tokyo Denki University. His current research interests are artificial intelligence, database and computing theory.