数理解析研究所講究録
第 1744 巻 2011 年 144-150

2010 年度冬の LA シンポジウム [28]

Cancel minimal linear grammars with a particular nonterminal symbol

Kaoru Fujioka * Hirofumi Katsuno †

1 Introduction

Among the variety of normal forms for phrase structure grammars ([1], [2], [4]), Geffert normal forms in [1] are unique in that each of them consists of context-free type productions with a fixed number of specific cancellation productions that replace a sequence of nonterminal symbols with the empty string ϵ.

In [3], Geffert normal forms are formalized into a grammar which has minimal linear type productions and a finite set of cancellation productions, called *cancel minimal linear grammar*. Within the framework of cancel minimal linear grammars, one of the Geffert's results ([1]) means that the cancel minimal linear grammar with two cancellation productions $AB \rightarrow \epsilon$ and $CC \rightarrow \epsilon$ generates any recursively enumerable language.

The generative powers of the cancel minimal linear grammars are examined in [3] especially with only one of the two cancellation productions above under the assumption of dealing with only ϵ-free languages. It has been shown that any language generated by the cancel minimal linear grammar with $AB \rightarrow \epsilon$ is context-free, and that any linear language can be generated by the grammar. Furthermore, the class of languages generated by the cancel minimal linear grammar with $CC \rightarrow \epsilon$ is showed to be a proper subset of the class of linear languages.

In this paper, we consider a particular nonterminal symbol C except S and examine the generative powers of a cancel minimal linear grammar with a unique cancellation production $C^m \rightarrow \epsilon$ for any $m \geq 1$. 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, the class of languages generated by those grammars is shown to be equivalent to the class of linear languages.

These results imply a new hierarchy of language classes using cancel minimal linear grammars[3].

2 Preliminaries

We assume the reader to be familiar with the rudiments in formal language theory from [4].

A *phrase structure grammar* (a *grammar* for short) is a quadruple $G = (N, T, P, S)$, where N is a set of *nonterminal symbols*, T is a set of *terminal symbols*, P is a set of *productions*, and S in N is the *initial symbol*. A production in P is of the form $\pi_1 \rightarrow \pi_2$, 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 $r : \pi_1 \rightarrow \pi_2 \in P$, then we 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 confusion, we simply write $\alpha_1 \Longrightarrow \alpha_2$, and we denote the reflexive and transitive closure of $\Longrightarrow$ by $\Longrightarrow^*$.

We define a *language* $L(G)$ generated by a grammar $G = (N, T, P, S)$ as follows: $L(G) = \{z \in T^* \mid S \Longrightarrow^* z\}$.

*Office for Strategic Research Planning, Kyushu University.
†Department of Science and Engineering, Tokyo Denki University.

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 deal with only ϵ-free languages.

A grammar $G = (N, T, P, S)$ is *linear* if each production in P is of the form $N_i \rightarrow \alpha$, 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 $N_1 \rightarrow uN_2$, $N_1 \rightarrow N_2u$, or $N_1 \rightarrow u$, 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 $N_1 \rightarrow uN_2$ or $N_1 \rightarrow u$ (resp. $N_1 \rightarrow N_2u$ or $N_1 \rightarrow u$), 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 the one 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 $S \rightarrow uSv$ or $S \rightarrow w$, where $u, v, w \in T^*$. Any language generated by such a grammar is called *minimal linear*.

Let *RE*, *LIN*, *REG*, and *ML* be the classes of recursively enumerable, linear, regular, and minimal linear languages, respectively.

Geffert [1] shows the following theorem for recursively enumerable languages.

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 $S \rightarrow \alpha_1 S \alpha_2$ or $S \rightarrow \alpha$, where $\alpha_1, \alpha_2, \alpha \in (T \cup N_C)^*$,*
- *$N_C = \{A, B, C\}$ and $P_C = \{AB \rightarrow \epsilon,\ CC \rightarrow \epsilon\}$.*

Motivated by this *Geffert normal form*, a new grammar is introduced as follows [3].

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 initial symbol.
(2) N_C is a finite set of nonterminal symbols except S.
(3) T is a finite set of terminal symbols.
(4) $\Omega = \{\Omega_i \mid 1 \leq i \leq n\}$, where $\Omega_i \in N_C^+$.
(5) $P = P_M \cup P_C$ is a finite set of productions, where
  (a) $P_M \subseteq \{S \rightarrow \alpha_1 S \alpha_2,\ S \rightarrow \alpha \mid \alpha_1, \alpha_2, \alpha \in (T \cup N_C)^*\}$,
  (b) $P_C = \{\Omega_i \rightarrow \epsilon \mid 1 \leq i \leq n\}$.

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.

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.*

An Ω-cml grammar G is called a terminal Ω-cml grammar, if any ml-production in P is one of the forms of a terminal production. A language L is called a terminal Ω-cml language if there is a terminal Ω-cml grammar that generates L.

The classes of terminal Ω-cml languages are denoted by t-CML_Ω.

The generative powers of some classes of terminal $\{AB\}$-cml grammars and terminal $\{C^2\}$-cml grammars are examined in [3] and the following theorem is the result concerning terminal $\{C^2\}$-cml grammars.

Theorem 2

1. $ML \subset t\text{-}CML_{\{C^2\}} \subset LIN$
2. *REG and* $t\text{-}CML_{\{C^2\}}$ *are incomparable.*

In this paper, we focus on a terminal $\{C^m\}$-cml grammar for any positive integer m.

3 Terminal $\{C^m\}$-cml languages

In this section, we consider the generative power of terminal $\{C^m\}$-cml grammars. The case $m = 1$ is simple, because $C \to \epsilon$ means that C can be canceled any time in derivations. Therefore, the following lemma is obvious.

Lemma 1 $t\text{-}CML_{\{C\}} = ML$.

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

3.1 Minimal linear type productions

In the following, for simplicity, if $i = 0$ then we regard C^i and C^{m-i} as ϵ in ml-productions of $\{C^m\}$-cml grammars. For example, the ml-production $S \to C^iuC^{m-i}S$ represents $S \to uS$ for $i = 0$.

In every $\{C^m\}$-cml grammar $G = (\{S, C\}, T, P, S)$, we may assume that any ml-production in P is 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 produce 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^*.

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. Let $P(t)$ be a set of terminal ml-production in P and $P(tf)$ be a set of terminal-free ml-production and the c-production, then

$$
\begin{aligned}
P(t) &= P(1) \cup P(2) \cup P(3) \cup P(4) \\
P(tf) &= P(5) \cup P(6) \cup \{r_C\}.
\end{aligned}
$$

In the following, we call a production in $P(tf)$ as a terminal-free production.

3.2 Terminal $\{C^m\}$-cml grammars and nondeterministic finite automatons

We show that for any terminal $\{C^m\}$-cml grammar G, there exists a nondeterministic finite automaton M_G such that $L(M_G)$ and $L(G)$ are closely related.

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 3 *For a terminal* $\{C^m\}$*-cml grammar* $G = (\{S, C\}, T, P, S)$, $M_G = (Q, \Sigma_G, \delta, q_{0,0}, \{q_0\})$ *is a nondeterministic finite automaton derived from* G, *where*

$$
\begin{aligned}
Q &= \{q_{i,j} \mid 0 \leq i, j < m\} \cup \{q_0\}, \\
\Sigma_G &= \{[u|v] \mid S \to C^iuC^kSC^lv^RC^j \in P(1) \cup P(2) \cup P(3)\} \cup \\
&\quad \{[u] \mid S \to C^iuC^j \in P(4)\}.
\end{aligned}
$$

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 $(0 \leq j < m)$, $\delta(q_{i,j}, [u|\epsilon]) \ni q_{k,l}$

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 $(0 \leq i < m)$, $\delta(q_{i,j}, [\epsilon|v]) \ni q_{k,l}$

with $k = (i + i') \bmod m$ *and* $j = (m - j') \bmod m$.

If $S \to C^{i'}uC^{j'}$ *is in* $P(4)$*, then*

$\delta(q_{i,j}, [u]) = \{q_0\}$

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.*

We note the following points about M_G in Definition 3.

1. Intuitively, a state $q_{i,j}$ ($0 \leq i, j < m$) in M_G corresponds to a derivation $S \Longrightarrow_G^* \tau_1 C^i S C^j \tau_2$ for some $\tau_1, \tau_2 \in (T \cup \{C^m\})^*$.
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 3.

Lemma 2 *If a string* $\alpha \in \Sigma_G^*$ *is in* $L(M_G)$*, then* α *is one of the forms:* $[u]$ *and* $[u_1|v_1] \cdots [u_n|v_n][u]$ $(n \geq 1)$.

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

Theorem 3 *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 $M_G = (Q, \Sigma_G, \delta, q_{0,0}, \{q_0\})$ derived from G.

We will show that if $\delta(q_{i,j}, [u_1|v_1] \cdots [u_n|v_n][u]) \ni q_0$ then there is a derivation $C^i S C^j \Longrightarrow^* u_1 \cdots u_n u v_n^R \cdots v_1^R$ by using the induction on n. Note that for the case $i = j = 0$, this implies Theorem 3.

Base step, $n = 0$: Assume that $\delta(q_{i,j}, [u]) \ni q_0$. 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, $C^i S C^j \overset{r}{\Longrightarrow} C^i C^{i'} u C^{j'} C^j \Longrightarrow^* u$ holds.

Induction step: For $n \geq 1$, assume that q_0 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 $\delta(q_{k,l}, [u_2|v_2] \cdots [u_n|v_n][u]) \ni q_0$. From the induction hypothesis, there is a derivation $C^k S C^l \Longrightarrow^* u_2 \cdots u_n u v_n^R \cdots v_2^R$.

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 $r : S \to C^{i'} u_1 C^k S C^l v_1^R C^{j'}$ in P with $i = (m - i') \bmod m$ and $j = (m - j') \bmod m$. Therefore, there is a derivation

$$\begin{aligned} C^i S C^j \overset{r}{\Longrightarrow} & \; C^i C^{i'} u_1 C^k S C^l v_1^R C^{j'} C^j \Longrightarrow^* u_1 C^k S C^l v_1^R \\ \Longrightarrow^* & \; u_1 u_2 \cdots u_n u v_n^R \cdots v_2^R v_1^R. \end{aligned}$$

□

Theorem 4 *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 $C^i S C^j \overset{\gamma}{\Longrightarrow} w$ 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 $\delta^*(q_{i,j}, [u_1|v_1] \cdots [u_n|v_n][u]) \ni q_0$ 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 4.

Base step, $n = 0$: Assume that there is a derivation $C^i S C^j \overset{\gamma}{\Longrightarrow} w$, 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 $\delta(q_{i,j}, [w]) \ni q_0$.

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

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

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 $\alpha = [u_1|v_1]\cdots[u_n|v_n][u']$, $\delta^*(q_{k,l}, \alpha) \ni q_0$, and $w' = u_1\cdots u_nu'v_n^R\cdots v_1^R$. Hence, $\delta^*(q_{i,j}, [u|v]\alpha) \ni q_0$ and $w = uu_1\cdots u_nu'v_n^R\cdots v_1^Rv^R$ hold. □

3.3 Linear languages and regular languages

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

Theorem 5 *For a given integer $m \geq 2$, every terminal $\{C^m\}$-cml language is linear.*

Proof For a terminal $\{C^m\}$-cml grammar G, consider a nondeterministic finite automaton $M_G = (Q, \Sigma, \delta, q_{0,0}, \{q_0\})$ derived from G. Based on M_G, construct a linear grammar $G_l = (N, T, P_l, N_{0,0})$, where

$$\begin{aligned} N &= \{N_{i,j} \mid q_{i,j} \in Q\}, \\ P_l &= \{N_{i,j} \to uN_{k,l}v^R \mid \delta(q_{i,j}, [u|v]) \ni q_{k,l}\} \cup \\ &\quad\ \{N_{i,j} \to u \mid \delta(q_{i,j}, [u]) \ni q_0\}. \end{aligned}$$

From Theorems 3 and 4, it is obvious that $L(G) = L(G_l)$. □

We will show that the class of terminal $\{C^m\}$-cml languages and the class of regular languages are incomparable.

Theorem 6 *For a given integer $m \geq 2$, t-$CML_{\{C^m\}}$ and REG are incomparable.*

Proof Since ML and REG are incomparable ([2]) and ML is included in t-$\mathrm{CML}_{\{C^m\}}$, it suffices to show that there exists a regular language that is not a terminal $\{C^m\}$-cml language.

Consider a regular language

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

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 $M_G = (Q, \Sigma_G, \delta, q_{0,0}, \{q_0\})$ be the nondeterministic finite automaton derived from G.

For each l $(0 \leq l \leq 2m^2)$, since $\{(a_l)^k \mid k \geq 0\}$ is a subset of L_r, it follows from Theorem 3 and $L_r = L(G)$ that there exist a state $\widehat{q_l} \in Q$ and integers $i_l, j_l \geq 0$ 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 $0 \leq l_1, l_2, l_3 \leq 2m^2$ 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 5 and 6.

Theorem 7 *For a given integer $m \geq 2$, t-$CML_{\{C^m\}} \subset$ LIN.*

4 $\{C^*\}$-cml languages

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

Definition 4 *A language L is a $\{C^*\}$-cml language (resp. terminal $\{C^*\}$-cml language) if there is some integer $m \geq 1$ 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 4 and Theorem 5, the following are obvious.

$$\cup_{m\geq 1}\text{t-CML}_{\{C^m\}} = \text{t-CML}_{\{C^*\}} \subseteq \text{LIN}.$$

Lemma 3 *A linear language is a terminal $\{C^*\}$-cml language.*

Proof Consider a 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 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 $C^qSC^q \Longrightarrow^*_{G'} z'$. Therefore, there is a derivation $C^pSC^p \overset{r'}{\Longrightarrow}_{G'} C^pC^{n-p}\tau C^qSC^yC^p \overset{\sigma_c}{\Longrightarrow}_{G'} \tau C^qSC^q \Longrightarrow^*_{G'} \tau z'$, 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 $\gamma : C^pSC^p \Longrightarrow^*_{G'} z$, where $0 \leq p < n$, $z \in T^+$, and only one ml-production occurs in γ. Then, the ml-production is $r : S \rightarrow C^{n-p}zC^{n-p}$. Since r is in P'_f, it follows from the construction of P' that $N_p \rightarrow z$ is in P. Therefore, there is a derivation $N_p \Longrightarrow_G z$.

Induction step: Consider a derivation γ : $C^pSC^p \overset{r}{\Longrightarrow}_{G'} \alpha \overset{\gamma_1}{\Longrightarrow}_{G'} z$, where r is an ml-production, ml-productions occur k times in γ_1, $0 \leq p < n$, 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 $S \rightarrow C^{n-p}\tau C^qSC^y$, $y = (n + q - p) \bmod n$, and $N_p \rightarrow N_q \in P$. Hence, the derivation γ is $C^pSC^p \overset{r}{\Longrightarrow}_{G'} C^pC^{n-p}\tau C^qSC^yC^p \overset{\gamma_1}{\Longrightarrow}_{G'} \tau z' = z$. Therefore, there is a derivation $\gamma_2 : C^qSC^q \overset{\gamma_2}{\Longrightarrow}_{G'} z'$ 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 3, we have the following theorem.

Theorem 8 *t-$CML_{\{C^\bullet\}}$=LIN.*

5 Concluding Remarks

In this paper, we considered the generative powers of terminal cancel minimal linear grammars with a unique nonterminal symbol except S. Figure 1 shows the results proved in this paper.

Geffert [1] shows other types of cml grammars, for example,

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

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

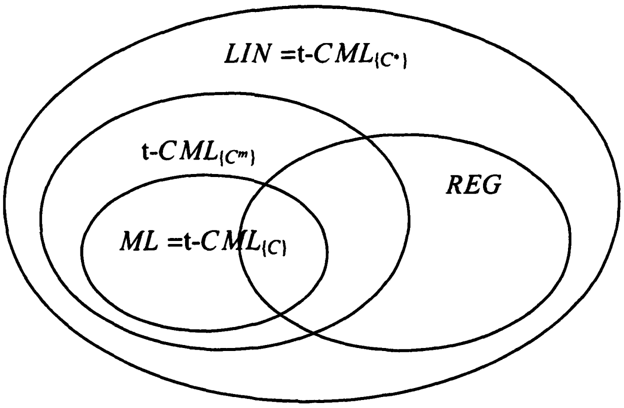

Fig. 1: Language hierarchy

References

[1] V.Geffert. Normal forms for phrase-structure grammars. *Theoretical Informatics and Applications*, RAIRO, **25**, **5**, pp.473–496, 1991.

[2] S.Okawa and S.Hirose. Homomorphic characterizations of recursively enumerable languages with very small language classes. *Theoretical Computer Science*, **250**, pp.55–69, 2001.

[3] K.Onodera. On the generative powers of some extensions of minimal linear grammars. *IEICE Trans. Inf.&SYS.*, **E90-D**, pp.895–904, 2007.

[4] G.Rozenberg and A.Salomaa, Eds. *Handbook of Formal Languages*. Springer, 1997.