数理解析研究所講究録 1375 巻 2004 年 208-213

On the generative power of an extension of minimal linear grammars

早稲田大学・教育学研究科　　小野寺　薫 (Kaoru Onodera)
Department of Mathematics, School of Education,
Waseda University

1 Introduction

Among a variety of normal forms for phrase structure (or type-0) grammars, Geffert normal forms are unique in that each of them consists of minimal linear type productions with a fixed number of specific cancellation productions. More specifically, we are interested in one of the Geffert normal forms in which besides minimal linear type productions, only two cancellation productions $AB \to \epsilon$ and $CC \to \epsilon$ are allowed.

Motivated from these forms, first we formalize Geffert normal forms into grammars with minimal linear type productions and a finite set of cancellation productions which we refer to as *cancel minimal linear grammars*. Then, within cancel minimal linear grammars, we consider the effects of restrictive use of the above two cancellation productions on the generative powers. That is, we examine the generative powers of two types of cancel minimal grammars with either $AA \to \epsilon$ (exclusively) or $AB \to \epsilon$.

We will show that cancel minimal linear grammars with the cancellation production $AA \to \epsilon$, only generate linear languages, while with the cancellation production $AB \to \epsilon$, they only generate context-free languages. Thus, a slight difference of cancellation productions has an effect on the generative powers. Their inclusion relations to the class of regular languages are also established.

2 Preliminaries

Let $G = (N, T, P, S)$ be a *minimal linear grammar*, where $N = \{S\}$ is a set of *nonterminal symbol*, T is a set of *terminal symbols*, S in N is the initial symbol, and P is a finite set of *minimal linear productions* of the forms, $S \to uSv$ or $S \to w$, where $u, v, w \in T^*$. A language L is a *minimal linear language* if there is a minimal linear grammar G such that $L = L(G)$, where $L(G) = \{w \in T^* \mid S \Longrightarrow_G^* w\}$.

We introduce a cancel minimal linear grammar as follows: a *cancel minimal linear grammar (cml grammar)* is a 4-tuple $G = (\{S\} \cup N_C, T, P, S)$, where T

and S are the same as before. Let N_C be a finite set of nonterminal symbols except for S. P is a finite set of productions and consists of *minimal linear type productions (ml-productions)* P_M and *cancellation productions (c-productions)* P_C, where

$$P_M = \{S \to uSv \mid u, v \in (T \cup N_C)^*\} \cup \{S \to w \mid w \in (T \cup N_C)^*\}, \text{ and}$$
$$P_C = \{\alpha \to \epsilon \mid \alpha \in N_C^*\}.$$

A language L is a *cancel minimal linear language (cml language)* if there is a cml grammar G such that $L = L(G)$. In a cml-grammar G, if $P_C = \{\alpha \to \epsilon,\ \beta \to \epsilon, \cdots, \gamma \to \epsilon\}$ holds, then we say that $L(G)$ is an $\{\alpha, \beta, \cdots, \gamma\}$*-cml language.*

For a derivation $S \overset{\sigma_1}{\Longrightarrow} \alpha$, if there exists a derivation σ_2 such that $\alpha \overset{\sigma_2}{\Longrightarrow} w \in T^*$, then α is called a *valid* string. When α is valid, the derivation σ_1 is called a *valid derivation.*

Consider a valid derivation $S \overset{\sigma_1}{\Longrightarrow} \alpha_1$. If there exists no string α_2 such that $\alpha_1 \overset{\sigma_2}{\Longrightarrow} \alpha_2$, where $\sigma_2 \in P_C^+$, then we say that α_1 is *irreducible.*

In what follows, we consider only ϵ-free languages. The classes of recursively enumerable, context-free, linear, minimal linear, $\{\alpha, \beta, \cdots, \gamma\}$-cancel minimal linear, and regular languages are denoted by RE, CF, LIN, ML, $CML_{\{\alpha,\beta,\cdots,\gamma\}}$, and REG respectively.

For the class of recursively enumerable languages, there exists the following theorem.

Theorem 1 (Geffert) [1] *Each recursively enumerable language L can be generated by a cml grammar with a set of cancellation productions P_C which is one of the following five sets:*

$$1: \{AB \to \epsilon,\ CD \to \epsilon\}, \quad 2: \{AB \to \epsilon,\ CC \to \epsilon\},$$
$$3: \{AA \to \epsilon,\ BBB \to \epsilon\}, \quad 4: \{ABBBA \to \epsilon\},$$
$$5: \{ABC \to \epsilon\}.$$

3 Main results

3.1 $\{AA\}$-cml languages

Firstly, we show some results concerning $\{AA\}$-cml grammars. With the c-production $AA \to \epsilon$, cml grammars can only generate linear languages.

To show the relationship with other language classes, we consider a linear language generated by G_1 which indicates the proper inclusion between the classes of linear languages and $\{AA\}$-cml languages:

$G_1 = (\{N_0, N_1, N_2, N_3, N_4\}, \{a, b, c, d, e, f\}, P_1, N_0)$, where

$$P_1 = \{ \ N_0 \to aN_0a,\ N_0 \to aN_1a,\ N_1 \to bN_1b,\ N_1 \to bN_2b,\ N_2 \to cN_2c,$$
$$N_2 \to cN_3c,\ N_3 \to dN_3d, N_3 \to dN_4d,\ N_4 \to eN_4e,\ N_4 \to efe\ \}.$$

Then, $L(G_1) = \{a^{k_1}b^{k_2}c^{k_3}d^{k_4}e^{k_5}fe^{k_5}d^{k_4}c^{k_3}b^{k_2}a^{k_1} \mid k_1, k_2, k_3, k_4, k_5 \geq 1\}$ which is proved to be not an $\{AA\}$-cml language.

On the other hand, the regular language $L_2 = \{a^{k_1}b^{k_2}c^{k_3}d^{k_4}e^{k_5} \mid k_1, k_2, k_3, k_4, k_5 \geq 1\}$ is not an $\{AA\}$-cml language.

The following $\{AA\}$-cml language $L(G_3)$ indicates the proper inclusion between the classes of minimal linear languages and $\{AA\}$-cml languages: $L(G_3) = \{a^n b^m a^n \mid m \geq 1,\ n \geq 0\}$, where $G_3 = (\{S, A\}, \{a, b\}, P_3, S)$, and $P_3 = \{S \to aSa,\ S \to SAb,\ S \to bA,\ S \to SAbA,\ S \to b,\ AA \to \epsilon\}$. It is easy to see that $L(G_3)$ is not minimal linear.

By these languages, we have the following theorem.

Theorem 2 *1. $ML \subset CML_{\{AA\}} \subset LIN$.*

2. REG and $CML_{\{AA\}}$ are incomparable.

3.2 $\{AB\}$-cml languages

We will show that $\{AB\}$-cml grammars can only generate context-free languages.

Let $G = (N, T, P, S)$ be an $\{AB\}$-cml grammar. Without loss of generality, we may assume that any ml-production in P is of the form $S \to B^i u A^j S B^k v A^l$ or $S \to B^i w A^l$, where $u, v \in T^*$, $w \in T^+$, $i, j, k, l \geq 0$.

We set $P = P_{M_1} \cup P_{M_2} \cup P_C$, where

$$P_{M_1} = \left\{ \begin{array}{l} r_{11} : S \to B^{i_{11}} u_{11} A^{j_{11}} S B^{k_{11}} v_{11} A^{l_{11}}, \\ \cdots, \\ r_{1p} : S \to B^{i_{1p}} u_{1p} A^{j_{1p}} S B^{k_{1p}} v_{1p} A^{l_{1p}}, \end{array} \right\}$$

$$P_{M_2} = \left\{ \begin{array}{l} r_{21} : S \to B^{i_{21}} w_{21} A^{l_{21}}, \\ \cdots, \\ r_{2q} : S \to B^{i_{2q}} w_{2q} A^{l_{2q}}, \end{array} \right\}$$

$$P_C = \{r_c : AB \to \epsilon\}.$$

Consider a derivation $S \overset{\gamma}{\Longrightarrow} w$, where γ be a derivation which uses t_{1k} times applications of r_{1k}, for each $1 \leq k \leq p$. At the last step, we use a production r_{2s} in P_{M_2} at most one time, and we also use the c-production some times in γ. We first examine a necessary condition of γ for w to be in $L = L(G)$.

Lemma 1 *On the number of nonterminal symbols A and B, the following equations hold with γ by w_{2s} in a production r_{2s} in P_{M_2}:*

$$\begin{pmatrix} j_{11} - i_{11} & \cdots & j_{1p} - i_{1p} \\ k_{11} - l_{11} & \cdots & k_{1p} - l_{1p} \end{pmatrix} \begin{pmatrix} t_{11} \\ \vdots \\ t_{1p} \end{pmatrix} = \begin{pmatrix} i_{2s} \\ l_{2s} \end{pmatrix} \cdots (1).$$

Now, we set $M = \begin{pmatrix} j_{11} - i_{11} & \cdots & j_{1p} - i_{1p} \\ k_{11} - l_{11} & \cdots & k_{1p} - l_{1p} \end{pmatrix}$ and the rank of M is r_M.

Obviously, $t_{11}, \cdots, t_{1p}$ should be integer solutions of equations (1). To solve these equations, firstly we consider the next equation,

$$M \begin{pmatrix} t_{11} \\ \vdots \\ t_{1p} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \cdots (2).$$

There exist $(p - r_M = \bar{M})$ vectors that are linearly independent, and the linear combination of those vectors is the solutions of (2). A solution vector v of (1) is represented as $\mathrm{v} = b_1\mathrm{v}_1 + \cdots + b_{\bar{M}}\mathrm{v}_{\bar{M}} + \mathrm{v}_t$, where $\mathrm{v}_1, \cdots, \mathrm{v}_{\bar{M}}$ are base vectors that satisfy (2), $b_1, \cdots, b_{\bar{M}}$ are integers, and v_t is a base vector which satisfies (1).

Now, we consider the vector $\mathrm{v}_t = (t_{t1}, \cdots, t_{tp})$. Then, there exist at most $\dfrac{(t_{t1} + \cdots + t_{tp})!}{t_{t1}! \cdots t_{tp}!}$ different irreducible derivations. For each derivation $S \overset{\gamma_e}{\Longrightarrow} x_eSy_e$, where $1 \leq e \leq \dfrac{(t_{t1} + \cdots + t_{tp})!}{t_{t1}! \cdots t_{tp}!}$, we can effectively check whether it is a valid derivation or not. Since it satisfies the equation (1), for a valid derivation, there exists some $r_{2s} \in P_{M_2}$, we eventually have a derivation, $S \overset{\gamma_e r_{2s} \gamma_c^*}{\Longrightarrow} w$, where $w \in T^*$. In this case, we say that the irreducible valid derivation is *compatible with* v_t.

Let R_t be a finite union of the set of all possible irreducible valid derivations compatible with the vector v, for all $\mathrm{v} \in \left\{\mathrm{v}_t + \sum_{i \in I} \mathrm{v}_i \mid I \subseteq \{1, \cdots, \bar{M}\}\right\}$.

Now, we consider a vector $\mathrm{v}_i = (t_{i1}, \cdots, t_{ip})$ which satisfies (2). By the similar way to v_t, we also effectively check whether it is a valid derivation or not, and for a valid derivation, we eventually have its irreducible form, $S \Longrightarrow^* B^iuA^iSB^lvA^l$, where $u, v \in T^*$ and $i, l \geq 0$. In this case, we say that the irreducible valid derivation is *compatible with* v_i.

Let R be a finite union of the set of all possible irreducible valid derivations compatible with the vector v_i, where $1 \leq i \leq \bar{M}$.

[Construction]

Let $G_4 = (\{S, A, B\}, \{a, b, c, d, e\}, P_4, S)$ be an $\{AB\}$-cml grammar, where

$$P_4 = \{ \; r_1 : S \to aASB^5,\ r_2 : S \to BbASB^3A^5,\ r_3 : S \to BcASB^2A, \\ r_4 : S \to BdASB^3A,\ r_5 : S \to BeA^5,\ \gamma_c : AB \to \epsilon \; \}.$$

By using an example $\{AB\}$-cml grammar G_4, we show how to construct a context-free grammar $G' = (V, T, P', N_{00})$ which satisfies $L(G') = L(G_4)$.

We construct nonterminal symbols in V and productions in P' based on R and R_t. From productions in P_4, we construct the following equation,

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 5 & -2 & 1 & 2 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \end{pmatrix} \cdots (1_E).$$

The solution of (1_E) is represented as $\mathrm{v} = b_1\mathrm{v}_1 + b_2\mathrm{v}_2 + \mathrm{v}_t$, where $\mathrm{v}_1 = (0, 1, 0, 1)$, $\mathrm{v}_2 = (0, 1, 2, 0)$ and $\mathrm{v}_t = (1, 0, 0, 0)$.

At first, we consider about R_t.

- For v_t corresponding to the valid derivation $S \overset{r_1}{\Longrightarrow} aASB^5 \overset{r_5}{\Longrightarrow} aABeA^5B^5 \overset{\gamma_c^*}{\Longrightarrow} ae$, we construct nonterminal symbols and productions, $N_{00} \to aN_{15}$, $N_{15} \to e$.

- For $v_t + v_1$ corresponding to a valid derivation $S \xRightarrow{r_1} aASB^5 \xRightarrow{r_2\gamma_c^*} abASB^3 \xRightarrow{r_4\gamma_c^*} abdASB^5 \xRightarrow{r_5\gamma_c^*} abde$, we construct new nonterminal symbols and productions, $N_{15} \to bN_{13}$, $N_{13} \to dN_{15}$.

  For a valid derivation $r_1r_4r_2r_5\gamma_c^*$, construct new nonterminal symbols and productions, $N_{15} \to dN_{17}$, $N_{17} \to bN_{15}$.

- For $v_t + v_2$ and $v_t + v_1 + v_2$, construct new nonterminal symbols and productions, $N_{13} \to cN_{14}$, $N_{14} \to cN_{15}$, $N_{14} \to dN_{16}$, $N_{15} \to cN_{16}$, $N_{16} \to bN_{14}$, and $N_{16} \to cN_{17}$. $N_{16} \to dN_{18}$, $N_{18} \to bN_{16}$, $N_{17} \to dN_{19}$, $N_{19} \to bN_{17}$, $N_{18} \to dN_{19}$, $N_{17} \to cN_{18}$.

Next, we consider about R.

- For v_1 corresponding to the derivations, $S \xRightarrow{r_4} BdASB^3A$, and $S \xRightarrow{r_2} BbASB^3A^5$, we construct derivations $X_{1j} \to dX_{1j+2}b$ for each $3 \le j \le 7$.
  For each $5 \le j \le 9$, $X_{1j} \to bX_{1j-2}d$,

- For v_2 corresponding to the a derivation $r_3r_2r_3\gamma_c^*$, we construct derivations $X_{1j} \to cX_{1j+1}bc$ for each $4 \le j \le 8$.
  For each $4 \le j \le 9$, $X_{1j} \to cbX_{1j-1}c$.

  For derivations $r_3r_3r_2\gamma_c^*$, construct derivations $X_{1j} \to cX_{1j+1}cb$ for each $3 \le j \le 8$.
  For each $3 \le j \le 7$, $X_{1j} \to ccX_{1j+2}b$.

  For derivations $r_2r_3r_3\gamma_c^*$, construct derivations $X_{1j} \to bX_{1j-3}cc$ for each $6 \le j \le 9$.
  For each $5 \le j \le 9$, $X_{1j} \to bcX_{1j-1}c$.

At last, we have a context-free grammar G', such that $L(G_4) = L(G')$, where $G' = (\{N_{00}, N_{13}, \cdots N_{19}, X_{13}, \cdots X_{19}\}, \{a, b, c, d, e\}, P', N_{00})$,

$$
\begin{aligned}
P' = \{\ & N_{00} \to aN_{15},\ N_{13} \to cN_{14} \mid dN_{15},\ N_{14} \to cN_{15} \mid dN_{16}, \\
& N_{15} \to e \mid bN_{13} \mid bdN_{15} \mid dbN_{15} \mid cN_{16} \mid dN_{17}, \\
& N_{16} \to bN_{14} \mid cN_{17} \mid dN_{18},\ N_{17} \to bN_{15} \mid cN_{18} \mid dN_{19}, \\
& N_{18} \to bN_{16} \mid dN_{19},\ N_{19} \to bN_{17}\ \} \\
\cup \{\ & N_{1j} \to X_{1j}N_{1j},\ X_{1j} \to X_{1k}X_{1j} \mid \epsilon, \text{ where } 3 \le j \le 9,\ 3 \le k \le j\} \\
\cup \{\ & X_{1j} \to dX_{1j+2}b, \text{ where } 3 \le j \le 7\ \} \\
\cup \{\ & X_{1j} \to bX_{1j-2}d, \text{ where } 5 \le j \le 9\ \} \\
\cup \{\ & X_{1j} \to cX_{1j+1}bc, \text{ where } 4 \le j \le 8\ \} \\
\cup \{\ & X_{1j} \to cbX_{1j-1}c, \text{ where } 4 \le j \le 9\ \} \\
\cup \{\ & X_{1j} \to cX_{1j+1}cb, \text{ where } 3 \le j \le 8\ \} \\
\cup \{\ & X_{1j} \to ccX_{1j+2}b, \text{ where } 3 \le j \le 7\ \} \\
\cup \{\ & X_{1j} \to bX_{1j-2}cc, \text{ where } 5 \le j \le 9\ \} \\
\cup \{\ & X_{1j} \to bcX_{1j-1}c, \text{ where } 5 \le j \le 9\ \}.
\end{aligned}
$$

From the above argument, demonstrated by an example grammar G_4, we conclude that an $\{AB\}$-cml language L is a context-free language.

In order to show the relationship with other language classes, we know that an $\{AB\}$-cml language $L(G_4)$ is not minimal linear. Further, a context-free language $L_5 = \{a^m b^m c^n d^n \mid m, n \geq 1\}$ indicates the proper inclusion between the classes of context-free languages and $\{AB\}$-cml languages.

As for the relationship with regular languages, it is possible to show that any regular language can be generated by an $\{AB\}$-cml grammar. Then, we have the following theorem.

Theorem 3 $LIN \subset CML_{\{AB\}} \subset CF$.

4 Conclusion

In this paper, we considered the generative powers of $\{AA\}$-cml grammars and $\{AB\}$-cml grammars. There are many possible variations from Geffert normal forms in Theorem 1, which include, for example, $\{AB, A\}$-cml, $\{AB, AA\}$-cml languages for type 2, $\{AAB\}$-cml languages for type 5. The status of all these language families in Chomsky hierarchy remains open, and we are now working on.

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