INFORMATION AND CONTROL 28, 30–34 (1975)

A Universal Context-Free Grammar

TAKUMI KASAI

*Research Institute for Mathematical Science, Kyoto University,
Kitashirakawa, Kyoto, Japan*

In this report we show that, for each alphabet Σ, there exists a context-free grammar G which satisfies the property that for each context-free language $L \subseteq \Sigma^*$ a regular control set C can be found such that $L_C(G) = L$.

The notion of control set was first introduced by Ginsburg and Spanier (1968), and farther investigated by Moriya (1973a, 1973b), Salomaa (1970) and Mayer (1972). The reader is referred to Salomaa (1973) for background material and additional details.

The language generated by a grammar G with control set C, denoted by $L_C(G)$, is the set of those words generated by leftmost derivations in G whose corresponding string of productions is an element of C. A context-free grammar G over an alphabet Σ is said to be *universal* if for every context-free language L over Σ^*, there exists a regular control set C such that $L_C(G) = L$. In this paper we show that there exists a universal context-free grammar G for each alphabet Σ.

First we shall need some definitions.

DEFINITION. Let $G = (N, \Sigma, P, S)$ be a context-free grammar with the set of nonterminal symbols N, the set of terminal symbols Σ, the set of productions P and the initial symbol S. We conventionally denote $N \cup \Sigma$ by V. Let

$$\Pi\colon w_0 \overset{\pi_1}{\Rightarrow} w_1 \overset{\pi_2}{\Rightarrow} \cdots \overset{\pi_k}{\Rightarrow} w_k$$

be a leftmost derivation in G, where in the transition from w_{i-1} to w_i $(1 \leqslant i \leqslant k)$ the production π_i is applied. Then the word $\pi_1\pi_2 \cdots \pi_n$ is called the *associate* of Π. If Π is a derivation of length zero (i.e., $k = 0$), then the associate of Π will be considered to be the empty word ϵ.

For x and y in V^* and α in P^*, the notation $x \overset{\alpha}{\Rightarrow} y$ means that there exists a leftmost derivation

$$x = w_0 \overset{\pi_1}{\Rightarrow} w_1 \overset{\pi_2}{\Rightarrow} \cdots \overset{\pi_k}{\Rightarrow} w_k = y$$

such that $\alpha = \pi_1 \pi_2 \cdots \pi_n$. Thus, $x \overset{\epsilon}{\Rightarrow} x$ for all x in V^*. Let C be a subset of P^*. Let

$$L(G) = \{w \text{ in } \Sigma^* \mid S \overset{\alpha}{\Rightarrow} w,\ \alpha \text{ in } P^*\},$$

$$A(G) = \{\alpha \text{ in } P^* \mid S \overset{\alpha}{\Rightarrow} w,\ w \text{ in } \Sigma^*\}, \text{ and}$$

$$L_C(G) = \{w \text{ in } \Sigma^* \mid S \overset{\alpha}{\Rightarrow} w,\ \alpha \text{ in } C\}.$$

The set $L(G)$ is the context-free language generated by G. The set $A(G)$ will be called the *associate language* of G. $L_C(G)$ is called the *language generated by G with control set C.*

Notation. Let Σ be a given alphabet. In the rest of the paper we shall assume that Σ is fixed. Let c, $\bar{c}$, d, $\bar{d}$ be symbols not in Σ, and let $\Delta = \Sigma \cup \{c, \bar{c}, d, \bar{d}\}$. Let $\sim$ be the binary relation on Δ^* defined by

$$xc\bar{c}y \sim xy, \quad xd\bar{d}y \sim xy, \quad xay \sim xy$$

for all x, y in Δ^* and a in Σ. Let $\overset{*}{\sim}$ be the reflexive transitive closure of $\sim$. Let

$$D = \{w \text{ in } \Delta^* \mid w \overset{*}{\sim} \epsilon\}.$$

Note that if $\Sigma = \phi$, then D is a Dyck language. Let h be the homomorphism defined on Δ^* by $h(c) = h(\bar{c}) = h(d) = h(\bar{d}) = \epsilon$ and $h(a) = a$ for each a in Σ.

We now prove a stronger version of Chomsky and Stanley's theorem which will be used to demonstrate the main result of this paper.

LEMMA. For every context-free language $L \subseteq \Sigma^*$, there exists a regular set $R \subseteq \Delta^*$ such that $L = h(D \cap R)$.

Proof. Suppose that ϵ is not in L. Without loss of generality we may assume that $L = L(G)$ for some grammar $G = (\{X_1, \ldots, X_n\}, \Sigma, P, X_1)$, where each production of P is of the form $X_i \rightarrow a$, a in Σ, or $X_i \rightarrow X_j X_k$ (Chomsky, 1959). Let g be the homomorphism of P^* into Δ^* defined by $g(X_i \rightarrow a) = \bar{c}\bar{d}^i\bar{c}a$ and $g(X_i \rightarrow X_j X_k) = \bar{c}\bar{d}^i\bar{c}cd^kccd^jc$. Let

$$R = cdc\{g(\pi) \mid \pi \text{ in } P\}^*.$$

It can be proved by induction on n that for $\pi_1, \dots, \pi_n$ in P.

$$X_1 \overset{\pi_1\cdots\pi_n}{\Rightarrow} tX_{i_1}X_{i_2}\cdots X_{i_m}, \qquad t \text{ in } \Sigma^*,$$

if and only if

$$cdcg(\pi_1 \cdots \pi_n) \overset{*}{\sim} cd^{i_m}c \cdots cd^{i_2}ccd^{i_1}c$$

and

$$h(g(\pi_1 \cdots \pi_n)) = t.$$

Thus, w is in $L = L(G)$ if and only if $X_1 \overset{\alpha}{\Rightarrow} w$, w in Σ^*, for some α in P^*. This occurs if and only if $cdcg(\alpha)$ is in D and $h(g(\alpha)) = w$. Hence

$$\begin{aligned} L &= h(cdcg(P^*) \cap D) \\ &= h(R \cap D). \end{aligned}$$

Now suppose that ϵ is in L. Let $R' = R \cup \{\epsilon\}$. Then

$$\begin{aligned} h(R' \cap D) &= h(R \cap D) \cup h(\epsilon) \\ &= (L - \{\epsilon\}) \cup \{\epsilon\} = L. \end{aligned}$$

THEOREM *There exists a context-free grammar G with the property that for each context-free language $L \subseteq \Sigma^*$ a regular control set C can be found such that $L_C(G) = L$.*

Proof. Let $G = (\{X, Y\}, \Sigma, P, X)$, where P is defined below. Let

$$\begin{aligned} P_a &= \{Z \to aZ \mid Z \text{ in } \{X, Y\}\}, \text{ for each } a \text{ in } \Sigma, \\ P_c &= \{X \to XX,\ Y \to XY\}, \qquad P_{\bar c} = \{X \to \epsilon\}, \\ P_d &= \{X \to YX,\ Y \to YY\}, \qquad P_{\bar d} = \{Y \to \epsilon\}, \end{aligned}$$

and let $P = \bigcup_{x \text{ in } \Delta} P_x$. Let f be the homomorphism from P^* into Δ^* defined by $f(\pi) = x$ if π is in P_x.

We now show that $f(A(G)) = D\bar{c}$. First we show that:

(a) If $Z \overset{\alpha}{\Rightarrow} wZ$, Z in $\{X, Y\}$, w in Σ^*, α in P^*, then $f(\alpha) \overset{*}{\sim} \epsilon$.

The proof of (a) will be by induction on the length of α. If $|\alpha| = 0$, then $\alpha = \epsilon$, and (a) is trivially satisfied. Now suppose (a) is true for all α with $|\alpha| \leqslant n$ and consider a derivation $Z \overset{\pi}{\Rightarrow} tZ \overset{\alpha}{\Rightarrow} wZ$, t in $N \cup \Sigma$. Such a derivation can be of one of the three following forms.

(a-i) $Z \overset{\pi}{\Rightarrow} XZ \overset{\beta}{\Rightarrow} uXZ \overset{\bar\pi}{\Rightarrow} uZ \overset{\gamma}{\Rightarrow} uvZ$, where u and v are in Σ^*, $\pi = Z \to XZ$ and $\bar\pi = X \to \epsilon$. From the inductive hypothesis,

$f(\beta) \overset{*}{\sim} \epsilon$ and $f(\gamma) \overset{*}{\sim} \epsilon$. Since $f(\pi) = c$ and $f(\bar{\pi}) = \bar{c}$, we have $f(\pi\beta\bar{\pi}\gamma) = cf(\beta)\,\bar{c}f(\gamma) \overset{*}{\sim} c\bar{c} \sim \epsilon$.

(a-ii) $Z \overset{\pi}{\Rightarrow} YZ \overset{\beta}{\Rightarrow} uYZ \overset{\bar{\pi}}{\Rightarrow} uZ \overset{\gamma}{\Rightarrow} uvZ$, where u and v are in Σ^*. An argument analogous to (i) shows that $f(\pi\beta\bar{\pi}\gamma) = df(\beta)\,\bar{d}f(\gamma) \overset{*}{\sim} \bar{d}d \sim \epsilon$.

(a-iii) $Z \overset{\pi}{\Rightarrow} aZ \overset{\alpha}{\Rightarrow} awZ$, where a is in Σ and w is in Σ^*. From the inductive hypothesis $f(\alpha) \overset{*}{\sim} \epsilon$. Since $f(\pi) = a \sim \epsilon$, we have $f(\pi\alpha) \overset{*}{\sim} \epsilon$.

Thus, the statement (a) is valid for all α in P^*. Now let α be in $A(G)$. Then there exists β in P^* such that $\alpha = \beta\pi$ and $X \overset{\beta}{\Rightarrow} wX \overset{\pi}{\Rightarrow} w$, where $\pi = X \rightarrow \epsilon$ and w is in Σ^*. By (a), $f(\beta)$ is in D. Hence $f(\alpha) = f(\beta)\,\bar{c}$ is in $D\bar{c}$. Thus $f(A(G)) \subseteq D\bar{c}$.

Let x be in Δ^*. It can be proved by induction on the length of x that if $x \overset{*}{\sim} \epsilon$, then for each Z in $\{X, Y\}$, there exists α in P^* such that $f(\alpha) = x$, $Z \overset{\alpha}{\Rightarrow} wZ$ for some w in Σ^*. The details are left for the reader.

Thus, for each x in D, there exists α in P^* such that $X \overset{\alpha}{\Rightarrow} wX \overset{\pi}{\Rightarrow} w$, w in Σ^*, $\pi = X \rightarrow \epsilon$, and $f(\alpha) = x$. Thus, $x\bar{c}$ is in $f(A(G))$, from which $D\bar{c} \subseteq f(A(G))$.

By the previous lemma, for each context-free language $L \subseteq \Sigma^*$, there exists a regular set $R \subseteq \Delta^*$ such that $L = h(D \cap R)$. Let $C = f^{-1}(R\bar{c})$. Then

$$\begin{aligned} L &= h(D \cap R) = h(D\bar{c} \cap R\bar{c}) \\ &= h(f(A(G)) \cap R\bar{c}) = hf(A(G) \cap f^{-1}(R\bar{c})) \\ &= hf(A(G) \cap C). \end{aligned}$$

Since $X \overset{\alpha}{\Rightarrow} hf(\alpha)$ for each α in $A(G)$, we have

$$L = hf(A(G) \cap C) = L_C(G).$$

Since $L_C(G)$ is context-free for every regular control set C (Ginsburg and Spanier, 1968), we have the following result.

COROLLARY. *There exists a context free grammar G such that $\{L_C(G) \mid C$ is regular$\}$ is identical to the class of context-free languages over Σ^*.*

RECEIVED: May 31, 1974

REFERENCES

CHOMSKY, N. (1963), Formal properties of grammars, *in* "Handbook of Mathematical Psychology," (D. Luce, R. Bush, and E. Galanter, Eds.), John Wiley & Sons, Inc., New York.

CHOMSKY, N. (1959), On certain formal properties of grammars, *Inform. Contr.* **2**, 137–167.

GINSBURG, S., AND SPANIER, E. H. (1968), Control sets on grammars, *Math. Systems Theory* **2**, 159–177.

KASAI, T. (1970), An hierarchy between context-free and context-sensitive languages, *J. Comput. System Sci.* **4**, 492–508.

MORIYA, E. (1973), Associate languages and derivational complexity of formal grammars and languages, *Inform. Contr.* **22**, 139–162.

MORIYA, E. (1973b), Some remarks on state grammars and matrix grammars, *Inform. Contr.* **23**, 48–57.

MAYER, O. (1972), Some restrictive devices for context-free grammars, *Inform. Contr.* **20**, 69–92.

SALOMAA, A. (1970), Periodically time-variant context-free grammars, *Inform. Contr.* **17**, 294–311.

SALOMAA, A. (1973), "Formal Languages," Academic Press, New York and London.