INFORMATION AND CONTROL 29, 252–267 (1975)

Phrase Structure Grammars Generating Context-Free Languages

ROBERT L. CANNON, JR.

Department of Mathematics and Computer Science,
University of South Carolina, Columbia, South Carolina 29208

For a phrase structure grammar G an algebraic approach is used for representing the structural derivations of the grammar. This representation yields the canonical derivations of elements of $L(G)$. It is shown that if all "right-canonical" derivations of all elements of $L(G)$ are such that the number of "nonrightmost" derivations between "rightmost" derivations is bounded, then $L(G)$ is context-free.

INTRODUCTION

In a recent paper, Book (1973) remarked that there is no convenient specification of a structural description of generation by a context-sensitive grammar. He has noted the use of a derivation tree as a structural description for generation by a context-free grammar and the lack of a similar mechanism for describing context-sensitive generation.

An algebraic approach, which yields a unique algebraic expression over a set of symbols from the alphabet of the grammar, is presented here. The expression yields not only a structural description of a derivation but also the canonical derivation associated with the description. In the algebraic representation it is possible in one linear expression to observe the contextual interaction of symbols as a string is generated by the grammar. As an indication of the utility of this approach, it is used to describe a condition under which the language generated by a phrase structure grammar[1] will be context-free.[2]

[1] A *phrase structure grammar* is a system $G = \langle V_N, V_T, P, S \rangle$, where $V_N \cap V_T = \varnothing$, V_N is a finite set of *nonterminal* symbols, V_T a set of *terminal* symbols, and $V = V_N \cup V_T$ is the alphabet of G. For Λ the empty string, V^+ the closure of V under catenation, and $V^* = V^+ \cup \{\Lambda\}$, $P \subseteq V^+ \times V^*$ is the set of *productions* of G. The string $\pi\sigma\omega$ derives the string $\pi\tau\omega$, written $\pi\sigma\omega \Rightarrow \pi\tau\omega$, if there exists a production $\sigma \to \tau$ in P. The reflexive transitive closure of $\Rightarrow$ is written as $\overset{*}{\Rightarrow}$. The *language generated by* G is $L(G) = \{x \mid x \in V_T^*, S \overset{*}{\Rightarrow} x\}$ (cf. Hopcroft and Ullman, 1969).

[2] A phrase structure grammar is *context-free* if $P \subseteq V_N \times V^*$.

1. Algebraic Preliminaries

For a phrase structure grammar we wish to obtain the structural descriptions generated by the grammar. A number of different approaches have been taken both for context-free grammars (Chomsky and Schützenburger, 1961; Weiss, Magó, and Stanat, 1973; Thatcher, 1967) and phrase structure grammars (Griffiths, 1968; Loeckx, 1970; Eickel and Loeckx, 1972; Hart, 1974).

As a means of demonstrating the existence of ambiguous derivations for a context-free grammar, Chomsky and Schützenberger (1961) presented a formal power series representation for the derivations of a context-free grammar. In the power series representation, each element of V_T^* occurs with a coefficient that indicates the degree of ambiguity with which the grammar generates a string. Also for context-free grammars, Weiss, Magó, and Stanat (1972) represented algebraically not only the existence of an element x in $L(G)$ but actually the sequence(s) of derivations by which x is generated. Corresponding to $x \in L(G)$ is a formal sum such that each term in the formal sum represents a derivation sequence for x.

In presenting a structural description of a phrase structure grammar a more complex representation becomes necessary. For a context-free grammar G, $L(G)$ may be generated by derivations in which the leftmost terminal is always rewritten. For a phrase structure grammar G, if $L(G)$ is not context-free, then G cannot generate $L(G)$ by only leftmost derivations (Evey, 1963; Matthews, 1963). Thus, a canonical form for derivation by a phrase structure grammar cannot be leftmost derivation. The definition of canonical derivation to be used here is one that has appeared in several equivalent forms (Griffiths, 1968; Loeckx, 1970). It requires that a derivation be as near leftmost (or rightmost), as possible in a manner to become more explicit later.

In the structural description to be presented, each production of a grammar will be represented by an element of the free monoid[3] $\mathscr{W} = \langle (V \cup \overline{V} \cup \{[\,,]\})^*,$ catenation, $\Lambda \rangle$[4] with the convention that $\overline{AB} \equiv \bar{B}\bar{A}$. As an example the production $AB \rightarrow CxDy$ will be represented by

$$[AB\overline{CxDy}] \equiv [AB\bar{y}\bar{D}\bar{x}\bar{C}] \in \mathscr{W}.$$

Elements of $\mathscr{W}$ will become polynomials in a semiring[5] $\mathscr{R}(\mathscr{W})$ of polynomials such that (1) each polynomial is a formal sum (under $+$) of terms;

[3] A *monoid* $\langle A, \cdot, 1 \rangle$ is closed under the associative operation $\cdot$ and has identity 1.

[4] $\overline{V} = \{\bar{v} \mid v \in V\}$.

[5] A *semiring* is an algebraic system $\langle S, +, \cdot, 0 \rangle$ such that $\langle S, +, 0 \rangle$ is a commutative monoid, $\langle S, \cdot \rangle$ is closed under the associative operation $\cdot$, and the operator $\cdot$ distributes over $+$.

(2) each term is of the form $c\rho$, where c is in the Boolean semiring $\mathscr{B}$[6] of coefficients; (3) $b\zeta + c\zeta = (b + c)\zeta$, $(b\eta) \cdot (c\zeta) = (bc)(\eta\zeta)$, $b, c \in \mathscr{B}$, $\eta, \zeta \in \mathscr{W}$; and (4) addition and multiplication of polynomials are performed in the usual manner consistent with (3).

All coefficients of elements of $\mathscr{R}(\mathscr{W})$ are either *1* or *0*. By convention the terms with coefficient *0* will not be written and in the other terms the *1* will not explicitly appear. If p is a polynomial of $\mathscr{R}(\mathscr{W})$ the use of $\alpha \in p$ will indicate that α is a term of p.

As an example the productions $S \to a$, $S \to XY$, $SX \to aY$ of a grammar would yield the polynomial $[A] + [S\bar{a}] + [S\bar{Y}\bar{X}] + [SX\bar{Y}\bar{a}] \in \mathscr{R}(\mathscr{W})$ associated with S.

For G a context-free grammar and I an index set[7] for the productions of P, a sequence of indices from I would be sufficient for describing the canonical (leftmost) derivations of a word $x \in L(G)$. As mentioned earlier, however, a non-context-free grammar G cannot generate $L(G)$ only by the use of leftmost derivations. Therefore, a sequence of indices from I is insufficient for describing generation by a phrase structure grammar. Additional information about the position of the string next to be rewritten must also be given. This paper gives an algebraic approach to describing that position and of observing the use of context in rewriting a symbol.

Terms in a polynomial in $\mathscr{R}(\mathscr{W})$ will represent potential derivations from a grammar G. Brackets in the term will assist in determining the position of the left side of a production when the left side is rewritten. If, after the brackets are removed, the remaining term cancels in the half-group D_V[8] generated by $V \cup \bar{V}$, then $S \overset{*}{\Rightarrow} x$.

As an example, the string

$$[\bar{S}[S\bar{C}\bar{B}\bar{A}[AB\bar{q}[qC\bar{r}\bar{q}]\,\bar{A}[A\bar{p}]]]] \in \mathscr{R}(\mathscr{W})$$

when debracketized yields

$$\bar{S}S\bar{C}\bar{B}\bar{A}AB\bar{q}qC\bar{r}\bar{q}\bar{A}A\bar{p} = \bar{r}\bar{q}\bar{p} = \overline{pqr} \in D_V$$

and will correspond to the "right canonical" derivation

$$\begin{aligned} S &\Rightarrow ABC & S &\to ABC \\ &\Rightarrow AqC & AB &\to Aq \\ &\Rightarrow Aqr & qC &\to qr \\ &\Rightarrow pqr & A &\to p. \end{aligned}$$

[6] $\mathscr{B} = \langle\{0, 1\}, +, \cdot, 0\rangle$ is a semiring such that $1 + x = 1 \cdot 1 = 1$, $0 \cdot x = 0 + 0 = 0$ for $x \in \{0, 1\}$.

[7] $I = \{0, 1, \ldots, n - 1\}$; $P = \{p_0, p_1, \ldots, p_{n-1}\}$.

[8] D_V is the half-group $\langle (V \cup \bar{V})^*, \cdot, \Lambda\rangle$ with the relation $\bar{v}v = \Lambda$ for all $v \in V$. The Dyck set on the alphabet V is a subhalf-group of D_V.

2. A Structural Description and Canonical Form

For $G = \langle V_N, V_T, P, S\rangle$, a phrase structure grammar, and $V = V_N \cup V_T$ the alphabet of G, let $\mathscr{W}$ be the monoid $\langle V \cup \bar{V} \cup \{[\,,\,]\})^*$, catenation, $\Lambda\rangle$, $[\,,\,] \notin V$. Define

$$h\colon V \to \mathscr{R}(\mathscr{W})$$

such that

1. $[\Lambda]$ is a term in $h(x)$
2. $[X\alpha\bar{\beta}]$ is a term in $h(x)$ for all $X \in V$ such that $X\alpha \to \beta$ is in P
3. There are no other terms in $h(X)$.[9]

As an example, for the grammar

$$\begin{aligned} G &= \langle V_N, V_T, P, S\rangle \\ V_N &= \{S, A, B, C, D\} \\ V_T &= \{x, y, z\} \\ P &= S \to ABCD,\ A \to x,\ BD \to yD,\ C \to \Lambda,\ D \to z\}, \end{aligned} \tag{1}$$

the mapping h would be defined as follows:

$$\begin{aligned} h\colon S &\mapsto \Lambda + [S\bar{D}\bar{C}\bar{B}\bar{A}] \\ A &\mapsto \Lambda + [A\bar{x}] \\ B &\mapsto \Lambda + [BD\bar{y}] \\ C &\mapsto \Lambda + [C] \\ D &\mapsto \Lambda + [D\bar{z}] \\ x &\mapsto \Lambda \\ y &\mapsto \Lambda \\ z &\mapsto \Lambda. \end{aligned}$$

For $\alpha_0\bar{X}_0\alpha_1\bar{X}_1 \cdots \alpha_m\bar{X}_m\alpha_{m+1}$ in $\mathscr{W}$, $X_i \in V$,

$$\alpha_i \in (V \cup \bar{V} \cup \{[\,,\,]\})^*$$

$$\alpha_i \neq [\beta, \beta \in (V \cup \{[\,,\,]\})^*,$$

[9] For a context-free grammar in Greibach normal form, Greibach (1973) constructs terms in an expression such that all terms represent productions which share the leftmost symbol of the right side in common, rather than the leftmost symbol of the left side, as given here.

define

$$\delta: \mathscr{W} \to \mathscr{R}(\mathscr{W})$$

by

$$\delta: \alpha_0\overline{X}_0\alpha_1\overline{X}_1 \cdots \alpha_m\overline{X}_m\alpha_{m+1} \mapsto \sum_{i=0}^{m}\left(\left(\sum_{j=0}^{i-1}\alpha_j\overline{X}_j\right)\alpha_i\overline{X}_i h(X_i)\left(\sum_{j=i+1}^{m}\alpha_j\overline{X}_j\right)\alpha_{m+1}\right).$$

Thus, δ inserts $h(X_i)$ into the string immediately to the right of $\overline{X}_i$. This is done for each occurrence of a symbol from $\overline{V}$. Clearly, δ can be extended to a homomorphism

$$\delta: \mathscr{R}(\mathscr{W}) \to \mathscr{R}(\mathscr{W}).$$

As an example for the grammar in (1),

$$\delta: [\bar{S}] \mapsto [\bar{S}] + [\bar{S}[S\bar{D}\bar{C}\bar{B}\bar{A}]].$$

A map

$$\phi: V \cup \overline{V} \cup \{[\ ,\ ]\} \to D_V$$

is defined by

$$\phi: x \mapsto \begin{cases} x & \text{if } x \in V \\ \Lambda & \text{if } x \in \{[\ ,\ ]\}. \end{cases}$$

Thus, ϕ erases the brackets and maps elements of $V \cup \overline{V} \cup \{[\ ,\ ]\}$ into the half-group D_V. ϕ may be extended to a homomorphism of $\mathscr{R}(\mathscr{W})$ into $\mathscr{R}(D_V)$. Moreover, for any polynomial α in $\mathscr{R}(\mathscr{W})$, $\phi(\alpha)$ allows left cancellation in D_V of the elements of V by the elements of $\overline{V}$.

Continuing the example above

$$\phi\delta: [\bar{S}] \mapsto \bar{S} + \bar{D}\bar{C}\bar{B}\bar{A}, \qquad (\bar{D}\bar{C}\bar{B}\bar{A} = \overline{ABCD}),$$

and this polynomial represents two derivations from S: the null derivation $S \overset{*}{\Rightarrow} S$ and the derivation $S \Rightarrow ABCD$.

In like manner

$$\begin{aligned}\delta^2: [\bar{S}] \mapsto\ & [\bar{S}] + [\bar{S}[S\bar{D}\bar{C}\bar{B}\bar{A}]] + [\bar{S}[S\bar{D}[D\bar{z}]\ \bar{C}\bar{B}\bar{A}]] \\ & + [\bar{S}[S\bar{D}\bar{C}[C]\ \bar{B}\bar{A}]] + [\bar{S}[S\bar{D}\bar{C}\bar{B}[BD\bar{D}\bar{y}]\ \bar{A}]] \\ & + [\bar{S}[S\bar{D}\bar{C}\bar{B}\bar{A}[A\bar{x}]]] \qquad (2)\end{aligned}$$

$$\phi\delta^2: [\bar{S}] \mapsto \bar{S} + \bar{D}\bar{C}\bar{B}\bar{A} + \bar{z}\bar{C}\bar{B}\bar{A} + \bar{D}\bar{B}\bar{A} + \bar{D}\bar{C}D\bar{D}\bar{y}\bar{A} + \bar{D}\bar{C}\bar{B}\bar{x}. \qquad (3)$$

TABLE I

Summands in $\delta^2([\bar{S}])$ and Their Associated Derivations

Summands in $\delta^2([\bar{S}])$	Associated summands in $\phi\delta^2([\bar{S}])$	Associated derivation
$[\bar{S}]$	$\bar{S}$	$S \overset{*}{\Rightarrow} S$
$[\bar{S}[S\bar{D}\bar{C}\bar{B}\bar{A}]]$	$\bar{D}\bar{C}\bar{B}\bar{A}$	$S \Rightarrow ABCD$
$[\bar{S}[S\bar{D}[D\bar{z}]\bar{C}\bar{B}\bar{A}]]$	$\bar{z}\bar{C}\bar{B}\bar{A}$	$S \Rightarrow ABCD \Rightarrow ABCz$
$[\bar{S}[S\bar{D}\bar{C}[C]\bar{B}\bar{A}]]$	$\bar{D}\bar{B}\bar{A}$	$S \Rightarrow ABCD \Rightarrow ABD$
$[\bar{S}[S\bar{D}\bar{C}\bar{B}[BD\bar{D}\bar{y}]\bar{A}]]$	$\bar{D}\bar{C}D\bar{D}\bar{y}\bar{A}$	None
$[\bar{S}[S\bar{D}\bar{C}\bar{B}\bar{A}[Ax]]]$	$\bar{D}\bar{C}\bar{B}\bar{x}$	$S \Rightarrow ABCD \Rightarrow xBCD$

TABLE II

Summands in $\delta^3([S])$ and Their Associated Derivations

Summands in $\delta^3([\bar{S}])$	Associated summands in $\phi\delta^3([\bar{S}])$	Associated derivation
$[\bar{S}]$	$\bar{S}$	$S \overset{*}{\Rightarrow} S$
$[\bar{S}[S\bar{D}\bar{C}\bar{B}\bar{A}]]$	$\bar{D}\bar{C}\bar{B}\bar{A}$	$S \Rightarrow ABCD$
$[\bar{S}[S\bar{D}[D\bar{z}]\bar{C}\bar{B}\bar{A}]]$	$\bar{z}\bar{C}\bar{B}\bar{A}$	$S \Rightarrow ABCD \Rightarrow ABCz$
$[\bar{S}[S\bar{D}\bar{C}[C]\bar{B}\bar{A}]]$	$\bar{D}\bar{B}\bar{A}$	$S \Rightarrow ABCD \Rightarrow ABD$
$[\bar{S}[S\bar{D}\bar{C}\bar{B}[BD\bar{D}\bar{y}]\bar{A}]]$	$\bar{D}\bar{C}D\bar{D}\bar{y}\bar{A}$	None
$[\bar{S}[S\bar{D}\bar{C}\bar{B}\bar{A}[A\bar{x}]]]$	$\bar{D}\bar{C}\bar{B}\bar{x}$	$S \Rightarrow ABCD \Rightarrow xBCD$
$[\bar{S}[S\bar{D}[D\bar{z}]\bar{C}[C]\bar{B}\bar{A}]]$	$\bar{z}\bar{B}\bar{A}$	$S \Rightarrow ABCD \Rightarrow ABCz \Rightarrow ABz$
$[\bar{S}[S\bar{D}[D\bar{z}]\bar{C}\bar{B}[BD\bar{D}\bar{y}]\bar{A}]]$	$\bar{z}\bar{C}D\bar{D}\bar{y}\bar{A}$	None
$[\bar{S}[S\bar{D}[D\bar{z}]\bar{C}\bar{B}\bar{A}[A\bar{x}]]]$	$\bar{z}\bar{C}\bar{B}\bar{x}$	$S \Rightarrow ABCD \Rightarrow ABCz \Rightarrow xBCz$
$[\bar{S}[S\bar{D}\bar{C}[C]\bar{B}[BD\bar{D}\bar{y}]\bar{A}]]$	$\bar{D}\bar{y}\bar{A}$	$S \Rightarrow ABCD \Rightarrow ABD \Rightarrow AyD$
$[\bar{S}[S\bar{D}\bar{C}[C]\bar{B}\bar{A}[A\bar{x}]]]$	$\bar{D}\bar{B}\bar{x}$	$S \Rightarrow ABCD \Rightarrow ABD \Rightarrow xBD$
$[\bar{S}[S\bar{D}\bar{C}\bar{B}[BD\bar{D}\bar{y}]\bar{A}[A\bar{x}]]]$	$\bar{D}\bar{C}D\bar{D}\bar{y}\bar{x}$	None

The relationship between the summands in (3) and derivations for G is given in Table I.

If δ is applied once more to $[\bar{S}]$, then

$$\begin{aligned}\delta^3\colon [\bar{S}] \mapsto \delta^2([\bar{S}]) &+ [\bar{S}[S\bar{D}[D\bar{z}]\ \bar{C}[C]\ \bar{B}\bar{A}]] \\ &+ [\bar{S}[S\bar{D}[D\bar{z}]\ \bar{C}\bar{B}[BD\bar{D}\bar{y}]\ \bar{A}]] + [\bar{S}[S\bar{D}[D\bar{z}]\ \bar{C}\bar{B}\bar{A}[A\bar{x}]]] \\ &+ [\bar{S}[S\bar{D}\bar{C}[C]\ \bar{B}[BD\bar{D}\bar{y}]\ \bar{A}]] + [\bar{S}[S\bar{D}\bar{C}[C]\ \bar{B}\bar{A}[A\bar{x}]]] \\ &+ [\bar{S}[S\bar{D}\bar{C}\bar{B}[BD\bar{D}\bar{y}]\ \bar{A}[A\bar{x}]]]. \qquad (4)\end{aligned}$$

Table II shows for each term in (4) the associated term in D_V and the associated derivation.

Note that the left brackets of a summand in $\mathscr{R}(\mathscr{W})$ indicate application of a production and that these may be interpreted from left to right. An example from (4) is the following:

$$\begin{array}{l}[\bar{S}[S\bar{D}\bar{C}[C]\bar{B}[BD\bar{D}\bar{y}]\bar{A}]] \\ \quad\ |\quad\ \diagdown\quad\ \diagdown \\ S \Rightarrow ABCD \Rightarrow ABD \Rightarrow AyD\end{array} \qquad (5)$$

Thus, in (5) the symbol C is erased before application of the production $BD \Rightarrow yD$.

A problem arises with the term $[\bar{S}[S\bar{D}\bar{C}[BD\bar{D}\bar{y}]\ \bar{A}]]$ in (2). The expression implies the rewriting of BD as Dy prior to the erasure of C. Incorrect use of context becomes apparent when the expression is mapped into D_V, for

$$\phi\colon [\bar{S}[S\bar{D}\bar{C}\bar{B}[BD\bar{D}\bar{y}]\ \bar{A}]] \mapsto \bar{D}\bar{C}D\bar{D}\bar{y}\bar{A}.$$

The presence of a nonbarred symbol in a term of a polynomial in $\mathscr{R}(D_V)$ will indicate an invalid term in $\mathscr{R}(\mathscr{W})$. This situation is remedied by erasing such terms from the polynomial. The mapping ϵ will accomplish that erasure. Define

$$\epsilon\colon \mathscr{R}(D_V) \to \mathscr{R}(D_V)$$

such that

$$\epsilon\colon \alpha \mapsto \begin{cases}\alpha & \text{if } \alpha \in \overline{V}^* \\ \Lambda & \text{otherwise.}\end{cases}$$

Lastly, define

$$\theta\colon \mathscr{R}(\mathscr{W}) \to \mathscr{R}(\mathscr{W})$$

such that for ξ a summand in $\mathscr{R}(\mathscr{W})$

$$\theta\colon \xi \mapsto \begin{cases} \delta(\xi) & \text{if } \epsilon\phi\delta(\xi) = \phi\delta(\xi) \\ \Lambda & \text{otherwise.} \end{cases}$$

The mapping θ performs an iteration of δ and erases from $\mathscr{R}(\mathscr{W})$ any element that does not yield a string of symbols over $\overline{V}$ after removing the brackets and using the left-cancellation property of the half-group D_V. This requirement is similar to that imposed upon a "quick-cancel" derivation of Savitch (1973). Thus,

$$\begin{aligned}
\theta([\bar{S}]) &= [\bar{S}] + [\bar{S}[S\bar{D}\bar{C}\bar{B}\bar{A}]] \\
\theta^2([\bar{S}]) &= \theta([\bar{S}]) + [\bar{S}[S\bar{D}[D\bar{z}]\,\bar{C}\bar{B}\bar{A}]] + [\bar{S}[S\bar{D}\bar{C}[C]\,\bar{B}\bar{A}]] \\
&\quad + [\bar{S}[S\bar{D}\bar{C}\bar{B}\bar{A}[A\bar{x}]]] \\
\theta^3([\bar{S}]) &= \theta^2([\bar{S}]) + [\bar{S}[S\bar{D}[D\bar{z}]\,\bar{C}[C]\,\bar{B}\bar{A}]] + [\bar{S}[S\bar{D}[D\bar{z}]\,\bar{C}\bar{B}\bar{A}[A\bar{x}]]] \\
&\quad + [\bar{S}[S\bar{D}\bar{C}[C]\,\bar{B}[BD\bar{D}\bar{y}]\,\bar{A}]] \\
\theta^4([\bar{S}]) &= \theta^3([\bar{S}]) + [\bar{S}[S\bar{D}[D\bar{z}]\,\bar{C}[C]\,\bar{B}\bar{A}[A\bar{x}]]] \\
&\quad + [\bar{S}[S\bar{D}\bar{C}[C]\,\bar{B}[BD\bar{D}[D\bar{z}]\,\bar{y}]\,\bar{A}]] \\
&\quad + [\bar{S}[S\bar{D}\bar{C}[C]\,\bar{B}[BD\bar{D}\bar{y}]\,\bar{A}[A\bar{x}]]] \\
\theta^5([\bar{S}]) &= \theta^4([\bar{S}]) + [\bar{S}[S\bar{D}\bar{C}[C]\,\bar{B}[BD\bar{D}[D\bar{z}]\,\bar{y}]\,\bar{A}[A\bar{x}]]].
\end{aligned}$$

The new summand introduced into $\theta^5([\bar{S}])$ is a rightmost-except-for-context derivation of xyz from S, represented as follows:

$$\begin{array}{l}
[\bar{S}[S\bar{D}\bar{C}[C]\bar{B}[BD\bar{D}[D\bar{z}]\,\bar{y}]\bar{A}[A\bar{x}]]] \\
\quad\;\; | \qquad \diagdown \qquad \diagdown \qquad \diagdown \qquad \diagdown \\
S \Rightarrow ABCD \Rightarrow ABD \Rightarrow AyD \Rightarrow Ayz \Rightarrow xyz.
\end{array}$$

The mapping θ, when iterated, yields all the sentential forms that can be generated by a grammar. Moreover, the summands in $\mathscr{R}(\mathscr{W})$ satisfy a canonical form for describing generation. The form may be characterized by the restriction that all derivations are rightmost except that any symbol to the right of any one being rewritten or used for context is allowed to remain in a sentential form as long and only as long it is required for context in a subsequent derivation. This is the definition given originally by Griffiths (1968) and in an equivalent form by Loeckx (1970).

It will first be shown that iteration of θ yields exactly the sentential forms which can be generated by a grammar.

THEOREM 1. *Let* $G = \langle V_N, V_T, P, S\rangle$ *be a grammar. A string* β *in* $(V \cup \overline{V})^+$ *is a sentential form for* G *if and only if there exists a summand* α *in* $\mathscr{R}(\mathscr{W})$ *with* $\phi(\alpha) = \bar{\beta}$ *and an integer* $n (n > 0)$ *such that* α *is a summand of* $\theta^k([\bar{S}])$ *for all* $k \geqslant n$.

Proof. Assume that α is a summand in $\theta^k([\bar{S}])$ for all $k \geqslant n$. The proof is by induction on n.

Basis. Let α be a summand in $\theta^k([\bar{S}])$ for every $k \geqslant 1$. Then

$$\alpha = [\bar{S}[S\bar{\beta}]],$$

representing the production $S \to \beta$. Note that

$$\phi(\alpha) \cdot \beta = \bar{\beta}\beta = \Lambda.$$

Thus, the sentential form β is represented by the term α.

Induction. Assume α is a summand in $\theta^k([\bar{S}])$, for every $k \geqslant n + 1$. Then there is a string β in V^+ such that

$$\phi(\alpha) \cdot \beta = \Lambda.$$

Moreover, there is a string γ in $\theta^n([\bar{S}])$ such that $\delta(\gamma) = \alpha$. By the inductive assumption γ represents a sentential form. Thus there is a string ζ in V^+ such that $\phi(\gamma) \cdot \zeta = \Lambda$, and $S \overset{*}{\Rightarrow} \zeta$. The string γ may be written as

$$\gamma = \mu_0\overline{X}_0\mu_1\overline{X}_1 \cdots \mu_m\overline{X}_m\mu_{m+1},$$

where $X_i \in V$, $0 \leqslant i \leqslant m$, $\mu_j \in (V \cup \{[\,,]\})^*$, $\mu_j \neq [\nu$, $0 \leqslant j \leqslant m + 1$. Now

$$\delta(\gamma) = \sum_{i=0}^{m} \left(\left(\sum_{j=0}^{i-1} \mu_j\overline{X}_j \right) \mu_j\overline{X}_i h(X) \left(\sum_{j=i+1}^{m} \mu_j\overline{X}_j \right) \mu_{m+1} \right),$$

and thus

$$\alpha = \sigma\overline{X}h(X)\tau, \qquad \sigma, \tau \in \mathscr{W}.$$

But α is an element of $\theta(\gamma)$ only if $\phi\delta(\gamma)$ is in $\overline{V}^*$. This implies that the elements of V that were inserted into the string γ have been canceled by the elements of $\overline{V}$ which were already in the string γ. Thus, the proper context was available for rewriting the sentential form represented by γ. Thus, if α is in $\theta(\gamma)$, then all contextual requirements have been satisfied, so that there exists ξ in V^+ such that $\phi(\alpha) \cdot \xi = \Lambda$, and $S \overset{*}{\Rightarrow} \xi$. Thus, ξ is a sentential form.

To prove the converse, assume β to be a sentential form for G. It must be shown that for some integer n there is a summand α in $\theta^n([\bar{S}])$ such that $\phi(\alpha) \cdot \beta = \Lambda$ and α is a summand of $\theta^k([\bar{S}])$ for all $k \geqslant n$. The proof will be by induction upon the length of the derivation

$$S \Rightarrow \beta_1 \Rightarrow \cdots \Rightarrow \beta_n = \beta.$$

Basis. If $S \Rightarrow \beta$, then $S \rightarrow \beta$ is a production of the grammar. Thus $S\bar{\beta}$ is a summand of $h(S)$, so that $[\bar{S}[S\bar{\beta}]]$ is a summand of $\delta([\bar{S}])$ and $\phi\delta([\bar{S}]) = \bar{S}S\bar{\beta} = \bar{\beta}$. Letting $\alpha = [\bar{S}[S\bar{\beta}]]$, then α is a summand of $\theta^1([\bar{S}])$. Since Λ is a summand of $h(X)$ for all X in V, α is then a summand of $\theta^k([\bar{S}])$ for all $k \geqslant 1$.

Induction. Let γ be a sentential form. There is a sequence $S \Rightarrow \gamma_1 \Rightarrow \gamma_2 \Rightarrow \cdots \Rightarrow \gamma_k \Rightarrow \gamma$ of derivations for γ. Assume the proposition true for all derivations of length $\leqslant k$. By the inductive assumption there is a term ζ and an integer n such that $\phi(\zeta) \cdot \gamma_k = \Lambda$ and ζ is a summand of $\theta^j([\bar{S}])$ for all $j \geqslant n$. The derivation $\gamma_k \Rightarrow \gamma$ may be written as

$$\gamma_k = \mu X \sigma \nu = \gamma \qquad \mu, \nu, \sigma \in V^*, \quad X \in V,$$

where $X\sigma \rightarrow \tau$ is a production from P. Since $\phi(\zeta) \cdot \gamma_k = \Lambda$, it follows that $\phi(\zeta) = \bar{\nu}\bar{\sigma}\bar{X}\bar{\mu}$. There is a summand $[X\sigma\bar{\tau}]$ in $h(X)$ corresponding to the production $X\sigma \rightarrow \tau$ used to rewrite γ_k as γ. Thus, for the summand ζ in $\delta^n([\bar{S}])$,

$$\begin{aligned} \phi(\delta(\zeta)) &= \bar{\nu}\bar{\sigma}\bar{X}X\sigma\bar{\tau}\bar{\mu} \\ &= \bar{\nu}\bar{\tau}\bar{\mu}, \end{aligned}$$

so that

$$\begin{aligned} \phi(\delta(\xi)) \cdot \gamma &= \bar{\nu}\bar{\tau}\bar{\mu}\mu\tau\nu \\ &= \Lambda. \end{aligned}$$

Because Λ is a summand of $h(X)$ for all $X \in V$, any summand in $\theta^n([\bar{S}])$ is in $\theta^k([\bar{S}])$ for all $k \geqslant n$. ∎

Next it will be shown that the algebraic representation includes not only all the derivations of a word in $L(G)$ but also exactly the ones that are canonical in the sense of Griffiths (1968).

DEFINITION. A sequence of derivations such that if all derivations in a derivation sequence are of the form $\pi\sigma\omega \Rightarrow \pi\tau\omega = \pi'\sigma'\omega' \Rightarrow \pi'\tau'\omega'$ then $|\omega| < |\omega'| - |\sigma'|$[10] is called a *right-canonical* derivation sequence.

[10] $|\alpha|$ is the length of the string α.

THEOREM 2. *The algebraic expression α for a sentential form β represents the set of right-canonical derivations of β.*

Proof. Consider a term $\xi\zeta\overline{X}\eta$ in α for $\xi, \zeta, \eta \in \mathscr{W}$, $x \in V$, $\eta \neq [\rho, \rho \in \mathscr{W}$. The mapping δ when applied to this term will substitute $h(X)$ immediately to the right of $\overline{X}$. Assume moreover that ξ is the longest substring of the term such that no symbols in ξ cancel symbols in $h(X)$. Thus $\phi(\xi) = \omega, \phi(\zeta\overline{X}) = \sigma$.

Now consider a second application of δ (under the assumption that the term $\xi\zeta\overline{X}h(X)\eta$ represents a valid derivation sequence). Thus, it may be assumed that $\xi\zeta\overline{X}h(X)\eta = \xi'\zeta'\overline{Y}\eta'$ with $\xi', \zeta', \eta' \in \mathscr{W}$, $Y \in V, \eta' \neq [\rho, \rho \in \mathscr{W}$, and ξ' satisfying the same condition with respect to $h(Y)$ as satisfied by ξ with respect to $h(X)$. Application of δ yields $\xi'\eta'\overline{Y}h(Y)\eta'$. For this term, $\phi(\xi') = \bar{\omega}'$ and $\phi(\eta'\overline{Y}) = \bar{\sigma}'$.

The requirement that the derivation be canonical is thus $|\phi(\xi)| < |\phi(\xi')| + |\phi(\zeta'\overline{Y})|$. This is equivalent to $|\xi| < |\xi'| + |\xi'\overline{Y}|$ since ϕ only serves to erase pairs of symbols. If $|\xi| \geqslant |\xi'| + |\zeta'\overline{Y}|$, then $\overline{X}$ is already present as a symbol in ξ. Thus, δ must also have substituted $h(X)$ at the right of $\overline{X}$, and such a term must also be present in α. Any substitution to the left of $\overline{Y}$ which is valid implies application of the production represented by $h(X)$ prior to application of the production represented by $h(Y)$. Thus, it is impossible for any term in α to represent a derivation which is not right-canonical. ∎

The algebraic representation has thus yielded an expression that gives the set of right-canonical derivations for any sentential form that may be derived from a phrase structure grammar.

3. GRAMMARS GENERATING ONLY CONTEXT-FREE LANGUAGES

The use of context in generation of a non-context-free language is not well understood. Book (1973) discusses a number of such efforts. They can be classified as studies of the use of context for "passing messages," restrictions upon the form of an arbitrary grammar, or restrictions upon the manner in which the rewriting rules of a grammar may be applied. New results in this paper fall into the first category.

Evey (1963) and Matthews (1963) showed that if the condition that a derivation be "rightmost" ("leftmost") is relaxed to the extent that productions of the grammar may be applied within some fixed distance of the rightmost (leftmost) nonterminal symbol in a sentential form, then the grammar still generates a context-free language. A result given here for a grammar is

that if in a right-canonical derivation sequence the number of "nonrightmost" derivations performed between "rightmost" derivations is less than some positive integer k, then $L(G)$ is context-free.

DEFINITION. Let $G = \langle V_N, V_T, P, S \rangle$ be a grammar. Let α, a term in a polynomial in $\mathscr{R}(\mathscr{W})$, represent a sentential form for G. Let y represent an occurrence of a symbol $\bar{x}$ in α, $\bar{x} \in \overline{V}$. For the set of natural numbers N, a function $d: \overline{V} \to N$ is defined such that if y represents the occurrence of a symbol $\bar{x}$ in α, $\bar{x} \in \overline{V}$, then

$$\begin{aligned} d(y) &= 0 \quad \text{if the occurrence of } \bar{x} \text{ represented by } y \text{ is not canceled in } \phi(\alpha) \\ &= 2+ \text{ the number of symbols separating the occurrence of } \bar{x} \text{ represented by } y \text{ and the occurrence of the symbol which cancels it in } \phi(\alpha). \end{aligned}$$

For x, a symbol in α, $\bar{x} \in \overline{V}$, the *scope* of $\bar{x}$ in α is

$$\rho_\alpha(\bar{x}) = \max\{d(y) \mid y \text{ represents an occurrence of the symbol } \bar{x} \text{ in } \alpha\}.$$

The *scope* of α in G is

$$\rho_G(\alpha) = \max\{\rho_\alpha(\bar{x}) \mid \bar{x} \text{ a symbol in } \alpha,\ \bar{x} \in \overline{V}\}.$$

The *scope* of G is

$$\rho(G) = \max\{\rho_G(\alpha) \mid \alpha \text{ a summand in } \mathscr{R}(\mathscr{W}) \text{ representing a sentential form}\}.$$

The scope of a grammar is the maximum number of symbols separating the occurrence of a symbol x and the symbol that cancels it in a term representing a sentential form of the grammar. This maximum is over all terms representing sentential forms of G.

LEMMA 1. *Let n be a positive integer. Let α be a term of a polynomial in $\mathscr{R}(\mathscr{W})$ representing a sentential form for G. If $\rho_G(\alpha) = 2n$, then for any symbol from V involved in any application of a rewriting rule of G, the distance of that symbol from the rightmost symbol ever again to be rewritten or used for context in the application of a rewriting rule is less than or equal to $n - 1$.*

Proof. By induction on n.

Basis. For $n = 1$, α is of the form

$$\cdots \bar{X}X \cdots \bar{X}X \cdots \bar{X}X \cdots \qquad X \in V.$$

If α had this form, then all rules of G would be of the form $X \rightarrow \xi$. The grammar would be context-free, and the canonical derivation would always rewrite the rightmost symbol. Thus, the distance of the symbol being rewritten from the rightmost symbol would be zero.

Induction. Assume the lemma true for $n = k$. Thus $\rho_G(\alpha) = 2k$. Consider a substring of α of the form

$$\underbrace{\cdots \bar{Z} \cdots Z \cdots}_{2k \text{ symbols}}.$$

The symbol $\bar{Z}$ represents the rightmost symbol ever again to be involved in the application of a rewriting rule of G. A production of the form

$$Y_1 \cdots Y_{k-1} Z \rightarrow \zeta$$

would be represented by

$$\underbrace{\bar{Z}\bar{Y}_{k-1} \cdots \bar{Y}_1 Y_1 \cdots Y_{k-1} Z \bar{\zeta}}_{2k \text{ symbols}}.$$

Clearly the distance between Y_1 and Z is $k - 1$.

If two additional symbols are added, then in the worst case the substring of α would be of the form

$$\bar{Z}\bar{Y}_{k-1} \cdots \bar{Y}_1 \bar{W} W Y_1 \cdots Y_{k-1} Z \bar{\zeta},$$

and the distance between W and Z would be k.

Thus, distance is bounded by scope. ▮

THEOREM 3. *Let $G = \langle V_N, V_T, P, S \rangle$ be a grammar. If there exists a positive integer k such that for all right-canonical derivations of elements of $L(G)$, $\rho(G) < k$, then $L(G)$ is context-free.*

Proof. For an arbitrary grammar G, if the distance of the symbols being rewritten is a bounded distance from the rightmost symbol ever again to be rewritten, then $L(G)$ is context-free (Evey, 1963; Matthews, 1963). By Lemma 1, bounded scope implies that any symbol involved in a step of a derivation is a bounded distance from the rightmost symbol ever again to be rewritten. Thus, $L(G)$ is context-free. ▮

The above condition is sufficient but not necessary. As an example consider the grammar.

$$G = \langle V_N, V_T, P, S \rangle$$
$$V_N = \{S, X, Y\}$$
$$V_T = \{c\}$$
$$P = \{S \to c,\ S \to XY,\ XY \to XYY,\ X \to c,\ cY \to cc\}.$$

The scope of G is unbounded, yet $L(G) = \{c^n \mid n \geqslant 1\}$.

As a corollary to Theorem 3, it can now be shown that if the number of "nonrightmost" derivations between "rightmost" derivations is bounded then the right-canonical derivations of G generate a context-free language.

DEFINITION. Let $G = \langle V_N, V_T, P, S \rangle$ be a grammar. Let $w \in L(G)$. In a right-canonical derivation

$$S \overset{*}{\Rightarrow} \alpha \Rightarrow \beta \overset{*}{\Rightarrow} w$$

for w, a step in the derivation $\alpha \Rightarrow \beta$ is *rightmost* if $\alpha = \eta\sigma\theta$, $\beta = \eta\tau\theta$, η, $\theta \in V^*$, $\sigma, \tau \in V^+$, and there does not exist a symbol $z \in V$ in the string θ that is either rewritten or used as context in any later application of a rewriting rule. A derivation in the sequence $S \overset{*}{\Rightarrow} w$ is *nonrightmost* if it is not rightmost.

COROLLARY. *Let $G = \langle V_N, V_T, P, S \rangle$ be a grammar. If there exists a positive integer k such that for each $w \in L(G)$ and each right-canonical derivation for w the number of nonrightmost derivations occurring between rightmost derivations is less than k, then $L(G)$ is context-free.*

Proof. In the algebraic representation for a right-canonical derivation any symbol X which is rewritten or used for context appears as $\cdots \overline{X} \cdots X \cdots$. Any symbol to the right of X involved in a later rewriting would appear as

$$\cdots \overline{Y} \cdots \overline{X} \cdots X \cdots Y.$$

The expression can be broken up into sections, each of which represents a return to the rightmost symbol of the sentential form

$$\cdots \mid \overline{Y} \cdots Y \mid \cdots \mid \overline{Z} \cdots Z \mid \cdots.$$

If each of these "sections" has bounded scope then the scope of the grammar is bounded.

Assume that the number of symbols used to represent any derivation of G is no more than j. By hypothesis the number of nonrightmost derivations between rightmost derivations is no more than k. Thus, each "section"

contains no more than $j(k+1)$ symbols, and $\rho(G) \leqslant j(k+1)$. Hence $L(G)$ is context-free. ∎

Theorem 3 and Lemma 1 give yet another means of studying the manner in which information is passed around in the sentential forms of a derivation sequence. The restriction to the right-canonical form of any derivation serves, of course, to limit the distance of the symbols being rewritten from the leftmost or rightmost symbols ever again to be involved in a derivation. Still, it is now clear that there cannot be arbitrarily many nonrightmost derivations between rightmost derivations if the right-canonical derivations of a grammar are to generate a context-free language.

Whether an arbitrary number of steps of a derivation may "interact" (Book, 1973) is still not known. It is hoped that the algebraic representation given here will provide a basis for further investigation of that question.

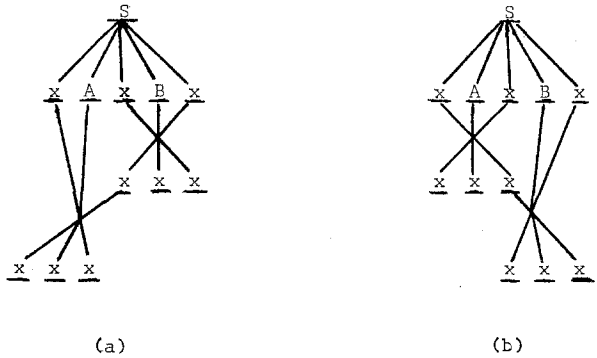

FIG. 1. Syntactical graphs for two canonical derivations of *xxxxx*.

RECEIVED: August 28, 1974; REVISED: April 18, 1975

REFERENCES

BAKER, B. S. (1974), Non-context-free grammars generating context-free languages, *Inform. Contr.* **24**, 231–246.

BOOK, R. (1973), On the structure of context-sensitive grammars, *Int. J. Computer Inform. Sci.* **2**, 129–139.

CHOMSKY, N. AND SCHÜTZENBERGER, M. (1961), The algebraic theory of context-free languages, *in* "Computer Programming and Formal Systems" (Braffort, Ed.), pp. 118–161, North-Holland, Amsterdam.

EICKEL, J. AND LOECKX, J. (1972), The relation between derivations and syntactical structures in phrase-structure grammars, *J. Computer Sys. Sci.* **6**, 267–282.

EVEY, R. (1963), "The Theory and Application of Pushdown Store Machines," Doctoral Dissertation, Harvard University.

GREIBACH, S. (1973), The hardest context-free language, *SIAM J. Computing* **2**, 304–310.

GRIFFITHS, T. V. (1968), Some remarks on derivations in general rewriting systems, *Inform. Contr.* **12**, 27–54.

HART, J. M. (1974), Acceptors for the derivation languages of phrase-structure grammars, *Inform. Contr.* **25**, 75–92.

HOPCROFT, J. E. AND ULLMAN, J. D. (1969), "Formal Languages and Their Relation to Automata," Addison-Wesley, Reading, MA.

LOECKX, J. (1970), The parsing for general phrase-structure grammars, *Inform. Contr.* **16**, 443–464.

MATTHEWS, G. H. (1963), Discontinuity and asymmetry in phrase structure grammars, *Inform. Contr.* **6**, 137–146.

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

SAVITCH, W. J. (1973), How to make arbitrary grammars look like context-free grammars, *SIAM J. Computing* **2**, 174–182.

THATCHER, J. (1967), Characterizing derivation trees of context-free grammars through a generalization of finite automata theory, *J. Computer Sys. Sci.* **1**, 317–322.

WEISS, S. F., MAGÓ, G. A., AND STANAT, D. F. (1973), Algebraic parsing techniques for context-free grammars, *in* "Automata, Languages and Programming" (M. Nivat, Ed.), pp. 493–498, North-Holland, Amsterdam.