JOURNAL OF COMPUTER AND SYSTEM SCIENCES 6, 217–232 (1972)

A Homomorphism Theorem for Weighted Context-Free Grammars

DONALD F. STANAT

Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina 27514

Received April 26, 1971

Productions of a context-free grammar can be given coefficients from semirings, inducing weights for both derivations in the grammar and strings over the terminal alphabet. For a weighted context-free grammar in Greibach normal form, the weight of any string, as well as the set of derivations of the string, may be determined from the image under a homomorphism which maps each terminal symbol to a polynomial. The definition of the homomorphism is a straightforward function of the productions. Some examples of interesting semiring structures are included.

INTRODUCTION

One of the most handsome results of the algebraic approach to formal language theory was established by Shamir [1967]. For a context-free grammar in Chomsky normal form, Shamir exhibited a noniterative procedure for determining the number of derivations of a string over the terminal alphabet of the grammar. The method consists of establishing a homomorphism from the free monoid generated by the terminal alphabet to the multiplicative structure of a ring of polynomials. If there are k derivations of x from A, then, after applying certain cancellation rules, the coefficient of A in the image of x is k.

The proof given by Shamir is based on a construction of Gaifman given in Bar-Hillel, Gaifman, and Shamir [1960] and is rather unwieldy. The purpose of this note is to exhibit a form of the theorem using a Greibach normal form and allowing weighted productions. The Greibach normal form results in an extremely straightforward construction. We will use weighted grammars in our development because in terms of the construction it costs no more to do so and because it enables us to do much more than simply count derivations.

1. MATHEMATICAL PRELIMINARIES AND NOTATION

A *semigroup* is formally presented as an ordered pair (e.g., $\langle S, \cdot \rangle$) where the first element denotes the set and the second the closed associative binary operation.

Similarly, a *monoid* is presented as a triple consisting of a set, an operation and a two-sided identity.

We denote the *free monoid generated by* A by $\langle A^*, \text{concatenation}, \Lambda\rangle$ and the *free semigroup generated by* A by $\langle A^+, \text{concatenation}\rangle$. We denote the length of $x \in A^*$ or $x \in A^+$, by $| x |$.

A *semiring* is an algebraic system $\mathbf{S} = \langle S, +, \cdot, 0\rangle$ such that

$$\langle S, +, 0\rangle \text{ is a commutative monoid;}$$
$$\langle S, \cdot\rangle \quad \text{is a semigroup;}$$

the operation $\cdot$ distributes over $+$:

$$a \cdot (b + c) = a \cdot b + a \cdot c,$$
$$(a + b) \cdot c = a \cdot c + b \cdot c.$$

A *semiring* is *commutative* if the operation $\cdot$ is commutative. A *semiring with identity* is a system $\langle S, +, \cdot, 0, 1\rangle$ where $\langle S, +, \cdot, 0\rangle$ is a semiring and $\langle S, \cdot, 1\rangle$ is a monoid.

In this paper we will assume all our semirings to have the following properties:

(a) they are commutative,
(b) they have an identity,
(c) the additive identity is a multiplicative zero; that is, $s \cdot 0 = 0 \cdot s = 0$.

In fact, this gives us very nearly a ring, but we don't have any use for additive inverses and so will not assume their existence. To eliminate repetition, we will use $\mathbf{S}$ to denote an arbitrary semiring $\langle S, +, \cdot, 0, 1\rangle$.

For an arbitrary set V, we define $\overline{V} = \{\overline{v} \mid v \in V\}$. The *free half-group generated by* V, $\mathbf{H}(V)$, is defined to be the monoid generated by $V \cup \overline{V}$ together with the relation $\bar{a}a = 1$, where 1 is the monoid identity and a is any element of V. Note that in $\mathbf{H}(V)$ the elements of V are right inverses but not left inverses of the corresponding elements of $\overline{V}$.

For any function $\mathbf{f} : A \to S$ where S is the set of elements of a semiring $\mathbf{S}$, the *support of* $\mathbf{f}$, $\text{Supp}(\mathbf{f})$, is defined to be the set $\{x \mid \mathbf{f}(x) \neq 0\}$.

For a given set V and semiring $\mathbf{S}$, a *power series over the noncommuting variables in* V is a function $p : V^* \to S$, and can be expressed as a formal sum

$$p = \sum_{x \in V^*} \langle p, x\rangle x,$$

where $\langle p, x\rangle = p(x) \in S$. For power series p and q, the sum $p + q$ and product $p \cdot q$ are defined by setting

$$\langle p + q, x\rangle = \langle p, x\rangle + \langle q, x\rangle,$$
$$\langle p \cdot q, x\rangle = \sum_{\substack{y, z \in V^* \\ x = yz}} \langle p, y\rangle \cdot \langle q, z\rangle.$$

If p is a power series with finite support, then p is a *polynomial.* If $\mathrm{Supp}(p)$ is a singleton set, then p is a *monomial.*

THEOREM 1.1. *The set of all power series over V with coefficients from a semiring* $\mathbf{S}$ *forms a semiring under the operations $+$ and $\cdot$: the semiring of power series over V with coefficients from* $\mathbf{S}$. *The set of all polynomials over V with coefficients from* $\mathbf{S}$ *also forms a semiring under these operations, the semiring of polynomials over V with coefficients from* $\mathbf{S}$. *Both semirings have a multiplicative identity (the polynomial 1Λ) and the additive identity, which is the power series with all coefficients equal to zero, is a multiplicative zero.*

Remark. We will, in fact, deal with polynomials over the noncommuting variables of the set $V \cup \overline{V}$. This will associate a coefficient in $\mathbf{S}$ with each element of the set $(V \cup \overline{V})^*$. We will then map the elements of $(V \cup \overline{V})^*$ to the elements of $\mathbf{H}(V)$ by treating the elements of V as right inverses of the corresponding elements of $\overline{V}$.

For a free half-group $\mathbf{H}(V) = \langle U, \circ, 1\rangle$, a power series p over $\mathbf{H}(V)$ with coefficients from a semiring $\mathbf{S}$ is a function $p : U \to S$, and can be expressed as a formal sum

$$p = \sum_{x \in U} \langle p, x\rangle x,$$

where $\langle p, x\rangle = p(x) \in S$. The sum and product of two polynomials p and q over a free half-group $\mathbf{H}(V)$ may be defined by setting

$$\langle p + q, x\rangle = \langle p, x\rangle + \langle q, x\rangle,$$

$$\langle p \cdot q, x\rangle = \sum_{\substack{y,z \in U \\ x = y \circ z}} \langle p, y\rangle \cdot \langle q, z\rangle.$$

In the sequel we will not distinguish between the half-group operation and the semigroup operation and we write $y \circ z$ as yz.

THEOREM 1.2. *The set of all polynomials over $\mathbf{H}(V)$ with coefficients from* $\mathbf{S}$ *forms a semiring under the operations $+$ and $\cdot$.*

Note that the set of power series over $\mathbf{H}(V)$ does not form a semiring, since the product of two power series may not be defined.

THEOREM 1.3. *Let* $\mathbf{h}$ *be the homomorphism from the free monoid generated by $V \cup \overline{V}$ to the free half-group generated by V, where* $\mathbf{h}$ *is the identity map on the set $V \cup \overline{V}$. Define a map* $\mathbf{g}$ *from the polynomials over the noncommuting variables in $V \cup \overline{V}$ to the polynomials over $\mathbf{H}(V)$, where* $\mathbf{g}$ *is defined by setting*

$$\mathbf{g}(cx) = c\mathbf{h}(x), \qquad c \in S,\ x \in (V \cup \overline{V})^*$$

$$\mathbf{g}(c_1 x + c_2 y) = \mathbf{g}(c_1 x) + \mathbf{g}(c_2 y).$$

Then **g** *is a homomorphism from the semiring of polynomials over the noncommuting variables in* $V \cup \overline{V}$ *to the semiring of polynomials over the free half-group* $\mathbf{H}(V)$.

TERMINOLOGY. We will call the map **h** of Theorem 1.3 the *canonical homomorphism* from the free monoid generated by $V \cup \overline{V}$ to the free half-group $\mathbf{H}(V)$. We will call the map **g** of Theorem 1.3 the *canonical homomorphism* from the semiring of polynomials over the noncommuting variables of $V \cup \overline{V}$ to the semiring of polynomials over the free halfgroup $\mathbf{H}(V)$.

Notation. In the sequel we will use the letters w, x, y, z to denote elements of V^* considered as either elements of the free monoid generated by V or the free monoid generated by $V \cup \overline{V}$. The corresponding capital letters W, X, Y, Z denote elements of $\mathbf{H}(V)$ which are equal to w, x, y, z when considered as sequences of symbols. As a result, $|w| = |W|$ and $\mathbf{h}(w) = W$, where **h** is defined in Theorem 1.3. The notation $\overline{W}, \overline{X}, \overline{Y}, \overline{Z}$ will be used to represent the string obtained by reversing the sequence of elements W, X, Y, Z respectively and barring the individual letters

$$\overline{A} = A,$$
$$\overline{XA} = \overline{A}\,\overline{X}.$$

The symbols $\bar{w}, \bar{x}, \bar{y}$ and $\bar{z}$ will denote the corresponding strings in the free monoid generated by $V \cup \overline{V}$. Consequently, $|\bar{w}| = |\overline{W}|$ and $\mathbf{h}(\bar{w}) = \overline{W}$.

2. WEIGHTED GRAMMARS AND DERIVATIONS

DEFINITION 2.1. A *weighted phrase-structure grammar* (wpsg) **G** *over a semiring* $\mathbf{S} = \langle S, +, \cdot, 0, 1\rangle$ is a system $\mathbf{G} = \langle V, V_T, P, A_1\rangle$ where

V is a finite set (the *alphabet*),

$V_T \subset V$ is the set of *terminal symbols*,

$A_1 \in V - V_T$ is called the initial symbol, or *axiom*, of **G**,

$P : (V - V_T)^+ \times V^* \to S$ is a *production function* of finite support.

Notation. We denote the set $V - V_T$ by V_N and refer to it as the set of nonterminal symbols.

DEFINITION 2.2. If $\mathrm{Supp}(P) \subset V_N \times V^+$ then **G** is a *weighted context-free grammar* (wcfg).

TERMINOLOGY. If $(x, y) \in \mathrm{Supp}(P)$ then (x, y) is called a production of **G**, and $P(x, y)$ is called the weight of the production. The set $\{(x, y) \mid (x, y) \in \mathrm{Supp}(P)\}$ will

be called the set of productions of **G**. We adopt a modification of a definition of of Griffiths [1968] of a derivation in a wcfg **G**.

DEFINITION 2.3. Let $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ be a wcfg over **S**. If

$$\mathbf{r} = (A_i, x) \in \mathrm{Supp}(P),$$

$w \in V^*$, $w = uA_iz$ and $| u | = k$, then we say $\mathbf{r}(k+1)$ is *applicable* to w and define $w\mathbf{r}(k+1)$ to be uxz. We call $k+1$ the index of application of $\mathbf{r}$ (or more correctly, this application of $\mathbf{r}$). We also say that $\mathbf{r}(k+1)$ rewrites (the symbol instance of) A_i as the string x. If $\mathbf{r} = (A_i, x)$, we will sometimes write $w\mathbf{r}(k+1)$ as $w((A_i, x)(k+1))$.

DEFINITION 2.4. A *derivation of z from x* is a sequence x, $\mathbf{r}_1(i_1)$, $\mathbf{r}_2(i_2)$,..., $\mathbf{r}_n(i_n)$, where

(1) $\mathbf{r}_j \in \mathrm{Supp}(P)$,
(2) $\mathbf{r}_j(i_j)$ is applicable to $x\mathbf{r}_1(i_1)\mathbf{r}_2(i_2) \cdots \mathbf{r}_{j-1}(i_{j-1})$,
(3) $z = x\mathbf{r}_1(i_1)\mathbf{r}_2(i_2) \cdots \mathbf{r}_n(i_n)$.

The *weight* of a derivation x, $\mathbf{r}_1(i_1)$, $\mathbf{r}_2(i_2)$,..., $\mathbf{r}_n(i_n)$ is $\prod_{i=1}^{n} P(\mathbf{r}_i)$ if $n > 0$. For $n = 0$ the weight of the derivation is defined to be 1. The *length* of the derivation is n.

Remark 2.1. For every string x there is a derivation of x from x of length 0.

TERMINOLOGY. We will call $\mathbf{r}_j(i_j)$ the j-th production of the derivation x, $\mathbf{r}_1(i_1)$, $\mathbf{r}_2(i_2)$,..., $\mathbf{r}_j(i_j)$,..., $\mathbf{r}_t(i_t)$. In doing so we knowingly confuse the production with an instance of its application.

DEFINITION 2.5. A derivation s, $\mathbf{r}_1(i_1)$, $\mathbf{r}_2(i_2)$,..., $\mathbf{r}_n(i_n)$ is a *canonical derivation* if $m < k$ implies $i_m \leqslant i_k$.

Remark 2.2. It is easy to show that if **d** is a canonical derivation from x to y and $y \in V_T^*$, then **d** corresponds to the usual notion of a left-most derivation.

DEFINITION 2.6. Let $\mathbf{d} = x$, $\mathbf{r}_1(i_1)$, $\mathbf{r}_2(i_2)$,..., $\mathbf{r}_t(i_t)$ be a derivation of z in **G**. We define $\mathbf{d}(k)$ to be the sequence of production applications

$$\mathbf{r}_1(k + i_1), \mathbf{r}_2(k + i_2), \ldots, \mathbf{r}_t(k + i_t).$$

If $u = wxy$, $| w | = k$ then $\mathbf{d}(k)$ is *applicable* to u and u, $\mathbf{d}(k)$ is a derivation of wzy. We will also write $u\mathbf{d}(k) = wzy$.

3. Languages and Equivalences of Grammars

We can now define the weight of a word $x \in V_T^*$ relative to a wpsg $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ over a semiring $\mathbf{S}$.

Definition 3.1. Let $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ be a wcfg over a semiring $\mathbf{S}$, and for any derivation $\mathbf{d}$ in $\mathbf{G}$, denote the weight of $\mathbf{d}$ by $\omega(\mathbf{d})$. For any word $x \in V_T^*$ let D_x be the set of all canonical derivations of x from A_1 in $\mathbf{G}$. The *weight of* x, denoted by $\omega(x)$, is defined to be 0 if D_x is empty. If D_x is not empty, then $\omega(x) = \sum_{\mathbf{d} \in D_x} \omega(\mathbf{d})$ if this sum converges in a sense appropriate to the semiring $\mathbf{S}$; otherwise $\omega(x)$ is undefined.

Examples. Let $\mathbf{G} = \langle \{A, a\}, \{a\}, P, A \rangle$ and let $\mathbf{S}$ be the set of nonnegative real numbers under ordinary addition and multiplication. If $P(A, A) = P(A, a) = 1$, there are an infinite number of derivations of the string a and each derivation has a weight of 1. Therefore $\omega(a)$ is undefined if we use the usual definition of convergence. If $P(A, A) = P(A, a) = 1/2$, for each positive integer n there is a derivation of the string a of length n and weight $(1/2)^n$. Consequently, $\omega(a) = 1$.

Definition 3.2. Let $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ be a wcfg and for $x \in V_T^*$ let D_x denote the set of all derivations of x from A_1. The *language generated by* $\mathbf{G}$, $L(\mathbf{G})$, is defined to be $\{x \mid x \in V_T^* \text{ and } D_x \neq \varnothing\}$.

Remark 3.1. The weight of one derivation may be the additive inverse of another. Consequently grammars exist such that $x \in L(\mathbf{G})$ but $\omega(x) = 0$.

Remark 3.2. The weight of a string x may be considered to be a generalized notion of ambiguity. If all productions have a weight of 1 then all derivations will have a weight of 1. Consequently $\omega(x)$ will equal the number of distinct canonical derivations of x if $\mathbf{S}$ is the semiring of nonnegative integers with plus and times.

Definition 3.3. Let $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ and $\mathbf{G}' = \langle V', V_T', P', A_1' \rangle$ be wcfg's over a semiring $\mathbf{S}$. Then $\mathbf{G}$ and $\mathbf{G}'$ are said to be *weakly equivalent* if and only if $L(\mathbf{G}) = L(\mathbf{G}')$. $\mathbf{G}$ and $\mathbf{G}'$ are *strongly equivalent* if they are weakly equivalent and for all $x \in V_T^*$, $\omega(x) = \omega'(x)$, where $\omega(x)$ is the weight of x in $\mathbf{G}$ and $\omega'(x)$ its weight in $\mathbf{G}'$.

Definition 3.4 (Griffiths). Let $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ and $H = \langle W, W_T, P', A_1' \rangle$ be wcfg's. Then H is an *extension* of $\mathbf{G}$ if and only if there is an effectively calculable injection E,

$$E: V^* \times V^* \to W^* \times W^*,$$

such that $E(w, x) = (y, z)$ implies there is an effectively calculable bijection from

canonical derivations from w to x to canonical derivations from y to z which preserves the weights of the derivations. The function E is called a projection function.

PROPOSITION 3.1. *If* **G** *and* **H** *are* wcfg*'s over a semiring* **S**, *then*

[**H** *is an extension of* **G** *with the projection function equal to the identity function and* $L(\mathbf{H}) = L(\mathbf{G})$]
$\Rightarrow$ [**H** *and* **G** *are strongly equivalent*]
$\Rightarrow$ [**H** *and* **G** *are weakly equivalent*].

The proof follows from Definitions 3.1, 3.2, 3.3, and 3.4.

4. NORMAL FORMS FOR GRAMMARS

In order to establish the domain of applicability of our later theorems, we state the following results without proof. The first asserts that we can restrict ourselves to grammars in which all productions rewrite a nonterminal symbol as either a single terminal symbol or as a string of nonterminal symbols.

PROPOSITION 4.1. *Let* $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ *be a* wcfg *over* **S**. *There exists a* wcfg $\mathbf{G}' = \langle V', V_T, P', A_1 \rangle$ *over* **S** *such that*

(a) **G**′ *is an extension of* **G** *with the identity projection function.*
(b) $L(\mathbf{G}) = L(\mathbf{G}')$.
(c) $\mathrm{Supp}(P') \subset (V_N \times V_N^+) \cup (V_N \times V_T)$.

The construction of **G**′ is as follows: If $V_T = \{a_1, a_2, \ldots, a_n\}$, let $M = \{a_1', a_2', \ldots, a_n'\}$, and set $V_N' = V_N \cup M$. For each a_i' define $P'(a_i', a_i) = 1$. For all $(x, y) \in \mathrm{Supp}(P)$, denote by y' the string obtained by replacing every occurrence of a_i in y by a_i' for each $a_i \in V_T$. Set $P'(x, y') = P(x, y)$. The proof using this construction is by induction.

DEFINITION 4.1. A wcfg $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ is in *Chomsky normal form* if every member of $\mathrm{Supp}(P)$ is of one of the following forms:

$$(A_i, A_jA_k),$$
$$(A_i, a),$$

where $A_i, A_j, A_k \in V_N$ and $a \in V_T$.

PROPOSITION 4.2. *Let* $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ *be a* wcfg *over* $\mathbf{S}$. *If there is no derivation of positive length in* $\mathbf{G}$ *of* A_i *from* A_i *where* $A_i \in V_N$, *then there exists a* wcfg $\mathbf{G}' = \langle V', V_T, P', A_1 \rangle$ *such that*

(a) $\mathbf{G}$ *and* $\mathbf{G}'$ *are strongly equivalent.*

(b) $\mathbf{G}'$ *is in Chomsky normal form.*

The construction is a straightforward modification of the traditional one. Notice that we can assume the function P obeys the restrictions of P' in Proposition 4.1. Furthermore, for $A_i, A_j \in V_N$ the set of derivations from A_i to A_j is finite and contains only derivations of length less than or equal to the cardinality of V_N. Hence for any $A_i \in V_N$ and $a \in V_T$ we can sum the weights of all derivations from A_i to a in $\mathbf{G}$ and set $P'(A_i, a)$ equal to the result. This will eliminate all productions which are elements of $V_N \times V_N$. For productions which are elements of $V_N \times V_N^+$ but not in the proper form, each production is broken up into a sequence of productions in the usual way. For example, if $P(A, BCDE) = c$, then $P'(A, K_1E) = P'(K_1, K_2D) = 1$ and $P'(K_2, BC) = c$. The proof using this construction is by induction.

COROLLARY 4.1. *Let* $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ *be a* wcfg *over* $\mathbf{S}$. *If* $P(A_i, A_j) = 0$ *for all* $A_i, A_j \in V_N$, *then there exists a* wcfg $\mathbf{G}' = \langle V', V_T, P', A_1 \rangle$ *such that*

(a) $\mathbf{G}'$ *is an extension of* $\mathbf{G}$ *with the identity projection function.*

(b) $L(\mathbf{G}) = L(\mathbf{G}')$.

(c) $\mathbf{G}'$ *is in Chomsky normal form.*

The proof follows from the fact that if $P(A_i, A_j) = 0$ then there is at most one derivation from any $A_i \in V_N$ to any $a \in V_T$. As a result, the construction for Proposition 4.2 will suffice to prove this corollary.

DEFINITION 4.2. A wcfg $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ is in *Greibach normal form* if and only if every element of Supp(P) is one of the following forms:

$$(A_i, aA_jA_k),$$
$$(A_i, aA_j),$$
$$(A_i, a),$$

where $A_i, A_j, A_k \in V_N$ and $a \in V_T$.

PROPOSITION 4.3. *Let* $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ *be a* wcfg *over* $\mathbf{S}$ *such that* $P(A_i, A_j) = 0$ *for all* $A_i, A_j \in V_N$. *Then there exists a* wcfg $\mathbf{G}' = \langle V', V_T, P', A_1 \rangle$ *such that*

(1) $\mathbf{G}'$ *is in Greibach normal form.*

(2) $\mathbf{G}'$ *is an extension of* $\mathbf{G}$ *with the identity projection function.*

(3) $L(\mathbf{G}) = L(\mathbf{G}')$.

It is simpler to prove a similar theorem for a grammar $\mathbf{G}^\circ$ which is strongly equivalent to **G**. The proof in this case is a modification of that given in Hopcroft and Ullman [1969]. Proposition 4.3 asserts the existence of a grammar which satisfies the stronger condition of being an extension of **G**. The proof of this assertion requires a construction similar to that used in Greibach [1965] and using the methods described in Book *et al.* [1971]. In either case the modifications to the classical proofs are relatively straightforward. When a sequence of productions is constructed in $\mathbf{G}'$ to replace a single production in **G**, the last production of the sequence is given the weight of the production in **G** and all other productions of the sequence are given the weight of 1.

5. The Homomorphism Theorem

Let $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ be a wcfg over **S** in Greibach normal form. Denote by **V** the free monoid generated by $V \cup \overline{V}$ and by $\mathbf{H}(V)$ the free half-group generated by V. Denote by $R_{\text{pol}}(\mathbf{V})$ the ring of polynomials with coefficients from **S** and noncommuting variables from $V \cup \overline{V}$. Denote by $R_{\text{pol}}(\mathbf{H})$ the ring of polynomials with noncommuting variables from the free half-group $\mathbf{H}(V)$ with coefficients from **S**. Denote by **h** the canonical homomorphism from **V** to $\mathbf{H}(V)$, and by **g** the canonical homomorphism from $R_{\text{pol}}(\mathbf{V})$ to $R_{\text{pol}}(\mathbf{H})$.

Definition 5.1. For $x \in V_T^+$, define the set τ_x of monomials of $R_{\text{pol}}(\mathbf{V})$ as follows:

$$\text{For } a \in V_T, \qquad cA_i \in \tau_a \leftrightarrow P(A_i, a) = c,$$

$$cA_i\bar{A}_j \in \tau_a \leftrightarrow P(A_i, aA_j) = c,$$

$$cA_i\overline{A_jA_k} \in \tau_a \leftrightarrow P(A_i, aA_jA_k) = c.$$

$$\text{For } x \in V_T^+,\ a \in V_T, \quad \tau_{xa} = \{c_1 \cdot c_2t_1t_2 \mid c_1t_1 \in \tau_x \text{ and } c_2t_2 \in \tau_a\}.$$

Furthermore, we define τ to be the disjoint union of all τ_x, $x \in V_T^+$. By disjoint union, we imply that if $x \neq y$ then the monomials of τ_x are to be treated as distinct from those of τ_y.

Definition 5.2. Let **D** be the set of all canonical derivations **d** in **G** from $x \in V_N^+$ to $y \in V^+$ such that every symbol of x is rewritten by a production of **d**.

Recall that W, X, Y, Z denote strings over the alphabet V in the free half-group $\mathbf{H}(V)$, while w, x, y, z denote the same sequences of symbols in **V**, the free monoid generated by $V \cup \overline{V}$. Similarly, $\overline{W}, \overline{X}, \overline{Y}$, and $\overline{Z}$ denote strings over the alphabet $\overline{V}$ in $\mathbf{H}(V)$, where $\overline{W}$ is obtained from W by reversing the sequence of symbols in W and barring the individual symbols, and $\overline{w}, \overline{x}, \overline{y}$, and $\overline{z}$ denote the same strings considered as elements of **V**.

LEMMA 5.1. *Let $\tau' \subset \tau$ be the set of elements ct such that $\mathbf{h}(t) = X\overline{Y}$ for some $X, Y \in V_N{}^*$. There is a bijection α from τ' to $\mathbf{D}$ such that if $ct \in \tau_w$ and $\mathbf{h}(t) = X\overline{Y}$ then $\mathbf{d} = \alpha(ct)$ is a derivation of wy from x and $\omega(\mathbf{d}) = c$.*

Proof. The proof is by induction on the length of w. We note that since elements of $\overline{V}$ are left inverses of elements of V and not right inverses in the free half-group, and since every element of τ_a has an element of V on the left, it follows that every element ct of τ_w is such that $t \in V_N(V_N \cup \overline{V}_N)^*$ and therefore $h(t) \in V_N(V_N \cup \overline{V}_N)^*$.

We first establish the assertion in the case that $| w | = 1$. Then $cA_i \in \tau_a$ if and only if $P(A_i, a) = c$. Consequently we set $\alpha(cA_i) = A_i, ((A_i, a), 1)$.

Similarly if $cA_i\bar{A}_j \in \tau_a$ then $\alpha(cA_i\bar{A}_j) = A_i, ((A_i, aA_j), 1)$. If $cA_i\overline{A_jA_k} \in \tau_a$, then $\alpha(cA_i\overline{A_jA_k}) = A_i, ((A_i, aA_jA_k), 1)$. Note that in each case the weight of the derivation is equal to the coefficient of the element of τ_a.

Suppose the assertion is established for all strings of $V_T{}^+$ of length less than some $m \geqslant 2$. Let $| w | = m - 1$ and consider the string wa. If ct is an element of τ_{wa} then $ct = c_1t_1 \cdot c_2t_2$ where $c_1t_1 \in \tau_w$ and $c_2t_2 \in \tau_a$. Furthermore, if $\mathbf{h}(t)$ is of the general form $W\overline{Z}$ then one of the following cases must hold:

Case	$\mathbf{h}(t_1)$	$\mathbf{h}(t_2)$
1	X	A_i
2	X	$A_i\bar{A}_j$
3	X	$A_i\overline{A_jA_k}$
4	$X\overline{A_iY} = X\overline{Y}\bar{A}_i$	A_i
5	$X\overline{A_iY} = X\overline{Y}\bar{A}_i$	$A_i\bar{A}_j$
6	$X\overline{A_iY} = X\overline{Y}\bar{A}_i$	$A_i\overline{A_jA_k}$.

Consider the first case. By the induction hypothesis if c_1t_1 is an element of τ_w and $\mathbf{h}(t_1) = X$, then if follows that for some derivation $\mathbf{d} \in \mathbf{D}$, $\alpha(c_1t_1) = \mathbf{d}$, $\omega(\mathbf{d}) = c_1$, and $x\mathbf{d}(0) = w$, that is, $\mathbf{d}$ is a derivation of w from x. Then we set $\alpha(ct)$ equal to the derivation $xA_i, \mathbf{d}(0), ((A_i, a), | w | + 1)$ and we note that the assertion is satisfied. The second and third cases are analogous. For example, in the third case $\alpha(ct)$ is set equal to the derivation $xA_i, \mathbf{d}(0), ((A_i, aA_jA_k), | w | + 1)$, where $\mathbf{d}$ denotes the same derivation used for the first case.

Consider the fourth case. By the induction hypothesis, since c_1t_1 is an element of τ_w and $\mathbf{h}(t_1) = X\overline{A_iY}$, it follows that $\alpha(c_1t_1) = \mathbf{d}$, where $\mathbf{d}$ is a leftmost derivation of wA_iy from x: $x\mathbf{d}(0) = wA_iy$. Then we set $\alpha(ct) = x, \mathbf{d}(0), ((A_i, a), | w | + 1)$,

which is a derivation of the string way from x. Furthermore, the weight of $\alpha(ct)$ is equal to c and therefore the derivation satisfies the assertion.

The fifth and sixth cases are similar, and we treat only the sixth. By the induction hypothesis, since c_1t_1 is an element of τ_w and $\mathbf{h}(t_1) = X\overline{A_i}\overline{Y}$, it follows that $\alpha(c_1t_1) = \mathbf{d}$, where $\mathbf{d}$ is a derivation of wA_iy from x and $\omega(\mathbf{d}) = c_1$. Note that $c_2 = P(A_i, aA_jA_k)$. Set $\alpha(ct) = x$, $\mathbf{d}(0)$, $((A_1, aA_jA_k), |w| + 1)$, which is a derivation of waA_jA_kY from x of weight $c_1 \cdot c_2 = c$. This derivation satisfies the assertion since $\mathbf{h}(t) = \mathbf{h}(t_1)\,\mathbf{h}(t_2) = X\overline{Y}\overline{A_i}A_i\overline{A_j}\overline{A_k} = XA_jA_kY$.

To complete the proof we need to show that α is bijective. It is straightforward to show by induction on the length of w that α is injective.

In order to show that α is surjective we proceed by induction on the length of the derivation. If $\mathbf{d} \in \mathbf{D}$ is of length 1, then $\mathbf{d}$ must consist of a single production applied to a string of length one. If the terminal symbol produced by the production is $a \in V_T$, then the derivation will be the image of some element of τ_a.

Suppose every derivation $\mathbf{d} \in \mathbf{D}$ of length no greater than n is the image of some monomial $ct \in \tau'$. Consider a derivation $\mathbf{d} \in \mathbf{D}$ of length $n+1$, $\mathbf{d} = x, \mathbf{r}_1(i_1), \mathbf{r}_2(i_2),\ldots, \mathbf{r}_n(i_n), \mathbf{r}_{n+1}(i_{n+1})$. Since $\mathbf{d}$ is a canonical derivation, the derivation

$$\mathbf{d}_1 = x, \mathbf{r}_1(i_1),\ldots, \mathbf{r}_n(i_n)$$

is also a canonical derivation with some weight $\omega(\mathbf{d}_1) = c_1$, and there are $u \in V_T^+$, $A_i \in V_N$, and $z \in V_N^*$ such that $x\mathbf{r}_1(i_1) \cdots \mathbf{r}_n(i_n) = uA_iz$. We must treat treat two cases.

Case 1. Suppose $\mathbf{d}_1 \in \mathbf{D}$. Then by the induction hypothesis there is some term $c_1t_1 \in \tau_u$, such that $\alpha(c_1t_1) = \mathbf{d}_1$ and $\mathbf{h}(t_1) = X\overline{A_iZ}$. Suppose $\mathbf{r}_{n+1} = (A_i, ay)$ and $P(A_i, ay) = c_2$. Then $\mathbf{d}$ is a derivation of $uayz$ and $\omega(\mathbf{d}) = c_1 \cdot c_2$. Since $c_2\underline{A_i}\bar{y} \in \tau_a$, $c_1 \cdot c_2t_1A_i\bar{y} \in \tau_{ua}$, and $\mathbf{h}(t_1A_i\bar{y}) = X\overline{A_iZ}A_i\overline{Y} = X\overline{Z}\overline{A_i}A_i\overline{Y} = X\overline{Z}\overline{Y} = XYZ$. By the construction of α, $\alpha(c_1 \cdot c_2t_1A_i\bar{y}) = \mathbf{d}$.

Case 2. Suppose $\mathbf{d}_1 \notin \mathbf{D}$. Then $x = wA_i$, where A_i is not rewritten by $\mathbf{d}_1$, and $z = A$. But A_i is rewritten by $\mathbf{d}$, hence all of w must be rewritten by $\mathbf{d}_1$ and there must be a derivation of u from w in $\mathbf{D}$; in fact, $w\mathbf{d}_1(0) = u$. By the induction hypothesis, there is some term $c_1t_1 \in \tau_u$ such that $\alpha(c_1t_1) = \mathbf{d}_1$, $c_1 = \omega(\mathbf{d}_1)$ and $\mathbf{h}(t_1) = W$. Suppose $\mathbf{r}_{n+1} = (A_i, ay)$ and $P(A_i, ay) = c_2$. Then $\mathbf{d}$ is a derivation of the string way from xA_i and $\omega(\mathbf{d}) = c_1 \cdot c_2$. Furthermore, τ_{ua} contains the product of c_1t_1 and $c_2A_i\bar{y}$. Then $\mathbf{h}(t_1A_i\bar{y}) = WA_i\overline{Y}$, and by the construction of α, $\alpha(c_1 \cdot c_2t_1A_i\bar{y}) = \mathbf{d}$. Thus we have shown that if $\mathbf{d} \in \mathbf{D}$, then $\mathbf{d}$ is contained in the image of α for some argument from τ'.

On the basis of Lemma 5.1 we can now prove a form of Shamir's homomorphism theorem.

THEOREM 5.1. *Let* $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ *be a* wcfg *in Greibach normal form over the semigroup* $\mathbf{S} = \langle S, +, \cdot, 0, 1 \rangle$. *If* $\mathbf{g}$ *is the canonical homomorphism from* $R_{\text{pol}}(\mathbf{V})$ *to* $R_{\text{pol}}(\mathbf{H})$, *define the homomorphism* $\mathbf{f}$ *as follows*:

$$\mathbf{f} : V_T^+ \to R_{\text{pol}}(\mathbf{H}),$$

$$\mathbf{f}(a) = \sum_{T \in \tau_a} \mathbf{g}(T) \qquad a \in V_T,$$

$$\mathbf{f}(aw) = \mathbf{f}(a)\,\mathbf{f}(w) \qquad a \in V_T,\ w \in V_T^+.$$

Then for any $x \in V_T^+$, *the coefficient of* A_1 *in the polynomial* $\mathbf{f}(x) \in R_{\text{pol}}(\mathbf{H})$ *is equal to* $\omega(x)$.

Proof. From Lemma 5.1 we know that there is a bijection between derivations $\mathbf{d}$ of x from A_1 such that $\omega(\mathbf{d}) = c$ and terms $T \in \tau_x$ such that $\mathbf{g}(T) = cA_1$. Clearly $\mathbf{f}(x) = \sum_{T \in \tau_x} \mathbf{g}(T)$. It follows that the coefficient of A_1 in $\mathbf{f}(x)$ will be a sum of a set of elements of $\mathbf{S}$ such that each summand is the weight of a distinct canonical derivation of x.

Remark 5.1. We have used what Greibach calls "standard 2-form" for our definition of Greibach normal form. In fact, if we permit productions of the form

$$(A_i, aA_jA_k \cdots A_m)$$

there is no essential change to Lemma 5.1, Theorem 5.1, or their proofs. The set τ_a would be defined so that it would include the term

$$P(A_i, aA_jA_k \cdots A_m)\, A_i\overline{A_jA_k \cdots A_m}$$

for each production of the above form.

The reader will note (the author wishes to thank the referee for suggesting the substance of the following remarks) that the construction used in the proof of Lemma 5.1 is essentially the same as specification of the single-state push-down automaton which recognizes a language by empty store on the basis of a grammar in Greibach normal form (see Hopcroft and Ullman [1969], p. 75). Any sequence of transitions which results in the recognition of a word by this single-state push-down automaton corresponds to a specific product of monomials which cancels to the element A_1 in the half-group $\mathbf{H}(V)$. Furthermore, the product of the first k monomials will be equal to the symbol A_1 followed by a string corresponding to the reversed contents of the push-down store of the automaton after reading the first k elements of the input string. Thus, there is a natural correspondence between the contents of the push-down store at any point in its recognition procedure and the product of the initial factors of the monomial product.

We have used the Greibach normal form in our proof of the homomorphism theorem and as a consequence the construction of the map **g** is considerably simpler than in the original proof of Shamir [1967]. Shamir's proof assumed the grammar to be in Chomsky normal form; however, in that proof as in ours, the elements of τ_a for $a \in V_T^*$ were of the form $B\bar{Q}$, where $B \in W$ and $\bar{Q} \in \overline{W}^*$ are elements of a half-group $\mathbf{H}(W)$. Furthermore, the degree of ambiguity was preserved by the homomorphism in Shamir's theorem as in ours. Given the existence of a homomorphism which preserves ambiguity and such that the elements of τ_a are of the form $B\bar{Q}$, we can conclude the existence of an ambiguity-preserving Greibach normal form. The construction of the grammar $\mathbf{G}'$ in Greibach normal form is straightforward: if $B\bar{Q} \in \tau_a$ for some $a \in V_T$, then the productions of $\mathbf{G}'$ will include $B \to aQ$.

In summary, we can assert that the following three assertions are equivalent statements about the same characteristic of context-free grammars (let $\mathbf{G} = \langle V, V_T, P, A_1 \rangle$ and $x \in V_T^+$):

(1) For every context-free grammar **G** there exists a grammar $\mathbf{G}'$ over V_T in Greibach normal form which preserves the ambiguity of every string x.

(2) For every context-free grammar **G** there is a homomorphism **g** from V_T^+ to the semiring of polynomials with integer coefficients over some half-group $\mathbf{H}(W)$ such that

(a) every element of τ_a for $a \in V_T$ is a monomial of the form $B\bar{Q}$ where $B \in W$ and $\bar{Q} \in \overline{W}^*$.

(b) the coefficient of some distinguished $A \in W$ in $\mathbf{g}(x)$ is the degree of ambiguity of x in **G**.

(3) For every context-free grammar **G** there exists a single-state push-down automaton without Λ-transitions which accepts $L(\mathbf{G})$ by empty store, and such that the number of distinct machine traces which result in x being accepted is equal to the ambiguity of x in **G**.

This paper has established that (1) implies (2); our remarks above indicate that (2) implies (1). The proof that (1) implies (3) may be found in Hopcroft and Ullman ([1969], p. 75), and the proof that (3) implies (1) may easily be established by reversing the methods in the proof that (1) implies (3).

EXAMPLE. To illustrate the import of the theorem, consider the wcfg $= \langle \{A, B, a, b\}, \{a, b\}, P, A \rangle$ where

$$\begin{aligned} P(A, aAB) &= 3, \\ P(A, aA) &= 2, \\ P(A, a) &= 1, \\ P(B, bB) &= 5, \\ P(B, b) &= 1, \end{aligned}$$

and the semigroup is taken to be the real numbers under ordinary multiplication and division. Constructing the map $\mathbf{f}$ defined in Theorem 5.1, we have

$$\mathbf{f}(a) = 3A\bar{B}\bar{A} + 2A\bar{A} + A,$$
$$\mathbf{f}(b) = 5B\bar{B} + B.$$

Consider the word $aaabb \in V_T^*$. The value of $\omega(aaabb)$ is the coefficient of A in $\mathbf{f}(aaabb)$. By definition,

$$\mathbf{f}(aaabb) = (3A\bar{B}\bar{A} + 2A\bar{A} + A)(3A\bar{B}\bar{A} + 2A\bar{A} + A) \times (3A\bar{B}\bar{A} + 2A\bar{A} + A)(5B\bar{B} + B)(5B\bar{B} + B).$$

There will be 108 terms in the product before cancellation rules are applied, a fact which makes it clear that in some sense the homomorphism theorem simply gives us a way of burying the combinatorics in the algebra. However, we can take advantage of the fact that inverses are only one-sided to rule out many of the 108 terms. We do this by noting if two unbarred symbols occur in a partial product after application of the cancellation rules, no further multiplicative factors on the right can result in a product of cA. By applying this fact to $\mathbf{f}(aaabb)$ we can show that it will suffice to consider the product

$$(3A\bar{B}\bar{A} + 2A\bar{A})(3A\bar{B}\bar{A} + 2A\bar{A} + A)(A)(5B\bar{B} + B)(B)$$

which only has twelve terms. From this product we can conclude that the products which will cancel to A will be

$$3A\bar{B}\bar{A} \cdot 3A\bar{B}\bar{A} \cdot A \cdot B \cdot B = 9A,$$
$$3A\bar{B}\bar{A} \cdot 2A\bar{A} \cdot A \cdot 5B\bar{B} \cdot B = 30A,$$
$$2A\bar{A} \cdot 3A\bar{B}\bar{A} \cdot A \cdot 5B\bar{B} \cdot B = 30A.$$

Each of these products correspond to a distinct canonical derivation of *aaabb*, and the productions used in each derivation may be found by simply inspecting the sequence of factors in each product.

6. Semirings

In addition to probabilistic automata, several developments have been described in the literature in which weights are associated in a variety of ways with either automaton transitions or the productions of formal grammars (Salomaa [1969], Santos [1968, 1969], Lee and Zadeh [1969]). Although the operations used in these systems differ from one to another, they have all used the real numbers as the set of coefficients and the

operations have all been chosen so that the structure forms a commutative semiring in which the additive identity is a multiplicative zero. It would seem that much of the work may be fit into a uniform treatment using the general notion of semirings.

Recall that we require the following characteristics of our operations (which we will denote $+$ and $\cdot$):

(1) $+$ and $\cdot$ are associative and commutative binary operations.

(2) $\cdot$ distributes over $+$.

(3) $+$ and $\cdot$ both have identity elements (denoted 0 and 1 respectively).

(4) The identity for $+$ is a zero for $\cdot$.

In this section we will give several interesting examples of semiring structures over the real numbers. These structures will be specified by a pair of operations and a subset of the reals. For each pair of operations, the order of presentation will be $(+, \cdot)$. With the semiring descriptions we will include some comments about the effects of using them with weighted grammars.

(1) (max, min) The weight of each derivation will be the weight of the production of smallest weight used in the derivation. The weight of a word will be the weight of the derivation of that word with largest weight. Note that the set of semiring elements can be defined to be any subset of the reals which contains a least upper bound and a greatest lower bound. Since the production function has finite support, derivations will be limited to a finite set of values, and consequently the weight of a word is always defined. The identity for max will be the greatest lower bound and strictly less than any of the "nonzero" production weights; the identity for min is the least upper bound.

(2) (min, max) The weight of each derivation will be the weight of the production of greatest weight used in the derivation. The weight of a word is the weight of the derivation of that word which has the smallest weight. The semiring elements can be defined as above. The identity for max and min are the same as in (1).

(3) (max, $\cdot$) where $\cdot$ denotes ordinary multiplication over the real numbers. The identity for max is 0; the identity for $\cdot$ is 1. The set of semiring elements must include 0 and 1 and be closed under multiplication, e.g. [0, 1] or the nonnegative integers. The weight of a derivation is the product of the weights of the productions used in the derivation; the weight of a word is the maximum over all weights of derivations of that word.

(4) (min, $+$) where the set of semiring elements is contained in some closed interval $[0, K]$ and $+$ is defined over the elements as follows:

$$x + y = \min(x + y, K).$$

The identity for min is K and the identity for $+$ is 0. The value K is an upper bound for the weights of the productions. The weight of a derivation is the sum of the weights

of its productions up to the maximum value of K; the weight of a word is the minimum over the weights of all derivations of a word.

(5) (min, $\odot$) where the set of semiring elements is contained in a closed interval $[1, K]$ such that $K \geqslant 1$ and the operation $\odot$ is defined for all x, y in the interval as follows:

$$x \odot y = \min(x \cdot y, K) \text{ (where } \cdot \text{ denotes ordinary multiplication).}$$

The identity for $\odot$ is 1 and K is the identity for min.

(6) (max, $\odot$) where the set of semiring elements is contained in an interval $[0, K]$ such that $K \geqslant 1$, and the operation $\odot$ is defined as in the preceding example. The identity for max is 0 and for $\odot$ is 1.

It would be convenient (for such things as flow problems) if such pairs as (+, max) or (+, min) provided a semiring structure. However, the distributive law fails for these pairs.

As can be seen by the above structures, the semiring of coefficients is not restricted to one which will count derivations, but in fact may be chosen to perform any of a variety of calculations.

References

Bar-Hillel, Y., C. Gaifman, and E. Shamir, On categorial and phrase-structure grammars, *Bull. Res. Counc. Isr., Sect. F* **9** (1960), 1–16.

Book, R., S. Even, S. Greibach, and G. Ott, Ambiguity in graphs and expressions, *IEEE Trans. Computers* **c-20** (1971), 149–153.

Chomsky, N., and M. Schutzenberger, The algebraic theory of context-free languages, *in* "Computer Programming and Formal Systems" (P. Braffort and D. Hirschbert, Eds.), North Holland, Amsterdam, 1963.

Greibach, S., A new normal theorem for context-free phrase-structure grammars, *J. ACM* **12** (1965), 42–52.

Griffiths, T. V., Some remarks on derivations in general rewriting systems, *Information and Control* **12** (1968), 27–54.

Hopcroft, J. E., and J. D. Ullman, "Formal Languages and Their Relations to Automata," Addison-Wesley, Reading, Massachusetts, 1969.

Lee, E. T., and L. A. Zadeh, Note on fuzzy languages, *Inform. Sci. (New York)* **1** (1969), 421–434.

Salomaa, Arto, Probabilistic and weighted grammars, *Information and Control* **15** (1969), 529–544.

Santos, Eugene S., Maximum automata, *Information and Control* **13** (1968), 363–377.

Santos, Eugene S., Maximin sequential-like machines and chains, *Math. Systems Theory* **3** (1969), 300–309.

Shamir, Eliahu, A representation theorem for algebraic and context-free power series in noncommuting variables, *Information and Control* **11** (1967), 239–254.