JOURNAL OF COMPUTER AND SYSTEM SCIENCES 33, 423–455 (1986)

A Representation Theorem of Infinite Dimensional Algebras and Applications to Language Theory*

GÜNTER HOTZ

Fachbereich Angewandte Mathematik und Informatik, Universität des Saarlandes, 6600 Saarbrücken 11, Germany

Received February 27, 1984; revised August 7, 1985

The algebraic theory we present here continues the earlier work of several authors. The leading idea is to develop a machine and production free language theory. The interest in such a theory is supported by the hope that the proofs in such a theory need fewer case discussions, which often lead to errors, and that a view which is free from nonessentials of language theory will lead to a progress in the direction of our problems. Even if the theory is in an early stage, the attempt pays out in a machine free definition of $LL(k)$ and $LR(k)$ languages, which leads easily to generalizations of non-deterministic $LL(k)$ and $LR(k)$ languages with the same space and time complexity behaviour. Furthermore, we are able to show that this theory is not restricted to the context-free languages but also applies to the whole Chomsky hierarchy. Our theory is in a sense dual to the theory of formal power series as introduced by M. Schützenberger.

MOTIVATION

Quite a few main problems of language theory are still open. Such problems include the problem of the different complexities of the word problem of the class of context-free languages, the equivalence problem of deterministic c.f. languages, the search for more and stronger invariants of grammars under language preserving transformations, and questions concerning natural generalizations of known fundamental results on context-free languages.

Formal language theory cannot be looked upon as a new discipline any more and progress in the direction of the open problems has become rare. Therefore one may assume that the intuition arising from the combinatorial language theory and other theories which are not very new any more is no longer stimulating new ideas for research.

The aim of this paper is to embed language theory rigorously into an algebraic framework stemming from the representation theory of finite dimensional algebras. I expect that from this point of view additional intuition can be gained. One can

* This research has been supported by the DFG contract Ho 251/10-1.

0022-0000/86 $3.00

not expect that a single theory offers a "Königsweg" to the solution of any problem, but combining insight from different theories may yield the progress we are looking for.

Our theory is in a certain sense dual to the theory of formal power series, as the reader will see. We transform context-free grammars in such a way that these grammars can be looked upon as the multiplication rule of an infinite associative algebra over a semiring R. We find a representation of this algebra $\mathscr{A}$ in the ring $R\langle Z^{(*)}\rangle$, the ring of polynomials with coefficients from R and monomials from the polycyclic monoid $Z^{(*)}$ generated by certain alphabet Z.

This representation can easily be proved to be correct. It plays the crucial role in the whole paper. But even more important that the possibility to gain all these results within such little space seems to me the fact that standard questions of representation theory of algebras become applicable to language theory. The representation yields an easy access to the understanding of the $LL(k)$ and $LR(k)$ languages and leads to natural generalizations of these classes. As I have shown in [11], these ideas allow to generalize the theorem of Greibach on hardest c.f. languages under homomorphic reductions for r.e. sets, for c.s. languages, and for the intersection closure of s.f. languages.

Fundamental Definitions

Let X be a set and X^* be the free monoid generated by X. The empty word is $1 \in X^*$ and $|u|$ means the length of $u \in X^*$. For monoids M and semirings R the semiring of the finite sums

$$p = \sum_{m \in M} \alpha_m \cdot m$$

where $\alpha_m \in R$ is $R\langle M\rangle$.

Often we write $\alpha_m = \langle p, m\rangle$; $\langle p, m\rangle \neq 0$ holds only for finitely many elements $m \in M$. We always assume that R has a multiplicative unit, which we identify with $1 \in M$.

The syntactic monoid $X^{(*)}$ of the Dyck language $D(X)$ over X is of special importance for our theory. This monoid, called polycyclic monoid by Perrot [16], can be defined as follows.

Take a set X', which is bijectively equivalent to X via the bijection $x \to \bar{x}$. We require that $X \cap X' = \varnothing$ and fix a symbol $0 \notin X \cup X'$. Then we form the quotient of the free monoid $(X \cup X' \cup \{0\})^*$ by the relation system

$$x \cdot \bar{x} = 1, \qquad x \cdot \bar{y} = 0, \qquad 0 \cdot z = z \cdot 0 = 0 \qquad \text{for} \quad x, y \in X, z \in X \cup X' \cup \{0\}.$$

Sometimes we write x^{-1} (resp x^1) instead of $\bar{x}$ (resp. x).

Furthermore we use context-free grammars $G=(X, T, P, S)$ with $X \cap T=\varnothing$, $P \subset X \times X^2 U X \times T$, and $S \in X$. Consequently we have no ε-productions and $1 \notin L(G)$, $L(G)$ being the language generated by G. Moreover we assume G to be free from superfluous variables. This means, that for $x \in X$ there exist derivations f and g such that

$$S \xrightarrow{f} uxv \xrightarrow{g} w \qquad \text{and} \qquad w \in T^*.$$

Finally we require that S does not appear on the right-hand side of any production $q \in P$.

It is usual to write P also as an equation system

$$x=\sum \alpha_{x,y} \cdot u \qquad \text{for} \quad x \in X$$

where $\alpha_{x,u}=1$ if $(x, u) \in P$ and $\alpha_{x,u}=0$ in all other cases.

Schützenberger [2] has shown, that this makes sense in the following way: The equation system can be solved by a system of formal power series. $L(G)$ can be looked upon as the support of the power series belonging to S. The coefficient of the word w in the series gives the multiplicity of w relative to G, that is to say, the number of essentially different derivations of w from S.

We assign an equation system to the grammar in a dual way, by writing the quadratic terms on the left side and the corresponding linear terms as sums on the right side. To be concrete, we study equation systems of the form

$$x \cdot y=\sum_{z \in X} \alpha^z_{x,y} \cdot z, \qquad t=\sum_{z \in X} \alpha^z_t \cdot z \qquad (x, y \in X; t \in T)$$

with $\alpha^z_{x,y}$, $\alpha^z_t \in \{0, 1\}$, and

$$\alpha^z_{x,y}=1 \Leftrightarrow (z, xy) \in P,$$
$$\alpha^z_t=1 \Leftrightarrow (z, t) \in P.$$

These relations are similar to the multiplication rules of finite dimensional algebras over a ring R. In general such an equation system does not define an associative algebra. But with a simple trick we get an associative algebra from this idea.

We assign to G a new alphabet $\bar{X}$ by setting

$$X_l=\{(x, l) \mid (z, xy) \in P\},$$
$$X_r=\{(y, r) \mid (z, xy) \in P\},$$
$$\bar{X}=X_l \cup X_r.$$

For (x, l) (resp. (y, r)) we often write shorter x_l (resp. y_r). Now we define the grammar $\bar{G} = (\bar{X}, T, \bar{P}, S_r)$ by

$$\begin{aligned}\bar{P} = &\{(x_i, y_l z_r) \mid (x, yz) \in P,\ x \neq S,\ i \in \{1, r\}\} \\ &\cup \{(S_r, x_l z_r) \mid (S, xz) \in P\} \\ &\cup \{(x_i, t) \mid i \in \{l, r\},\ (x, t) \in P\}.\end{aligned}$$

Obviously $L(G) = L(\bar{G})$. Now we assign to G the following equation system:

$$x \cdot y = \sum \alpha^z_{x,y} \cdot z \qquad \text{for} \quad x \in X_1,\ y \in Y_r \tag{$\mathscr{R}_G$}$$

where

$$\alpha^z_{x,y} = \begin{cases} 1 & \text{of} \quad (z, xy) \in \bar{P} \\ 1 & \text{if} \quad z = S_r,\ x = S_l,\ y = S_r \\ 0 & \text{in all other cases.} \end{cases}$$

Because our proofs will not become harder, we generalize ($\mathscr{R}_G$), to the following situation: X_l and X_r are any two alphabets with $X_l \cap X_r = \varnothing$. We put $\bar{X} = X_l \cup X_r$. Moreover there are two mappings

$$\delta' : X_l \times X_r \to R\langle \bar{X}^* \rangle$$

and

$$\eta' : T \to R\langle \bar{X}^* \rangle$$

with

$$\delta'(x, y) = \sum_{z \in \bar{X}} \alpha^z_{x,y} \cdot z \qquad \text{for} \quad (x, y) \in X_l \times X_r,$$

$$\eta'(t) = \sum_{z \in X} \alpha^z_t \cdot z \qquad \text{for} \quad t \in T.$$

We extend δ' to $\bar{X}^*$ by defining

$$\delta(u) = \begin{cases} u & \text{for} \quad u \in X_r^* \cdot X_l^* \\ \delta'(x, y) & \text{for} \quad u = xy \in X_l \cdot x_r \\ u_1 \delta(xy)\, u_2 & \text{for} \quad u_1 \in X_r^* \cdot X_l^* \quad \text{and} \quad xy \in X_l \cdot X_r, \end{cases}$$

where $u = u_1 xy u_2$.

Now we extend δ linearly to $R\langle \bar{X}^* \rangle$; η is the corresponding extension of η' to $R\langle T^* \rangle$. The equation system

$$xy = \delta(xy) \qquad \text{for} \quad xy \in X_l \cdot X_r \tag{$\mathscr{R}$}$$

is the generalization of the system $(\mathscr{R}_G)$.

We now assign an associative algebra $\mathscr{A}_R(\delta)$ to $(\mathscr{R})$. For this purpose we iterate δ and finally form the transitive closure δ^* of δ. This means that $\delta \circ \delta^* = \delta^*$. Now one easily proves

LEMMA 1. $\delta^*(uv) = \delta^*(\delta^*(u) \cdot \delta^*(v))$.

Proof. The argument is an induction on the length $|uv|$ of uv. If $|uv| \leqslant 2$ there is nothing to prove. The lemma obviously holds also for $uv \in X_r^* \cdot X_l^*$. Suppose $uv \notin X_r^* \cdot X_l^*$. This means that there exists a decomposition

$$uv = w_1 xy w_2 \qquad \text{such that } w_1 \in X_r^* \cdot X_l^* \text{ and } xy \in X_l \cdot X_r.$$

Then we have

$$\delta(uv) = w_1 \left(\sum \alpha_{xy}^z \cdot z\right) w_2.$$

Because each of the words of that decomposition has a length $< n$, we are allowed to apply the induction hypothesis.

We discuss two cases:

Case 1. xy is totally part of u or part of v. We assume the first situation: $u = u_1 xy u_2$. Then we have

$$\delta(uv) = u_1 \circ \left(\sum \alpha_{xy}^z \cdot z\right) \circ u_2 \cdot v.$$

By induction we conclude

$$\delta^*\left(u_1 \left(\sum \alpha_{x,y}^z \cdot z\right) u_2 v\right) = \delta^*\left(\delta^*\left(u_1 \left(\sum \alpha_{x,y}^z \cdot z\right) u_2\right) \cdot \delta^*(v)\right)$$

$$= \delta^*(\delta^*(u) \cdot \delta^*(v)).$$

Therefore our lemma holds in this case.

Case 2. $u = u_1 x$, $v = y v_1$, and $u_1 x \in X_r^* \cdot X_l^*$, $x \in X_l^*$, $y \in X_r$. Then we have

$$\delta(uv) = u_1 \left(\sum_{\bar{X}} \alpha_{x,y}^z \cdot z\right) v_1$$

$$= \left(u_1 \left(\sum_{X_r} \alpha_{x,y}^z \cdot z\right) v_1 + u_1 \left(\left(\sum_{X_l} \alpha_{x,y}^z \circ z\right) v_1\right)\right.$$

We apply the induction hypothesis to this expression indicated by the parentheses:

$$\begin{aligned}
\delta^*(uv) &= \delta^*\left(\delta^*\left(u_1 \sum_{X_l} \alpha^z_{x,y} \cdot z\right) \cdot \delta^*(v_1)\right) \\
&\quad + \delta^*\left(\delta^*(u_1) \cdot \delta^*\left(\left(\sum_{X_r} \alpha^z_{x,y} \cdot z\right) \cdot v_1\right)\right) \\
&= \delta^*\left(\delta^*(u_1)\left(\sum_{X_l} \alpha^z_{x,y} \cdot z\right) \cdot \delta^*(v_1)\right) \\
&\quad + \delta^*\left(\delta^*(u_1)\left(\sum_{X_r} \alpha^z_{x,y} \cdot z\right) \circ \delta^*(v_1)\right) \\
&= \delta^*(\delta^*(u_1) \cdot xy \cdot \delta^*(v_1) \\
&= \delta^*(\delta^*(u_1 x) \cdot \delta^*(yv_1)).
\end{aligned}$$

The last relation holds because of

$$\delta^*(u_1 x) = \delta^*(u)\, x \qquad \text{and} \qquad \delta^*(yv_1) = y\delta^*(v_1).$$

This proves Case 2 and our Lemma 1 has been proved.

Now we define the operation "$\circ$" on $R\langle \bar{X}^* \rangle$ by setting

$$u \circ v := \delta^*(uv).$$

From this it follows

$$\begin{aligned}
(u \circ v) \circ w &= \delta^*(\delta^*(uv) \cdot w) = \delta^*(\delta^*(uv)\, \delta^*(w)) = \delta^*(uvw), \\
u \circ (v \circ w) &= \delta^*(u \cdot \delta^*(vw)) = \delta^*(\delta^*(u)\, \delta^*(vw)) = \delta^*(uvw).
\end{aligned}$$

Therefore the following theorem holds:

THEOREM 1. *$\mathscr{A}_R(\delta) := (R\langle \bar{X}^* \rangle, +, \circ)$ is an associative algebra and*

$$\delta^*: (R\langle \bar{X}^* \rangle, +, \cdot) \to (R\langle \bar{X}^* \rangle, +, \circ)$$

is an algebra homomorphism.

If δ comes from the grammar G we write $\mathscr{A}_R(G)$ too. We extend this algebra to include the terminals. For this purpose we use the defined mapping η and extend η to $(\bar{X} \cup T)^*$ by setting $\eta(x) = x$ for $x \in \bar{X}$. Now for $u, v \in (\bar{X} \cup T)^*$ we define

$$u \circ v = \delta^*(\eta(uv)).$$

The associative algebra that we get by this construction is called $\bar{\mathscr{A}}_R(G)$.

For $u_1 \circ u_2 \circ \cdots \circ u_n$ we write again $u_1 u_2 \cdots u_n$. In this case it is not clear which product we mean. We write

$$u_1 u_2 \cdots u_n [\mathscr{A}_R(G)]$$

if the product is in $\mathscr{A}_R(G)$. Analogously we proceed with other algebras.

The following concerns the questions:

How are the algebras $\mathscr{A}_R(G)$ structured?

What information does $\mathscr{A}_R(G)$ contain about $L(G)$?

What is the structure of $\mathscr{A}_R(G)$ if G is determistic?

The following section is dedicated to the first question.

A Representation Theorem for $\mathscr{A}_R(\delta)$

We are going to show that for each algebra $\mathscr{A}_R(\delta)$ there exists a nontrivial representation $\varphi: \mathscr{A}_R(\delta) \to R\langle X^{(*)}\rangle$. We will show that the algebra $R\langle X^{(*)}\rangle$ plays a similar role for our algebras and for the finite dimensional algebras as does the matrix ring for the finite dimensional case. It is clear that $R\langle X^{(*)}\rangle$ is a special case of our algebras $\mathscr{A}_R(\delta)$. The following lemma shows that $R\langle X^{(*)}\rangle$ has a very simple algebraic structure.

Lemma 2. *$\mathscr{A}_D = R\langle X^{(*)}\rangle$ contains only trivial two sided ideals if R = boolean ring of two elements or* card $X = \infty$. *Ideals $\mathfrak{A}$ of $\mathscr{A}_D$ here are considered to be trivial, if there exists an ideal $\mathfrak{A}'$ of R such that $\mathfrak{A} = \mathfrak{A}'\langle X^{(*)}\rangle$.*

Proof of Lemma 2. Let $\mathfrak{A} \subset \mathscr{A}_D$ be a two-sided ideal, that means that $\mathscr{A}_D \mathfrak{A} \mathscr{A}_D \subset \mathfrak{A}$ holds. We study several cases:

(1) Let be $\alpha \in R$ and $\alpha \cdot \bar{u}v$ with $u, v \in X^*$ in $\mathfrak{A}$. Then it follows $\alpha \in \mathfrak{A}$.

(2) $p = \alpha \bar{u} v + q \in \mathfrak{A} \Rightarrow p' = \alpha + q' \in \mathfrak{A}$. $q' = uqv$.

(3) $p = \alpha + \beta \bar{u} v + q \in \mathfrak{A}$.

(a) $u\bar{v} = 0 \Rightarrow up\bar{v} = \beta + q'$. $up\bar{v}$ has one summand less then p.

(b) $u\bar{v} \neq 0$. We may assume $u\bar{v} = u' \in X^*$, $u' \neq 1$. We have

$$up\bar{v} = \alpha u' + \beta + uq\bar{v}.$$

Choose $y \in X$, $y \neq$ last letter of u':

$$up\bar{v}\bar{y} = \beta\bar{y} + uq\bar{v}\bar{y}; \qquad \text{this means one summand less.}$$

(4) From (1), (2), and (3) it follows:

$$\langle p, u\rangle = \alpha, \qquad p \in \mathfrak{A} \Rightarrow \alpha \in \mathfrak{A}.$$

Let be $\mathfrak{A}' = \mathfrak{A} \cap R$, then therefore $\mathfrak{A} = \mathfrak{A}'\langle X^{(*)}\rangle$ holds as we have claimed.

We now show that each finite dimensional algebra $\mathscr{A}$ over R has a non-trivial representation in $\mathscr{A}_D$. Let Z be a finite basis of $\mathscr{A}$ over R and $\mathscr{A}$ be given by the relations

$$x \cdot y = \sum_{z \in Z} \alpha_{x,y}^{z} \cdot z, \qquad \alpha_{x,y}^{z} \in R.$$

We define $\varphi: \mathscr{A} \to \mathscr{A}_D$ by

$$\varphi(y) := \sum_{z,u \in Z} \bar{z} \alpha_{z,y}^{u} \circ u \qquad \text{for} \quad y \in Z.$$

This defines φ uniquely ($\bar{z}$ is the inverse of z in $Z^{(*)}$).

THEOREM 2. *φ is an algebra homomorphism. If $\mathscr{A}$ contains a multiplicative unit, then φ is injective.*

Proof. It is sufficient to show that the relation

$$\varphi(y_1) \cdot \varphi(y_2) = \varphi(y_1 y_2) \qquad \text{holds for } y_1, y_2 \in Z.$$

We calculate straightforwardly and get

$$\begin{aligned} \varphi(y_1) \cdot \varphi(y_2) &= \sum_{z_1,u_1,z_2,u_2} \bar{z}_1 \alpha_{z_1,y_1}^{u_1} \cdot u_1 \cdot \bar{z}_2 \cdot \alpha_{z_2,y_2}^{u_2} \cdot u_2 \\ &= \sum_{z_1,u_1,u_2} \bar{z}_1 \alpha_{z_1,y_1}^{u_1} \alpha_{u_1,y_2}^{u_2} \cdot u_2 \\ &= \sum_{z_1,u_2} \bar{z}_1 \left(\sum_{u_1} \alpha_{z_1,y_1}^{u_1} \alpha_{u_1,y_2}^{u_2} \right) \cdot u_2. \end{aligned}$$

Now we apply $(z_1 y_1) y_2 = z_1 (y_1 y_2)$. Because R is elementwise commutable with Z, we get

$$\begin{aligned} \sum_{z_1,u_2} \bar{z}_1 \left(\sum_{u_1} \alpha_{y_1,y_2}^{u_1} \cdot \alpha_{z_1,u_1}^{u_2} \right) \cdot u_2 &= \sum_{u_1} \alpha_{y_1,y_2}^{u_1} \varphi(u_1) \\ &= \varphi(y_1 \cdot y_2). \end{aligned}$$

Thus the first part of our theorem has been proven.

Let

$$u = \sum_{y \in Z} \beta_y \circ y \qquad \text{and} \qquad \varphi(u) = 0.$$

Then it follows

$$\varphi(u) = \sum_{x,\, z \in Z} \bar{z} \cdot \sum_{y \in Z} \beta_y a_{z,y}^{x} \cdot x = 0;$$

therefore we have

$$\sum_{y \in Z} \alpha^x_{z,y} \beta_y = 0 \qquad \text{for} \quad x, z \in Z. \tag{1}$$

Let now $v \in \mathscr{A}$ such that

$$v = \sum_{v \in Z} \gamma_y \cdot y.$$

We form

$$v \cdot u = \sum_{y_1, y_2} \gamma_{y_1} \cdot \beta_{y_2} y_1 y_2 = \sum_{y_1, x} \gamma_{y_1} \left(\sum_{y_2} \alpha^x_{y_1, y_2} \beta_{y_2} \right) \cdot x.$$

Because of (1) we conclude that

$$v \cdot u = 0 \qquad \text{for all} \quad v \in \mathscr{A}.$$

We choose $v = 1$ and have $u = 0$. This proves the second part of our theorem.

Without proof we give another representation for the case of matrix rings.

THEOREM 3. *Let $\mathscr{A}$ be a finite dimensional ring of quadratic matices $(a_{z,y})_{z,y \in Z}$. Then*

$$\varphi(a) = \sum_{z, y \in Z} \bar{z} a_{z,y} \cdot y$$

is a monomorphism from $\mathscr{A}$ into $\mathscr{A}_D$.

Now we come to the main result of this section. To construct the representation $\varphi: \mathscr{A}_R(\delta) \to \mathscr{A}_D$ we first define a suitable alphabet for $\mathscr{A}_D$. For $u \in \bar{X}$ and $x \in X_r$ (remember $\bar{X} = X_1 \cup X_r$), we define

$$[u: x] = \begin{cases} 0 & \text{if for all } w \in \bar{X}^*, \quad \langle \delta^*(uw), x \rangle = 0, \\ 1 & \text{for } u = x, \\ \text{free variable in all other cases.} \end{cases}$$

Clearly from $[u: x] \neq 0$ and $u \in X_r$ it follows that $u = x$. We set

$$Z = \{ [u: x] \mid [u: x] \neq 1, 0; u \in \bar{X}, x \in X_r \}$$

and

$$\mathscr{A}_D = R\langle Z^{(*)} \rangle.$$

For $z \in \bar{X}$ we define

$$\varphi'(z) = \sum_{\substack{y, v, u, x \\ [y: x] \in Z}} \alpha^u_{y,v} \overline{[y: x]} [u: x] [z: v].$$

THEOREM 4. *There exists an uniquely defined extension of φ' to an algebra homomorphism $\varphi: \mathscr{A}_R(\delta) \to \mathscr{A}_D$.*

Proof. $\bar{X}$ generates $\mathscr{A}_R(\delta)$ and therefore there does not exist more than one homomorphic extension of φ' to $\mathscr{A}_R(\delta)$. To show that such an extension exists it is sufficient to show that for the linear extension φ of φ' holds

$$\varphi(z_1) \cdot \varphi(z_2) = \varphi(z_1 z_2) \qquad \text{where} \quad z_1 \in X_l,\ z_2 \in X_r.$$

By a straightforward calculation we get

$$\varphi(z_1) \cdot \varphi(z_2)$$

$$= \sum_{\substack{y_1, v_1, u_1, x_1, \\ y_2, v_2, u_2, x_2 \\ [y_1 : x_1] \in Z \\ [y_2 : X_2] \in Z}} \alpha^{u_1}_{y_1, v_1} \overline{[y_1, x_1]}[u_1 : x_1][z_1 : v_1]\, \alpha^{u_2}_{y_2, v_2} [\overline{[y_2 : x_2]}[u_2 : x_2][z_2 : v_2]$$

$$= \sum_{\substack{y_1, v_1, u_1, x_1 \\ v_2, u_2}} \alpha^{u_1}_{y_1, v_1}\, \alpha^{u_2}_{z_1, v_2} \overline{[y_1 : x_1]}[u_1 : x_1][u_2 : v_1][z_2 : v_2].$$

For $z_2 \neq v_2$ we have $[z_2 : z_2] = 0$ because $z_2 \in X_r$. Thus there remains only the case $z_2 = v_2$, that means $[z_2 : v_2] = 1$. We use the commutativity of R and have

$$\varphi(z_1) \cdot \varphi(z_2) = \sum_{u_2} \alpha^{u_2}_{z_1, z_2} \cdot \sum \alpha^{u_1}_{y_1, v_1} \overline{[y_1 : x_1]}[u_1 : x_1][u_2 : v_1]$$

$$= \sum \alpha^{u_2}_{z_1 z_2} \varphi(u_2) = \varphi(z_1 \cdot z_2).$$

Historical Remark. Nivat [15] in this thesis uses a homomorphism ψ, which formally looks like our homomorphism φ. But ψ is a mapping

$$\psi: R\langle X^* \rangle \to R\langle H(B) \rangle,$$

where $H(B)$ is the free half group generated by B. The main difference comes from the different domains of φ and ψ. Nivat uses ψ to prove the representation theorem of Shamir. But for his proof he needs the normal form theorem of Greibach, which follows from the existence like the theorem of Shamir [19] of φ. The reason is that $\mathscr{A}_R(G)$ contains a lot of information about G, but $R\langle X^* \rangle$ none at all. The reader may find more detailed informations on this subject in the book of Salomaa [18].

As we will show later, one can derive a representation of $L(G)$ from φ by a grammar in Greibach normal form. The size of the grammar corresponds to the size of φ. We define

$$|\mathscr{A}_R(\delta)| = \sum_{x, yz \in \bar{X}} |\alpha^z_{x, y}|$$

with

$$|\alpha| = \begin{cases} 1 \in \mathbb{N} & \text{for } \alpha \neq 0 \\ 0 \in \mathbb{N} & \text{else.} \end{cases}$$

For $p \in \mathscr{A}$ we put

$$|p| = \sum_{w \in Z^{(*)}} |\langle p, w \rangle|.$$

We define the size $|\varphi|$ of φ by

$$|\varphi| = \sum_{z \in \bar{X}} |\varphi(z)|.$$

One easily proves

LEMMA 3. $|\varphi| \leqslant |\mathscr{A}_R(\delta)| \cdot |\bar{X}|^2$, *where* $|\bar{X}|$ *is the number of elements of* $\bar{X}$.

INVARIANTS OF THE TRANSFORMATION $G \to \bar{G}$

We return to grammars and study which properties of G remain unchanged when passing from G to $\bar{G}$ as we did in Section 1.

The set of derivations of G of words into other words using G will be denoted by $\mathscr{F}$. If $f \in \mathscr{F}$, then $Q(f)$ is the word on which the derivation starts and $Z(f)$ is the result of the derivation f. If $f, g \in \mathscr{F}$ and $Q(f) = Z(g)$, then $f \circ g$ is the derivation which one gets by applying first g and then f. Obviously $Q(f \circ g) = Q(g)$, $Z(f \circ g) = Z(f)$, and "$\circ$" is associative. The empty derivation belonging to the word w is 1_w. We have $1_{Z(f)} \circ f \circ 1_{Q(f)} = f$. In the case $Q(f) = w$, $Z(f) = v$ we write also

$$w \xrightarrow{f} v.$$

If we have

$$w_1 \xrightarrow{f_1} v_1 \qquad \text{and} \qquad w_2 \xrightarrow{f_2} v_2,$$

we may form the derivation

$$w_1 \cdot w_2 \xrightarrow{f_1 \times f_2} v_1 \cdot v_2.$$

This leads to an additional associative operation on $\mathscr{F}$. The unit belonging to "x" is 1_x. Both operations are connected by the property

$$(f_1 \circ g_1) \times (f_2 \circ g_2) = (f_1 \times f_2) \circ (g_1 \times g_2)$$

if the left side is defined. $(\mathscr{F}, (X \cup T)^*, Q, Z, O, x)$ forms a free monoidal category which in [7] has been called free x- category and syntactical category in [1]. The elements of $\mathscr{F}$ are trees or words over the derivation trees in the case of context-free grammars. The trees of the production set P generate $\mathscr{F}$. $\bar{\mathscr{F}}$ is the category belonging to $\bar{G}$. The structure preserving mappings are called x functors. An x functor consists of two mappings (φ_1, φ_2). φ_1 is a monoid homomorphism from the monoid of the domain category into the monoid of the target category. φ_2 maps the derivation set into the derivation set. Moreover we use the abbreviations:

$$\mathrm{Mor}_{\mathscr{F}}(w, v) = \{f \in \mathscr{F} \mid Q(f) = w, Z(f) = v\},$$

$$\mathrm{mult}_G(w) = \mathrm{card}\, \mathrm{Mor}_{\mathscr{F}}(S, w).$$

The multiplicity of w over G tells us in how many essentially different ways w may be derived from S using G.

LEMMA 4. *For* $w \in T^*$

$$\mathrm{mult}_G(w) = \mathrm{mult}_{\bar{G}}(w).$$

Proof. To prove this lemma we construct the x functor $\varphi = (\varphi_1, \varphi_2)$ from $\bar{\mathscr{F}}$ onto $\mathscr{F}$ which deletes the indices r, l, in $\bar{G}$. Thus we define

$$\varphi_1(x, i) = x \qquad \text{for} \quad x \in X \quad \text{and} \quad i \in \{l, r\}$$

and for $f \in \bar{P}$

$$\varphi_2(f) = f' \Leftrightarrow \varphi_1(Q(f)) = Q(f'), \qquad \varphi_1(Z(f)) = Z(f').$$

This defines uniquely an x functor from $\bar{\mathscr{F}}$ into $\mathscr{F}$. Obviously $\varphi_2(\bar{P}) = P$.

We now show for $x_i \in \bar{X}$ that the restriction

$$\varphi_2 \mid \mathrm{Mor}_{\bar{\mathscr{F}}}(x_i, (\bar{X} \cup T)^*) \to \mathrm{Mor}_{\mathscr{F}}(x, (X \cup T)^*)$$

is bijective. From this fact our lemma follows immediately. The proof is an induction on the number $|f|$ of knots of the trees of f.

Our claim is true for all f such that $Q(f) = x_i$ and $|f| = 1$. Inductively we assume, that it holds for

$$\varphi_2 \mid \{f \in \mathrm{Mor}_{\bar{\mathscr{F}}}(x_i, (\bar{X} \cup T)^*) \mid |f| \leqslant n\} \to \{f \in \mathrm{Mor}_{\mathscr{F}}(x, (X \cup T)^*) \mid |f| \leqslant n\}.$$

It is clear that

$$|f| = |\varphi_2(f)| \qquad \text{for} \quad f \in \bar{\mathscr{F}}.$$

Let be $|f| = n + 1$ and $Q(f) = x_i$. We decompose

$$f = (1_u \times h \times 1_v) \circ g$$

such that $h \in \bar{P}$ and $|u|$ being minimal with this condition. This determines h uniquely. From

$$(1_u \times h \times 1_v) \circ g = (1_u \times h \times 1_v) \circ g'$$

it follows that $g = g'$, i.e., g is uniquely determined by this condition [7].

Because of

$$\varphi_2(f) = (1_{\varphi_1(u)} \times \varphi_2(h) \times 1_{\varphi_1(v)}) \circ \varphi_2(g)$$

and $|\varphi_1(u)| = |u|$ we see that $\varphi_2(f)$ has exactly one co-image. This proves our lemma.

Now we will show that the $LL(k)$ and $LR(k)$ properties of G do not change when passing from G to $\bar{G}$. For this purpose we introduce the following notation. We call $f \in \mathscr{F}$ *u-left-prime* for $u \in (X \cup T)^*$, iff from $f = (1_u \times h) \circ g$ it follows that $g = f$.

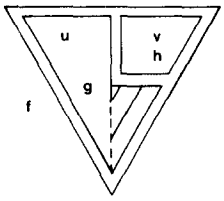
The definition *u-right-prime* is symmetric to the foregoing definition.

One easily shows

LEMMA 5. *For each $f \in \mathscr{F}$, u prefix of $Z(f)$, there exists exactly one decomposition $f = (1_u \times h) \circ g$ such that g is* u-left-prime.

Adopting the notation of this lemma we call g the *u-left-prime* factor of f and h the *v-right-base* of f if $Z(f) = u \cdot v$. We write

$$g = \text{left-prime}(u, f), \qquad h = \text{right-base}(v, f).$$



This figure should explain the definitions. We also use the notions which we get from this definition by changing "left" into "right" and "right" into "left."

We now give a definition of $LR(k)$ which is equivalent to the definition [5, p. 502] and of $LL(k)$ which is equivalent to the one given by Lewis and Stearns [6]. The reader should remember that we assume G to be in Chomsky *NF*, and G without ε-productions: G is a $LR(k)$ grammar (resp. $LL(k+1)$ grammar) for $k = 0, 1, \ldots$, if the following holds:

For all $f, f' \in \mathscr{F}$ (resp. $f, f' \in \text{Mor}_{\mathscr{F}}(S, T^*)$) for $LR(k)$ with $Z(f) = u \cdot v$ and $Z(f') = u \cdot v'$

we have

$$\text{left-base}\ (u, f) = \text{left-base}\ (u, f')$$

if $Q(f) = Q(f') = S$ and $\text{First}_k(v) = \text{First}_k(v')$ resp. for $LL(k+1)$

$$\text{left-prime}\ (u, f) = \text{left-prime}\ (u, f')$$

if $Q(f) = Q(f') \in X$ and $\text{First}_k(v) = \text{First}_k(v')$).

Remember that we assume that S never appears on the right-hand side of any production. Hence [5, p. 525] our $LR(0)$ grammars produce only $ALR(O)$-language, i.e., strict derministic languages.

LEMMA 6. *If G is a $LL(k)$ (resp. $LR(k)$) grammar, then $\bar{G}$ is a $LL(k)$ (resp. $LR(k)$) grammar.*

Proof. To prove this lemma we use the x functor defined in the proof of Lemma 4. Let be $f: x_i \to uv$ any derivation tree of $\mathscr{F}$ and $x_i \in \bar{X}$. We define

$$h = \text{left-base}(u, f) \qquad \text{and} \qquad g = \text{left-prime}(u, f).$$

Then we have

$$h' = \varphi_2(h) = \text{left-base}(\varphi_1(u), \varphi_2(f))$$

and

$$g' = \varphi_2(g) = \text{left-prime}(\varphi_1(u), \varphi_2(f)).$$

Now x_i and g' determine g uniquely as shown in Lemma 4. Now let G be a $LL(k)$ grammar. Then g' is uniquely determined by $\varphi_1(x_i)$ and $\varphi_1(u)$. $\text{First}_k \varphi_1(v)$. Thus x_i and $u \cdot \text{First}_k(v)$ determine g' and thereby g uniquely. This means that $\bar{G}$ is a $LL(k)$ grammar.

Now we study the case that G is a $LR(k)$ grammar. By the same argumentation as before we see that h is uniquely determined by h' and $Q(h)$.Using the $LR(k)$ property we see that $Q(f) = S_r$ and $u \cdot \text{First}_k v$ determine h' uniquely. If we are able to show that $Q(h)$ is uniquely determined by $u \cdot \text{First}_k v$, then it follows that $\bar{G}$ has the $LR(k)$ property. For this purpose it is sufficient to show that $Q(h) \in X_1^*$ holds. Therefore let

$$f = (h \times 1_v) \circ g,$$

where by definition of h as left-base of f the factor g is v-right-prime. Suppose $Q(h) \in X_1^*$. Then there exists a decomposition

$$Q(h) = q_1 x_1 x_r q_2$$

and we have

$$Z(g) = q_1 x_1 x_r q_2 v.$$

This contradicts the assumption that g is v-right-prime. Thus we have $Q(h) \in Q_1^*$, as we wished to show.

The last result in our proof will be used in a later part of this paper. Therefore we formulate it as

LEMMA 7. *If f is v-right-pime, then $u \in X_1^*$ for $Z(f) = u \cdot v$. If h is u-left-base of f, then $Q(h) \in X_1^*$. The lemma remains true if we exchange the words left and right.*

CONNECTIONS BETWEEN $L(G)$, $\mathscr{A}_R(G)$, AND φ

In this section we work out the general relations between $L(G)$ and $\mathscr{A}_R(G)$ and our representation φ. A first information is given by

THEOREM 5. $w \in L(G) \Leftrightarrow \langle \eta(w), S_r \rangle \neq 0$ *for* $\chi(R) = 0$ ($\chi(R) =$ *characteristic of* R),

$$\text{mult}_G(w) = \langle \eta(w), S_r \rangle [\mathscr{A}_{\mathbb{N}}(G)].$$

(Remember: "[ ]" contains the algebra "in relation" to which the relations is to be understood.)

Proof. As we have shown in Lemma 4, we may use $\bar{G}$ instead of G. The proof is by induction on the length $|w|$ of w. We show a somewhat more general result:

$$\text{mult}_{\bar{G}}(x_i, w) = \langle \eta(w), x_i \rangle \qquad \text{for} \quad w \in T^*, x_i \in \bar{X},$$

where

$$\text{mult}_{\bar{G}}(x_i, w) = \text{card Mor}_{\bar{\mathscr{F}}}(x_i, w).$$

The theorem is obvious for $|w| = 1$. Let $f: x_i \to w$ be a derivation and $|w| > 1$. Then we may decompose

$$f = (f_1 \times f_2) \circ p, \qquad p \in \bar{P}.$$

Hence, we have

$$\text{mult}_{\bar{G}}(x_i, w) = \sum_{\substack{w_1 \cdot w_2 = w \\ w_1 \neq 1, w_2 \neq 1 \\ \langle \delta^*(y_l z_r), x_i \rangle = 1}} \text{mult}_{\bar{G}}(y_l, w_1) \cdot \text{mult}_{\bar{G}}(z_r, w_2).$$

By the induction hypothesis

$$\begin{aligned} \text{mult}_{\bar{G}}(x_i, w) &= \sum_{\substack{w_1 \cdot w_2 = w \\ w_1 \neq 1, w_2 \neq 1}} \langle \eta(w_1), y_l \rangle \cdot \langle \eta(w_2), z_r \rangle \cdot \langle y_l z_r, x_i \rangle \\ &= \langle \eta(w), x_i \rangle \end{aligned}$$

and the proof is complete.

In the following we use the definition

$$(u) = u + \mathscr{A}_R(G).$$

(u) is the additive residue class of u.

COROLLARY TO THEOREM 5. *For* $R = \mathbb{B} =$ *boolean ring with two elements we have*

$$L(G) = \eta^{-1}(S_r).$$

Now we study how the representation φ transforms the residue class (S_r). If we define $\langle w, s\rangle := \langle \delta^*(w), s\rangle$ for $w \in \bar{X}^*$, we get

LEMMA 8. *For $z_0 z_1 \cdots z_n \in \bar{X}^*$ with $z_0 \neq s$ we have*

$$\langle z_0 z_1 \cdots z_n, s\rangle = \langle \varphi(z_1 \cdots z_n), \overline{[z_0 : s]}\rangle.$$

Proof. The proof is an induction on n.

Basis. We have

$$\langle z_0 z_1, s\rangle = \alpha^s_{z_0, z_1}.$$

Since

$$\varphi(z_1) = \sum \alpha^u_{y,v}\, \overline{[y : x]}[u : x][z_1 : v],$$

we get

$$\langle \varphi(z_1), \overline{[z_0 : s]}\rangle = \sum_{\overline{[y:x]}[u:x][z_1:v] = \overline{[z_0:s]}} \alpha^u_{y,v}.$$

Hence the sum is only to be taken over the cases

$$y = z_0, \qquad x = s, \qquad u = x, \qquad z_1 = v.$$

Therefore

$$\langle \varphi(z_1), \overline{[z_0 : s]}\rangle = \alpha^s_{z_0, z_1}$$

and the basis is complete.

Induction step. By

$$\varphi(z_1 \cdots z_n) = \sum_{u,y,v,x} \alpha^u_{y,v}\, \overline{[y : x]}[u : x][z_1 : v]\, \varphi(z_2 \cdots z_n).$$

we get

$$\begin{aligned}
\langle \varphi(z_1 \cdots z_n), \overline{[z_0 : s]}\rangle &= \sum_{u,v} \alpha^u_{z_0,v} \langle [u : s][z_1 : v]\, \varphi(z_2 \cdots z_n), 1\rangle \\
&= \sum_{u,v} \alpha^u_{z_0,v} \sum_{j=2}^{n-1} \langle \varphi(z_2 \cdots z_j), \overline{[z_1 : v]}\rangle \cdot \langle \varphi(z_{j+1} \cdots z_n), \overline{[u : s]}\rangle \\
&\quad + \sum_v \alpha^s_{z_0,v} \langle \varphi(z_2 \cdots z_n), \overline{[z_1 : v]}\rangle \\
&\quad + \sum_u \alpha^u_{z_0,z_1} \langle \varphi(z_2 \cdots z_n), \overline{[u : s]}\rangle.
\end{aligned}$$

By the induction hypothesis this sum is equal to

$$\sum_{u,v} \alpha^{u}_{z_0,v} \sum_{j=1}^{n} \langle z_1 \cdots z_j, v\rangle \langle uz_{j+1} \cdots z_n, s\rangle, \tag{2}$$

where $z_{n+1} = 1$. On the other hand

$$\langle z_0 z_1 \cdots z_n, s\rangle = \sum_{k=0}^{n-1} \sum_{y_0,y_1} \alpha^{s_0}_{y_0,y_1} \langle z_0 \cdots z_k, y_0\rangle \langle z_{k+1} \cdots z_n, y_1\rangle. \tag{3}$$

We prove by induction that

$$\langle z_0 \cdots z_k, y_0\rangle = \sum_{\substack{j=1 \\ u,v}}^{k} \alpha^{u}_{z_0,v} \langle z_1 \cdots z_j, v\rangle \langle uz_{j+1} \cdots z_k, y_0\rangle.$$

For $k = 1$ we have

$$\langle z_0 z_1, y_0\rangle = \sum_{u,v} \alpha^{u}_{z_0,v} \langle z_1, v\rangle \langle u, y_0\rangle = \alpha^{y_0}_{z_0,z_1}.$$

Therefore our claim holds for $k = 1$.

We assume the claim to be correct for $k < n$ and apply it to (3). We get

$$\begin{aligned}
\langle z_0 \cdots z_n, s\rangle &= \sum_{\substack{k=1 \\ y_0,y_1}}^{n-1} \alpha^{s}_{y_0,y_1} \sum_{\substack{j=1 \\ u,v}}^{k} \alpha^{u}_{z_0,v} \langle z_1 \cdots z_j, v\rangle \langle uz_{j+1} \cdots z_k, y_0\rangle \\
&\quad \times \langle z_{k+1} \cdots z_n, y_1\rangle. \\
&\quad + \sum_{y_0,y_1} \alpha^{s}_{y_0,y_1} \langle z_0, y_0\rangle \langle z_1 \cdots z_n, y_1\rangle \\
&= \sum_{\substack{j=1 \\ u,v}}^{n-1} \alpha^{u}_{z_0,v} \langle z_1 \cdots z_j, v\rangle \sum_{\substack{k=1 \\ y_0,y_1}}^{n-1} \alpha^{s}_{y_0,y_1} \langle uz_{j+1} \cdots z_k, y_0\rangle \\
&\quad \times \langle z_{k+1} \cdots z_n, y_1\rangle \\
&\quad + \sum_{v} \alpha^{s}_{z_0,v} \langle z_1 \cdots z_n, v\rangle \\
&= \sum_{\substack{j=1 \\ u,v}}^{n} \alpha^{u}_{z_0,v} \langle z_1 \cdots z_j, v\rangle \langle uz_{j+1} \cdots z_n, s\rangle.
\end{aligned}$$

Hence, our claim is true for $k = n$. This result and (2) prove our lemma.

LEMMA 9. *Using the notation of Lemma* 8 *we get that*

$$\langle z_1 \cdots z_n, S_r \rangle = \langle \varphi(z_1 \cdots z_n), \overline{[S_l : S_r]} \rangle.$$

Proof. Lemma 8 implies that

$$\langle S_l z_1 \cdots z_n, S_r \rangle = \langle \varphi(z_1 \cdots z_n), \overline{[S_l : S_r]} \rangle.$$

By the definition of $\mathscr{A}_R(G)$ we get

$$\langle S_l z_1 \cdots z_n, S_r \rangle = \langle z_1 \cdots z_n, S_r \rangle$$

and the proof is complete.

If we now compose the homomorphisms η and φ, we get a homomorphism $h = \varphi \circ \eta$ from T^* into $R\langle X^{(*)} \rangle$. This leads us to a representation theorem for c.f. languages which is nearly the theorem of Shamir ([19]; cf. [15]). Shamir uses the half group $H(X)$, instead of $X^{(*)}$, i.e., he does not use the relations $x \cdot \bar{y} = 0$ for $x \neq y$.

THEOREM 6 (Shamir). *For each c.f. language* $L \subset T^*$ *there exists a monoid homomorphism* $h: T \to R\langle Z^{(*)} \rangle$ *and an additive residue class* (\$) *such that* $L = h^{-1}((\$))$.

Proof. The proof follows from Lemma 9 and Theorem 5 by choosing $\$ = \overline{[S_l : S_r]}$. Each polycyclic monoid $Z^{(*)}$ can be embedded by a monomorphism into $\{x_1, x_2\}^{(*)}$. This embedding can even be done in a way that $\overline{[S_l : S_r]}$ will always be mapped onto the same element $a_0 \in \{x_1, x_2\}^{(*)}$. We extend this embedding to a ring homomorphism from $R\langle Z^{(*)} \rangle$ into $R\langle \{x_1, x_2\}^{(*)} \rangle$ and put it behind h. Let $\bar{h}$ be the resulting homomorphism. Then the following corollary holds.

COROLLARY TO THEOREM 6. *For each c.f. language* $L \subset T$, *there exists a homomorphism*

$$\bar{h}: T \to R\langle \{x_1, x_2\}^{(*)} \rangle$$

such that

$$L = \bar{h}^{-1}((a_0)).$$

In this form this result was first given in [8], where it was derived from the theorem of Chomsky and Schützenberger, which is an algebraic version of the theorem of Greibach about a hardest language under homomorphic reduction [3]. One gets this language from the representation given above by forming the c.f. language of the expressions consisting of products of polynomials of $R\langle \{x_1, x_2\} \rangle$. The theorem of Greibach and the above representation have been found independently of the theorem of Shamir. For a long time no attention was paid to the theorem of Shamir outside of the French School, because its complexity theoretic

aspects were overlooked. As shown in [11] one can construct similar representations for r.e., c.s., d.c.s., and other classes of languages. It seems to be possible to construct a language which is hardest in the category of homomorphic reductions for each complexity class given by a time bound $T(n)$.

We show that it is as easy to prove the theorem of Chomsky and Schützenberger from our Theorem 5 and Lemma 9 as in the case of the theorem of Shamir. For this purpose we somewhat change the definition of h, but in such a way that Lemma 9 remains applicable.

We define a homomorphism $g: T^* \to R\langle(\tilde{Z} \cup \tilde{T})^*\rangle$ by ($\tilde{Z} := Z \cup \bar{Z}$, $\tilde{T} := T \cup \bar{T}$),

$$g(t) = \sum_{z \in X} \alpha_t^z \sum \alpha_{y,v}^u \overline{[y: x]}\, t\bar{t}[u: x][z: v] \qquad \text{for} \quad t \in T.$$

We notice that the difference of g and h exists in two points: The co-domain is different and the product $t\bar{t}$ has been inserted between $\overline{[y: x]}$ and $[u: x][z: v]$.

Let $\bar{g}$ be the prolongation of g to a homomorphism from T^* into $R\langle(\tilde{Z} \cup \tilde{T})^{(*)}\rangle$ which we obtain by applying first g and then the canonical mapping from $R\langle(\tilde{Z} \cup \tilde{T})^{(*)}\rangle$.

Then the following corollary is obvious.

COROLLARY TO LEMMA 9. $\langle \bar{g}(w), \overline{[S_l: S_r]}\rangle = \langle h(w), \overline{[S_l: S_r]}\rangle$.

We now define a regular set over $\tilde{Z} \cup \tilde{T}$,

$$\mathrm{REG} = [S_l: S_r] \cdot \{v \mid \exists (t \in T)\langle g(t), v\rangle \neq 0\}^*.$$

Let $(\tilde{Z} \cup \tilde{T})$ be the Dyck-language over $\tilde{Z} \cup \tilde{T}$ and $\sigma: (\tilde{Z} \cup \tilde{T})^* \to T^*$ the monoid homomorphism with

$$\sigma(z) = \varepsilon \qquad \text{for} \quad z \in \tilde{Z},$$
$$\sigma(\bar{t}) = \varepsilon \qquad \text{for} \quad t \in T,$$
$$\sigma(t) = t \qquad \text{for} \quad t \in T.$$

From Lemma 9 we get

THEOREM 7 (Chomsky-Schützenberger). $L(G) = \sigma(\mathrm{REG} \cap D(\tilde{Z} \cup \tilde{T}))$.

To conclude this section we construct a grammar in Greibach normal form for $L(G)$. We define

$$\tilde{P} = \{[y: x] \to t[z: v][u: x] \mid \alpha_t^z \cdot \alpha_{y,v}^u \neq 0\}$$

and

$$\tilde{G} = (Z, T, \tilde{P}, [S_1: S_r]).$$

Obviously $\tilde{G}$ is in Greibach normal form. We can prove

THEOREM 8. *$L(G) = L(\tilde{G})$ and more precisely*

$$\mathrm{mult}_G(w) = \mathrm{mult}_{\tilde{G}}(w) \qquad \textit{for} \quad w \in T^*.$$

The size of $|G|$ and $|\tilde{G}|$ are related by

$$|\tilde{G}| \leqslant 32 \cdot |P_N| \cdot |P_T| \cdot |X|,$$

where $P = P_N \cup P_T$, P_N the set of non-terminal and P_T the set of terminal productions.

Proof. We define a homomorphism $h_1 : T^* \to R\langle \tilde{Z}^* \rangle$ by

$$h_1(t) = \sum_{z \in X} \alpha_t^z \sum_{\substack{y,u,v \\ x, x \neq y}} \alpha_{y,v}^u \overline{[y:x]}[u:x][z:v] \qquad \text{for} \quad t \in T.$$

We use the canonical mapping

$$\mu : R\langle \tilde{Z}^* \rangle \to R\langle Z^{(*)} \rangle.$$

We write

$$h_1(w) = \sum_{m \in \tilde{Z}^*} \alpha_m \cdot m \qquad \text{where} \quad \alpha_m = \langle h_1(w), m \rangle.$$

Remember that $\alpha_z^t, \alpha_{y,v}^u \in \{0, 1\}$ because we start with δ originating from a grammar. Because the co-domain of h_1 is $R\langle \tilde{Z}^* \rangle$ we know that $\alpha_m \in \{0, 1\}$ for $m \in \tilde{Z}^*$. We put

$$W_1(w) = \{m \in \tilde{Z}^* \mid \alpha_m \neq 0,\ \mu(m) = \overline{[S_1 : S_r]}\},$$
$$W_2(w) = \tilde{Z}^{(*)} - W_1(w).$$

Then we can write

$$h_1(w) = \sum_{m \in W_1(w)} \alpha_m \cdot m + \sum_{m \in W_2(m)} \alpha_m \cdot m$$

and

$$\langle \mu \circ h_1(w), \overline{[S_l : S_r]} \rangle = \sum_{m \in W_1(w)} \alpha_m.$$

Because

$$\mathrm{mult}_G(w) = \langle \eta(w), S_r \rangle = \langle \varphi \circ \eta(w), \overline{[S_l : S_r]} \rangle$$

it follows that

$$\mathrm{mult}_G(w) = \sum_{m \in W_1(w)} \alpha_m.$$

Now we assign a unique derivation over $\tilde{P}$ to each $m \in W_1(w)$. For this purpose we generalize W_1 in such a way that any element of Z may be taken instead of $[S_1 : S_r]$. Therefore we define

$$W_1(w, \bar{z}) = \{m \in \bar{Z}^* \mid \langle h_1(w), m \rangle = 1, \mu(m) = \bar{z}\} \qquad \text{for } w \in T^* \text{ and } z \in Z.$$

We construct a bijective mapping from $W_1(w, \bar{z})$ onto

$$\mathrm{Mor}_{\mathscr{F}}(z, w) \qquad \text{where } \mathscr{F} \text{ is associated with to } G.$$

We take $\bar{z}ab \in W_1(w, \bar{z})$ and $w = t_0 \cdot w'$ and we assume

$$\langle h_1(t_0), \bar{z}ab \rangle = 1, \qquad a, b \in Z.$$

Since $\mu(m) = \bar{z}$, there exists a decomposition $w' = w_2 \cdot w_3$ such that

$$\mu(h_1(w_2)) = \bar{b} \qquad \text{and} \qquad \mu(h_1(w_3)) = \bar{a}.$$

Thus

$$|\tilde{G}| \leqslant 2\,|\tilde{P}_T| + 4\,|\tilde{P}_N| \leqslant 4\,|\tilde{P}|.$$

Now

$$|\tilde{P}| = \sum_{\substack{t \in T \\ z \in \bar{X}}} \alpha_t^z \sum_{u, y, v, x} \alpha_{y,v}^u \leqslant \left(\sum_{\substack{t \in T \\ z \in \bar{X}}} \alpha_t^z \right) \left(\sum_{u, y, v \in \bar{X}} \alpha_{y,v}^u \right) \cdot |\bar{X}|.$$

This means

$$|\tilde{P}| \leqslant |\bar{P}_t| \cdot |\bar{P}_N| \cdot |\bar{X}|,$$

where $\bar{P}$ belongs to $\bar{G}$. Hence

$$|\tilde{P}| \leqslant 8\,|P_T| \cdot |P_N| \cdot |X|$$

and

$$|\tilde{G}| \leqslant 32 \cdot |P_T| \cdot |P_N| \cdot |X|,$$

which had to be proved.

Remark. From this theorem we get immediately

$$|\tilde{G}| \leqslant \tfrac{16}{3}\,|G|^2 \cdot |X| < \tfrac{8}{3}\,|G|^3.$$

For large production systems, this means.

$$|P_T| = O(|T| \cdot |X|),\ P_N = O(|X|^3).$$

We have for $|T| < |X|$ and $\varepsilon > 0$ that

$$|\tilde{G}| \leqslant O(|T| \cdot |X|^5) \leqslant O(|G|^{2+\varepsilon}).$$

Syntactical Congruences

In this section we transfer the syntactical congruences to our algebra $\mathscr{A}_R(G)$ and we study how these congruences relate under our representation $\varphi: \mathscr{A}_R(G) \to R\langle Z^{(*)}\rangle$. In connection with this following lemma plays a central role.

Lemma 10. *For $w \in \bar{X}^*$ let exist an $u \in Z^*$ such that $\langle \varphi(w), [z_0 : t_0]\, u\rangle = \alpha \neq 0$. Then there exists $w' \in X_r^*$ such that $\langle z_0 w w', x_0\rangle \geqslant \alpha$.*

Proof. The proof is by induction on $n = |u|$. The case $n = 0$ follows from Lemma 8. Now assume the lemma is true for all u' with $|u'| \leqslant n$:

$$u = u_1[y : x], \qquad [y : x] \neq 0, 1, |u_1| = n.$$

Then there exist $v_1, v_2, \ldots, v_m \in X_r$ such that

$$\langle y v_1 v_2 \cdots v_m, x\rangle = \beta \neq 0.$$

By Lemma 8 we get

$$\langle \varphi(v_1 \cdots v_m), \overline{[y : x]}\rangle = \beta.$$

Therefore

$$\langle \varphi(w v_1 \cdots v_m), \overline{[z_0 : x_0]}\, u \overline{[y : x]}\rangle \geqslant \alpha \cdot \beta > 0.$$

Thus we have

$$\langle \varphi(w v_1 \cdots v_m), \overline{[z_0 : x_0]}\, u_1\rangle > 0.$$

Now the claim of the lemma follows inductively.

For $L \subset T^*$ we define

$$u =_r v(L) \Leftrightarrow \forall_w\ (uw \in L \Leftrightarrow vw \in L) \text{ as usual,}$$

$$=_r (L) \text{ is the } \textit{syntactical right congruence.}$$

For an easy formulation of the following results we extend our alphabet Z by a

new element $\dashv$. We call the new alphabet Z again and we use the abbreviation $\$ = \dashv \cdot [S_l: S_r]$.

The idea is to annulate words in $\varphi(\mathscr{A}(G))$ which have not the form $\overline{[S_1: S_r]} \cdot Z^*$ by multiplying them from the left by \$. *Remember* $\$ \cdot \bar{z} = 0$ for $z \in Z$ and $z \neq [S_l: S_r]$ and $\$[S_l: S_r] \cdot \bar{z} = 0$ for all $z \in Z$.

THEOREM 9. $w =_r O(L) \Leftrightarrow \$h(w) = O$.

Here h is the homomorphism of Theorem 6.

Proof. We assume $\$ \cdot h(w) \neq 0$. Applying Lemma 10 we find w' such that $\langle S_1 \eta(ww'), S_r \rangle \neq 0$, and by Lemma 9 we have $ww' \in L$. Therefore $w \neq_r O(L)$.

On the other hand, if there exists a word w' such that $w \cdot w' \in L$, then by Lemma 9, $\langle S_1 \eta(ww'), S_r \rangle \neq 0$ and therefore $\$ \cdot h(w) \neq 0$. This proves our theorem.

This theorem yields a procedure to decide $w =_r O(L)$ for L being a context-free language. Now we transfer the right congruence to $\mathscr{A}_R(G)$ by defining

$$p =_r p'(L) \Leftrightarrow \bigvee_{q \in \mathscr{A}_R(G)} (\langle p \cdot q, S_r \rangle = 0 \Leftrightarrow \langle p' \cdot q, S_r \rangle = 0).$$

for $p, p' \in \mathscr{A}_R(G)$.

In a symmetrical way we define the *left congruence* $=_l(L)$. We easily see, that for $R = \mathbb{B}$ or $R = \mathbb{N}$ these definitions define congruence relations, but this is not true for $R = \mathbb{Z}$ or R being a field. The same holds for the following definition of the syntactical equivalence modulo L:

$$p = p'(L) \Leftrightarrow \bigvee_{q, q' \in \mathscr{A}R(G)} (\langle q \cdot p \cdot q', S_r \rangle = 0 \Leftrightarrow \langle q \cdot p' \cdot q', S_r \rangle = 0).$$

The quotient of $\mathscr{A}_R(G)$ by the syntactical congruence yields the *syntactical algebra* $\mathscr{A}_R(G)/(L)$.

Because the syntactical monoid is hard to compute even for c.f. languages, this holds for $\mathscr{A}_R(G)/(L)$ too. Therefore it is of interest to look for algebras between $\mathscr{A}_R(G)$ and $\mathscr{A}_R(G)/(L)$. We put

$$\mathfrak{A}_r(L) = \{p \in \mathscr{A}_R(G) \mid p =_r O(L)\}$$

and

$$\mathfrak{A}(L) = \{p \in \mathscr{A}_R(G) \mid p = O(L)\}.$$

Obviously we have

LEMMA 11. $\mathfrak{A}_r(L)$ *is a right ideal.*

$\mathfrak{A}(L)$ *is a 2-sided ideal.*

Immediately we get a

COROLLARY TO THEOREM 9. *The word problem* $w \in \mathfrak{A}_r(L)$ *is decided by* $\$ \cdot \varphi(w)$ *for* $R = \mathbb{N}$ *or* $R = \mathbb{B}$.

Now $\varphi^{-1}(O)$ is a two sided ideal of $\mathscr{A}(G)$ and $\varphi^{-1}(O) \subset \mathfrak{A}_r$. Therefore one may ask whether $\varphi^{-1}(O)$ has an interesting syntactical property. Obviously we also have $\varphi^{-1}(O) \subset \mathfrak{A}(L)$.

One may ask whether it is possible to prolongate φ to a homomorphism $\psi: \mathscr{A}(G) \to R\langle Y^{(*)}\rangle$ with a suitable Y, such that $\psi^{-1}(O) = \mathfrak{A}(L)$. Because of Lemma 2 one cannot do this by a homomorphism from $R < Z^{(*)}\rangle$ into $R\langle Y^{(*)}\rangle$. But it is possible that such a prolongation from $\varphi(\mathscr{A}(G))$ into a suitable $R\langle Y^{(*)}\rangle$ exists because $1 \notin \varphi(\mathscr{A}(G))$.

Presumedly such a homomorphism does not exist, because each semigroup homomorphism from $Z^{(*)}$ in $R\langle Y^{(*)}\rangle$ which is induced by transformations $[y: x] \to \sum q[y: x]\, q'$ maps the elements $[y: x] \cdot [z: v]$ for $v \neq x$ into O. Therefore it remains an

Open Question. Do there exist non-trivial representations of $\mathscr{A}_R(G)/\mathfrak{A}(L)$ in $R\langle Y^*\rangle$?

Answering this question is of practical interest too, because a section u of a program of a language L is syntactically incorrect if $u = O(L)$. By evaluating $\$ \cdot \varphi(w)$ we are able to find the shortest syntactically incorrect prefix of a program u L. The representation of $\mathscr{A}_R(G)/\mathfrak{A}(L)$ that we are looking for would do the same for the shortest syntactically incorrect sections of a program.

One could object that the evaluation of our ring homomorphisms is not trivial. That is true if we wish to do it in a most efficient way. But there are several important problems that are reducible to this problem. We seize the opportunity and point out some additional problems which seem to be important.

The syntactical congruence of a language $L(G)$ does not reflect the structure of G very strongly, as the weak equivalence of two languages $L(G) = L(G')$ does not say much about relations between G and G'. One of the most important applications of language theory is to describe the syntax of programming or natural languages. The semantics of these languages depends strongly on the grammars G, which generate the syntax. Therefore it appears to me that the grammars deserve more interest than the languages. Languages are just one of different properties of grammars. If the grammars G and G' describe the syntax of two programming languages and if $L(G) = L(G')$ then these languages are not necessarily equal as programming languages. This leads to the question of formulating structural equivalences between grammars. Different equivalences of this kind have been defined but only one of them, the "strong" equivalence, is well known. These equivalences will be reflected by the existence of certain homomorphisms and products between our algebras $\mathscr{A}_R(G)$. We will come back to this problem later on. Here we only give a definition of a *finer syntactical congruence*, which is identical to the normal one in the case of unambiguous grammars.

For $p, p' \in \mathscr{A}_R(G)$ we define that p is *syntactically congruent* to p' modulo G:

$$p = p'(G) \Leftrightarrow \underset{q,q' \in \mathscr{A}_R(G)}{\forall} (\langle qpq', S_r \rangle = \langle qp'q', S_r \rangle).$$

We see that the O-classes in both congruences (L) and (G) are the same.

The word problem for the quotient algebra $\mathscr{A}(G)/(G)$ is closely related to the equivalence problem in the case of unambiguous grammars. Therefore these algebras are, as one may assume, hard to compute. It is clear that in connection with this a lot of interesting questions arise. For R being a field we have

$$p = p'(G) \Leftrightarrow p = p'\mathscr{A}_R(G)/\mathfrak{A}.$$

Therefore in this case $\mathscr{A}_R(G)/(G)$ is the syntactical algebra Reutenauer [17] has associated to the formal power series belonging to the grammar G. We think that it is very important to study each of these cases. Restricting to $R = \mathbb{Z}$ or R being a field makes important practical questions disappear from the theory.

Unambiguous Grammars, $LL(k)$ Grammars

In this section we assume always $R = \mathbb{N}$ and therefore we write $\mathscr{A}(G)$ for $\mathscr{A}_R(G)$. By definition for unambiguous grammars holds

$$\langle w, S_r \rangle \leqslant 1 \qquad \text{for} \quad w \in T^*.$$

Because of Lemma 10 this is equivalent to

$$\langle \$ \cdot \varphi(u), a \rangle \leqslant 1 \qquad \text{for} \quad u \in \bar{X}^* \quad \text{and} \quad a \in Z^*.$$

If we check the proof of Lemma 10, we see that the following lemma is true.

Lemma 12. *Let G be an unambiguous c.f. grammar and $w \cdot w' \in L(G)$. Then there exists exactly one monomial $a \in Z$ such that*

$$\alpha = \langle \$ \cdot h(w), a \rangle$$

$$\alpha' = \langle h(w'), \bar{a} \rangle$$

and $\alpha = \alpha' = 1$. Here $\overline{x_1 \cdots x_k} = \overline{x_k} \cdots \overline{x_1}$.

In the following we assume G to be a $LL(k)$ grammar unless the converse is explicitly stated. We are interested in studying $\mathscr{A}(G)$ and our representation in the case of $LL(k)$ grammars. As we have shown in lemma 7, it follows from f being u-left prime and $Z(f) = u \cdot v$ that $v \in X_r^*$. In $\mathscr{A}(G)$ we then have $\langle uv, Q(f) \rangle = 1$ if $Q(f) \in \bar{X}^*$. We call $v \in X_r^*$ an *almost right inverse* of u if there exists $z \in \bar{X}$ such that $\langle uv, z \rangle \neq 0$.

LEMMA 13. *Let card* $X = m \geqslant 2$. *For each* $u \in (\bar{X} \cup T)^*$ *where there exist at most* $2 \cdot m^{k+2}$ *elements* $v \in X_r^*$ *which are almost right inverses of* u *if* G *is* $LL(k)$.

Proof. Let be $v \in X_r^*$ and $\langle uv, y \rangle = 1$. Then we can find $f: y \to uv$. Because $v \in X_r^*$, f is u left prime. G is $LL(k)$ and hence the derivation tree f, and hereby v, is uniquely determined by $u \cdot \mathrm{First}_k(v)$ and y. Therefore there exist only m^{k+1} different words of length $\leqslant k$. Therefore there exist at most $2 \cdot m^{k+2}$ elements which are almost right inverses of u. We define

$$|p| = \sum_{u \in \overline{Z}Z^*} \langle p, u \rangle \qquad \text{for} \quad p \in R\langle Z^{(*)} \rangle.$$

$|p|$ is the sum of the coefficients of the monomial of p which contain an inverse to an element of Z in the first place and none elsewhere.

LEMMA 14. *Let be* $u \in \bar{X}^*$. *Then*

$$|\varphi(u)| \leqslant m^{k+3}.$$

Proof. Let be $|w|_\rho = w$ and $\langle \varphi(u), w \rangle = 0$, $w = \overline{[z: x]}\, w'$, and $w' \in Z^*$. By Lemma 10 we find $v \in X_r^*$ such that $\langle zuv, x \rangle \neq 0$. Now there exist not more than m^{k+1} elements $v \in X_r^*$ such that $\langle zuv, x \rangle \neq 0$ as shown in Lemma 14. There do not exist two different monoms $\overline{[z: x]}\, w_1'$ and $\overline{[z: x]}\, w_2'$ which have the same v as "right inverse." From this we could conclude $\langle zuv, x \rangle \geqslant 2$, which is in contradiction to the unambiguity of G. Therefore we have indeed $|\varphi(n)| \leqslant m^{k+3}$.

LEMMA 15. *Let be* $u \in (\bar{X} \cup T)^*$ *and* $[y_0: x_0] \in Z$. *If* $\dashv \cdot [y_0: x_0]\, \varphi(u) \neq 0$ *then there exists a decomposition* $u = u_1 \cdot u_2$ *and* $w \in Z^*$ *such that*

$$\dashv \cdot [y_0: x_0]\, \varphi(u) = w \cdot f(u_2), \qquad |u_2| \leqslant k.$$

Proof. By Lemma 10 it follows from $\dashv \cdot [y_0: x_0]\, \varphi(u) \neq 0$ that there exists $q \in X_r^*$ such that $\langle \varphi(u \cdot q), \overline{[y_0: x_0]} \rangle = 1$. Therefore we find $f: x_0 \to y_0 uq$ in $\mathscr{F}$. We decompose $u = u_1 \cdot u_2$ such that $u_1 = 1$ for $|u| \leqslant k$ and $|u_2| = k$ in the other cases. Now let g be the uniquely determined $y_0 u_1$—left prime factor of f. G is $LL(k)$ and therefore g is uniquely determined by x_0 and $y_0 u$. Therefore in $\dashv \cdot [y_0: x_0]\, \varphi(u_1)$ there exists exactly one monom w which will not be made to be 0 by multiplication with $\varphi(u_2)$. Therefore we have $\dashv \cdot [y_0: x_0]\, \varphi(u) = w \cdot \varphi(u_2)$, as the lemma claims.

From this directly follows

THEOREM 10. *The word problem* $w \in L(G)$, $G \in LL(k)$ *can be decided in linear space and linear time by multiplying out* $\$ \cdot \varphi(u)$ *sequentially from left to right.*

The method described in this theorem applied even to $LR(k)$ languages would generally lead to exponentially growing space complexity.

The converse of Theorem 10 is not true. There exist c.f. grammars G for non-deterministic languages such that their word problem can be decided by sequentially multiplying out from left to right in linear space and linear time.

DEFINITION. We call this class of c.f. languages $SMLR(N)$ iff

$$|\$ \cdot \varphi(u)| \leqslant N \qquad \text{for all} \quad u \in T^*.$$

Because of Lemma 15 and these remarks the following theorem is obvious:

THEOREM 11. (1) *The word problem for $SMLR(N)$ can be decided in linear time and linear space.*

(2) $LL(k) \subset SMLR(m^{k+3})$

(3) $SLLR = \bigcup_N SMLR(N)$ *is closed under* "$\bigcup$".

OPEN PROBLEMS. 1. Is it decidable for a context-free grammar G whether $G \in SMLR(N)$ for fixed N?

2. Is it decidable whether $L(G) = L(G')$ for $G, G' \in SMLR(N)$?

This section shows that by our theory we get a purely algebraic definition of the $LL(k)$ languages. In the next section we will show that this remains true for $LR(k)$ languages.

THEOREM 12. $\varphi \in SMLR(N)$ *is recursively undecidable.*

Proof. We show that this question can be reduced to the post correspondence problem [20]. Let $(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n) \in X^* \times X^*$. The correspondence problem is to determine whether or not there exists a sequence of natural numbers $i_1, i_2, \ldots, i_m \in \{1, \ldots, n\}$ such that

$$\alpha_{i_1} \cdots \alpha_{i_m} = \beta_{i_1} \cdots \beta_{i_m}.$$

Let S, S_1, S_2, A, B be new symbols, i.e., symbols not in X. We form the polynomials

$$p_i = \bar{A}\alpha_i A + \bar{B}\beta_i B, \qquad p_i' = \bar{A}\alpha_i + \bar{B}\beta_i$$

for $i = 1, \ldots, n$ and

$$q_j = \bar{x}_j \qquad \text{for} \quad x_j \in X,$$

$$r = \bar{S}(S_1 + S_2).$$

We ask where there exists a product

$$f \in \left\{ p_i, p_i', q_j, r \;\middle|\; \begin{matrix} i = 1, \ldots, n \\ j = 1, \ldots, m \end{matrix} \right\}^*$$

such that

$$|S(A+B)f| \geqslant 2.$$

Obviously this holds iff the correspondence problem has a solution.

In the *SMLR*-case the monomials of p are of length $\leqslant 3$. One reduces the general case to this special case by decomposing

$$p_i = p_{i,1} \cdots p_{i,l_i}$$

where the $p_{i,1_y}$ have degree $\leqslant 3$. Let

$$q = \bar{A}a_1 \cdots a_1 A + \bar{B}b_1 \cdots b_r B \qquad \text{and} \qquad a_i, b_y \in X, l \geqslant r.$$

Let $A_1, A_2, \ldots, A_{l-1}$ and $B_1, \ldots, B_{l-1}$ be new symbols. We define

$$\begin{aligned} q_1 &= \bar{A}a_1 A_1 + \bar{B}b_1 B_1, \\ q_i &= \bar{A}_{i-1} a_i A_i + \bar{B}_{i-1} b_i B_i \qquad \text{for} \quad i = 2, \ldots, l-1, \\ q_l &= \bar{A}_{l-1} a_l A + \bar{B}_{l-1} b_l B. \end{aligned}$$

Where $b_{r+1} = \cdots = b_1 = 1$. We see that $q = q_1 \cdot q_2 \cdot \cdots \cdot q_l$. By applying this decomposition to each p_i, p_i', $i = 1, \ldots, n$ with sets of new variables whose intersection is pairwise disjoint we get a reduction that shows that

$$|S(A+B)f| \geqslant 2$$

remains undecidable even if we restrict our question to the case degree $(p_i) \leqslant 3$. It remains open whether there exists a grammar G such that φ_G defines our polynomials.

$LR(k)$-Grammars

Here we derive similar results as in the foregoing section. The only difference comes in by the substitution of $R\langle X^{(*)}\rangle$ by $\mathscr{A}_R(G)$ mod $\mathfrak{A}_r(L)$ in the characterization of $LR(k)$. We get a first information by the following

Lemma 16. *Let G be a $LR(k)$-grammar and $u \in T^*$. If $u = \tilde{u}_1 + \cdots + \tilde{u}_m \ \mathfrak{A}_r(L))$ with $\tilde{u}_i \neq_r O(L)$, then $m \leqslant (|\bar{X}| + 1)^k$ holds.*

Proof. From $\tilde{u}_i \neq_r O(L)$ it follows that there exists v such that $u \cdot v \in L(G)$. Let be $f: S_r \to u \cdot v$ the derivation of $u \cdot v$ from S_r. Then $u \cdot \mathrm{First}_k(v)$ determines uniquely an v-leftbase g of f. Then $Q(g) = \tilde{u}_i$. This means that $u\ \mathrm{First}_k(v)$ uniquely determines the index i by the condition $\tilde{u}_i$. $\mathrm{First}_k(v) \neq_r O$. Now there are at most $(|\bar{X}| + 1)^k$ words f' of length $\leqslant k$, which select an index i by the condition $\tilde{u}_i \cdot v' \neq_r O(L)$. Therefore $m \leqslant (|\bar{X}| + 1)^k$ as claimed by the lemma.

This lemma does not yet characterize $LR(k)$-grammars. But going a second time through the proof of Lemma 16, we see that $\tilde{u}_1,\ldots,\tilde{u}_m$ have a common prefix, which is uniquely determined by u. We see this from the decomposition $u=u_1\cdot u_2$ such that $|u_2|=k$. Therefore one implication of the following theorem is true.

THEOREM 13. *The c.f. grammar G is of type $LR(k)$ iff for each $u\in T^*$: $u=\tilde{u}\cdot p$, $\tilde{u}\in X_1^*$, $p=\tilde{u}_1+\cdots+\tilde{u}_{m_{\mathfrak{A}}}\tilde{u}_i X_1^*\cup\{S_R\}$ and $|\tilde{u}_i|\leqslant k$.*

To prove this theorem completely it is sufficient to show that the word problem $w\in L(G)$ can be decided by a deterministic pda. We will not prove this here because it is a simple consequence of the following theorem, which is concerned with a more general class of c.f. languages. We generalize $LR(k)$ as before $LL(k)$ in the following

DEFINITION. The c.f. grammar G is in the class $BSLR(N)$ iff for all $u\in T^*$

$$u=\alpha_1\cdot\tilde{u}_1+\cdots+\alpha_m\cdot\tilde{u}_m(\mathfrak{a}_r(L)),\qquad L=L(G),\ \tilde{u}_i\in X_1^*\cup\{S_R\},\ \alpha_i\cdot\in R$$

$R=\mathbb{N}$ implies that

$$\sum_{i=1}^{m}\|\alpha_i\|\leqslant N,\ \|\alpha_i\|=\begin{cases}1 & \text{if } \alpha_i\neq 0\\ 0 & \text{if } \alpha_i=0.\end{cases}$$

The letters BS come from bounded size and LR from the use of the right congruence $=_r(L)$.

THEOREM 14. *The word problem $w\in L(G)$ for $G\in BSLR(N)$ can be decided sequentially in time $O(|w|)$.*

Proof. First we give the idea of the proof. For each of the words u_i we have to compute $\$\varphi(\tilde{u}_i)$ to decide $\tilde{u}_i=_r O(L)$. This computation can be done sequentially because $\tilde{u}_i\in X_1^*$. But to compute $\$\varphi(u_i\cdot\eta(t))$ is more difficult because $\tilde{u}_i\cdot z$ can produce several words in X_1^* which are of very different length. This could lead to a n^2 algorithm. We overcome this difficulty by computing for each prefix v of u_i all possible results of $v\cdot z$ for $z\in X_r$ in advance. It will happen in these computations that we get the same word u_i in different ways. Therefore we have to check this or use a data structure which makes this checking superfluous.

To prove our theorem we define two new functions. For $f\in R\langle Z^{(*)}\rangle$ we define

$$\operatorname{suffix}(f)=\{z\in Z\mid\exists_v\langle f, vz\rangle\neq 0\}.$$

To each $x\in X_1$ we assign a mapping $\psi(x)\colon 2^Z\to 2^Z$ by

$$\psi(x)(z)=\{y\in Z\mid\exists_v\langle\varphi(x),\bar{z}vy\rangle\neq 0\},$$

and

$$\psi(x)(Z')=\bigcup_{z\in Z'}\psi(x)(z)\qquad\text{for}\quad Z'\subset Z.$$

It follows immediately that

$$\text{suffix}(\$\varphi(ux)) = \psi(x)(\text{suffix}(\$\varphi(u)).$$

We use this property to compute $u \cdot t = \tilde{v}_1 + \cdots + \tilde{v}_n(\mathfrak{A}_r)$. It holds in $\mathscr{A}_R(G)$ for $\tilde{u}_i = u_i' \cdot x_i$ that

$$u \cdot t = \sum_{i=1}^{m} \tilde{u}_i \cdot \eta(t) = \sum_i \sum_{z \in \bar{X}} \tilde{u}_i \cdot z \cdot \alpha_z^t$$

$$= \sum_i \sum_{z \in X_l} \tilde{u}_i \cdot z \cdot \alpha_z^t + \sum_i \sum_{\substack{z \in X_r \\ y \in \bar{X}}} u_i' y \alpha_{x_i, z}^y$$

$$= \sum_{\substack{i \\ \psi(z)\ (\text{suffix}(\tilde{u}_i)) \neq \varnothing}} \sum_{z \in X_l} \tilde{u}_i \cdot z \cdot \alpha_z^t + \sum_i \sum_{\substack{z \in X_r \\ y \in \bar{X}}} u_i' \cdot y \alpha_{x_i \cdot z}^y.$$

This relation is recursive because the second sum is of the same kind as the whole sum. The recursion could run $|u|$ steps, which would lead to a $|u|^2$ algorithm. To use this relation more efficiently, we construct a tree-like data structure which represents $\tilde{u}_1 + \cdots + \tilde{u}_m$ by a tree and which contains feedback edges to shorten the recursion.

DEFINITION. The tree $T(u)$. $T(u)$ is an oriented tree. The root of the tree is 1.The other vertices of the tree are $\{v \mid \exists_i v \text{ prefix of } \tilde{u}_i\}$. The set of edges is defined by

$$\{(v, x) \mid vx \text{ prefix of } \tilde{u}_i, x \in X_l\}.$$

v is the initial endpoint of (v, x) and vx the terminal endpoint of (v, x). We label the vertices of $T(u)$ by

$$\mu(v) = \text{suffix}(\$\varphi(v)).$$

From our recursive relation it follows that

$$\mu(vx) = \psi(x)(\mu(v)).$$

This means that μ can be sequentially computed on the tree.

Now we introduce backward edges in $T(u)$. There exists a backward edge from v_1 to v_2 iff

$$v_2 \text{ is prefix of } v_1, \qquad v_1 \neq v_2,$$

and if

$$v_1 = v_2 \cdot v \text{ then there exist } x \in X_r \text{ and } z \in X_l$$
$$\text{such that } \langle vx, z \rangle \neq 0.$$

We denote this edge by (v_1, v_2, x, z). v_1 is the initial endpoint and v_2 the terminal endpoint and $\langle x, z \rangle$ is the "label" of (v_1, v_2, x, z). The number of backward edges from v_2 is bounded by $|X_r| \cdot N$; otherwise we get a contradiction to the assumption $G \in BSLR(N)$. We have $u \in L(G)$ iff S_r is vertex in $T(u)$. To prove our theorem it is therefore sufficient to show that $T(u \cdot t)$ can be constructed in constant time from $T(u)$. To this end we look at the vertex u_i:

(a) Let $\langle \eta(t), z \rangle \neq 0$ and $z \in X_1$. By computing $\psi(z)\, \mu(u_i)$ we decide whether (u_i, z) is an edge in $T(u \cdot t)$. The time for this computation depends only on G, not on $|ut|$.

(b) Let $\langle \eta(t), z \rangle \neq 0$ and $z \in X_r$. We examine the backward edges from u_i whether there are some with the label $\langle x, z \rangle$. If (v_1, v_2, x, z) is a backward edge then (v_2, x) is an edge in $T(u \cdot t)$. We have to consider at most

$$|X_r| \cdot |X_1| \cdot N^2$$

edges. This number again depends only on G.

(c) We have to compute the new backward edges for $T(u \cdot t)$. Let vx be a new vertex in $T(ut)$. Then for all $y \in X_r$ we compute

$$vxy = v \cdot \sum_{z \in X} \alpha^z_{x,y} \cdot z.$$

This we can do as before under (b) by using the backward edges from v. Again we need not more than $N^2 \cdot |X|^2$ steps.

(d) It is not necessary to delete the edges of $T(y)$ which do not appear in $T(ut)$ explicity by keeping a list of the "leaves" of the tree. Notice that "leaves" here mean the vertices, representing one of the u_i.

Generalizations to Non-c.f. Languages

So far we did the first steps in developing our theory for the case of c.f. languages. But the restriction to the c.f. case from our point of view is quite unnatural. We only used very special residual classes $a + R\langle Z^{(*)} \rangle$ in the representations of our languages. It is natural to allow this residual classes not only for monomials but for any polynomial $q \in R\langle Z^{(*)} \rangle$. We assume $R = \mathbb{N}$ or $R = \mathbb{B}$. One easily proves that for each monoid homomorphism

$$h\colon T^* \to R\langle Z^{(*)} \rangle$$

and $a \in Z^{(*)}$ the language

$$L = h^{-1}(a + R\langle Z^{(*)} \rangle)$$

is c.f. We now assume $a_1, a_2 \in Z^{(*)}$ and

$$h_i: T^* \to R\langle Z^{(*)}\rangle, \qquad i=1,2$$

to be monoid homomorphisms $Z \geqslant 2$, then

$$L_i = h_i^{-1}(a_i + R\langle Z^{(*)}\rangle), \qquad i=1,2$$

are c.f. languages. We assume $y_1, y_2 \notin Z$ and form $Y = Z \cup \{Y_1, Y_2\}$. Then

$$h = \bar{y}_1 h_1 y_1 + \bar{y}_2 h_2 y_2$$

defines a monoid homomorphism from T^* in $R\langle Y^{(*)}\rangle$. We easily see

$$L_1 \cap L_2 = h^{-1}(y_1 a_1 y_1 + y_2 a_2 y_2 + R\langle Y^{(*)}\rangle).$$

By standard coding tricks we transform h into a monoid homomorphism g from T^* into $R\langle Z^{(*)}\rangle$. With suitable monomials b_1, b_2 we have

$$L_1 \cap L_2 = g^{-1}(b_1 + b_2 + R\langle Z^{(*)}\rangle).$$

We generalize this result easily to the case of the intersection of k c.f. languages:

THEOREM 15. *Let Z be an alphabet with two or more elements. For each $k \in \mathbb{N}$ we find a polynomial $q \in R\langle Z^{(*)}\rangle$ with k monomials such that holds: For each set k c.f. languages $L_1, L_2, \ldots, L_k$ and $L = L_1 \cap L_2 \cap \cdots \cap L_k$ there exists a monoid homomorphism*

$$h: T^* \to R\langle Z^{(*)}\rangle \qquad \textit{such that } L = h^{-1}(q + R\langle Z^{(*)}\rangle.$$

With standard tricks as in the proof of the hardest language theorem of Greibach for c.f. languages we construct from Theorem 15 for each k a hardest language in the intersection closure of k c.f. languages under homomorphic reduction.

ACKNOWLEDGMENTS

The author wishes to thank J. Berstel, M. Harrison, C. Reutenauer, and J. Sakarovitch for helpful discussions. The comments of B. Becker and Thomas Kretschmer concerning an earlier version of this paper led to improvements of several proofs.

REFERENCES

1. D. BENSON, Syntax and semantix: A categorial view, *Inform. Control* **17** (1970), 145–160.
2. N. CHOMSKY AND M. P. SCHÜTZENBERGER, The algebraic theory of contextfree languages *in* "Computer Programming and Formal Systems," (P. Braffort and S. Hirschberg, Eds.), pp. 116–161, North-Holland, Amsterdam, 1970.

3. S. A. GREIBACH, The hardest context-free language, *SIAM J. Comp.* **2** (1970).
4. S. A. GREIBACH, Erasable context-free languages, *Inform. Control* **4** (1975).
5. M. A. HARRISON, "Introduction to Formal Language Theory," Addison-Wesley, Reading, Mass., 1978.
6. P. M. LEWIS, AND R. E. STEARNS, Syntax-directed transduction, *J. Assoc. Comput. Mach.* **15** (1968), 465–488.
7. G. HOTZ, Eine Algebraisierung des Syntheseproblemes von Schaltkreisen, *Elektron. Informationsverarb. Kybernet.* (1965), 185–231.
8. G. HOTZ, Der Satz von Chomsky-Schützenberger und die schwerste kontextfreie Sprache von S. Greibach, *Astérisque* **38–39** (1976), 105–115.
9. G. HOTZ, Normal-form transformations of context-free grammars, *Acta Cybernet.* **4**, Fasc. 1 (1978), 65–84.
10. G. HOTZ, A representation theorem for infinite dimensional associative algebras and applications in language theory," *in* "Proceedings, 9ième École de Printemps d'informatique théorétique," Murol, 1981.
11. G. HOTZ, About the Universality of the Ring $R\langle X^{(*)}\rangle$, *in* "Actes du Séminaire d'Informatique Théoretique Université Paris VI et VII," pp. 203–218, 1981–1982.
12. G. HOTZ AND K. ESTENFELD, "Formale Sprachen," Bibliograph Inst., Mannheim, 1981.
13. D. E. KNUTH, On the translation of languages from left to right, *Inform. Control* **8** (1965), 607–639.
14. D. E. KNUTH, Top-down-syntax analysis, *Acta Inform.* **1** (1971), 79–110.
15. M. NIVAT, Transductions des Languages des Chomsky, *Ann. Inst. Fourier* **18** (1968).
16. J. F. PERROT, "Contribution à l'étude des monoides syntactiques et des Certains groupes associés aux automates finis," Thèse Sc. Math., Univ. Paris VI, 1972.
17. C. REUTENAUER, "Series rationalles et algèbres syntactiques," Thèse de Doctorat d'État à Université Paris VI, pp. 1–209, 1980.
18. A. SALOMAA AND M. SOITTOLA, "Automata-Theoretic Aspects of Formal Power Series," Springer-Verlag, Berlin/New York, 1978.
19. E. SHAMIR, A representation theorem for algebraic and context-free power series in non-computing variables, *Inform. Control* **11** (1967), 239–254.
20. E. L. POST, Formal reductions of the general combinatorial decision problem, *Amer. J. Math.* **65** (1943), 197–215.