# Normal-form transformations of context-free grammars

By G. Hotz

To the memory of Professor László Kalmár

## Introduction

Each context-free grammar $G$ can be transformed into a Chomsky-normalform (CNF) and into a Greibach-normalform (GNF) without changing the languages generated by the grammars. Our interest does not concern the invariance of the languages under such transformations but the ambiguity of the grammars, the multiplicity of words relative to the grammars and relations between pairs of grammars. Syntactical transformations of languages are induced by the grammars. Therefore, it should be of interest, if certain syntactical transformations between languages transform in a natural manner with the normal form transformations. The role of monoid homomorphisms in connection with rational transformation is played by functors between the syntactical categories of grammars in connection with tree transformations.

In this paper we define three different transformations $\tau_1$, $\tau_2$ and $\tau_3$ of grammars in CNF into GNF. $\tau_1$ produces productions with one terminal and at most two non-terminals in the range of the productions. $\tau_2$ and $\tau_3$ generate productions $p$ with maximally two resp. three non-terminals and one terminal on each side of the range $(p)$.

$\tau_1$ has been considered for the first time in a technical report 1967 by S. Greibach. One finds it again in [Gr] (1975). Implicitely the construction is contained in [Ho 2] (1974) too. $\tau_2$ and $\tau_3$ seem to be studied here the first time.

Geller, Harrison and Havel showed in [Ge—Ha], that for each $LR(k)$ language there exist a $LR(k')$ grammar in GNF with $k'=k$ for $k\geqq 1$ and that there exist $LR(0)$ languages for which one has always $k'\geqq 1$. But they did not use the simple transformation $\tau_1$.

We show that $\tau_1$, $\tau_2$ and $\tau_3$ preserve unambiguity and do not increase multiplicities. But there exist grammars for which the multiplicity decreases. Non $LR(k)$ grammars may be transformed into $LR(k)$ grammars.

We show that functors between the syntactical categories of the grammars $G_1$ and $G_2$ are transformed into functors between the syntactical categories between the grammars $\tau_1(G_1)$ and $\tau_1(G_2)$.

With the same methods we show in a following paper, that $\tau_1$ preserves $LL(k)$ for all $k$ and $LR(k)$ for $k \geqq 1$ and that $LR(0)$ is transformed into $LR(1)$. The proofs for both properties are nearly identical. From this paper we use the unambiguity lemma for the existence of well formed decompositions of morphisms (classes of derivations) in products of $(t, 1)$-prime derivations. $\tau_2$ and $\tau_3$ may destroy the $LR$ and $LL$ properties. This means that transformations inverse to $\tau_2$ and $\tau_3$ may eventually transform non $LR(k)$ grammars into such grammars.

Because until now we do not know much about transformations which transform certain grammars of $LR(k)$ languages into $LR(k)$ grammars the relations $\tau_2^{-1}, \tau_3^{-1}$ may be of interest.

For certain transformations from general context-free grammars into CNF the $LR$-invariance has been showed by [Be] (1976) and [Sch] (1973).

We use the notation of $x$-categories or syntactical categories as defined in [Ho—Cl]. An introduction in related questions the reader may find in [A—Ull] or [Sa].

## Definitions and preliminaries

In the following $T$ is the terminal and $Z$ the variable alphabet, and $S$ is the axiom of the context-free grammar $G$. We assume that the set $P$ of productions of $G$ is in Chomsky normal form. This means that for $f \in P$ we have

$$f = (z, z_1 z_2) \quad \text{or} \quad f = (z, t),$$

where, as always in this paper, $z, z_1, z_2$ are in $Z$ and $t$ in $T$. We assign to $G$ the free $x$-category $F(G)$; that means that we wish to calculate with derivations of $G$, or — more precisely formulated — we wish to calculate with the classes of inessentially different derivations of $G$. We write

$$w \xrightarrow{f} u \quad \text{and} \quad D(f) = w, \quad C(f) = u$$

if $f$ is a derivation class from $w$ to $u$. $w$ is the domain and $u$ the codomain of $f$.

From

$$w \xrightarrow{f} u \quad \text{and} \quad w' \xrightarrow{g} u'$$

we from

$$ww' \xrightarrow{f \times g} uu',$$

the class of derivations we get from $f$ and $g$ by doing the derivations $f$ and $g$ in parallel.

We form

$$\begin{tikzcd} u \arrow[r, "f"] \arrow[dr, dashed, "h"'] & v \arrow[d, "g"] \\ & w \end{tikzcd}$$

$$h = g \circ f$$

by executing first $f$ and then $g$ if $C(f)=D(g)$.

For $F(G)$ we also write $F(P)$ where $P$ is the production set of $G$, and we write

$$w \xrightarrow[P]{} v$$

if there exists $f\in F(P)$ such that

$$w \xrightarrow{f} v$$

holds.

Now we study as in [Ho 2] special derivations which are related to canonical derivations of words $uw$ with $u\in T^*$ and $w\in Z^*$. These special derivations will be used to construct the productions in our normal form grammars.

**Definition.** A derivation $f$ in Chomsky normalform

$$z \xrightarrow{f} uwv,\ u\in T^*,\ v\in T^*,\ w\in Z^*$$

is called *$(u, v)$-prime* if from

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

it follows that $g=1_w$.

This means that $f$ is $(u, v)$-prime if $f$ is a shortest derivation which generates from $z$ a word which begins with the terminal symbols $u$ and ends with the terminal symbols $v$ and has only nonterminal symbols (possibly none) between.

As we will see later, of special interest are the cases

1. $u=1,\ v\in T$,
2. $u\in T,\ v=1$,
3. $u\in T,\ v\in T$.

Let

$$B(z, u, v) = \{w\in Z^* |\ \text{there exists } z \xrightarrow{f} uwv,\ f\ (u, v)\text{-prime}\}.$$

In [Ho 2] we showed that $B(z, u, 1)$ is a regular set for all $u\in T^*$. By symmetry arguments it follows that $B(z, 1, v)$, too, is a regular set for $v\in T^*$.

For $f$ $(u, v)$-prime $u, v\in T$ we have a decomposition

$$f=(1_u\times 1_w\times g)\circ h, \quad w\in Z^*$$

such that $h$ is $(u, 1)$-prime and $g$ is $(1, v)$-prime.

On the other hand $f$ is $(u, v)$-prime for all $(u, 1)$-prime $h$ and $(1, v)$-prime $g$.

We define for $L\subset Z^*$ and $x\in Z^*$

$$L_x=\{w\in Z^* | wx\in L\}$$

and

$${}_xL=\{w\in Z^* | xw\in L\}.$$

With this notation we have

$$B(z, t, r)=\bigcup_{y\in Z}[B(z, t, 1)]_y B(y, 1, r) \quad \text{for} \quad t, r\in T.$$

Now, the relation $w_1\equiv w_2 \Leftrightarrow L_{w_1}=L_{w_2}$ is the well known syntactical congruence (i.e., left invariant equivalence relation). For regular sets $L$ there are only a finite number of these congruence classes where each class is also a regular set. From this we conclude that $[B(z, t, 1)]_y$ is a regular set and thus that $B(z, t, r)$ is regular for all $t, r \in T \cup \{1\}$.

**Lemma 1.** The set

$$B = \{{}_x[B(z, t, r)]_y | z \in Z,\ x \in Z^*,\ y \in Z^*\}$$

is a finite set of regular sets for all $t, r \in T \cup \{1\}$.

*Proof.* We know from the above discussion that $B(z, t, r)$ is regular. We know also that

$$A = \{[B(z, t, r)]_y | y \in Z^*,\ z \in Z\}$$

is a finite set of regular sets.

Now as one sees immediately

$$[{}_pL_q]_s = {}_pL_{sq}, \quad {}_s[{}_pL_q] = {}_{ps}L_q.$$

Therefore we conclude from the finiteness of the set $A$, that $B$ is also finite and from the regularity of the elements of $A$, that the elements of $B$ are regular. This finishes our proof.

**Lemma 2.** For

$$u, v \in T^*, \quad z \in Z \quad \text{and} \quad t, r \in T$$

it follows that

$$B(z, ut, rv) = \underbrace{\qquad}_{x, y \in Z} B(x, t, 1)_x[B(z, u, v)]_y B(y, 1, r)$$

$$\cup \underbrace{\qquad\qquad}_{x \in B(z, u, v) \cap Z} B(x, t, r).$$

$B(z, ut, rv)$, then, is regular.

*Proof.* Let $f$ be a $(ut, rv)$-prime derivation with $D(f) = z$. Then $f$ can be decomposed into

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

such that $h$ is $(u, v)$-prime.

Now we discuss the two cases corresponding to $D(g) \in Z^m$ for $m \geqq 2$ and $D(g) \in Z$; these are the only two possibilities, since $G$ is in Chomsky normal form.

1. For this case $g$ can be decomposed into

$$g = g_1 \times 1_w \times g_2.$$

Here $g_1$ is $(t, 1)$-prime and $g_2$ is $(1, r)$-prime. Otherwise $f$ would not be $(ut, rv)$-prime. Let $D(g_1) = x$, $D(g_2) = y$. Then we have

$$C(h) = uxwyv, \quad \text{where} \quad xwy \in B(z, u, v).$$

Further, let $C(g_1) = tw_1$ and $C(g_2) = w_2r$. Then we have $w_1 \in B(x, t, 1)$, $w_2 \in B(y, 1, r)$ and for $C(f) = ut\tilde{w}rv$ the following holds:

$$\tilde{w} = w_1ww_2 \in B(x, t, 1) \cdot {}_x[B(z, u, v)]_y \cdot B(y, 1, r).$$

2. For this case $g$ can be decomposed into

$$g = (1_t \times 1_{w_1} \times g_2) \circ g_1$$

with $g_1(t, 1)$-prime and $g_2(1, r)$-prime. Then for $C(g_1) = tw_1y$, $C(g_2) = w_2r$, $C(f) = utwrv$ and $C(h) = uxv$ we have for $y \in Z$

$$w = w_1 \cdot w_2 \in [B(x, t, 1)]_y \cdot B(y, 1, r) \subset B(x, t, r).$$

From case 1 and case 2 we conclude that the left part of the equation in our lemma is contained in the right part. This completes the proof in one direction.

The inclusion in the other direction follows directly from the following facts. For $g_1(t, 1)$-prime and $g_2(1, r)$-prime and $h(u, v)$-prime, then if the product

$$f = (1_u \times g_1 \times 1_w \times g_2 \times 1_v) \circ h$$

is defined where $w \in Z^*$, $f$ is $(ut, rv)$-prime. This means that each $B(x, t, 1)\ [B(z, u, v)]_y \cdot B(y, 1, r)$ is contained in $B(z, ut, rv)$. For $x \in B(z, u, v)$ it is clear that $B(x, t, r) \subset B(z, ut, rv)$. This completes the proof of the lemma.

This lemma nearly gives us a recursive equation for calculating the sets $B(z, u, v)$. The importance of these sets follows from the obvious

**Theorem 1.**

$$w = u \cdot v \in L \Leftrightarrow 1 \in B(S, u, v).$$

This means that the word problem $w \in L$ can be reduced to the problem of whether 1 is in a regular set. We are here not interested in developing this direction further, however. For our purposes of constructing a normal form grammar we do not need a complete recursive definition of the $B(z, u, v)$.

To construct the productions of our normal form grammars we will use the $(u, v)$-prime derivations of the free $x$-category $F(G)$ for the special case that $u, v \in T$ rather than $T^*$. If, for example, $S \xrightarrow{f} uwv$ is such a $(u, v)$-prime derivation, we could include a production of the form $S \to uRv$ in one of our normal form production systems, $\tilde{P}$, where $R = B(S, u, v)$. Then for any such new variable $R$ we would also have to introduce productions of the form

$$R \to t \cdot B(R, t, r) \cdot r$$

into $\tilde{P}$ representing the class of all $(t, r)$-prime derivations from the set $R \subset Z^+$ in $P$. Here $B(R, t, r)$ is a simple extension of the definition of $B(z, t, r)$:

$$B(R, t, r) = \{\tilde{w} \mid w \xrightarrow{f} t\tilde{w}r \text{ is } (t, r)\text{-prime and } w \in R\}.$$

To see that this process of constructing productions for $\tilde{P}$ can be continued with the $B(R, t, r)$ sets we give the following lemma which one can easily prove.

**Lemma 3.** For $t, r \in T$ and $R \subset Z^+$

$$B(R, t, r) = \underbrace{\qquad}_{x, y \in Z} B(x, t, 1)_x [R]_y B(y, 1, r)$$

$$\cup \underbrace{\qquad}_{x \in R \cap Z} B(x, t, r).$$

This lemma gives us a way to factor the set $B(R, t, r)$ into the regular sets we introduced earlier. Thus we are now able to generate all codomain sets of derivations $f\in F(G)$, where each such $f$ is a product in "$\circ$" and "$\times$" of prime derivations, from the domain sets

$$_p[B(z, t, r)]_q, \quad t, r\in T\cup\{1\}, \quad t\cdot r\neq 1$$

where length $(p)=$length $(q)$ for $p, q\in Z^*$. From Lemma 1 it follows that $_p[B(z, t, r)]_q$ is a finite set of regular sets. More precisely formulated, from Lemma 3 we can select the following classes of derivations from $F(P)$ for all $p, q\in Z^*$, length $(p)=$length $(q)$ and $x, y\in Z$, $u, v\in T$, and $t, r\in T\cup\{1\}$, $t\cdot r\neq 1$:

$$S\to t\cdot B(S, t, r)\cdot r, \tag{$P'1$}$$

$$_p[B(z, t, r)]_q\to u\cdot B(x, u, 1)\cdot {}_{px}[B(z, t, r)]_{yq}\cdot B(y, 1, v)\cdot v, \tag{$P'2$}$$

$$_p[B(z, t, r)]_q\to u\cdot B(x, u, v)\cdot v, \quad x\in {}_p[B(z, t, r)]_q\cap Z. \tag{$P'3$}$$

Each of the classes $(P'1)$, $(P'2)$ and $(P'3)$ represents an entire set of derivations generated from the choices of $p, q, r, t, u, v, w, x$, and $y$. Clearly, many of these choices will lead to empty sets. However, it is evident that each of these classes is finite (since $B(z, t, r)$ is a regular set, the congruence relation established by $_p[B(z, t, r)]_q$ for all $p, q\in Z^*$ is of finite index). Therefore, we can use these derivations as the basis for constructing the productions $\tilde{P}$ of our normal form grammars. Before constructing such a normal form production system, however, we must convince ourselves that *every* derivation class in $F(P)$ can be decomposed as above.

## Well formed decompositions of derivations

Now we consider normal forms of derivations for any context-free grammar $G$ in Chomsky normal form using the $x$-categorical expressions which define derivations. We show that each class $f$ of derivations has exactly one normal form derivation which we will call *well formed* (w.f.).

**Definition.** A decomposition $f=f_n\circ\ldots\circ f_1$ with $D(f)\in Z$ and

$$f_i=f_{i,1}\times\ldots\times f_{i,m_i} \quad \text{for} \quad i=1, \ldots, n$$

is *well formed* if conditions $(W1)$, $(W2)$ and $(W3)$ hold.

See Fig. 1 for $(W1)$ and Fig. 2 for $(W2)$.

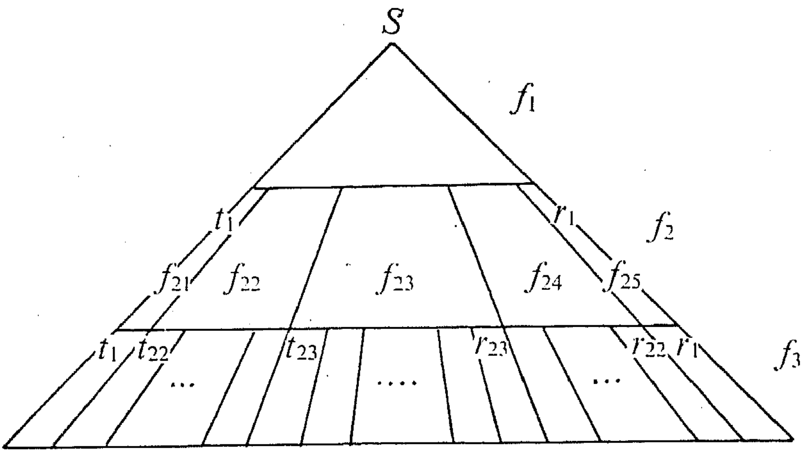
$(W1)$ $D(f_{i,l})\in Z\cup T$.
If $D(f_{i,l})=t\in T$ then $f_{i,l}=1_t$.
If $D(f_{i,l})\in Z$ then $f_{i,l}$ is $(t, r)$-prime with $t, r\in T\cup\{1\}$ and $t\cdot r\neq 1$.

$(W2)$ Let

$$f_{i+1}\circ f_i=F_1\times\ldots\times F_{m_l}$$

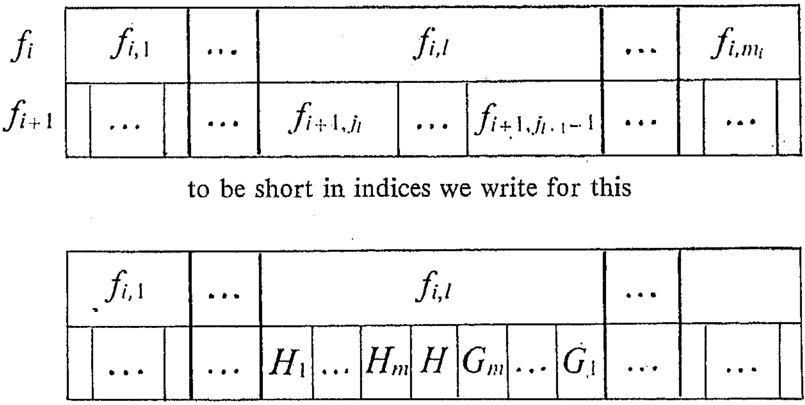
be the uniquely determined decomposition with $D(F_i)\in Z\cup T$ for $i=1, \ldots, m_l$, and

$$F_l=(H_1\times\ldots\times H_m\times H\times G_m\times\ldots\times G_1)\circ f_{i,l}$$



$f_{21} = 1_{t_1}$, $f_{25} = 1_{r_1}$,
$f_{22}$ $(t_{22}, 1)$-prime, $f_{24}$ $(1, r_{22})$-prime, $f_{23}$ $(t_{23}, r_{23})$-prime

*Fig. 1*



to be short in indices we write for this

with $H=1$ for $j_{l+1}-j_l=o(2)$

*Fig. 2*

be the uniquely defined decomposition with $D(H_i)\in Z\cup T$, $D(G_i)\in Z\cup T$, $D(H)\in Z\cup\{1\}$ and $f_{i,l}(t_1, r_1)$-prime. Then it follows that $H_1=1_{t_1}$, $G_1=1_{r_1}$ for $t_1, r_1\in T\cup\{1\}$, $t_1\cdot r_1\neq 1$ and

$$H_i \text{ is } (t_i, 1)\text{-prime}, \quad t_i\in T,$$

$$G_i \text{ is } (1, r_i)\text{-prime}, \quad r_i\in T$$

for $i=2, \dots, m$, and $H=1$ for length $(C(f_{i,l}))$ even; for length $(C(f_{i,l}))$ odd $H\in P$ (the set of productions of the underlying grammar) with $C(H)\in T^+$, or $H$ is $(t_{m+1}, r_{m+1})$-prime with $t_{m+1}, r_{m+1}\in T$.

$(W3)$ $f_1\in P$ and $f_1$ is terminal, or $f_1$ is $(t, r)$-prime with $t, r\in T$.

**Lemma 4.** Let $F(P)$ be the free $x$-category generated by the context-free production system $P$ in Chomsky normal form. For each $f\in F(P)$ there exists exactly one (w.f.)-decomposition of $f$ if $D(f)\in Z$ and $C(f)\in T^+$.

*Proof.* Let

$$z\xrightarrow{f} w, \quad z\in Z, \quad w\in T^+.$$

If $f\in P$, then $f$ is terminal and $f$ is a unique (w.f.)-decomposition of itself.

Now assume $f\notin P$. Then we can write $w=tw'r$ with $t, r\in T$ and $w'$ uniquely determined by $t$ and $r$. Therefore there exists a unique decomposition

$$f=(1_t\times h\times 1_r)\circ f_1$$

such that $f_1$ is $(t, r)$-prime. We decompose

$$h=H_2\times\dots\times H_k\times H\times G_k\times\dots\times G_2$$

such that $D(H_i)\in Z$, $D(G_i)\in Z$ and $H=1$ or $D(H)\in Z$ for $k\geqq 2$.

Again this decomposition is unique, if it is possible. If it is not possible to decompose $h$ in this way then $h=1$ and $f=f_1$; that is, one has

$$f=f_1=(g_1\times g_2)\circ g_3$$

with

$$z\xrightarrow{g_3} z_1 z_2, \quad z_1\xrightarrow{g_1} t, \quad z_2\xrightarrow{g_2} r$$

productions in $P$, and our lemma holds in this case.

Now with $h$ decomposed as shown,

$$z_i\xrightarrow{H_i} w_i, \quad y_i\xrightarrow{G_i} v_i, \quad w_i, v_i\in T^+$$

for $i=2, \dots, k$ and

$$x\xrightarrow{H} u, \quad u\in T^+$$

for $H\neq 1$.

For $i=2, \dots, k$ we can decompose the derivations uniquely as

$$H_i=(1_{t_i}\times h_i)\circ f_{2,i},$$

$$G_{k+2-i}=(g_i\times 1_{r_i})\circ f_{2,k+i}$$

in the case that $H=1$. Here $f_{2,i}$ is $(t_i, 1)$-prime and $f_{2,k+i}$ is $(1, r_i)$-prime. In the case $H\neq 1$ we have the same decomposition for $H_2, \dots, H_k$ but

$$H = (1_t \times h' \times 1_{r'}) \circ f_{2,k+1}, \quad t', r' \in T$$

and

$$G_{k+2-i} = (g_i \times 1_{r_i}) \circ f_{2,k+1+i}$$

for $i=2, \dots, k$, where $f_{2,k+1}$ is $(t', r')$-prime and $f_{2,k+1+i}$ is $(1, r_i)$-prime.

We now iterate this construction by applying it to the $h_i$, $g_i$ and $h'$ in the same way as we did to $h$, and so on.

After a finite number of steps we get the uniquely determined (w.f.)-decomposition of $f$.

## The first normal form transformation

Using the result of lemma 4 we now derive a production system from the relations $(P'1)$, $(P'2)$ and $(P'3)$.

We write

$$B \to B'$$

for $B, B' \subset (Z \cup T)^*$ iff for each $w \in B'$ there exists a $u \in B$ such that $u \to w$ in the usual sense holds. It follows directly that

$$B \to \{u\}$$

for $u \in B$. For simplicity we identify $u$ with $\{u\}$. Using this relation and the transitive closure property of derivations one has

$${}_p[B(z, t, r)]_q \to s$$

for $x \in {}_p[B(z, t, r)]_q$ and $x \to s$.

Let $V'$ be an alphabet and $v$ a mapping into $V'$ which is defined on

$$U = \{(z, t, r, p, q) | z \in Z;\ t, r \in T \cup \{1\},\ t \cdot r \neq 1,\ p, q \in Z^*, \text{length}(p) = \text{length}(q)\}$$

such that

$$v(z, t, r, p, q) = v(z', t'\ r', p', q')$$

iff

$${}_p[B(z, t, r)]_q = {}_{p'}[B(z', t', r')]_{q'}.$$

From lemma 1 we know that such an alphabet $V'$ and such a mapping $v$ can be constructed effectively and that $V'$ is finite. Let $\tilde{V} = V' \cup \{S\}$ be the nonterminal alphabet of our normal form. For $v(z, t, r, 1, 1)$ we write simply $v(z, t, r)$.

Using $(P', 1)$ we construct the productions $(\tilde{P}, 1)$ as
$(\tilde{P}, 1)$ $S \to u$ for $u \in T \cup T^2$ and $S \overset{*}{\to} u$.

$$S \to t \cdot v(S, t, r) \cdot r \quad \text{for} \quad B(S, t, r) \neq \emptyset.$$

We define $(\tilde{P}, 2)$ as follows.

$(\tilde{P}, 21)$ $\quad v(z, t, r, p\ \ q) \to u \cdot v(x, u, 1) \cdot v(z, t, r, px, yq) \cdot v(y, 1, w) \cdot w$

for $\quad B(x, u, 1) \cap Z^+ \neq \emptyset, \quad B(y, 1, w) \cap Z^+ \neq \emptyset, \quad {}_{px}[B(z, t, r)]_{yv} \cap Z^+ \neq \emptyset,$

$(\tilde{P}, 22)$ $v(z, t, r, p, q) \to u \cdot v(x, u, 1) \cdot v(y, 1, w) \cdot w$
for $1 \in {}_{px}[B(z, t, r)]_{yq}$, $B(x, u, 1) \cap Z^+ \neq \emptyset$, $B(y, 1, w) \cap Z^+ \neq \emptyset$,

$(\tilde{P}, 23)$ $v(z, t, r, p, q) \to u \cdot v(z, t, r, px, yq) \cdot v(y, 1, w) \cdot w$
for $1 \in B(x, u, 1)$, ${}_{px}[B(z, t, r)]_{yq} \cap Z^+ \neq \emptyset$, $B(y, 1, w) \cap Z^+ \neq \emptyset$,

$(\tilde{P}, 24)$ $v(z, t, r, p, q) \to u \cdot v(x, u, 1) \cdot v(z, t, r, px, yq) \cdot w$
for $1 \in B(y, 1, w)$, $B(x, u, 1) \cap Z^+ \neq \emptyset$, ${}_{px}[B(z, t\ r)]_{yq} \cap Z^+ \neq \emptyset$,

$(\tilde{P}, 25)$ $v(z, t, r, p, q) \to u \cdot v(x, u, 1) \cdot w$
for $1 \in {}_{px}[B(z, t, r)]_{yq} \cdot B(y, 1, w)$, $B(x, u, 1) \cap Z^+ \neq \emptyset$,

$(\tilde{P}, 26)$ $v(z, t, r, p, q) \to u \cdot v(z, t, r, px, yq) \cdot w$
for $1 \in B(x, u, 1) \cdot B(y, 1, w)$, ${}_{px}[B(z, t, r)]_{yq} \cap Z^+ \neq \emptyset$,

$(\tilde{P}, 27)$ $v(z, t, r, u, v) \to u \cdot v(y, 1, w) \cdot w$
for $1 \in B(x, u, 1) \cdot {}_{px}[B(z, t, r)]_{yq}$, $B(y, 1, w) \cap Z^+ \neq \emptyset$,

$(\tilde{P}, 28)$ $v(z, t, r, p, q) \to u \cdot w$
for $1 \in B(x, u, 1) \cdot {}_{px}[B(z, t, r)]_{yq} \cdot B(y, 1, w)$.

We set

$$(\tilde{P}, 2) = \bigcup_{i=1}^{8} (\tilde{P}, 2i).$$

We now define the productions $(\tilde{P}, 3)$.

$(\tilde{P}, 3)$ $v(z, t, r, p, q) \to u \cdot v(x, u, w) \cdot w$
for $x \in {}_{p}[B(z, t, r)]_{q} \cap Z$ and $B(x, u, w) \cap Z^+ \neq \emptyset$,
$v(z, t, r, p, q) \to u \cdot v$
for $x \in {}_{p}[B(z, t, r)]_{q} \cap Z$, $1 \in B(x, u, w)$,
$v(z, t, r, p, q) \to u$
for $x \in {}_{p}[B(z, t, r)]_{q} \cap Z$, $(x, u) \in P$.

We define

$$\tilde{P} = (\tilde{P}, 1) \cup (\tilde{P}, 2) \cup (\tilde{P}, 3)$$

and

$$\tilde{G} = (\tilde{V} \cup T,\ T,\ \tilde{P},\ S).$$

We write

$$\tilde{G} = \tau_3(G).$$

$\tau_3$ is our first normal form transformation.

Let

$$L = L(G), \quad \tilde{L} = L(\tilde{G})$$

be the languages generated by the grammars $G$ and $\tilde{G}$, respectively.

**Lemma 5.**

$$\tilde{L} \subset L.$$

*Proof.* We construct a functor from the free $x$-category $F(\tilde{P})$ into the monoidal category of the relations

$$B \to B' \quad \text{for} \quad B, B' \subset (Z \cup T)^*$$

which are induced by the production set $P$.

Let $\mathfrak{A}$ be the power set of $(Z \cup T)^*$; then we define the monoid homomorphism

$$\varphi_1 \colon (\tilde{V} \cup T)^* \to \mathfrak{A}$$

by setting

$$\varphi_1(t) = \{t\} \quad \text{for} \quad t \in T,$$

$$\varphi_1(S) = \{S\},$$

$$\varphi_1(v(z, t, r, p, q)) = {}_p[B(z, t, r)]_q$$

$$\text{for} \quad v(z, t, r, p, q) \in \tilde{V}.$$

We will write $t$ for $\{t\}$ and $S$ for $\{S\}$. For each $f \in \tilde{P}$ we define

$$\varphi_2'(f) = (\varphi_1(D(f)), \varphi_1(C(f))).$$

One can easily check that for $f \in \tilde{P}$

$$\varphi_1(D(f)) \to \varphi_1(C(f)).$$

We extend $(\varphi_1, \varphi_2')$ to the functor $\varphi = (\varphi_1, \varphi_2)$ which is determined uniquely by $(\varphi_1, \varphi_2')$. We have then for $S \xrightarrow{f} w$, $w \in T^*$

$$S = \varphi_1(S) \xrightarrow{\varphi_2(f)} \varphi_1(w) = w,$$

and therefore from the definition of $B \to B'$ for sets we have

$$S \to w$$

in the usual sense.

This means that $w \in L$ for all $w \in \tilde{L}$, and thus $\tilde{L} \subset L$.

**Lemma 6.**

$$L \subset \tilde{L}.$$

This lemma will be proved in two parts.

*Part 1.* A derivation step $f_s \colon Z^+ \to (Z \cup T)^+$ is called a *w.f. derivation step* iff $f_s = H_2 \times \ldots \times H_m \times H \times G_m \times \ldots \times G_2$ is a decomposition of $f_s$ into prime derivations in the usual sense (e.g. see ($W2$)). How, let $w_1 \ldots w_n \in B(p, t, r)$ for $w_i, p \in Z$ and $t, r \in T \cup \{1\}$ such that $t \cdot r \neq 1$, and let $f_s$ be a w.f. derivation step with $D(f_s) = w_1 \ldots w_n$. Then we can construct $\tilde{f}_s$ with $D(\tilde{f}_s) = v(p, t, r)$ such that

$$\tilde{\varphi}(C(\tilde{f}_s)) = \varphi(C(f_s))$$

where $\tilde{\varphi}$ and $\varphi$ are two homomorphisms which forget the nonterminals in a string and are constant on terminals.

To show this we must examine the two cases for $n$ even and $n$ odd. For $n$ even we have

$$w_1 \dots w_n = x_1 \dots x_m y_m \dots y_1$$

where $m=n/2$, $x_i=w_i$, $y_i=w_{n-i+1}$, and $n \geqq 2$. For $n$ odd we have similarly for $m \geqq 1$

$$w_1 \dots w_n = x_1 \dots x_m z y_m \dots y_1.$$

Since the proof for these two cases is similar, we will show the result only for the case that $n$ is odd.

Here we have for $w_1 \dots w_n \in B(p, t, r)$

$$w_1 \dots w_n = x_1 \dots x_m z y_m \dots y_1 \xrightarrow{f_s = H_1 \times \dots \times H_m \times H \times G_m \times \dots \times G_1}$$

$$t_1 x_{1_1} \dots x_{1_{n_1}} \dots t_m x_{m_1} \dots x_{m_{n_m}} t_{m+1} z_1 \dots z_j r_{m+1} y_{m_1} \dots y_{m_{l_m}} r_m \dots y_{1_1} \dots y_{1_{l_1}} r_1.$$

We then construct $\tilde{f}_s$ as shown below for $v(p, t, r)$, the variable corresponding to $B(p, t, r)$, using the rules $\tilde{P}$.

$$v(p, t, r) \xrightarrow{\tilde{f}_{s,1}}$$

$$t_1 v(x_1, t_1, 1) v(p, t_1, r_1, x_1, y_1) v(y_1, 1, r_1) r_1 \xrightarrow{\tilde{f}_{s,2}} \dots \xrightarrow{\tilde{f}_{s,m}}$$

$$t_1 v(x_1, t_1, 1) \dots t_m v(x_m, t_m, 1) v(p, t_1, r_1, x_1 \dots x_m, y_m \dots y_1)$$

$$v(y_m, 1, r_m) r_m \dots v(y_1, 1, r_1) r_1 \xrightarrow{f_{s,m+1}}$$

$$t_1 v(x_1, t_1, 1) \dots t_m v(x_m, t_m, 1) t_{m+1} v(z, t_{m+1}, r_{m+1}) r_{m+1}$$

$$v(y_m, 1, r_m) r_m \dots v(y_1, 1, r_1) r_1.$$

Now, set $\tilde{f}_s = \tilde{f}_{s,m+1} \circ \tilde{f}_{s,m} \circ \dots \circ \tilde{f}_{s,1}$.

Clearly $\tilde{\varphi}(C(\tilde{f}_s)) = \varphi(C(f))$. Further, for each string of "isolated" nonterminals in $C(f_s)$ (a string of nonterminals with a terminal on both ends) there is a corresponding $v$ variable in $C(\tilde{f}_s)$ in the same location. For example, for $t_k x_{k,1} \dots \dots x_{k,n_k}$ in $C(f_s)$, where $x_{k,1} \dots x_{k,n_k} \in B(x_k, t_k, 1)$, we have $t_k v(x_k, t_k, 1)$ in $C(\tilde{f}_s)$ in the same location in terms of the terminal symbols in $C(f_s)$ and $C(\tilde{f}_s)$.

If one of the $x_i$, $z$, or $y_i$ — for example $x_i$ — is rewritten to a single terminal followed by no nonterminal string, this corresponds to the fact the $1 \in B(x_i, t_i, 1)$ and thus that the corresponding $v$ variable also does not appear in $C(\tilde{f}_s)$. Therefore the isolated nonterminal strings and the $v$ variables correspond exactly. Using these results the lemma can now be easily proved.

*Part 2.* For $w \in L$ and $S \xrightarrow{f} w$ with $f \in F(P)$, let

$$f = f_n \circ \dots \circ f_1$$

be the unique (w.f.)-decomposition of $f$. Then we can construct $\tilde{f} \in F(\tilde{P})$ such that $S \xrightarrow{\tilde{f}} w$.

We will prove this inductively by showing that for each $f_1, f_2, \ldots, f_n$ we can find $\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_n$ such that $\varphi(C(f_i))=\tilde{\varphi}(C(\tilde{f}_i))$ for all $i$, where $\tilde{f}=\tilde{f}_n \circ \ldots \circ \tilde{f}_1$, and such that the isolated variable strings in $C(f_i)$ correspond exactly to the $v$ variables in $C(\tilde{f}_i)$.

For $f_1$ we have

$$S \xrightarrow{f_1} tx_1 \ldots x_m r$$

and

$$S \xrightarrow{\tilde{f}_1} tv(S, t, r)r,$$

which clearly satisfies our conditions (if $x_1 \ldots x_m$ is empty in $C(f_1)$, $v(S, t, r)$ does not appear in $C(\tilde{f}_1)$).

Assume that this is true for $f_k \circ \ldots \circ f_1$ and $\tilde{f}_k \circ \ldots \circ \tilde{f}_1$ for $n > k > 1$ to show for the case $k+1$, we look at the partial derivation $f_{k+1} \circ f_k \circ \ldots \circ f_1$. We know from the induction hypothesis that $\tilde{\varphi}(C(\tilde{f}_k))=\varphi(C(f_k))$ and further that the isolated nonterminal strings in $C(f_k)$ correspond exactly with the $v$ variables in $C(\tilde{f}_k)$.

From what we proved in Part 1, then, the result should be clear. To each terminal in $C(f_k)$ we apply an identity derivation; to each string of isolated nonterminals in $C(f_k)$ we apply a w.f. derivation step $f_s$. The "$\times$" product of these identity derivations and w.f. derivation steps forms $f_{k+1}$ as one can easily see from the proof of Lemma 4 and figure 1.

Let

$$f_{k+1} = g_1 \times \ldots \times g_m$$

be this product. Then we construct

$$\tilde{f}_{k+1} = \tilde{g}_1 \times \ldots \times \tilde{g}_m$$

where $\tilde{g}_j$ is the identity derivation if $g_j$ is the identity, and $\tilde{g}_j$ is the corresponding $\tilde{f}_s$ derivation if $g_j$ is a "type $f_s$" derivation. Clearly, then, the conditions of our assumptions hold, and we have that $C(f)=C(\tilde{f})$ and thus that $L \subset \tilde{L}$.

Our proof gives a sharper result than stated in the lemma. $f \to \tilde{f}$ is a mapping. If this mapping is surjective, then the multiplicity of each element $w \in L$ relative to $G$ will not be greater relative to $\tilde{G}$.

Analysing the proof of Lemma 5, one also sees that given $\tilde{f} \in F(\tilde{P})$ one can find a $g \in F(P)$ such that $\tilde{g}=\tilde{f}$.

From Lemma 5 and Lemma 6 and the above remark we have

**Theorem 2.** For each language $L$ generated by a context-free grammar $G$ our transformation $\tau_3$ produces a context-free grammar $\tilde{G}=\tau_3(G)$ which generates $L$ and in which the productions are of the form

$$z \to tpr \quad \text{or} \quad z \to v$$

where $p \in Z \cup Z^2 \cup Z^3$, $v \in T \cup T^2$ and $t, r \in T$, $z \in Z$. $\tau_3$ does not increase the multiplicity of words.

**Corollary.** $\tau_3$ transforms unambiguous grammars into unambiguous grammars.

We now define two more transformations $\tau_1$ and $\tau_2$ for which the same theorems hold.

$\tau_1$ is the transformation into Greibach normal form from Chomsky normal form with at most two nonterminals in the right hand side of each production.

$\tau_2$ transforms each $G$ in Chomsky normal form into a grammar $\tilde{G}$ in which the productions are of the form

$$z \to tqr \quad \text{or} \quad z \to v$$

with $q \in Z \cup Z^2$, $v \in T \cup T^2$, $z \in Z$, $t, r \in T$.

We will only give the relations corresponding to $(P', 1)$, $(P', 2)$ and $(P', 3)$. From this the definitions of $\tau_1$ and $\tau_2$ and the proofs of the corresponding theorems will be obvious to the reader.

## The transformation $\tau_1$

Now we apply the methods which led us to the just proved normal form theorem to the recursive equation of theorem 2 in [Ho 2].

$$B(s, ut, 1) = \bigcup_{z \in Z} B(z, t, 1)_z[B(s, u, 1)].$$

Again, we can try to construct productions from these $B$ sets of the form

$$B(s, v, 1) \to tB(s, vt, 1)$$

for $s \in Z$, $v \in T^+$, $t \in T$. Factoring the right hand side, we have

$$B(s, v, 1) \to t \cdot B(z, t, 1) \cdot {}_z[B(s, v, 1)]$$

for $t \in T$, $z \in Z$, $v \in T$, $s \in Z$.

We introduce as before variables $x(z, t, p)$ which we assign to ${}_p[B(z, t, 1)]$. Here we get a production system

$$x(s, v, p) \to t \cdot x(z, t, 1) \cdot x(s, v, pz)$$

for $B(z, t, 1) \neq \emptyset$, ${}_{pz}[B(s, v, 1)] \neq \emptyset$, and $B(z, t, 1)$ and ${}_{pz}[B(s, 0, 1)] \neq \{1\}$ and

$$x(s, v, p) \to t \cdot x(z, t, 1),$$
$$x(s, v, p) \to t \cdot x(s, v, pz)$$

for $1 \in_{pz}[B(s, v, 1)]$ or $1 \in B(z, t, 1)$, respectively.

We define the first and the terminal productions as follows:

$$S \to t \cdot x(S, t, 1) \quad \text{for} \quad B(S, t, 1) \neq \emptyset;$$
$$S \to t \quad \text{for} \quad (S, t) \in P;$$
$$x(z, t, p) \to r \quad \text{for} \quad y \in_p [B(z, t, 1)] \quad \text{and} \quad (y, r) \in P.$$

This grammar we call $\tilde{G}$ and the transformation from $G$ to $\tilde{G}$ is the desired transformation $\tau_1$.

As in the case of $\tau_3$ one proves

**Theorem 3.** The transformation $\tau_1$ transforms context-free grammars $G$ in Chomsky normal form into grammars $\tilde{G} = \tau_1(G)$ which are in Greibach normal form. The productions of $\tilde{G}$ contain on the right hand side not more than two variables. The transformation $\tau_1$ does not increase the multiplicity of words.

## The transformation $\tau_2$

Let

$$\mathfrak{R} = \{{}_u[B(z, t, r)]_v | z \in Z,\ t \in T \cup \{1\},\ r \in T \cup \{1\},\ t \cdot r \neq 1,\ u \in Z^*,\ v \in Z^*\}.$$

As we have seen $\mathfrak{R}$ is a finite set. We derived from relations of the form

$$R \to tR_1 R_2 R_3 r$$

and

$$S \to tRr$$

a cubic normal form for the context-free languages.

Now from relations of a similar form

$$[R_1 R_2] \to tB(z, t, 1) \cdot {}_z[R_1 R_2]_y \cdot B(y, 1, r) \cdot r$$

we can derive a quadratic normal form from the fact that

$${}_z[R_1 R_2]_y = {}_z[R_1] \cdot [R_2]_y \cup \sigma(R_1) \cdot {}_z[R_2]_y \cup \sigma(R_2) \cdot {}_z[R_1]_y$$

where

$$\sigma(R) = \begin{cases} \emptyset & \text{for} \quad 1 \in R \\ \{1\} & \text{for} \quad 1 \in R. \end{cases}$$

If we now write

$$R_0 = B(z, t, 1), \quad R_1' = {}_z[R_1], \quad R_2' = [R_2]_y, \quad R_3 = B(y, 1, r),$$
$$\tilde{R}_1 = \sigma(R_2) \cdot {}_z[R_1]_y, \quad \tilde{R}_2 = \sigma(R_1) \cdot {}_z[R_2]_y$$

then we have the relations

$$[R_1 R_2] \to t \cdot [R_0 R_1'] \cdot [R_2' R_3] \cdot r,$$
$$[R_1 R_2] \to t \cdot [R_0 \tilde{R}_1] \cdot [R_3] \cdot r,$$
$$[R_1 R_2] \to t \cdot [R_0] \cdot [\tilde{R}_2 R_3] \cdot r.$$

$\mathfrak{R}$, the set of all valid $R$ sets, is closed under left and right divisions by construction, and from $\mathfrak{R}$ finite it follows that $\mathfrak{R} \cup \mathfrak{R} \cdot \mathfrak{R}$ is also finite.

If we now choose variables $v(R_1)$ and $v(R_2, R_2)$ for $R_1, R_2 \in \mathfrak{R}$ just as we did in developing our cubic normal form we get a production system $\tilde{P}$ of the type

$$y \to txzr,$$
$$y \to txr,$$
$$y \to tr,$$
$$y \to t,$$

or $y, x, z$ nonterminals and $t, r$ terminals.

This is the transformation $\tau_2$.

As in the cubic case we have the following

**Theorem 4.** The normal form transformation

$$G \xrightarrow{\tau_2} \tilde{G}$$

defined for grammars in Chomsky normal form has the property $L(G)=L(\tilde{G})$. The transformation does not increase the multiplicity of the words $w \in L$.

The proof is completely analogous to the proof of theorem 2 and is therefore left to the reader.

## Functorial properties of the normal form transformations

Let $F(P_i)=((Z_i \cup T)^*, \mathfrak{M}_i, D, C)$ for $i=1,2$ be two $x$-categories generated by the context-free production systems $P_1$ and $P_2$ in Chomsky normal form. Further let $\varphi=(\varphi_1, \varphi_2)$ be an $x$-functor from $F(P_1)$ to $F(P_2)$.

This means that

$$\varphi_1 \colon (Z_1 \cup T)^* \to (Z_2 \cup T)^*$$

is a monoid homomorphism and that

$$\varphi_2 \colon \mathfrak{M}_1 \to \mathfrak{M}_2$$

fulfills

$$\varphi_2(f \circ g) = \varphi_2(f) \circ \varphi_2(g)$$

if $f \circ g$ is defined, and

$$\varphi_2(f \times g) = \varphi_2(f) \times \varphi_2(g).$$

Also for identities $1_w$ we have

$$\varphi_2(1_w) = 1_{\varphi_1(w)}.$$

We restrict ourselves to the case $\varphi_1(T) \subset T$ and $\varphi_1(Z_1^*) \subset Z_2^*$. From this follows

$$\text{length}\,(w) = \text{length}\,(\varphi_1(w)) \quad \text{for} \quad w \in T^*.$$

We have no derivation

$$1 \to u$$

for $u \neq 1$. Because we are in Chomsky normal form we have no superfluous variables. This means for each $z \in Z$ there exists

$$z \xrightarrow{f} w, \ w \in T^+;$$

therefore $\varphi_1(z)=1$ would be a contradiction. From this and the fact that $\varphi_1$ is length preserving on $T^*$ it also follows that $\varphi_1(Z_1) \subset Z_2$. Thus, since we are using Chomsky normal form, we have

$$\varphi_1(P_1) \subset P_2.$$

Now let

$$z \xrightarrow{f} tur$$

be a $(t, r)$-prime derivation. Then

$$\varphi_1(z) \xrightarrow{\varphi_2(f)} \varphi_1(t)\,\varphi_1(u)\,\varphi_1(r)$$

is $(\varphi_1(t), \varphi_1(r))$-prime. Therefore

$$\varphi_1(B(z, t, r)) \subset B(\varphi_1(z), \varphi_1(t), \varphi_1(r)).$$

For $R \subset Z^*$, $x, y \in Z$ the identity

$$\varphi_1({}_x[R]_y) = {}_{\varphi_1(x)}[\varphi_1(R)]_{\varphi_1(y)}$$

holds since $\varphi_1(Z_1) \subset Z_2$.

Let $\mathfrak{R}_i$ be the set of our sets ${}_p[B(z, t, r)]_q \neq \emptyset$ belonging to $G_1$ and $G_2$ respectively. Then for the variables $v(z, t, r, p, q)$ we can write $v(R)$ for certain $R \in \mathfrak{R}_i$. Then we have for the set $\tilde{Z}_i$ of variables of $\tilde{G}_i$

$$\tilde{Z}_i = \{v(R) | R \in \mathfrak{R}_i\}, \quad i = 1, 2.$$

Now $\varphi_1$ induces a mapping

$$\varphi_1 \colon \mathfrak{R}_1 \to \mathfrak{R}_2.$$

Using this we can define the monoid homomorphism

$$\tilde{\varphi}_1 \colon (\tilde{Z}_1 \cup T)^* \to (\tilde{Z}_2 \cup T)^*$$

by setting

$$\tilde{\varphi}_1(t) = t \quad \text{for} \quad t \in T$$

and

$$\tilde{\varphi}_1(v(R)) = v(\varphi_1(R)) \quad \text{for} \quad R \in \mathfrak{R}_1.$$

It is clear, then, that the following diagram commutes

$$\begin{tikzcd}
O(F(P_1)) \arrow[r, "\varphi_1"] \arrow[d, "\tau"'] & O(F(P_2)) \arrow[d, "\tau"] \\
O(F(\tilde{P}_1)) \arrow[r, "\tilde{\varphi}_1"] & O(F(\tilde{P}_2))
\end{tikzcd}$$

for $\tau = \tau_1, \tau_2, \tau_3$, where 0 is the object set of the given categories.

We can now define the function $\varphi_2'$ which maps the productions of $\tilde{P}_1$ to productions in $\tilde{P}_2$ by setting

$$\varphi_2'(z, q) := (\tilde{\varphi}_1(z), \tilde{\varphi}_1(q))$$

for $(z, q) \in \tilde{P}_1$.

Extending $(\tilde{\varphi}_1, \varphi_2')$ to the $x$-functor $(\tilde{\varphi}_1, \tilde{\varphi}_2)$ we have proved the following

**Theorem 5.** Let $\tau$ be one of our normal form transformations $\tau_1, \tau_2, \tau_3$ and let $\varphi = (\varphi_1, \varphi_2)$ be a functor from $F(P_1)$ to $F(P_2)$, where $P_1$ and $P_2$ are in Chomsky normal form with $\varphi_1(T) \subset T$ and $\varphi_1(Z_1^*) \subset Z_2^*$. Then there exists a natural transformation of $\varphi$ to a functor $\tilde{\varphi}$ from $F(\tau(P_1))$ to $F(\tau(P_2))$.

The theorem states in other words that the diagramm

$$\begin{tikzcd}
F(P_1) \arrow[r, "\varphi"] \arrow[d, "\tau"'] & F(P_2) \arrow[d, "\tau"] \\
F(\tilde{P}_1) \arrow[r, dotted] & F(\tilde{P}_2)
\end{tikzcd}$$

has a solution $\tilde{\varphi}$. This means that the $\tau_i$ induce a functor between the functor categories of the $x$-categories $F(P)$, $P$ in Chomsky normal form, and the functor categories $F(\tilde{P})$ with $\tilde{P}$ in one of the three normal forms 1, 2 or 3.

## Transformations of linear languages

We have seen that the transformations $\tau_i$ do not increase the multiplicity of words. Therefore the question arises whether an $LR(k)$-grammar $G$ is transformed into $LR(k')$-grammar $\tilde{G}$ by our transformations $\tau_i$. We are not able to solve this problem here, but we show that $\tau_1$ transforms one sided linear grammars into minimal linear grammars. This means that in this case $\tau_1$ transforms non-$LR(k)$-grammars into $LR(0)$-grammars. $\tau_1$ here corresponds to the reduction of finite automata.

Let $P$ be a left-linear grammar where productions are of the type

$$z \to z' \cdot t, \quad z \to t$$

for $z, z' \in Z$ and $t \in T$, where $Z$ is the variable alphabet, and $T$ is the terminal alphabet. We transform these productions into Chomsky normal form by introducing the variable alphabet $X = \{(x, t) \mid t \in T\}$ where $x$ is a fixed symbol.

We define

$$P_C = \{(z, z' \cdot (x, t)) | (z, z' \cdot t) \in P\}$$

$$\cup \{((x, t), t) | t \in T\} \cup \{(z, t) | (z, t) \in P\}.$$

$P_C$, then, is in Chomsky normal form and the grammars $G = (Z \cup T, T, P, S)$ and $G' = (Z \cup X \cup T, T, P_C, S)$ generate the same language $L$.

Now we apply our transformation $\tau_1$ to $P_C$. We have for $z \in Z$ and $(x, t) \in X$

$$B(z, t, 1) \subset X^*$$

and

$$B((x, t), u, 1) \subset \{1\}.$$

From this follows

$$_z[B(y, t, 1)] = \emptyset$$

for $z \in Z$ and $y \in Z \cup X$.

Therefore our relations which define $\tilde{P}_C$ have the form

$$_p[B(y', t, 1)] \to u \cdot {}_{py}[B(y', t, 1)]$$

for $p \in X^*$, $y = (x, u) \in X$, and $y' \in Z$.

Now let

$$\varphi \colon X^* \to T^*$$

be the monoid isomorphism defined by

$$\varphi(x, t) = t.$$

Then

$$\varphi({}_p[B(y, t, 1)]), \quad y \in Z, \quad t \in T, \quad p \in X^*$$

defines the syntactical congruence classes of $L$ (i.e. the left invariant equivalence relations). This means that $\tau_1$ transforms $P_C$ into a minimal grammar for $L$.

We therefore have the following

**Theorem 6.** $\tau_1$ transforms left linear grammars — represented in Chomsky normal form as shown — into minimal right linear grammars.

**Corollary.** $\tau_1$ transforms certain non-$LR(k)$-grammars into $LR(0)$-grammars. There exist grammars such that under the transformation $\tau_1$ the multiplicity of words decreases properly.

One can easily prove similar results for the transformations $\tau_2$ and $\tau_3$.

From our theorem about the multiplicity of words it follows that the transformations $\tau_i$ transform an $LR(k)$-grammar $G$ into an unambiguous grammar $G$. $\tau_2$ and $\tau_3$ do not preserve the $LL(k)$ and $LR(k)$ property of grammars, but $\tau_1$ does preserve it as we can show [Ho 3].

## A normal form for the Chomsky—Schützenberger theorem

Using our normal form transformations $\tau_2$ and $\tau_3$ one easily derives a normal form for the theorem of Chomsky—Schützenberger.

Let

$$X_k = \{x_1, \dots, x_k, x_1^{-1}, \dots, x_k^{-1}\}$$

where $x_i, x_i^{-1}$ are bracket pairs and $D_k$ the corresponding Dyck language over $X_k$. The well known theorem states that for each context-free language $L \subset T^*$ there exists an alphabet $X_k$, a standard regular event $R$, and a homomorphism $\varphi\colon X_k^* \to T^*$ with $\varphi(X_k) \subset T \cup \{1\}$ such that

$$L = \varphi(D_k \cap R).$$

Using our normal forms and following the well known proof of this theorem one finds the normal form of

**Theorem 7.** For each context-free language $L \subset T^*$ one can find $X_k$, $\varphi$, and $R$ such that $L = \varphi(D_k \cap R)$ and from $\varphi(w) \in T$ and the existence of $u, v$ such that $uwv \in R$ it follows length $(w) \leqq 3$.

From this theorem we arrive at the theorem of S. Greibach [Gr] about a hardest context-free language as it was proved in [Ho 1].

## Abstract

We discuss three normal form transformations $\tau_1$, $\tau_2$ and $\tau_3$ of grammars $G$ which are in Chomsky normal form into grammars $G_1$, $G_2$ and $G_3$ respectively. $G_1$ is in Greibach normal form with nonterminal productions restricted to $z \to tp$ such that $t \in T$ and $p \in Z^+$ and length $(p) \leqq 2$. The nonterminal productions of $G_2$ and $G_3$ are of the form $z \to tpr$ such that $t, r \in T$ and $p \in Z^+$, length $(p) \leqq 2$ or length $(p) \leqq 3$, respectively. It is shown that these transformations do not increase the multiplicity of words in the generated languages. Furthermore we show that certain functorial relations between languages are preserved under these transformations. The restriction of $\tau_1$ to one sided linear grammars produces the minimal grammars. $\tau_2$ and $\tau_3$ do not preserve the $LR(k)$ property of grammars. $\tau_1$ preserves $LL(k)$ for $k \geqq 0$ and $LR(k)$ for $k > 1$, $LR(0)$ may be transformed into $LR(1)$ as we show in the following paper.

UNIVERSITÄT DES SAARLANDES
D-66 SAARBRÜCKEN

## References

[A—ULL] AHO, V. and J. D. ULLMAN, *The theory of parsing, translation, compiling,* vol. 1.

[BE] BENSON, D. B., *Some properties of normal form grammars,* Comp. Sc. Dept. Washington State Univ., TR CS 75 24, July 1976.

[CH—SCH] CHOMSKY, N. and M. P. SCHÜTZENBERGER, *The algebraic theory of context-free languages, computer programming and formal systems,* North-Holland Publ., 1970, pp. 116—161.

[GE—HA] GELLER, M. M., M. A. HARRISON, I. M. HAVEL, *Normal forms of deterministic grammars,* Discrete Math.

[GR] GREIBACH, S. A., The hardest context-free language, *SIAM J. Computing,* v. 2, 1973, pp. 304—310.

— Erasable context-free languages, *Information and Control,* v. 4, 1975, pp. 301—326.

[Ho 1] HOTZ, G., The theorem of Chomsky—Schützenberger and the hardest context-free language of S. Greibach, *Astérisque,* v. 38—39, 1976, pp. 105—115.

[Ho 2] HOTZ, G., Sequentielle Analyse kontextfreier Sprachen, *Acta Informatica,* v. 4, 1974, pp. 55—75.

[Ho 3] HOTZ, G., *LL*(*k*)- und *LR*(*k*)-Invarianz von kontextfreien Grammatiken unter einer Transformation auf Greibach-Normalform, *EIK,* No. 1—2, 1979.

[Ho—CL] HOTZ, G., V. CLAUS, Automatentheorie und formale Sprachen III, *BI-Informatik,* Hochschulskripten, Mannheim, 1972.

[KN] KNUTH, D. E., On the translation of languages from left to right, *Information and Control,* v. 8, 1965, pp. 607—639.

[MAU] MAURER, H. A., Theoretische Grundlagen der Programmiersprachen, *BI-Informatik,* Hochschulskripten, Mannheim, 1969.

[SA] SALOMAA, A., *Formal languages,* Academic Press, N. Y and London, 1973.

[SCH] SCHAUERTE, R., *Transformationen von LR*(*k*)*-Grammatiken,* Diplomarbeit, Göttingen, 1973.

*(Received March 22, 1978)*