INFORMATIQUE THÉORIQUE ET APPLICATIONS

R. PARCHMANN
J. DUSKE

The structure of index sets and reduced indexed grammars

Informatique théorique et applications, tome 24, n° 1 (1990), p. 89-104

<http://www.numdam.org/item?id=ITA_1990__24_1_89_0>

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 24, n° 1, 1990, p. 89 à 104)

THE STRUCTURE OF INDEX SETS AND REDUCED INDEXED GRAMMARS (*)

by R. PARCHMANN (1) and J. DUSKE (1)

Communicated by J. BERSTEL

Abstract. – *The set of index words attached to a variable in derivations of indexed grammars is investigated. Using the regularity of these sets it is possible to transform an indexed grammar in a reduced from and to describe the structure of left sentential forms of an indexed grammar.*

Résumé. – *On étudie l'ensemble des mots d'index d'une variable dans les dérivations d'une grammaire d'index. La rationalité de ces ensembles peut être utilisée pour transformer une grammaire d'index en forme réduite, et pour décrire la structure des mots apparaissant dans les dérivations gauches d'une grammaire d'index.*

1. INTRODUCTION

In this paper we will further investigate indexed grammars and languages introduced by Aho [1] as an extension of context-free grammars and languages. This family of languages has many properties of the context-free languages and it is interesting to note how many of them can be carried over and which properties are specific to the family of indexed languages. Standing between the context-free and the context-sensitive family of languages, the indexed languages have much more in common with the context-free family.

One nice property of context-free languages is the fact that they can be generated by reduced grammars, which implies that each sentential form can produce a terminal word. In the case of indexed grammars there is the difficulty that a derivation of a terminal word from a variable depends on an attached index word, which can be arbitrary long. Hence it is important to investigate the structure of index words attached to a variable in various forms of derivations of an indexed grammar.

(*) Received November 1987, final version in June 1988.
(1) Institut für Informatik, Universität Hannover, D-3000 Hannover, West Germany.

Informatique théorique et Applications/Theoretical Informatics and Applications
0988-3754 90/01 89 16/$3.60/

In the second section we will investigate the set $\mathrm{TERM}_G(A)$, for every variable A. This is the set of all index words γ such that $A\gamma$ can produce a terminal word. Furthermore we will investigate the set $\mathrm{INDEX}_G(A)$ of all index words appearing in sentential forms attached to the variable A. We will show that for each indexed grammar G these two sets are regular and effectively constructable. We will relate these results to previous results of [9].

In the following section we will use the regularity of the sets $\mathrm{INDEX}_G(A)$ and $\mathrm{TERM}_G(A)$ for constructing a reduced form of an indexed grammar. The construction is a special case of the more general concept of "regular look-ahead" in [3] or the similar construction of a predicting machine (*see* [5]) or of a grammar transformation for indexed grammars in [8].

The final section gives a result concerning the structure of the variable part of sentential forms obtained by leftmost derivations of indexed grammars. These sentential forms obviously have a close relation to the pushdown lists of an IPDA (indexed pushdown automaton) introduced in [7]. In the context-free case these sets are regular [4], but in the indexed case we will show that the corresponding sets are context-free.

2. INDEX SETS

In this chapter we investigate the set of index words attached to a variable in derivations of an indexed grammar. In particular we will consider derivations starting from the start symbol of a grammar and derivations starting from a variable with attached index words leading to a terminal word.

First let us recall the definition of an indexed grammar as given in [2]:

DEFINITION 2.1: An *indexed grammar* is a 5-tuple $G=(N, T, I, P, S)$ where

(1) N, T, I are finite, pairwise disjoint sets; the sets of *variables*, *terminals*, and *indices* respectively;

(2) P is a finite set of pairs (Af, Θ), $A\in N$, $f\in I\cup\{e\}$, $\Theta\in(NI^*\cup T)^*$, the set of *productions;* $(A f, \Theta)$ is denoted by $A f\to\Theta$ and e denotes the empty word;

(3) $S\in N$, the start variable.

Let $\Theta=u_1 B_1 \beta_1 u_2 B_2 \beta_2 \ldots B_n \beta_n u_{n+1}$ with $u_i\in T^*$ for $i\in[1 : n+1]$, $B_j\in N$, and $\beta_j\in I^*$ for $j\in[1 : n]$ with $n\geqq 0$, be an element of $(NI^*\cup T)^*$, and let $\gamma\in I^*$. Then we set

$$\Theta : \gamma=u_1 B_1 \beta_1 \gamma u_2 B_2 \beta_2 \gamma \ldots B_n \beta_n \gamma u_{n+1}.$$

For Θ', $\Theta'' \in (\mathrm{NI}^* \cup \mathrm{T})^*$, we set $\Theta' \Rightarrow \Theta''$ iff $\Theta' = \Theta_1 A f \gamma \Theta_2$, $\Theta'' = \Theta_1 (\Theta : \gamma) \Theta_2$ with $\Theta_1, \Theta_2 \in (\mathrm{NI}^* \cup \mathrm{T})^*$ and $Af \to \Theta \in P$, $f \in I \cup \{e\}$.

$\overset{n}{\Rightarrow}$ is the n-fold product, $\overset{+}{\Rightarrow}$ is the transitive and $\overset{*}{\Rightarrow}$ is the reflexive, transitive closure of $\Rightarrow$.

The language $L(G)$ generated by an indexed grammar $G = (N, T, I, P, S)$ is the set $L(G) = \{ w \mid w \in T^*, S \overset{*}{\Rightarrow} w \}$. A language L is called an indexed language iff $L = L(G)$ for an indexed grammar G.

As for context-free grammars it is desirable to have indexed grammars with the property that each sentential form produces a terminal word. In the context-free case the reduction of a grammar yields a grammar with this property. The reduction process consists of two steps: (1) determination of all variables which produce a terminal word, and (2) determination of all variables which appear in a sentential form.

In the case of an indexed grammar, we have to consider the following difficulties:

(1) the derivation of a terminal word from a variable depends on its attached index word, and

(2) in sentential forms variables appear only with certain attached index words.

Hence we have to consider the following two sets of index words:

DEFINITION 2.2: Let $G = (N, T, I, P, S)$ be an indexed grammar and let $A \in N$. Then

(1) $\mathrm{TERM}_G(A) = \{ \gamma \mid \gamma \in I^*, A\gamma \overset{*}{\Rightarrow} w, w \in T^* \}$, and

(2) $\mathrm{INDEX}_G(A) = \{ \gamma \mid \gamma \in I^*, S \overset{*}{\Rightarrow} \Theta_1 A \gamma \Theta_2, \Theta_1, \Theta_2 \in (NI^* \cup T)^* \}$.

In this section we will show that these sets these sets are regular. For this it is convenient to have a form of an indexed grammar in which in each derivation step the length of an index word can increase or decrease at most by one. To be more precise we want the productions to be of the form $A \to \alpha$, $Af \to B$, or $A \to Bf$, where A, B are variables, f is an index and α is a word consisting of variables and terminals. A grammar of this form will be called *normal form grammar*. (In [1] a similar, but more restricted normal form is defined.)

To this end let $G = (N, T, I, P, S)$ be given. Construct the indexed grammar $G' = (N', T, I, P', S)$ in the following way:

Let $\pi : Af \to u_1 B_1 \gamma_1 u_2 B_2 \gamma_2 \ldots u_n B_n \gamma_n u_{n+1}$ be an arbitrary production in P.

(1) Replace π by $\pi' : A f \to u_1 \hat{B}_1^{(0)} u_2 \hat{B}_2^{(0)} \ldots u_n \hat{B}_n^{(0)} u_{n+1}$ and for all $i \in [1 : n]$ let $\pi_i' : \hat{B}_i^{(0)} \to B_i \gamma_i$.

(2) If $f \neq e$, then replace π' by $A f \to \hat{A}$ and $\hat{A} \to u_1 \hat{B}_1^{(0)} u_2 \hat{B}_2^{(0)} \ldots u_n \hat{B}_n^{(0)} u_{n+1}$.

(3) For each $i \in [1 : n]$ replace π_i' by the productions $\hat{B}_i^{(j)} \to \hat{B}_i^{(j+1)} f_{j+1}$ and $\hat{B}_i^{(n_i)} \to B_i$ with $f_{j+1} \in I$, $j \in [0 : n_i - 1]$ where $\gamma_i = f_{n_i} \ldots f_1$.

N' contains N and it is easy to see that the following lemma and corollary hold for G and G' :

LEMMA 2.3: *Let* $\Theta \in (NI^* \cup T)^*$, $A \in N$, *and* $\gamma \in I^*$. *Then we have* $A\gamma \overset{*}{\Rightarrow} \Theta$ *according to* G *iff* $A\gamma \overset{*}{\Rightarrow} \Theta$ *according to* G'.

This implies $L(G) = L(G')$ and furthermore we have

COROLLARY 2.4: $\mathrm{TERM}_G(A) = \mathrm{TERM}_{G'}(A)$ *for all* $A \in N$.

Since each sentential form according to G' can produce a sentential form according to G we have with Lemma 2.3:

COROLLARY 2.5: $\mathrm{INDEX}_G(A) = \mathrm{INDEX}_{G'}(A)$ *for all* $A \in N$.

Now we will show the regularity of $\mathrm{TERM}_G(A)$ and $\mathrm{INDEX}_G(A)$ with the aid of the Myhill-Nerode theorem (*see* e. g. [5]).

THEOREM 2.6: *Let* $G = (N, T, I, P, S)$ *be an indexed grammar and let* $A \in N$. *Then* $\mathrm{TERM}_G(A)$ *is regular.*

Proof: W.l.o.g. (*see* Corollary 2.4) let G be in normal form. Let $\tau : I^* \to \mathscr{P}(N)$ be defined as $\tau(\gamma) = \{ \mathrm{A} \mid \mathrm{A}\gamma^R \overset{*}{\Rightarrow} w,\ w \in T^* \}$.

[$\mathscr{P}(N)$ denotes the set of all subsets of N and if $\gamma = f_1 \ldots f_n$, then $\gamma^R = f_n \ldots f_1$.] Let R_τ be the following relation over I^*:

$$\text{for all } \gamma_1, \gamma_2 \in \mathrm{I}^* \text{ we have } (\gamma_1, \gamma_2) \in R_\tau \text{ iff } \tau(\gamma_1) = \tau(\gamma_2).$$

R_τ is an equivalence relation with finite index. Furthermore, R_τ is a right congruence. To prove this, let $(\gamma_1, \gamma_2) \in R_\tau$ and let $\gamma \in I^*$. Assume $A \in \tau(\gamma_1 \gamma)$, *i.e.*, there is a derivation $A\gamma^R \gamma_1^R \overset{*}{\Rightarrow} w$, $w \in T^*$. This derivation can be rearranged and then separated in an initial part which uses no indices of γ_1^R and a final part, which uses only indices of γ_1^R, *i.e.*, we have $A\gamma^R \overset{*}{\Rightarrow} w_1 B_1 w_2 \ldots w_k B_k w_{k+1}$ and $B_i \gamma_1^R \overset{*}{\Rightarrow} u_i$ with $B_i \in N$, $u_i \in T^*$, $w_j \in T^*$, $i \in [1 : k]$, $j \in [1 : k+1]$, $w_1 u_1 w_2 \ldots w_k u_k w_{k+1} = w$, and $k \geqq 0$. (Here we use the fact that G is in normal form.)

If $k=0$, then obviously we have $A \in \tau(\gamma_2 \gamma)$. If $k>0$, then we have, using $(\gamma_1, \gamma_2) \in R_\tau$ [*i.e.*, $\tau(\gamma_1)=\tau(\gamma_2)$], $B_i \gamma_2^R \overset{*}{\Rightarrow} u_i'$ with $u_i' \in T^*$, $i \in [1 : k]$.

Therefore there is a derivation

$$A \gamma^R \gamma_2^R \overset{*}{\Rightarrow} w_1 B_1 \gamma_2^R w_2 \ldots w_k B_k \gamma_2^R w_{k+1} \overset{*}{\Rightarrow} w_1 u_1' w_2 \ldots w_k u_k' w_{k+1} = w',$$

and hence $A \in \tau(\gamma_2 \gamma)$.

Finally we have

$$\bigcup_{\gamma \in I^*,\ A \in \tau(\gamma)} [\gamma]_\tau = \{ \gamma' \mid A \gamma'^R \overset{*}{\Rightarrow} w,\ w \in T^* \} = (\mathrm{TERM}_G(A))^R$$

where $[\gamma]_\tau$ denotes the equivalence class of R_τ containing γ. Since the family of regular sets is closed under reversal the theorem is proven. □

COROLLARY 2.7: *Let $G=(N, T, I, P, S)$ be an indexed grammar. Then the set* $\mathrm{EMPTY}_G(A)=\{ \gamma \mid A \gamma \overset{*}{\Rightarrow} e \}$ *is regular.*

Proof: Let P' be the set of all productions of P containing no terminal symbols and let $G'=(N, T, I, P', S)$. Then $\mathrm{EMPTY}_G(A)=\mathrm{TERM}_{G'}(A)$. □

THEOREM 2.8: *Let $G=(N, T, I, P, S)$ be an indexed grammar and let $A \in N$. Then* $\mathrm{INDEX}_G(A)$ *is regular.*

Proof: W.l.o.g. (*see* corollary 2.5) we can assume that G is in normal form.

Define $\sigma : I^* \to \mathscr{P}(N \times N)$ by

$$\sigma(\gamma)=\{ (A, B) \mid A, B \in N,\ A \overset{*}{\Rightarrow} \Theta_1 B \gamma \Theta_2,\ \Theta_1, \Theta_2 \in (NI^* \cup T)^* \},$$

and furthermore let R_σ be the following relation over I^* :

$$\text{for all } \gamma_1, \gamma_2 \in I^* \text{ we have } (\gamma_1, \gamma_2) \in R_\sigma \text{ iff } \sigma(\gamma_1)=\sigma(\gamma_2).$$

This is an equivalence relation with finite index, and we will show, that it is a right congruence. Let $(\gamma_1, \gamma_2) \in R_\sigma$ and $\gamma \in I^*$, and let $(A, B) \in \sigma(\gamma_1 \gamma)$, *i.e.*, there is a derivation $A \overset{*}{\Rightarrow} \Theta_1 B \gamma_1 \gamma \Theta_2$ with $\Theta_1, \Theta_2 \in (NI^* \cup T)^*$. In a corresponding derivation tree consider the path from the root A to the leaf $B \gamma_1 \gamma$. According to the special form of our grammar this path contains a node which is labeled by $C \gamma$ and the labels of all successor nodes on this path are of the form $D \gamma' \gamma$, $\gamma' \in I^*$. Hence there are derivations $A \overset{*}{\Rightarrow} \bar{\Theta}_1 C \gamma \bar{\Theta}_2$

and $C \overset{*}{\Rightarrow} \Theta'_1 B \gamma_1 \Theta'_2$, *i.e.*, $(C, B) \in \sigma(\gamma_1)$. Since $\sigma(\gamma_1) = \sigma(\gamma_2)$, there is a derivation $C \overset{*}{\Rightarrow} \tilde{\Theta}_1 B \gamma_2 \tilde{\Theta}_2$. Hence there exists the derivation $A \overset{*}{\Rightarrow} \bar{\Theta}_1 (\tilde{\Theta}_1 : \gamma) B \gamma_2 \gamma (\tilde{\Theta}_2 : \gamma) \bar{\Theta}_2$, but this means $(A, B) \in \sigma(\gamma_2 \gamma)$. With

$$\mathrm{INDEX}_G(A) = \bigcup_{\gamma \in I^*,\, (S, A) \in \sigma(\gamma)} [\gamma]_\sigma$$

the theorem is proved. □

Let $G = (N, T, I, P, S)$ be an indexed grammar and let $\tau : I^* \to \mathscr{P}(N)$ with $\tau(\gamma) = \{ A \mid A \gamma^R \overset{*}{\Rightarrow} w,\ w \in T^* \}$ be the function defined in the proof of theorem 2.6. Since the emptiness problem for indexed grammars is decidable (*see* [6], the proof in [1] is not correct) it is easy to show that τ is computable. To this end let $\gamma \in I^*$ and $A \in N$. Set $G_A = (N \cup \{ S' \}, T, I, P', S')$ with $P' = P \cup \{ S' \to A \gamma^R \}$. Then we have $A \in \tau(\gamma)$ iff $L(G_A) \neq \varnothing$.

Now it is possible to construct a deterministic finite automaton (DFA) which accepts $\mathrm{TERM}_G(A)^R$.

THEOREM 2.9: *Let $G = (N, T, I, P, S)$ be an indexed grammar and let $A \in N$. A* DFA *$\mathscr{A}$ with $L(\mathscr{A}) = \mathrm{TERM}_G(A)^R$ is effectively constructable.*

Proof: Determine a set $Z \subseteq \mathscr{P}(N)$ and a function $\delta : Z \times I \to Z$ as follows (**Q** denotes an initially empty queue):

Set $z_0 := \tau(e)$, $Z := \{ z_0 \}$, and $(z_0, e) \Rightarrow \mathbf{Q}$.
while Q *not empty*
  begin
  $\mathbf{Q} \Rightarrow (z, \gamma)$ { at this point $z = \tau(\gamma)$ *holds* }
  for all $f \in I$
    begin
    set $\delta(z, f) := \tau(\gamma f)$
    if $\tau(\gamma f) \notin Z$ **then**
      begin
      $Z := Z \cup \{ \tau(\gamma f) \}$
      $(\tau(\gamma f), \gamma f) \Rightarrow \mathbf{Q}$
      end
    end
  end

The algorithm terminates since $\mathscr{P}(N)$ is finite. Now set $\mathscr{A} = (Z, I, \delta, z_0, F)$ with $F = \{ z \mid z \in Z,\ A \in z \}$ and let $\hat{\delta} : Z \times I^* \to Z$ be the extended transition function of $\mathscr{A}$ defined as usual. We have $\hat{\delta}(z_0, \gamma) = \tau(\gamma)$, for $\hat{\delta}(z_0, e) = z_0 = \tau(e)$,

and, if $\gamma = \gamma' f, f \in I$, then

$$\hat{\delta}(z_0, \gamma) = \hat{\delta}(z_0, \gamma' f) = \delta(\hat{\delta}(z_0, \gamma'), f) = \delta(\tau(\gamma'), f) = \tau(\gamma' f) = \tau(\gamma).$$

(Here we use the fact R_τ is a right congruence.) Therefore we have

$$L(\mathcal{A}) = \{\gamma \mid \tau(\gamma) \in F\} = \{\gamma \mid A \in \tau(\gamma)\}$$
$$= \mathrm{TERM}_G(A)^R. \quad \square$$

A similar construction is possible for the set $\mathrm{INDEX}_G(A)$, but there is an easier way. We will construct a regular grammar generating this set.

THEOREM 2.10 *Let $G = (N, T, I, P, S)$ be an indexed grammar and let $A \in N$. A regular grammar G_A with $L(G_A) = \mathrm{INDEX}_G(A)$ is effectively constructable.*

Proof: We can assume that G is of the form used in the proof of Theorem 2.8. Now construct $G' = (N, T, I, P', S)$ in the following way:

If $C \to u_1 B_1 u_2 \ldots u_n B_n u_{n+1}$ with $u_j \in T^*$, $j \in [1 : n+1]$, $B_i \in N$, $i \in [1 : n]$, $n \geqq 0$, is in P, then the productions $C \to B_i$, $i \in [1 : n]$, are in P'. All productions in P of the form $C \to Bf$ or $Cf \to B$ are in P' too.

Now we will show that for all C, $B \in N$, it is decidable whether $C \overset{*}{\Rightarrow} B$ according to G' holds. To this end construct $G'' = (N \cup \{S'\}, T, I \cup \{\#\}, P'', S')$ with $P'' = P' \cup \{S' \to C\#, B\# \to e\}$. Obviously $C \overset{*}{\Rightarrow} B$ according to G' iff $L(G'') \neq \varnothing$ holds.

It is now possible to construct a regular gramar $G_A = (N, I, P_A, S)$, where P_A is defined as follows:

(1) If $C \to Bf$ is in P', then $C \to Bf$ is in P_A. (Note that f is a terminal symbol with respect to G_A in the second production.)

(2) If $C \overset{*}{\Rightarrow} B$ according to G', then $C \to B$ is a production in P_A.

(3) Furthermore $A \to e$ is in P_A.

Obviously $L(G_A) = \mathrm{INDEX}_G(A)$ holds.

Remark: In [9], given an indexed grammar $G = (N, T, I, P, S)$ and a variable $A \in N$, the notion

$$\mathrm{FLAGS}(A) = \{\gamma \mid \gamma \in I^*, S \overset{*}{\Rightarrow} wA\gamma\Theta, w \in T^*, \Theta \in (NI^* \cup T)^*\}$$

is introduced and it is stated that this set is regular if G contains no e-productions. But actually it is proved that the set $\mathrm{INDEX}_G(A)$ is regular.

The problem in this proof lies in the fact that a derivation $S \overset{*}{\Rightarrow} \Theta_1 A \gamma \Theta_2$, $\Theta_1, \Theta_2 \in (NI^* \cup T)^*$ can not necessarily be continued by $\Theta_1 \overset{*}{\Rightarrow} w$, $w \in T^*$.

Let us call in analogy to the notion $\mathrm{INDEX}_G(A)$ the set FLAGS (A) by $INDEX^l_G(A)$ (l means *left terminal*), *i.e.*,

$$\mathrm{INDEX}^l_G(A) = \{ \gamma \mid \gamma \in I^*, S \overset{*}{\Rightarrow} w A \gamma \Theta, w \in T^*, \Theta \in (NI^* \cup T)^* \}.$$

It is possible to show, using proof techniques as in Theorem 3.2 that this set is regular.

THEOREM 2.11: *Let* $G=(N, T, I, P, S)$ *be an indexed grammar and let* $A \in N$. *Then the set* $\mathrm{INDEX}^l_G(A)$ is regular.

Remark: A regular grammar for $\mathrm{INDEX}^l_G(A)$ is effectively constructable.

If we define $\mathrm{INDEX}^{lr}_G(A) = \{ \gamma \mid \gamma \in I^*, S \overset{*}{\Rightarrow} v A \gamma w, v, w \in T^* \}$, it can be shown in a similar way that this set is regular too.

It is easy to see that $\mathrm{TERM}_G(A) = \mathrm{TERM}_G(A) I^*$ holds for each grammar G and each variable A. Hence there are regular sets which are not of the form $\mathrm{TERM}_G(A)$. On the other hand we obviously have:

THEOREM 2.12: *Let* I *be an alphabet and* $R \subseteq I^*$ *a regular set. There exists an indexed grammar* G *such that* $R = \mathrm{INDEX}_G(A) = \mathrm{INDEX}^l_G(A) = \mathrm{INDEX}^{lr}_G(A)$ *holds for a variable* A.

3. REDUCED INDEXED GRAMMARS

A context-free grammar G is called reduced if every variable appears in a derivation of a terminal word, which is equivalent to

(1) each sentential form produces a terminal string, and

(2) each variable is reachable from the start symbol.

This is the motivation for the following definition:

DEFINITION 3.1: An indexed grammar $G=(N, T, I, P, S)$ is called *sentential form* (SF-) *reduced*, if $S \overset{*}{\Rightarrow} \Theta$ implies $\Theta \overset{*}{\Rightarrow} w$, $w \in T^*$. G is called *reduced* if it is SF-reduced and if for each $A \in N$ there is a derivation $S \overset{*}{\Rightarrow} \Theta_1 A \gamma \Theta_2$, $\Theta_1, \Theta_2 \in (NI^* \cup T)^*$, $\gamma \in I^*$.

Remark: If G is SF-reduced, we have $L(G) \neq \emptyset$.

The following theorem shows that each nonempty indexed language can be generated by an SF-reduced indexed grammar. We will use the same idea as in Theorem 2.11 but the construction of the grammar is quite different.

THEOREM 3.2: *Let* $G=(N, T, I, P, S)$ *be an indexed grammar with* $L(G)\neq\varnothing$. *Then an* SF-*reduced indexed grammar* $G'=(N', T, I', P', S')$ *with* $L(G)=L(G')$ *can effectively be constructed.*

Proof: W.l.o.g. we can assume that G is in normal form. Furthermore let $N=\{A_1, \ldots, A_r\}$ and $\mathscr{A}_i=(Z_i, I, \delta_i, z_i^0, F_i)$ be a DFA with $L(\mathscr{A}_i)=\mathrm{TERM}_G(A_i)^R$ for $i\in[1:r]$. (These DFA's can be effectively constructed, see Theorem 2.9.)

In a derivation $S\overset{*}{\Rightarrow}\Theta_1 A_i \gamma \Theta_2$, $\Theta_1, \Theta_2\in(NI^* \cup T)^*$, it is essential to know whether $\Theta_1\overset{*}{\Rightarrow} w$ with $w\in T^*$, holds, *i. e.*, whether a term $A_j\tilde{\gamma}$ ocurring in Θ_1 produces a terminal word. This is equivalent to the question whether $\hat{\delta}_j(z_j^0, \tilde{\gamma}^R)$ is in F_j. Hence we set

$$N'=N\times Z_1\times\ldots\times Z_r \qquad \text{and} \qquad S'=(S, z_1^0, \ldots, z_r^0).$$

Since it is possible to consume an index in a derivation, we have to save the states of the DFA's before producing this index, therefore we set

$$I'=I\times Z_1\times\ldots\times Z_r.$$

Now set $G'=(N', T, I', P', S')$, where P' is defined as follows:

(1) Let $A_{j_0}\to u_0 A_{j_1} u_1 \ldots u_{q-1} A_{j_q} u_q$ with $u_i\in T^*$ and $A_{j_i}\in N$, $i\in[0:q]$, $n\geqq 0$, be in P. Then for all $z_k\in Z_k$, $k\in[1:r]$, with $z_{j_i}\in F_{j_i}$ for $i\in[0:q]$ the production

$$(A_{j_0}, z_1, \ldots, z_r)\to u_0 (A_{j_1}, z_1, \ldots, z_r) u_1 \cdot \ldots \cdot u_{q-1} (A_{j_q}, z_1, \ldots, z_r) u_q \text{ is in } P'.$$

(2) Let $A_i\to A_j f$ be in P, then for all $z_k\in Z_k$, $k\in[1:r]$, with $z_i\in F_i$ and $\delta_j(z_j, f)\in F_j$ the production

$$(A_i, z_1, \ldots, z_r)\to (A_j, \delta_1(z_1, f), \ldots, \delta_r(z_r, f))(f, z_1, \ldots, z_r) \text{ is in } P'.$$

(3) Let $A_i f\to A_j$ be in P, then for all $z_k\in Z_k$, $k\in[1:r]$, with $z_j\in F_j$ and $\delta_i(z_i, f)\in F_i$ the production

$$(A_i, \delta_1(z_1, f), \ldots, \delta_r(z_r, f))(f, z_1, \ldots, z_r)\to (A_j, z_1, \ldots, z_r) \text{ is in } P'.$$

To compare derivations according to G and G', we need the following two functions:

(*a*) $\varphi : I^* \to I'^*$ with $\varphi(e) = e$ and

$$\varphi(f\,\gamma) = (f, \hat{\delta}_1(z_1^0, \gamma^R), \ldots, \hat{\delta}_r(z_r^0, \gamma^R))\,\varphi(\gamma), \qquad f \in I, \quad \gamma \in I^*.$$

(*b*) $h : I'^* \to I^*$ is a homomorphism with

$$h(f, z_1, \ldots, z_r) = f, \qquad f \in I, \quad z_k \in Z_k, \quad k \in [1:r].$$

First we will prove:

$A_i\gamma \overset{n}{\Rightarrow} w$, $w \in T^*$, according to G implies $(A_i, z_1, \ldots, z_r)\,\varphi(\gamma) \overset{n}{\Rightarrow} w$ where $z_k = \hat{\delta}_k(z_k^0, \gamma^R)$, $k \in [1:r]$ according to G'.

Let $n = 1$. Then $A_i \to w$ is in P, and $\gamma \in \mathrm{TERM}_G(A_i)$, *i. e.*, $\hat{\delta}_i(z_i^0, \gamma^R) \in F_i$, therefore $(A_i, z_1, \ldots, z_r) \to w$ with $z_k = \hat{\delta}_k(z_k^0, \gamma^R)$, $k \in [1:r]$ is in P'.

Now assume

$$A_i\gamma = A_{j_0}\gamma \quad \Rightarrow \quad u_0 A_{j_1} \gamma u_1 \ldots u_{q-1} A_{j_q} \gamma u_q \quad \overset{n}{\Rightarrow} \quad w$$

with $A_{j_l}\gamma \overset{*}{\Rightarrow} v_l$ for $l \in [1:q]$. We have $\gamma \in \mathrm{TERM}_G(A_{j_l})$ for $l \in [0:q]$, and therefore

$$(A_i, z_1, \ldots, z_r) \to u_0 (A_{j_1}, z_1, \ldots, z_r)\, u_1 \ldots (A_{j_q}, z_1, \ldots, z_r)\, u_q$$

with $z_k = \hat{\delta}_k(z_k^0, \gamma^R)$, $k \in [1:r]$, is in P', and with the induction hypothesis we have $(A_i, z_1, \ldots, z_r) \overset{n+1}{\Rightarrow} w$ according to G'.

Now assume $A_i\gamma \Rightarrow A_j f\,\gamma \overset{n}{\Rightarrow} w$. We have $\gamma \in \mathrm{TERM}_G(A_i)$ and $f\,\gamma \in \mathrm{TERM}_G(A_j)$, *i. e.*,

$$\hat{\delta}_i(z_i^0, \gamma^R) \in F_i$$

and

$$\hat{\delta}_j(z_j^0, \gamma^R f) = \delta_j(\hat{\delta}_j(z_j^0, \gamma^R), f) \in F_j,$$

therefore

$$(A_i, z_1, \ldots, z_r) \to (A_j, \delta_1(z_1, f), \ldots, \delta_r(z_r, f))\,(f, z_1, \ldots, z_r)$$

with $z_k = \hat{\delta}_k(z_k^0, \gamma^R)$, $k \in [1:r]$ is in P', and with the induction hypothesis and the definition of φ we have

$$(A_i, z_1, \ldots, z_r)\,\varphi(\gamma) \Rightarrow (A_j, \delta_1(z_1, f), \ldots, \delta_r(z_r, f))(f, z_1, \ldots, z_r)\,\varphi(\gamma)$$
$$= (A_j, \delta_1(z_1, f), \ldots, \delta_r(z_r, f))\,\varphi(f\,\gamma) \overset{n}{\Rightarrow} w$$

according to G'.

Let now $\gamma = f\,\alpha$ and $A_i\gamma = A_i f\,\alpha \Rightarrow A_j\alpha \overset{n}{\Rightarrow} w$ according to G. Then we have

$$\gamma = f\,\alpha \in \mathrm{TERM}_G(A_i) \quad \text{and} \quad \alpha \in \mathrm{TERM}_G(A_j),$$

i. e.,

$$\hat{\delta}_i(z_i^0, \gamma^R) = \delta_i(\hat{\delta}_i(z_i^0, \alpha^R), f) \in F_i \quad \text{and} \quad \hat{\delta}_j(z_j^0, \alpha^R) \in F_j,$$

therefore the production

$$(A_i, \delta_1(z_1', f), \ldots, \delta_r(z_r', f))(f, z_1', \ldots, z_r') \to (A_j, z_1', \ldots, z_r')$$

with $z' = \hat{\delta}_k(z_k^0, \alpha^R)$, $k \in [1:r]$, is in P' and with the induction hypothesis and the definition of φ we have

$$(A_i, \delta_1(z_1', f), \ldots, \delta_r(z_r', f))(f, z_1', \ldots, z_r'))\,\varphi(\alpha) \Rightarrow (A_j, z_1', \ldots, z_r')\,\varphi(\alpha) \overset{n}{\Rightarrow} w$$

according to G'.

In particular we have:

If $S \overset{*}{\Rightarrow} w$, $w \in T^*$, holds according to G, then $(S, z_1^0, \ldots, z_r^0) = S' \overset{*}{\Rightarrow} w$, according to G', hence $L(G) \subseteq L(G')$.

Furthermore it is easy to show:

$(A_i, z_1, \ldots, z_r)\,\varphi(\gamma) \overset{n}{\Rightarrow} w$ *with* $z_k = \hat{\delta}_k(z_k^0, \gamma^R)$, $k \in [1:r]$, *according to* G' *implies* $A_i\gamma \overset{n}{\Rightarrow} w$ *according to* G.

This implies: If $S' = (S, z_1^0, \ldots, z_r^0) \overset{*}{\Rightarrow} w$, $w \in T^*$, holds according to G', then $S \overset{*}{\Rightarrow} w$ according to G, hence $L(G') \subseteq L(G)$.

Now we have $L(G) = L(G')$. Note that G and G' are structural equivalent, *i. e.*, the derivation trees of terminal words according to G and G' are the same except for the labels at intermediate nodes.

Next we have to show that G' is SF-reduced. To this end we will first prove:

Let $S' \overset{n}{\Rightarrow} \Theta'$ *with* $\Theta' = \Theta'_1 B' \gamma' \Theta'_2, \Theta'_1, \Theta'_2 \in (NI^* \cup T)^*$, *where* $\gamma = h(\gamma')$, *and* $B' = (B, z_1, \ldots, z_r)$, *then* $z_k = \hat{\delta}_k(z_k^0, \gamma^R)$, $\gamma \in \mathrm{TERM}_G(B)$, *and* $\gamma' = \varphi(\gamma)$.

If $n = 0$, then $S' = B' = (S, z_1^0, \ldots, z_r^0)$ and $\gamma = h(\gamma') = e$. Since $L(G) \neq \varnothing$, we have $e \in \mathrm{TERM}_G(S)$.

Now let

$$S' \overset{n}{\Rightarrow} \Theta'' = \Theta'_1 A'_{j_0} \gamma' \Theta'_2 \Rightarrow \Theta'_1 u_0 A'_{j_1} \gamma' \ldots u_{q-1} A'_{j_q} \gamma' u_q \Theta'_2 = \Theta'.$$

The assumption holds for Θ'', in particular we have $A'_{j_0} = (A_{j_0}, z_1, \ldots, z_r)$ with $z_k = \hat{\delta}_k(z_k^0, \gamma^R)$, $\gamma = h(\gamma') \in \mathrm{TERM}_G(A_{j_0})$ and $\gamma' = \varphi(\gamma)$. The last production applied is

$$A'_{j_0} \to u_0 A'_{j_1} u_1 \ldots u_{q-1} A'_{j_q} u_q$$

with $A'_{j_l} = (A_{j_l}, z_1, \ldots, z_r)$, $l \in [1:q]$. We have $z_{j_l} \in F_{j_l}$ and hence $\gamma \in \mathrm{TERM}_G(A_{j_l})$, $l \in [1:q]$.

Now let $S' \overset{n}{\Rightarrow} \Theta'' = \Theta'_1 A'_i \gamma' \Theta'_2 \Rightarrow \Theta'_1 A'_j f' \gamma' \Theta'_2 = \Theta'$. The assumption holds for Θ'', in particular we have $A'_i = (A_i, z_1, \ldots, z_r)$ with $\gamma = h(\gamma') \in \mathrm{TERM}_G(A_i)$, $\gamma' = \varphi(\gamma)$, and $z_k = \hat{\delta}_k(z_k^0, \gamma^R)$. The last production applied is

$$(A_i, z_1, \ldots, z_r) \to (A_j, \delta_1(z_1, f), \ldots, \delta_r(z_r, f))(f, z_1, \ldots, z_r).$$

We have $\delta_j(z_j, f) = \hat{\delta}_j(z_j^0, \gamma^R f) \in F_j$, hence $f\gamma = h(f'\gamma') \in \mathrm{TERM}_G(A_j)$. Furthermore we have $\varphi(f\gamma) = f'\gamma'$ (*see* definition of φ).

Now let $S' \overset{n}{\Rightarrow} \Theta'' = \Theta'_1 A'_i f' \alpha' \Theta'_2 \Rightarrow \Theta'_1 A'_j \alpha' \Theta'_2 = \Theta'$. The assumption holds for Θ'', in particular we have $A'_i = (A_i, \tilde{z}_1, \ldots, \tilde{z}_r)$ with $\tilde{z}_k = \hat{\delta}_k(z_k^0, \alpha^R f)$, $f\alpha = h(f'\alpha') \in \mathrm{TERM}_G(A_i)$ and $f'\alpha' = \varphi(f\alpha) = (f, z_1, \ldots, z_r)\varphi(\alpha)$ with $z_k = \hat{\delta}_k(z_k^0, \alpha^R)$, $k \in [1:r]$. The last production applied is

$$(A_i, \tilde{z}_1, \ldots, \tilde{z}_r)(f, z_1, \ldots, z_r) \to (A_j, z_1, \ldots, z_r),$$

where $\tilde{z}_k = \delta_k(z_k, f)$, and $z_j \in F_j$, hence $\alpha \in \mathrm{TERM}_G(A_j)$.

Now let $S' \overset{*}{\Rightarrow} \Theta'_1 B' \gamma' \Theta'_2$, $\Theta'_1, \Theta'_2 \in (N' I'^* \cup T)^*$, where $B' = (B, z_1, \ldots, z_r)$ and $\gamma = h(\gamma')$. Since $\gamma \in \mathrm{TERM}_G(B)$, there is a derivation $B\gamma \overset{*}{\Rightarrow} w$ according to G, and since $z_k = \hat{\delta}_k(z_k^0, \gamma^R)$, $k \in [1:r]$ and $\varphi(\gamma) = \gamma'$ we have $B'\gamma' = (B, z_1, \ldots, z_r)\varphi(\gamma) \overset{*}{\Rightarrow} w$ according to G'. Hence G' is SF-reduced. □

Next we can prove

THEOREM 3.3: *Let $G=(N, T, I, P, S)$ be an indexed grammar with $L(G)\neq\varnothing$. Then an equivalent reduced indexed grammar can effectively be constructed.*

Proof: First construct using Theorem 3.2 an equivalent SF-reduced indexed grammar $G'=(N', T, I', P', S')$ for G. For each $A'\in N$ construct using Theorem 2.10 a regular grammar for $\mathrm{INDEX}_{G'}(A')$. Determine $N''=\{A' \mid \mathrm{INDEX}_{G'}(A')\neq\varnothing\}$ (note that we have $S'\in N''$) and set $G''=(N'', T, I', P'', S')$, where P'' consists exactly of those production of P' which only contain variables from N''. It is easy to see that G'' is equivalent to G and reduced. □

Remark: If G in the above theorem is an e-free indexed grammar, then G'' is e-free too.

Remark: In the context-free case, each production of a reduced grammar is applicable in a derivation of a terminal word. This is not necessarily true for a reduced indexed grammar $G=(N, T, I, P, S)$, because it is possible that a production of the form $A f\rightarrow B$ is not applicable in a derivation of a terminal word. This is equivalent to the fact that no word in $\mathrm{INDEX}_G(A)$ begins with f, *i. e.*, $\mathrm{INDEX}_G(A)\cap f I^*=\varnothing$. It is obviously possible to test this condition and to eliminate such a production (*see* also [8]).

4. THE STRUCTURE OF SENTENTIAL FORMS OF INDEXED GRAMMARS

It is well known (*see* [4]) that the set of strings which can appear on the pushdown list of a PDA is a regular set. In analogy, given a context-free grammar $G=(N, T, P, S)$, the set $\{\alpha \mid S\underset{l}{\overset{*}{\Rightarrow}} u\alpha,\ u\in T^*,\ \alpha\in N(N\cup T)^*\cup\{e\}\}$ is a regular language where $\underset{l}{\overset{*}{\Rightarrow}}$ denotes a leftmost derivation. A left linear grammar which generates this set can be constructed as follows:

The set of variables of this grammar is $N'=\{A' \mid A\in N\}$ and $N\cup T$ is the set of terminals. Now consider an arbitrary production

$$A_0\rightarrow u_0 B_1 u_1 \ldots B_i u_i \ldots B_q u_q$$

with $q>0$ of P. For each $i\in[1:q]$ such that $B_j\overset{*}{\Rightarrow} w_j$, $w_j\in T^*$ holds for $j\in[1:i-1]$ introduce the production $A'_0\rightarrow B'_i u_i B_{i+1}\ldots B_q u_q$. Furthermore,

for all $A \in N$, the production $A' \to A$ and, if $L(G) \neq \varnothing$, the production $S' \to e$, are introduced.

Using the fact that the sets $\mathrm{TERM}_G(A)$ are regular for an indexed grammar $G=(N, T, I, P, S)$, it is possible to show that in this case the corresponding set is context-free. To formalize this statement let us first introduce the homomorphism $\mu : (NI^* \cup T)^* \to (N \cup T)^*$ with $\mu(A\gamma) = A$, $A \in N$, $\gamma \in I^*$, and $\mu(a) = a$, $a \in T$. Now we can show

THEOREM 4.1: *Let $G=(N, T, I, P, S)$ be an indexed grammar. Then the set $M = \{ \mu(\Theta) \mid S \overset{*}{\underset{l}{\Rightarrow}} u\Theta,\ u \in T^*,\ \Theta \in NI^*(NI^* \cup T)^* \cup \{e\} \}$ is context-free.*

Proof: Let $N = \{ A_1, \ldots, A_r \}$ and let $\mathscr{A}_i = (Z_i, I, \delta_i, z_i^0, F_i)$ be the DFA's with $L(\mathscr{A}_i) = \mathrm{TERM}_G(A_i)^R$ for $i \in [1:r]$ and set $N' = N \times Z_1 \times \ldots \times Z_r$ and $I' = I \times Z_1 \times \ldots \times Z_r$ (*see* proof of Theorem 2.11).

In the sequel we need the function $\psi : I^* \times Z_1 \times \ldots \times Z_r \to I'^*$ defined by

$$\psi(e, z_1, \ldots, z_r) = e,$$

$$\psi(f\gamma, z_1, \ldots, z_r) = (f, \hat{\delta}_1(z_1, \gamma^R), \ldots, \hat{\delta}_r(z_r, \gamma^R))\,\psi(\gamma, z_1, \ldots, z_r)$$

for all $z_k \in Z_k$, $k \in [1:r]$, $f \in I$, and $\gamma \in I^*$.

We will now define a left linear indexed grammar $G' = (N'', T', I', P', S'')$ with $N'' = N' \cup \{S''\}$ (S'' is a new symbol), $T' = N \cup T$, and P' is defined as follows:

(1) $S'' \to (S, z_1^0, \ldots, z_r^0) = S'$ is in P'.

(2) If $L(G) \neq \varnothing$, then $S' \to e$ is in P'.

(3) For all $A \in N$ and for all $z_k \in Z_k$, $k \in [1:r]$, the production $(A, z_1, \ldots, z_r) \to A$ is in P'.

(4) Let $A_{j_0} f \to u_0 A_{j_1} \gamma_1 u_1 \ldots A_{j_q} \gamma_q u_q$ with $u_i \in T^*$, $A_{j_i} \in N$, $i \in [0:q]$, $f \in I \cup \{e\}$, $\gamma_j \in I^*$, $j \in [1:q]$, and $q > 0$ be in P. Then for all $i \in [1:q]$, and for all $z_k \in Z_k$, $k \in [1:r]$, with $\delta_{j_l}(z_{j_l}, \gamma_l^R) \in F_{j_l}$, $l \in [1:i-1]$, the production

$$(A_{j_0}, \hat{\delta}_1(z_1, f), \ldots, \hat{\delta}_r(z_r, f))\,\psi(f, z_1, \ldots, z_r) \to (A_{j_i}, \hat{\delta}_1(z_1, \gamma_i^R), \ldots, \hat{\delta}_r(z_r, \gamma_i^R))$$

$$\psi(\gamma_i, z_1, \ldots, z_r)\, u_i A_{j_i+1} \ldots A_{j_q} u_q \text{ is in } P'.$$

Obviously G' is a left linear indexed grammar and hence $L(G')$ is a context-free language [1].

An easy induction on n first shows:

$A\gamma \underset{l}{\overset{n}{\Rightarrow}} u\Theta$, $\Theta \in NI^*(NI^* \cup T)^*$ *according to* G *implies*

$$(A, \hat{\delta}_1(z_1^0, \gamma^R), \ldots, \hat{\delta}_r(z_r^0, \gamma^R))\psi(\gamma, z_1^0, \ldots, z_r^0) \overset{*}{\Rightarrow} \mu(\Theta)$$

according to G'.

In particular we have $S \underset{l}{\overset{*}{\Rightarrow}} u\Theta$, $u \in T^*$, $\Theta \in NI^*(NI^* \cup T)^*$ according to G implies $(S, z_1^0, \ldots, z_r^0) \overset{*}{\Rightarrow} \mu(\Theta)$ according to G'. Hence $M \subseteqq L(G')$. With another easy induction one can prove:

$$(A, \hat{\delta}_1(z_1^0, \gamma^R), \ldots, \hat{\delta}_r(z_r^0, \gamma^R))\psi(\gamma, z_1^0, \ldots, z_r^0) \overset{n}{\Rightarrow} w, \qquad w \in N(N \cup T)^*,$$

according to G' *implies* $A\gamma \underset{l}{\overset{*}{\Rightarrow}} u\Theta$, $u \in T^*$, $\Theta \in NI^*(NI^* \cup T)^*$, *according to* G *with* $\mu(\Theta) = w$.

In particular $(S, z_1^0, \ldots, z_r^0) \overset{*}{\Rightarrow} w$ according to G' implies $S \underset{l}{\overset{*}{\Rightarrow}} u\Theta$, $u \in T^*$, $\Theta \in NI^*(NI^* \cup T)^*$, according to G with $\mu(\Theta) = w$. This completes the proof. □

We will now show that it is not possible to substitute "context-free" by "regular" in this theorem.

Example: We will give an indexed grammar $G = (N, T, I, P, S)$ such that the set M investigated in the foregoing theorem is not regular. Set

$$N = \{S, A, B\}, \qquad T = \{a, b, c\}, \qquad I = \{f, g\},$$

and

$$P = \{S \to Ag,\ A \to Afc,\ A \to B,\ Bf \to aBb,\ Bg \to e\}.$$

We have

$$L(G) = \{a^n b^n c^n \mid n \geqq 0\}$$

and

$$M = \{S, e\} \cup \{Ac^i \mid i \geqq 0\} \cup \{Bb^j c^i \mid 0 \leqq j \leqq i\}.$$

The set M is not regular.

Remark: There is a correspondence between indexed grammars and indexed pushdown automata (IPDA) (*see* [7]). It is not difficult to see that the set of strings appearing on the pushdown list of such automata are context-free.

ACKNOWLEDGMENTS

The authors wish to thank the referees for their useful hints and suggestions.

REFERENCES

1. A. V. AHO, *Indexed Grammars*, J.A.C.M., Vol. 15, 1968, pp. 647-671.
2. J. DUSKE and R. PARCHMANN, *Linear Indexed Languages*, Theoret. Computer Sci., Vol. 32, 1984, pp. 47-60.
3. J. ENGELFRIET and H. VOGLER, *Look-Ahead on Pushdowns*, Inform. and Comput., Vol. 73, 1987, pp. 245-279.
4. S. A. GREIBACH, *A Note on Pushdown Store Automata and Regular Systems*, Proc. Amer. Math. Soc., Vol. 18, 1967, pp. 263-268.
5. J. E. HOPCROFT and J. D. ULLMAN, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, Reading, MA, 1979.
6. T. S. E. MAIBAUM, *Pumping Lemmas for Term Languages*, J. Comput. System Sci., Vol. 17, 1978, pp. 319-330.
7. R. PARCHMANN, J. DUSKE and J. SPECHT, *On Deterministic Indexed Languages*, Inform. and Control, Vol. 45, 1980, pp. 48-67.
8. R. PARCHMANN, J. DUSKE and J. SECHT, *Indexed* LL(*k*)*-Grammars*, Acta Cybernetica, Vol. 7, 1984, pp. 33-53.
9. R. W. SEBESTA and N. D. JONES, *Parsers for Indexed Grammars*, Internat. J. Comput. and Inform. Sci., Vol. 7, 1978, pp. 345-359.