ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

480

# ON CONTROL SETS INDUCED BY GRAMMARS

BY

TIMO JÄRVI

HELSINKI 1970
SUOMALAINEN TIEDEAKATEMIA

doi:10.5186/aasfm.1971.480

Communicated 11 September 1970 by ARTO SALOMAA.

KESKUSKIRJAPAINO
HELSINKI 1970

## Introduction

While studying different ways to restrict the use of the productions of a given grammar $G$, Salomaa [6] has introduced the notion of a *control language* $C$, which is a language over the productions of $G$. The language $L_C(G)$ generated by $G$ with $C$ as a control language is a subset of the language $L(G)$, consisting of words which possess at least one derivation whose string of productions belongs to $C$. In what follows we shall study the *control set induced by grammar* $G$, i.e., the particular control language $C_G$, whose every word is a string of productions of some derivation of a word in $L(G)$ and vice versa. This notion of a control set has been studied by Stotskiĭ [7]. Let us note that it differs from the control sets of Ginsburg and Spanier [3], because (i) it is determined by the grammar and not fixed independently and (ii) attention is not restricted to leftmost derivations.

## Definitions

We shall first recall some definitions about languages and grammars. We shall present the definitions in the form used in [4], where background material can be found, too.

In the definition of a phrase structure grammar $G = (I_N, I_T, X_0, F)$ the symbols have the following meanings: $I_N$ (the non-terminals) and $I_T$ (the terminals) are finite disjoint alphabets, $X_0$ (the initial letter) is in $I_N$, and $F$ is a finite set of ordered pairs $P \to Q$ (productions), where $P, Q \in (I_N \cup I_T)^*$, and $P$ contains at least one letter of $I_N$. ($I^*$ is the set of all words over the alphabet $I$.)

The phrase structure language $L(G)$ generated by the grammar $G$ is the set of words $P \in I_T^*$, for which there exists a sequence of words

(1) $$X_0 = P_0, P_1, P_2, \ldots, P_r = P,$$

called a derivation, in which

(2) $$P_{i-1} = Q_i R_i S_i, \; P_i = Q_i T_i S_i, \; R_i \to T_i \in F, \quad Q_i, S_i \in (I_N \cup I_T)^*, \; \forall\, i = 1, 2, \ldots, r.$$

Let the productions of $F$ be labelled by $f_1, f_2, \ldots, f_k$. Then we shall form the set $C_G$ of all finite strings of productions

$$f_{j_1} f_{j_2} \ldots f_{j_r} ,$$

which generate a derivation of some word of $L(G)$, i.e., there exists a sequence (1) such that in (2) $f_{j_i}$ is the production $R_i \to T_i$, for all $i = 1, 2, \ldots, r$. The set $C_G$ will be called the *control set induced by the gammar* $G$.

## Relations between the types of $G$ and $C_G$

We divide grammars and languages into types 0, 1, 2, and 3 as usual, cf. [4, p. 168].

**Theorem 1.** *If the grammar $G$ is of type 3, then the control set $C_G$ is of type 3, too.*

*Proof.* The proof of theorem 1 is trivial, cf. [7, p. 35].

We may extend theorem 1 to *non-terminal bounded* grammars, i.e., to type 2 grammars $G$, for which there exists a positive integer $n$ such that in the derivations (2) in no word $P_i$ there exist more than $n$ non-terminal letters:

**Theorem 1′.** *If $G$ is non-terminal bounded, then $C_G$ is of type 3.*

*Proof.* This theorem is an exercise in ref. [1, p. 51], and it is easily proved, e.g., by letting the different possible combinations of non-terminal letters be different states of a finite deterministic automaton.

**Theorem 2.** *There exists a type 2 language $L$ such that, for no type 2 grammar $G$ which generates $L$, the control set $C_G$ is of type 3.*

*Proof.* To prove this theorem we shall use the concept of the *index of a type 2 grammar* and *language* as defined in [5]. There exists a type 2 language $L$ with infinite index [5]. Let $G$ be a type 2 grammar such that $L(G) = L$. There must be at least one production which increases the number of non-terminal letters and which can be used an unlimited number of times in a derivation. On the other hand, there are productions which decrease the number of non-terminals by one. So we cannot represent the control set $C_G$ in a finite deterministic automaton, which proves that $C_G$ is not of type 3.

**Theorem 3.** *For every non-empty language $L$ of type 2, there is a grammar $G$ such that $L = L(G)$ and $C_G$ is not of type 2.*

*Proof.* Let $P$ be some word of $L$. Then there exists a type 2 grammar $G_1 = (I_N, I_T, X_1, F)$ such that $L(G_1) = L - \{P\}$. Now for the type 2 grammar $G_2 = (I'_N, I_T, X_2, F')$, where $I'_N = \{X_2, Y, Z\}$, $I'_N \cap (I_N \cup I_T) = \emptyset$, and $F' = \{g_1 = X_2 \to YX_2Z, g_2 = Y \to \lambda, \quad g_3 = Z \to \lambda, \quad g_4 =$

$X_2 \to P\}$, $L(G_2) = \{P\}$. Let us construct a type 2 grammar

$$G = (I_N \cup I'_N \cup \{X\}, I_T, X, F \cup F' \cup \{X \to X_1, X \to X_2\}),$$

where $X \notin I_N \cup I'_N \cup I_T$. Clearly, $L(G) = L$.

We shall use the following

**Lemma.** *For each type 2 grammar $G$ there exist integers $p$ and $q$ with the property that each word $P$, $lg(P) > p$, in $L(G)$ is of the form $ABCDE$, where $BD \neq \lambda$, $lg(BCD) \leq q$, and $AB^nCD^nE$ is in $L(G)$ for all $n \geq 1$.* [1, p. 84].

Next we shall study the control set $C_G$ induced by $G$. Let us suppose that $C_G$ is of type 2, and $H$ is such a type 2 grammar that $L(H) = C_G$. Let $p$ and $q$ be the integers of the lemma for the grammar $H$. The word $g_1^m g_2^m g_4 g_3^m$, $m \geq q, p/3$, is obviously in $C_G$, but is not representable in the form of the lemma. Consequently, $C_G$ is not of type 2.

The following theorem was established by Stotskiĭ [7], but because the reference might be rather unknown and we have been able to shorten the original proof, we shall prove the theorem here.

**Theorem 4.** *If $G$ is a grammar of type 1, then the control set $C_G$ is of type 1, too.*

*Proof.* Let $G = (I_N, I_T, X_0, F)$ be a type 1 grammar, where

$$F = \{f_j \mid j = (0, \text{ if } f_0 = X_0 \to \lambda \in F), 1, 2, \ldots, k\}.$$

Let us form a grammar $G' = (I'_N, I'_T, X'_0, F')$, where

$$I'_N = I_N \cup I_T \cup \{g_1, g_2, \ldots, g_k\} \cup \{\xi, f, X'_0\},$$

$I'_T = F \cup \{c\}$, and $F'$ consists of the following productions:

1) $X'_0 \to f\xi X_0$, (if $f_0 \in F$, we shall take an additional production $X'_0 \to f_0 f\xi$,)

2) $f\xi \to \xi f_j$,

3) $f_j x \to x f_j$, $\forall\, x \in I_N \cup I_T$,

4) $f_j P_j \to g_j Q_j$, where $f_j = P_j \to Q_j$,

5) $x g_j \to g_j x$, $\forall\, x \in I_N \cup I_T$,

6) $\xi g_j \to f_j f \xi$,

7) $f\xi \to cc$,

8) $cx \to cc$, $\forall\, x \in I_T$,

where in 2)—6) $j = 1, 2, \ldots, k$.

To get a word in $I'_T$ we must first use productions 1)—6) to get a word of the form $Ef\xi P$, where $P \in I_T^*$ and $E \in F^*$ is the string of productions which is used in the derivation of $P$ according to $G$. Then by using productions 7) and 8) we get the word $Ec^n$. The grammar $G'$ is *length increasing* and hence $L(G')$ is a type 1 language [4, pp. 200—201]. Because an integer $m$ satisfying

$$lg(Ec^n) \leq m\, lg(E), \forall\, Ec^n \in L(G')\,,$$

is easily found, the language $\{E \mid Ec^n \in L(G')\}$ is of type 1 [2, Theorem 1.3, p. 568]. On the other hand this language is just the control set $C_G$, and so we have proved the theorem.

## Note added in proof

Friant has shown in [8] that the conclusion of theorem 4 is valid even when $G$ is of type 0. The proof above can be modified as follows to include this case:

Let us add the letter $\eta$ to $I_N'$. If $\lg(P_j) > \lg(Q_j)$ in the productions 4), $Q_j$ is substituted by $Q_j' = Q_j \eta^k$, where $k = \lg(P_j) - \lg(Q_j)$. We shall include the production $c\eta \to cc$ in 8) and finally add the productions

9) $\eta x \to x\eta, \ \forall\, x \in I_N \cup I_T$

to $F'$.

Instead of the word $P$ we get in the derivation $P' \in I_T^* \cup \{\eta\}$, where $P'$ is the word $P$ plus possibly some extra letters $\eta$. Otherwise the proof remains the same.

## Acknowledgement

The author is indebted to Professor A. Salomaa for suggesting the subject and for helpful comments on this paper.

Institute for Applied Mathematics
University of Turku, Finland

## References


1] GINSBURG, S.: The Mathematical Theory of Context-Free Languages - McGraw-Hill, New York, 1966.

[2] GINSBURG, S. and GREIBACH, S. A.: Mappings which preserve context sensitive languages - Inform. Contr. 9 (1966), 563—582.

[3] GINSBURG, S. and SPANIER, E. H.: Control sets on grammars - Math. Systems Theory 2 (1968), 159—177.

[4] SALOMAA, A.: Theory of Automata - Pergamon Press, Oxford, 1969.

[5] —»— On the index of a context-free grammar and language - Inform. Contr. 14 (1969), 474—477.

[6] —»— On grammars with restricted use of productions - Ann. Acad. Sci. Fenn., Ser. A I 454 (1969).

[7] STOTSKIĬ, E. D. (Стоцкий, Э. Д.): О некоторых ограничениях на способ вывода в грамматиках непосредственных составляющих - Автоматизация перевода текста, Научно-техническая информация 7 (1967), 35—38.

[8] FRIANT, J.: Grammaires ordonnees — grammaires matricielles. MA — 101, Université de Montreal, Octobre 1968.



Printed December 1970