ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

518

AN EXTENSION FOR THE CONCEPT OF FINITE INDEX OF A CONTEXT-FREE GRAMMAR

BY

TIMO LEPISTÖ

HELSINKI 1972
SUOMALAINEN TIEDEAKATEMIA

doi:10.5186/aasfm.1972.518

ISBN 951—41—0058—1

Communicated 14 April 1972 by ARTO SALOMAA

KESKUSKIRJAPAINO
HELSINKI 1972

An extension for the concept of finite index of a context-free grammar

1. BRAINERD [1] and SALOMAA [7] have considered the concept of finite index of a context-free grammar. Salomaa extends the notion of index to include context-free languages and he also proves that the family of languages of finite index is properly included in the family of context-free languages. Problems related to the concept of finite index have also been considered by YNTEMA [9], NIVAT [6], GINSBURG and SPANIER [3] and GRUSKA [4]. In [5] we considered an extension of finite index to the case of ordered context-free grammars. In this paper we consider this extension in the case, where the relation by which the ordering is defined is empty and we thus have an ordinary context-free grammar and language. We shall show that our extension of the concept of finite index is so general that for every context-free language there exists a grammar which has this property and generates the language in question.

Let $G = (I_N, I_T, X_0, F)$ be a context-free grammar, where I_N is the set of nonterminals, I_T the set of terminals, $X_0 \in I_N$ is the initial symbol and F is the set of productions. For any word P, $\lg(X \mid P)$ denotes the number of occurrences of the letter X in P and $\lg P$ the length of the word P.

Let L be the language generated by G and let

(1) $$D\colon X_0 = P_0 \Rightarrow \cdots \Rightarrow P_r = Q$$

be a derivation according to G. By the length of a derivation we mean the number of times we have applied productions. Thus, the length of the derivation (1) equals r. If there exist a natural $u(i)$ and an integer j such that

(2) $$\begin{cases} \lg(X_i \mid P_j) < u(i) \\ \lg(X_i \mid P_{j+1}) \geqq u(i), \end{cases} \qquad (X_i \in I_N),$$

then we say that the derivation D goes through the point $u(i)$ with respect to X_i. We further say that the derivation D goes through $u(i)$ k_i times with respect to X_i, if there exist k_i distinct indices j for which the condition (2) holds. We say that a grammar G possesses the finite point property with respect to a set $S(\subset I_N)$ (f.p.p. S) iff for each

$X_i \in S$ there exist natural numbers $u(i)$ and v_i such that every word $Q(\in L)$ has a derivation according to G which goes through $u(i)$ with respect to X_i at most v_i times. We see immediately that if a grammar is of finite index, it also possesses f.p.p. I_N. If a grammar, on the contrary, possesses f.p.p. I_N, it may be impossible to assert any bound for the number of occurrences of a nonterminal. Therefore the condition that a grammar possesses f.p.p. I_N is not so strict as the condition that a grammar is of finite index. The following theorem shows, on the other hand, that the extension is an essential one:

Theorem. *For every context-free language L there exists a grammar $\bar{G} = (\bar{I}_N, I_T, \bar{X}_0, \bar{F})$ such that $L = L(\bar{G})$ and $\bar{G}$ possesses* f.p.p. $\bar{I}_N$. *More specifically, $\bar{I}_N = I_N \cup I'_N$ $(I_N \cap I'_N = \Phi)$ in such a way that $o(I_N) = o(I'_N)$ or $o(I_N) = o(I'_N) - 1$ if $\lambda \notin L$ or $\lambda \in L$ respectively and every word of L has a derivation according to $\bar{G}$ which goes through* 1 *at most once with respect to each nonterminal of I_N and through* 3 *zero times with respect to each nonterminal of I'_N.*

2. Before going to the proof of the above theorem we consider some preliminary concepts. Assume in the following that $\lambda \notin L$. Because for every context-free grammar there exists an equivalent grammar in the Chomsky normal form (cf. [2] and [8]) we may assume that G is in the Chomsky normal form. This means that all the productions of G are of the two forms $X \to YZ$ and $X \to a$, where X, Y, Z are nonterminals and a is a terminal letter. Denote

$$I'_N = \{\bar{X} \mid X \in I_N\}$$

and

$$F' = \{X \to \bar{Y}Z,\ X \to Y\bar{Z} \mid X \to YZ \in F\},$$

$$F'' = \{\bar{X} \to P \mid X \to P \in F \cup F',\ \lg P = 2\},$$

$$F''' = \{\bar{X} \to \bar{Y}\bar{Z},\ X \to \bar{Y}\bar{Z} \mid X \to YZ \in F\},$$

$$F^{(4)} = \{X \to \bar{X} \mid X \in I_N\}.$$

Consider the grammars

$$\bar{G}' = (\bar{I}_N, I_T, X_0, \bar{F}')$$

and

$$\bar{G} = (\bar{I}_N, I_T, \bar{X}_0, \bar{F}),$$

where $\bar{I}_N = I_N \cup I'_N$, $\bar{F}' = F \cup F' \cup F''$ and $\bar{F} = \bar{F}' \cup F''' \cup F^{(4)}$. The following lemma is obvious:

Lemma 1. *The grammars* G, $\bar{G}'$ *and* $\bar{G}$ *are equivalent.*

In a word P over $\bar{I}_N$ a nonterminal X or $\bar{X}$ may be in two states, namely, in the yes-state and in the no-state. A derivation can change the state according to the following rules. Let $P_1 \Rightarrow P_2$ be a step in a derivation according to some of the grammars $G, \bar{G}'$ and $\bar{G}$. If, for a nonterminal X, $\lg(X \mid P_1) = \lg(X \mid P_2)$, then X is in the same state in both words P_1 and P_2; if $\lg(X \mid P_1) < \lg(X \mid P_2)$, then X is in the yes-state in the word P_2 (and thus the state changes if X is in the no-state in the word P_1); finally, if $\lg(X \mid P_1) = \lg(X \mid P_2) + 1$, then X is in the no-state in the word P_2. Respectively we define the changes for a nonterminal $\bar{X}$. We assume that in the words X_0 and $\bar{X}_0$ (the initial symbols) the nonterminals X_0 and $\bar{X}_0$ are in the yes-state respectively. Thus, it should be noted that the state of X (or $\bar{X}_0$) in some word P depends on the derivation by which we get P from X_0 or $\bar{X}_0$. Let $P_1 \Rightarrow P_2$ be a derivation according to the grammar $G, \bar{G}'$ or $\bar{G}$. Let $S(P_1, P_2)$ be the subset of $\bar{I}_N$ such that

$$S(P_1, P_2) = \{X, \bar{X} \mid X \in I_N, \bar{X} \in I'_N, \lg(X \mid P_1) = \lg(X \mid P_2) + 1, \lg(\bar{X} \mid P_1) = \lg(\bar{X} \mid P_2) + 1\}.$$

We say that the derivation satisfies the condition A iff X (or $\bar{X}$) belongs to the set $S(P_1, P_2)$ only if X (or $\bar{X}$) is in the yes-state in the word P_1. In this case we denote

$$P_1 \underset{G,A}{\Rightarrow} P_2, \quad P_1 \underset{\bar{G}',A}{\Rightarrow} P_2, \quad P_1 \underset{\bar{G},A}{\Rightarrow} P_2, \tag{3}$$

where we have a derivation according to $G, \bar{G}'$ or $\bar{G}$ respectively. If some word P generates a word Q according to $G, \bar{G}'$ or $\bar{G}$ as follows:

$$P = P_0 \Rightarrow P_1 \Rightarrow \cdots \Rightarrow P_r = Q \quad (r \geqq 1),$$

where every step $P_i \Rightarrow P_{i+1}$ $(i = 0, 1, \ldots, r-1)$ satisfies the condition A, we say that the derivation $P \overset{*}{\Rightarrow} Q$ satisfies the condition A and denote analogously to (3)

$$P \underset{G,A}{\overset{*}{\Rightarrow}} Q, \quad P \underset{\bar{G}',A}{\overset{*}{\Rightarrow}} Q, \quad P \underset{\bar{G},A}{\overset{*}{\Rightarrow}} Q.$$

(The derivation $P \overset{*}{\Rightarrow} P$ is defined to satisfy the condition A.) We now prove

Lemma 2. *Let there exists a derivation*

$$\bar{X}_0 \underset{\bar{G}}{\overset{*}{\Rightarrow}} T_1 \bar{X} T_2$$

such that $\bar{X}$ *is in the yes-state in* $T_1\bar{X}T_2$ *and* T_1 *and* T_2 *are words over* $I_N \cup I_T$. *If there exists a derivation*

$$X \underset{G}{\overset{*}{\Rightarrow}} S_1 Y S_2 \; (X \, , \, Y \in I_N) \tag{4}$$

of length $\geqq 1$, *where* S_1 *and* S_2 *are words over* $I_N \cup I_T$, *then there exists a derivation*

$$T_1\bar{X}T_2 \underset{\bar{G}',A}{\overset{*}{\Rightarrow}} T_1 S_1' Y S_2' T_2 \tag{5}$$

of length $\geqq 1$ *such that*

$$S_1' Y S_2' \underset{G}{\overset{*}{\Rightarrow}} S_1 Y S_2 \, .$$

In (5) *we apply only productions the left-hand sides of which belong to* I_N'. *Every word, except* $T_1 S_1' Y S_2' T_2$ *of the derivation* (5) *contains exactly one nonterminal of* I_N' *and* $S_1' Y S_2'$ *is a word over* I_N. *If in the word* $T_1\bar{X}T_2$ *some nonterminal of* T_1 *or* T_2 *is in the yes-state, so it is in the word* $T_1 S_1' Y S_2' T_2$.

Proof. We prove lemma 2 by induction on the length of the derivation (4). Assume first that the derivation (4) is of the length 1. Then (4) is $X \underset{G}{\Rightarrow} YZ$ or $X \underset{G}{\Rightarrow} ZY$. Consider, for instance, the derivation

$$T_1\bar{X}T_2 \underset{\bar{G}',A}{\Rightarrow} T_1 Y Z T_2 \, .$$

We can see that this derivation satisfies the conditions of the derivation (5) for arbitrary T_1 , T_2 over $I_N \cup I_T$. Therefore the lemma is true in this case.

Assume now that the lemma is true for all words T_1 , T_2 over $I_N \cup I_T$, if the length of the derivation (4) is smaller than $n(\geqq 2)$. Consider a derivation (4) of the length n. Write it in the form

$$X \underset{G}{\Rightarrow} ZU \underset{G}{\overset{*}{\Rightarrow}} S_1 Y S_2 \, . \tag{6}$$

We can now conclude that there exist derivations

$$Z \underset{G}{\overset{*}{\Rightarrow}} S_1 Y S_3 \, , \; U \underset{G}{\overset{*}{\Rightarrow}} S_4 \, ,$$

where $S_2 = S_3 S_4$ or derivations

$$Z \underset{G}{\overset{*}{\Rightarrow}} S_5 \, , \; U \underset{G}{\overset{*}{\Rightarrow}} S_6 Y S_2 \, ,$$

where $S_1 = S_5 S_6$. Suppose, for instance, the preceding case; the other case can be treated analogously. Assume first that the length of the derivation

(7) $$Z \underset{G}{\overset{*}{\Rightarrow}} S_1 Y S_3$$

equals 0. Consequently $S_1 = S_3 = \lambda$, $Y = Z$ and $S_2 = S_4$. We thus have a derivation

(8) $$U \underset{G}{\overset{*}{\Rightarrow}} S_2 .$$

Consider the derivation

$$T_1 \bar{X} T_2 \underset{\bar{G}',A}{\Rightarrow} T_1 Y U T_2 .$$

We can see, by (8) and the inductive hypothesis, that this derivation satisfies the conditions of the derivation (5) in lemma 2.

Assume now that the length of the derivation (7) is $\geqq 1$. Because it must be $< n$, we can decide, by induction hypothesis, that if $\bar{Z}$ is in the yes-state in the word $T_1 \bar{Z} U T_2$, then there exists a derivation

(9) $$T_1 \bar{Z} U T_2 \underset{\bar{G}',A}{\overset{*}{\Rightarrow}} T_1 S_1' Y S_3' U T_2$$

of the length $\geqq 1$ such that $S_1' Y S_3' \overset{*}{\Rightarrow} S_1 Y S_3$ according to G. In (9) we apply only productions the left-hand sides of which belong to I_N'. Every word, except $T_1 S_1' Y S_3' U T_2$ contains exactly one nonterminal of I_N' and $S_1' Y S_3'$ is a word over I_N. If in the word $T_1 \bar{Z} U T_2$ some nonterminal of I_N is in the yes-state, then so it is in the word $T_1 S_1' Y S_3' U T_2$. In addition, all the nonterminals of $S_1' Y S_3'$ are in the yes-state in the word $T_1 S_1' Y S_3' U T_3$. Consider the derivation

(10) $$T_1 \bar{X} T_2 \underset{\bar{G}'}{\Rightarrow} T_1 \bar{Z} U T_2 \underset{\bar{G}'}{\overset{*}{\Rightarrow}} T_1 S_1' Y S_3' U T_2 .$$

This derivation satisfies the condition of the derivation (5) in lemma 2. Because $\bar{X}$ is in the yes-state in the word $T_1 \bar{X} T_2$, it follows that $\bar{Z}$ is in the yes-state in $T_1 \bar{Z} U T_2$ and the whole derivation (10) satisfies the condition A. Thus we can write (10) in the form

(10)′ $$T_1 \bar{X} T_2 \underset{\bar{G}',A}{\overset{*}{\Rightarrow}} T_1 S_1' Y S_3' U T_2 .$$

Further $S_1' Y S_3' U \overset{*}{\Rightarrow} S_1 Y S_3 U \overset{*}{\Rightarrow} S_1 Y S_3 S_4 = S_1 Y S_2$ according to G. Also we see immediately that all the productions we have applied in (10)′ start from nonterminals of I_N' and every word, except $T_1 S_1' Y S_3' U T_2$, contains one nonterminal of I_N'. By induction hypothesis, $S_1' Y S_3' U$ is a word over I_N. Let some nonterminal of I_N be in the yes-state in the word $T_1 \bar{X} T_2$. Then it is also in the yes-state in the word $T_1 \bar{Z} U T_2$ and therefore, by the induction hypothesis, also in the yes-state in the word $T_1S'_1YS'_3UT_2$. The nonterminal U is in the yes-state in word $T_1\bar{Z}UT_2$. By induction hypothesis it is in the yes-state in the word $T_1S'_1YS'_3UT_2$. Therefore it follows that all the nonterminals of $S'_1YS'_3U$ are in the yes-state in the word $T_1S'_1YS'_3UT_2$. Our lemma is thus established.

3. We now begin the proof of our above theorem. Assume as above that $\lambda \notin L(G)$ and G is in the Chomsky normal form. Consider a derivation (1) according to G. Let the word Q be fixed in the following way. Let there exist a derivation

$$X_0 \underset{G}{\overset{*}{\Rightarrow}} S_1YS_2$$

of length $\geqq 1$ such that $S_1YS_2 \overset{*}{\Rightarrow} Q$ according to G. By lemma 2, we then have a derivation

$$\bar{X}_0 \underset{\bar{G}',A}{\overset{*}{\Rightarrow}} S'_1YS'_2$$

such that

$$S'_1YS'_2 \underset{G}{\overset{*}{\Rightarrow}} S_1YS_2 \underset{G}{\overset{*}{\Rightarrow}} Q\ .$$

On the other hand, all the nonterminals of $S'_1YS'_2$ are in the yes-state in the word $S'_1YS'_2$. Let X be a nonterminal of $S'_1YS'_2$ such that $S'_1YS'_2$ is of the form T_1XT_2 (by lemma 2 , T_1 and T_2 are words over $I_N \cup I_T$) and there exists a derivation

$$X \underset{G}{\overset{*}{\Rightarrow}} S_3ZS_4$$

such that

$$T_1S_3ZS_4T_2 \underset{G}{\overset{*}{\Rightarrow}} Q\ . \tag{11}$$

By lemma 2, we thus have

$$\bar{X}_0 \underset{\bar{G},A}{\overset{*}{\Rightarrow}} T_1XT_2 \underset{\bar{G},A}{\Rightarrow} T_1\bar{X}T_2 \underset{\bar{G}',A}{\overset{*}{\Rightarrow}} T_1S'_3ZS'_4T_2\ ,$$

where

$$S'_3ZS'_4 \underset{G}{\overset{*}{\Rightarrow}} S_3ZS_4\ .$$

Hence, by (11),

$$T_1S'_3ZS'_4T_2 \underset{G}{\overset{*}{\Rightarrow}} Q\ .$$

Every nonterminal of $T_1S'_3ZS'_4T_2$, except possibly X, is in the yes-state in the word $T_1S_3'ZS_4'T_2$. We continue in this way to obtain a derivation

(12) $$\bar{X}_0 \overset{*}{\underset{\bar{G},A}{\Rightarrow}} P$$

such that $P \overset{*}{\Rightarrow} Q$ according to G and every nonterminal of P is in the no-state or if some nonterminal, say X, is in P in the yes-state and P is of the form $P = T_1XT_2$, then there exists no derivation of the form

$$X \overset{*}{\underset{G}{\Rightarrow}} S_1YS_2$$

of length $\geqq 1$ such that

$$T_1S_1YS_2T_2 \overset{*}{\underset{G}{\Rightarrow}} Q .$$

If X is in the yes-state in the word P, then the only applicable productions which start from X are of the form $X \to a$ (a is a terminal letter).

Assume that P is of the form T_1XT_2 and there exists a derivation

(13) $$X \overset{*}{\underset{G}{\Rightarrow}} BXC$$

of length $\geqq 1$ such that B and C are words over $I_N \cup I_T$ and

(14) $$T_1BXCT_2 \overset{*}{\underset{G}{\Rightarrow}} Q .$$

We then say that P has a cycle. Let $X_1, X_2, \cdots, X_n (n \geqq 2)$ be some distinct nonterminals of P such that P is of the form

(15) $$P = T_1X_1T_2X_2 \ldots T_nX_nT_{n+1} .$$

We also say that P has a cycle if there exist derivations

(16) $$\begin{cases} X_1 \overset{*}{\underset{G}{\Rightarrow}} B_1X_{i(2)}\,C_1 , \\ X_{i(2)} \overset{*}{\underset{G}{\Rightarrow}} B_{i(2)}\,X_{i(3)}\,C_{i(2)} , \\ \cdots\cdots \\ X_{i(n)} \overset{*}{\underset{G}{\Rightarrow}} B_{i(n)}X_1\,C_{i(n)} \end{cases}$$

such that $X_{i(j)} \neq X_{i(k)}$ if $j \neq k$ and $X_{i(2)}, \ldots, X_{i(n)}$ are the nonterminals $X_2, \ldots, X_n$ in some order, B's and C's are words over $I_N \cup I_T$ and, in addition,

(17) $$T_1X_1T_2 \cdots T_{i(j)}B_{i(j)}X_{i(j+1)}C_{i(j)}T_{i(j)+1} \cdots T_{n+1} \overset{*}{\underset{G}{\Rightarrow}} Q \,,$$

where j runs through the values $1, \ldots, n$ $(i(1) = i(n+1) = 1)$.

Assume that P has a cycle of the form (13). Let R be the last word in the derivation (12), where X is in the yes-state. Assume that R is of the form $N_1XN_2XN_3$. Without loss of generality we may assume that the occurrence of the nonterminal X between N_1 and N_2 disappears and the occurrence of X between N_2XN_3 remains in the derivation

(18) $$N_1XN_2XN_3 \overset{*}{\underset{\bar{G},A}{\Rightarrow}} P$$

which is a part of the derivation (12). Also we find that R cannot contain a nonterminal of I'_N, because in that case X would also be in the yes-state in the following word (by lemma 2) contradicting the choice of R. Because P is in this case of the form T_1XT_2 we may conclude that there exist derivations

(19) $$N_1XN_2 \overset{*}{\underset{\bar{G}}{\Rightarrow}} T_1 \,,\; N_3 \overset{*}{\underset{\bar{G}}{\Rightarrow}} T_2 \,.$$

The derivations (19) are obtained because in the derivation (18) we do not apply any production for the nonterminal X between N_2 and N_3. It should be noted that, by lemma 2, the only productions which we apply for X in the derivation (12) (and consequently in the derivation (18)) are of the form $X \rightarrow \bar{X}$. Because P has a cycle of the form (13) there exists, by lemma 2, a derivation

$$R = N_1XN_2XN_3 \underset{\bar{G},A}{\Rightarrow} N_1XN_2\bar{X}N_3 \overset{*}{\underset{\bar{G}',A}{\Rightarrow}} N_1XN_2B'XC'N_3 = R_1 \,.$$

Every nonterminal which is in the yes-state in the word R is also, by lemma 2, in the yes-state in the word R_1. Therefore we can apply the derivation (18) for R_1 and get, by (19),

$$R_1 \overset{*}{\underset{\bar{G},A}{\Rightarrow}} T_1B'XC'T_2 = P'_1 \,.$$

By lemma 2, it follows that

$$B'XC' \overset{*}{\underset{G}{\Rightarrow}} BXC \,.$$

Hence, by (14) $P'_1 \overset{*}{\Rightarrow} Q$ according to G. If it is possible, we now continue from P'_1 in the same way as in the derivation (12). We thus have a derivation

(20) $$\bar{X}_0 \underset{\bar{G},A}{\overset{*}{\Rightarrow}} P_1' \underset{\bar{G},A}{\overset{*}{\Rightarrow}} P' ,$$

where $\lg P' \geqq P_1' > \lg P$ and $P' \overset{*}{\Rightarrow} Q$ according to G.

Assume now that P' has a cycle of the form (16) and P' is thus of the form $P' = T_1X_1T_2X_2 \cdots T_nX_nT_{n+1}$. Because $P' \overset{*}{\Rightarrow} Q$ according to G, we can infer that

(21) $$T_1 \underset{G}{\overset{*}{\Rightarrow}} Q_1 , X_1 \underset{G}{\overset{*}{\Rightarrow}} Q_1' , T_2 \underset{G}{\overset{*}{\Rightarrow}} Q_2 , X_2 \underset{G}{\overset{*}{\Rightarrow}} Q_2' , \ldots , T_{n+1} \underset{G}{\overset{*}{\Rightarrow}} Q_{n+1}$$

such that $Q_1Q_1'Q_2Q_2' \cdots Q_{n+1} = Q$. Because of the relations (17), it follows that

(22) $$B_{i(j)}X_{i(j+1)}C_{i(j)} \underset{G}{\overset{*}{\Rightarrow}} Q_{i(j)}' \quad (j = 1 , 2 , \ldots , n) .$$

Let R be the word in the derivation (20) such that one of the nonterminals $X_1 , X_2 , \ldots , X_n$ is in the yes-state in R and in the other words of the derivation

(23) $$R \underset{\bar{G},A}{\overset{*}{\Rightarrow}} P'$$

(which is a part of the derivation (20)) all the nonterminals $X_1 , \ldots , X_n$ are in the no-state. Assume, for instance, that X_1 is in the yes-state in R. The case, where some other of the nonterminals $X_1 , \ldots , X_n$ is in the yes-state in R can be treated analogously. Because the only productions the left-hand sides of which belong to I_N and which we have possibly applied in (20) are of the form $X \to \bar{X}$, we can conclude that in the derivation (23) we have not applied any production for the nonterminals $X_2 , \ldots , X_n$ or for the nonterminal X_1 which remains in the derivation (23). Suppose that R is of the form

$$R = N_1'X_1N_1X_1N_2X_2N_3 \ldots X_nN_{n+1}$$

and the nonterminal X_1 between N_1' and N_1 disappears in the beginning of the derivation (23). From the above it now follows that

(24) $$\begin{cases} N_1'X_1N_1 \underset{\bar{G}}{\overset{*}{\Rightarrow}} T_1 , \\ N_i \underset{\bar{G}}{\overset{*}{\Rightarrow}} T_i \quad (i = 2 , \ldots , n + 1) . \end{cases}$$

Because X_1 is in the yes-state in R we have, by (16) and lemma 2,

$$R \underset{\bar{G},A}{\Rightarrow} N_1'X_1N_1\bar{X}_1 \cdots N_{n+1} \underset{\bar{G}',A}{\overset{*}{\Rightarrow}} N_1'X_1N_1B_1'X_{i(2)}C_1' \cdots N_{n+1} = R_1$$

It should be noted that it follows from the choice of R that $N_1'X_1N_1$ and $N_2 \cdots N_{n+1}$ are words over $I_N \cup I_T$ and we can apply lemma 2. Because in R_1 $X_{i(2)}$ is in the yes-state (by lemma 2) we get further, by (16) and lemma 2,

$$R_1 \underset{\bar{G},A}{\Rightarrow} N_1'X_1N_1B_1'X_{i(2)}C_1' \cdots \bar{X}_{i(2)} \cdots N_{n+1} \overset{*}{\underset{\bar{G}',A}{\Rightarrow}}$$

$$N_1'X_1N_1B_1'X_{i(2)}C_1' \cdots B_{i(2)}'X_{i(3)}C_{i(2)}' \cdots N_{n+1} = R_2 \,.$$

Continuing in the same way we finally get

$$R_{n-1} \overset{*}{\underset{\bar{G},A}{\Rightarrow}} N_1' \cdots B_{i(n)}'X_1C_{i(n)}' \cdots N_{n+1} = R_n \,.$$

By lemma 2, it follows that each nonterminal which is in the yes-state in R is also in the yes-state in R_n. Therefore we can apply the derivation (23) for R_n and we thus get, by (24).

$$R_n \overset{*}{\underset{\bar{G},A}{\Rightarrow}} T_1B_1'X_{i(2)}C_1' \cdots B_{i(n)}'X_1C_{i(n)}' \cdots T_{n+1} = P_1''$$

It further follows, by lemma (2), (21) and (22), that

$$P_1'' \overset{*}{\underset{G}{\Rightarrow}} T_1B_1X_{i(2)}C_1 \cdots T_{i(n)}B_{i(n)}X_1C_{i(n)} \cdots T_{n+1}$$

$$\overset{*}{\underset{G}{\Rightarrow}} T_1Q_1' \cdots T_{i(n)}Q_{i(n)}' \cdots T_{n+1}$$

$$\overset{*}{\underset{G}{\Rightarrow}} Q_1Q_1' \cdots Q_{i(n)}Q_{i(n)}' \cdots Q_{n+1} = Q \,.$$

If it is possible, we now continue from P_1'' in the same way as in the derivation (12). We thus have a derivation

$$\bar{X}_0 \overset{*}{\underset{\bar{G},A}{\Rightarrow}} P_1'' \overset{*}{\underset{\bar{G},A}{\Rightarrow}} P'' \,,$$

where $\lg P'' \geqq \lg P_1'' > \lg P' > \lg P$ and $P'' \overset{*}{\Rightarrow} Q$ according to G. In this way we continue eliminating one cycle after another. Because the length of the word Q is fixed we finally get a derivation

$$\bar{X}_0 \overset{*}{\underset{\bar{G},A}{\Rightarrow}} P^{(i)} \; (i \geqq 0 \,, P^{(0)} = P) \,, \tag{25}$$

where $P^{(i)}$ has no cycles. In each word of the derivation (25) there exists at most one nonterminal of I_N'.

In $P^{(i)}$ there exists a nonterminal X with the property that if Y runs through all the nonterminals of $P^{(i)}$ and $P^{(i)}$ is written in the form KYM, then there exist no derivations of the form

(26)
$$Y \overset{*}{\underset{G}{\Rightarrow}} BXC$$

of length $\geqq 1$ such that

$$KBXCM \overset{*}{\underset{G}{\Rightarrow}} Q .$$

We prove this statement indirectly. In fact, assume the contrary, for every nonterminal of $P^{(i)}$ there exists at least one derivation of the form (26). Let $X_1 , X_2 , \ldots , X_n$ be the distinct nonterminals of $P^{(i)}$ and let $P^{(i)}$ respectively be of the form $P^{(i)} = K_j X_j M_j , j = 1 , 2 , \ldots , n$. We then have a sequence

$$X_{j(t+1)} \overset{*}{\underset{G}{\Rightarrow}} B_{j(t)} X_{j(t)} C_{j(t)} \quad (t = 1 , 2 , 3 \ldots)$$

such that

(27)
$$K_{j(t+1)} B_{j(t)} X_{j(t)} C_{j(t)} M_{j(t+1)} \overset{*}{\underset{G}{\Rightarrow}} Q .$$

Because the number of the nonterminals X_j is finite, there must be two distinct values of t, say k and $m (k < m)$, such that $j(k) = j(m)$ and $X_{j(k)} = X_{j(m)}$. We then have

$$X_{j(m)} \overset{*}{\underset{G}{\Rightarrow}} B_{j(m-1)} X_{j(m-1)} C_{j(m-1)} ,$$

$$X_{j(m-1)} \overset{*}{\underset{G}{\Rightarrow}} B_{j(m-2)} X_{j(m-2)} C_{j(m-2)} ,$$

. . . .

$$X_{j(k+1)} \overset{*}{\underset{G}{\Rightarrow}} B_{j(k)} X_{j(k)} C_{j(k)}$$

such that (27) holds. This means that $P^{(i)}$ has a cycle which is impossible.

Let, for instance, X be the nonterminal with the above property. We then eliminate X from the word $P^{(i)}$ by applying all possible productions which start from X. We thus have a derivation

(28)
$$P^{(i)} \overset{*}{\underset{G}{\Rightarrow}} E$$

such that

(29)
$$E \overset{*}{\underset{G}{\Rightarrow}} Q$$

and the nonterminal X does not occur in the derivation (29). Let

(30)
$$E \overset{*}{\underset{\bar{G},A}{\Rightarrow}} E_1$$

be any derivation which we get in the same way as the derivation (12), in other words, by applying productions of the form (5) and $X \to \bar{X}$ such that

$$E_1 \underset{G}{\overset{*}{\Rightarrow}} Q .$$

Assume further that E_1 is a word over $I_N \cup I_T$. The eliminated nonterminal X cannot occur in any derivation of the form (30). In fact, assume that there exists a derivation

$$E \underset{\bar{G},A}{\overset{*}{\Rightarrow}} E_2 ,$$

where $X \in E_2$ such that $E_2 \overset{*}{\Rightarrow} Q$ according to $\bar{G}$. If E_2 contains a nonterminal of I'_N, we can form a derivation

$$E \underset{\bar{G},A}{\overset{*}{\Rightarrow}} E_2 \underset{\bar{G}}{\overset{*}{\Rightarrow}} E_3$$

such that $X \in E_3$, $E_3 \overset{*}{\Rightarrow} Q$ according to $\bar{G}$ and E_3 is a word over $I_N \cup I_T$. We now have, by lemma 1, a derivation $E \overset{*}{\Rightarrow} E_3 \overset{*}{\Rightarrow} Q$ according to G. This, however, leads to a contradiction.

Assume that E_1 has a cycle, for instance, of the form (16). Suppose that the nonterminals $X_1, X_2, \ldots, X_n$ are in the no-state in the every word of the derivation

$$P^{(i)} \underset{\bar{G}}{\overset{*}{\Rightarrow}} E_1 \tag{31}$$

Let, for instance, $P^{(i)}$ be of the form

$$P^{(i)} = N_1X_1N_2X_2 \cdots X_nN_{n+1}$$

and E_1 of the form

$$E_1 = T_1X_1T_2X_2 \cdots X_nT_{n+1} .$$

Because in the derivation (31) we do not apply any production for the nonterminals $X_i (i = 1, 2, \ldots, n)$, we may infer that

$$N_i \underset{\bar{G}}{\overset{*}{\Rightarrow}} T_i \quad (i = 1, 2, \ldots, n+1) . \tag{32}$$

Because E_1 is assumed to have a cycle of the form (16), the relations (17), (21) and (22) hold. By (32) and lemma 1, we may now replace T_i in (21) by $N_i (i = 1, 2, \ldots, n+1)$ and thus we see that $P^{(i)}$ has a cycle which is impossible. Therefore we may conclude that in the derivation (31) there exists a word, where some of the nonterminals $X_1, \ldots, X_n$ are in the yes-state. If this word exists in the derivation (28), then the word E also has this property. We thus see that a word of this kind can always be found in the derivation (30). This means that if we eliminate a cycle from E_1, we do not change the derivations (25) and (28) in any way.

We now continue by eliminating cycles from the words in the same way as before and we thus get a derivation

$$E \underset{\bar{G},A}{\overset{*}{\Rightarrow}} E^{(i)} ,$$

where $E^{(i)}$ does not contain any cycles. If $E^{(i)}$ contains nonterminals of I_N then there exists in $E^{(i)}$ a nonterminal Y with the property that if Z runs through all the nonterminals of $E^{(i)}$ and $E^{(i)}$ is written in the form KZM, then there exist no derivations of the form

$$Z \underset{G}{\overset{*}{\Rightarrow}} BYC$$

such that

$$KBYCM \underset{G}{\overset{*}{\Rightarrow}} Q .$$

We eliminate this nonterminal by applying all possible productions which start from Y. Thus

$$E^{(i)} \underset{G}{\overset{*}{\Rightarrow}} W$$

such that

(33) $$W \underset{G}{\overset{*}{\Rightarrow}} Q .$$

The eliminated nonterminals X and Y do not appear in the derivation (33). Continuing in this way we finally get a derivation

(34) $$\bar{X}_0 \underset{\bar{G}}{\overset{*}{\Rightarrow}} Q .$$

This derivation has the following properties:

(1) The number of occurrences of a nonterminal never decreases by more than one before it again increases, with the exception of a part of the derivation (34), in which the number of occurrences of this nonterminal monotonically decreases and the nonterminal wholly disappears from the derivation.

(2) Each word of the derivation (34) contains at most one nonterminal of I'_N.

In order to reach the proof of the theorem, we have still to modify the derivation (34) to some degree. For a nonterminal $X \in I_N$, there exist three possibilities:

(i) X does not appear in the derivation (34) at all.
(ii) X occurs in each word of the derivation (34) at most once.
(iii) In the derivation (34) there exists a word, where X occurs twice.

In case (i) we make no changes in the derivation (34).

Consider case (ii). Assume that in the derivation (34) X appears and respectively disappears at least two times. When X appears the first time, we have applied a production, where X is in the right-hand side and which starts from a nonterminal different from X, for instance, $\bar{Y} \to X\bar{Z}$. We now replace this production by $\bar{Y} \to \bar{X}\bar{Z}$. When X in the derivation (34) disappears the first time, we have applied a production which is of the form $X \to \bar{X}$. This production is now unnecessary, because in the place of X there is already $\bar{X}$. In this way we can modify the derivation (34) in such a way that X appears and disappears only once in the modified derivation.

Consider case (iii). In the same way as in the preceding case we can eliminate X every time when it appears, occurs only once and disappears before we reach a word, where X occurs twice. After these arrangements we see that the derivation (34) goes through 1 once with respect to X.

After the above modifications we have a modified derivation (34) which goes through 1 at most once with respect to $X \in I_N$. On the other hand, in this modified derivation the nonterminal $\bar{X}$ of I'_N may occur twice in some words. We see, in addition, that the modification does not affect any other nonterminals of $\bar{I}_N$.

We now choose another nonterminal Y of I_N and modify the derivation again in such a way that we have a derivation which goes through 1 at most once with respect to Y. Continuing in this way we finally get a derivation

$$\bar{X}_0 \overset{*}{\underset{\bar{G}}{\Rightarrow}} Q \tag{35}$$

which goes through 1 at most once with respect to every nonterminal of I_N. Further (35) goes through 3 zero times with respect to every nonterminal of I'_N. Because the word Q was arbitrary we can construct a derivation of this kind for every word of L.

At the beginning of the proof we assumed that $\lambda \notin L$. If $\lambda \in L$, we first form a λ-free context-free grammar G'' such that $L(G'') = L - \{\lambda\}$ (see for instance [8]). There exists a context-free grammar G equivalent to G'', which is in the Chomsky normal form. For this grammar G we perform the proof as above and form an equivalent grammar $\bar{G}$ which satisfies our theorem. Then we add to the set I'_N a nonterminal $\bar{X}'_0$ which will be a new initial symbol. To the set of the productions of $\bar{G}$ we add the productions $\bar{X}_0' \to \bar{X}_0$ and $\bar{X}_0' \to \lambda$. This new grammar clearly satisfies our theorem and generates the language L . Our theorem is thus established.

University of Oulu
Oulu, Finland

References

[1] BRAINERD, B.: An analog of a theorem about context-free languages. — Inform. Control 11, 561—567 (1968).
[2] CHOMSKY, N.: On certain formal properties of grammars. — Inform. Control 2, 137—167 (1959).
[3] GINSBURG, S. and SPANIER, E. H.: Derivation — bounded languages. — J. Comput. System Sci. 2, 228—250 (1958).
[4] GRUSKA, J.: A characterization of context-free languages. — J. Comput System Sci. 5, 353—364 (1971).
[5] LEPISTÖ, T.: On ordered context-free grammars. — In preparation.
[6] NIVAT, M.: Transductions des languages de Chomsky, — Unpublished Doctoral dissertation, Grenoble University (1967).
[7] SALOMAA, A.: On the index of a context-free grammar and language. — Inform. Control 14, 474—477 (1969).
[8] SALOMAA, A.: Formal languages. — In preparation.
[9] YNTEMA, M. K.: Inclusion relations among families of context-free languages — Inform. Control 10, 572—597 (1967).

Printed September 1972