Armin B. Cremers; Kenneth M. Weiss

Some further remarks on the index of context-free languages

*Kybernetika*, Vol. 9 (1973), No. 6, (461)--466

Persistent URL: http://dml.cz/dmlcz/125835

KYBERNETIKA — VOLUME 9 (1973), NUMBER 6

# Some Further Remarks on the Index of Context-Free Languages

A. B. Cremers, K. Weiss

Let $\varkappa$ be a complexity measure for grammars. The following problem is investigated: Do there exist such context-free languages $L$ that no context-free grammar generating $L$ can be minimal both according to $\varkappa$ and according to the index? For a set of well known complexity measures the answer is in the affirmative.

## 1. THE INDEX OF GRAMMARS AND LANGUAGES

Let $G = (N, T, P, S)$ be a context-free grammar (CFG), where $N$ is the set of nonterminal symbols, $T$ the set of terminal symbols, $P \subset N \times (N \cup T)^*$ the set of productions and $S$ in $N$ the start variable. Let $\varepsilon$ denote the empty word and $L(G)$ the language generated by $G$.

Following [1], we now define the index of $G$. Let $F$ be a derivation of a word $w$ in $(N \cup T)^*$ according to $G$:

$$F : S = w_0 \Rightarrow^* w_1 \Rightarrow^* \ldots \Rightarrow^* w_n = w .$$

We define

$$\operatorname{Ind}(F) = \max \{l(d(w_i)) \mid 0 \leqq i \leqq n\} ,$$

where $d(w)$ is the word obtained from $w$ by deleting all terminal symbols, and for a word $w$, $l(w)$ denotes the length of $w$;

$$\begin{aligned} \operatorname{Ind}(w) &= \min \{\operatorname{Ind}(F) \mid F \text{ is a derivation of } w \text{ according to } G\} ; \\ \operatorname{Ind}(G) &= \max \{\operatorname{Ind}(w) \mid w \text{ in } L(G)\} , \\ \operatorname{Ind}(L) &= \min \{\operatorname{Ind}(G) \mid L = L(G)\} . \end{aligned}$$

In [5] the existence of a context-free language (CFL) of infinite index is proved and in [3] a hierarchy of context-free languages is established with respect to the index. This gives rise to the question how this hierarchy is related to well known complexity hierarchies of context-free languages. To this end, we collate in Section 2 the definitions of several complexity measures for grammars, as introduced in [2]. In Section 3 we show that, for a CFG, the requirements of simplicity with respect to such a complexity measure and with respect to the index are in general in conflict.

## 2. COMPLEXITY MEASURES FOR GRAMMARS

Let $G = (N, T, P, S)$ be a CFG. A binary relation $\rhd$ on $N$ is defined as follows. For $A$, $B$ in $N$ the relation $A \rhd B$ holds, iff there exist $x$, $y$ in $(N \cup T)^*$ such that $A \to xBy$ is a production in $P$. Let $\rhd^*$ denote the reflexive and transitive closure of the relation $\rhd$. The nonterminal symbols $A$ and $B$ are said to be equivalent, shortly $A \equiv B$, iff both $A \rhd^* B$ and $B \rhd^* A$ holds. Each equivalence class of $N$ according to $\equiv$ is called grammatical level of $G$ (cf. [2]). For a grammatical level $Q$ of $G$, let

$$\text{Depth}(Q) = \text{card}(Q) .$$

A grammatical level $Q$ is termed nontrivial if $\text{Depth}(Q) > 1$.

We define

$$\begin{aligned}
\text{Depth}(G) &= \max\{\text{Depth}(Q) \mid Q \text{ is a grammatical level of } G\} ,\\
\text{Lev}(G) &= \text{the number of grammatical levels of } G ,\\
\text{NLev}(G) &= \text{the number of nontrivial grammatical levels of } G ,\\
\text{Var}(G) &= \text{card}(N) ,\\
\text{Prod}(G) &= \text{card}(P) .
\end{aligned}$$

Let $\varkappa_\gamma$ be a complexity measure defined for a class $\gamma$ of grammars and $L$ a language which can be generated by a grammar in $\gamma$.

Then we define

$$\varkappa_\gamma(L) = \min\{\varkappa_\gamma(G) \mid G \text{ in } \gamma,\ L = L(G)\} .$$

If a complexity measure $\varkappa$ is defined for all CFG's and CFL's, respectively, we mostly omit the subscript of $\varkappa$.

## 3. INCOMPATIBILITY OF THE INDEX AND GRUSKA'S COMPLEXITY MEASURES

Let $\varkappa$ be one of Gruska's complexity measures of Section 2. In the following, we study the question whether there are CFL's $L$ such that no CFG generating $L$ can be minimal both according to $\varkappa$ and according to the index. As it will be shown in this section, the answer to this question is in the affirmative for each complexity criterion of Section 2.

Let $\gamma$ denote a class of grammars and $\Gamma = \{L = L(G) \mid G \text{ in } \gamma\}$. For a complexity measure $\varkappa_\gamma$ defined on $\gamma$ and a language $L$ in $\Gamma$ let 

$$\varkappa_\gamma^{-1}(L) = \{G \in \gamma \mid L = L(G),\ \varkappa_\gamma(G) = \varkappa_\gamma(L)\}\,.$$

**Definition.** Two complexity measures $\varkappa_{\gamma,1}$ and $\varkappa_{\gamma,2}$ are said to be compatible iff

$$\varkappa_{\gamma,1}^{-1}(L) \cap \varkappa_{\gamma,2}^{-1}(L) \neq \emptyset$$

for each $L$ in $\Gamma$.

Let c and lin denote the class of all context-free grammars and the class of all linear grammars, respectively.

The proofs of the results in this section are based on the following consideration: Clearly, for each linear language $L$, Ind $(L) = 1$ holds; furthermore, Ind $(G) = 1$ iff $G$ is a linear grammar. Thence, in order to show that a complexity measure $\varkappa$ and Ind are incompatible, it is sufficient to construct a linear language $L$ such that for a nonnegative integer $n$ both $\varkappa_c(L) \leqq n$ and $\varkappa_{\mathrm{lin}}(L) > n$ holds.

**Theorem 1.**

(1) Var *and* Ind *are incompatible.*

(2) Lev *and* Ind *are incompatible.*

Proof. Let $R = \{b\}^*\, a\{b\}^*\, a\{b\}^*\, a\{b\}^*\, a$. $R$ is also written in the form

$$R = R_1 R_2$$

where $R_1 = \{b\}^*$ and $R_2 = a\{b\}^*\, a\{b\}^*\, a\{b\}^*\, a$ and

$$R = R_3 a R_4 a R_5 a R_6 a$$

where

$$R_i = \{b\}^*\,, \quad 3 \leqq i \leqq 6\,.$$

(1) $R$ is generated by the following grammar:

$$G_1 = (\{S, A\},\ \{a, b\},\ \{S \to AaAaAaAa,\ A \to bA,\ A \to \varepsilon\},\ S)\,.$$

Thence, Var $(R) \leqq 2$.

Next we show that $\mathrm{Var}_{\mathrm{lin}}(R) > 2$.

Assume $\mathrm{Var}_{\mathrm{lin}}(R) = 1$. Let $G_2$ be a linear grammar with only one variable $S$ generating $R$. For a word $x = x_1 x_2 \ldots x_n$ of arbitrary length, $x_i$ in $\{a, b\}$, let $q(x)$ denote the number of indices $i$ such that $x_i \neq x_{i+1}$. Clearly, for all $w$ in $R$, $q(w) \leqq 7$ holds.

If $S \to \beta_1 S \beta_2$ is a production in $G_2$, then $q(\beta_1) = q(\beta_2) = 0$. Otherwise, a word $\beta_1^8 w \beta_2^8$ could be generated which does not belong to $R$. But then we may conclude that $\beta_1$ is in $\{b\}^*$ and $\beta_2 = \varepsilon$. Thence, by productions of the form $S \to \beta_1 S \beta_2$ only $R_1$ is generated. Therefore, $R \neq L(G_2)$.

Assume $\mathrm{Var}_{\mathrm{lin}}(R) = 2$ and let $G_3$ be a linear grammar for $R$ with only two variables $S$ and $A$. If $S \equiv A$ then for all $\beta_1, \beta_2, \beta_3, \beta_4$ in $T^*$ with $S \Rightarrow^* \beta_1 A \beta_2$ and $A \Rightarrow^* \beta_3 S \beta_4$, $\beta_1\beta_3$ in $\{b\}^*$ and $\beta_2\beta_4 = \varepsilon$ holds. Furthermore, if $A \to \alpha_1 A \alpha_2$ and $S \to \gamma_1 S \gamma_2$ are productions of $G_3$, then $\alpha_1\gamma_1$ is in $\{b\}^*$ and $\alpha_2\gamma_2 = \varepsilon$. Thence, only $R_1$ is generated by the productions considered so far. If $S \not\equiv A$, then $R_4 a R_5 a R_6$ must be generated by productions of the form $A \to \alpha_1 A \alpha_2$ and $A \to \gamma$ where $\alpha_1, \alpha_2, \gamma$ in $T^*$; but this is impossible. Therefore, $\mathrm{Var}_{\mathrm{lin}}(R) > 2$.

(2) Since $R = L(G_1)$, $\mathrm{Lev}(R) \leqq 2$ holds.

We show that $\mathrm{Lev}_{\mathrm{lin}}(R) > 2$:

Clearly, $\mathrm{Lev}_{\mathrm{lin}}(R) > 1$. Assume $\mathrm{Lev}_{\mathrm{lin}}(R) = 2$. Then there is a linear grammar $G_4$ generating $R$. Let $N_1 = \{S = A_0, A_1, \ldots, A_n\}$ and $N_2 = \{B_1, \ldots, B_m\}$ be the equivalence classes of nonterminal symbols of $G_4$ according to $\equiv$. If $A_i \to \alpha A_j \beta$, $0 \leqq i \leqq n$, $1 \leqq j \leqq n$, is a production of $G_4$, then $\alpha$ is in $\{b\}^*$ and $\beta = \varepsilon$ holds. Thence, $R_4 a R_5 a R_6$ must be generated by productions whose left-hand sides are in $N_2$, i.e. productions of the form $B_i \to \alpha' B_j \beta'$ and $B_i \to \gamma$. Since $\alpha'\beta'$ must be in $\{b\}^*$ we get a contradiction. Therefore, $\mathrm{Lev}_{\mathrm{lin}}(R) > 2$.

**Theorem 2.** Depth *and* Ind *are incompatible.*

Proof. Let $R = \{\{b\}^* a\{b\}^* a\{b\}^* a\{b\}^* a\}^+ a$. $R$ is a regular language, therefore $\mathrm{Depth}(R) = 1$. (For a set of words $M$, $M^+$ denotes the $\varepsilon$-free catenation closure of $M$.)

In the following, we show that $\mathrm{Depth}_{\mathrm{lin}}(R) > 1$:

Assume that there is a linear grammar $G = (N, T, P, S)$ such that $\mathrm{Depth}(G) = 1$ and $R = L(G)$. Let $N = \{S = A_1, \ldots, A_n\}$. $G$ is a sequential grammar, i.e.

$$A_i \rhd^* A_j \quad \text{implies} \quad i \leqq j\,, \quad 1 \leqq i \leqq n\,.$$

At first we consider productions of the form

$$A_i \to \alpha_{ij} A_i \beta_{ij}\,,$$

$1 \leqq i \leqq n$, $1 \leqq j \leqq n_i$. Let $l_a(w)$ denote the number of occurences of $a$ in a word $w$.

**Assertion 1.** *For each production* $A_i \to \alpha_{ij} A_i \beta_{ij}$ *there is a nonnegative integer* $q$ *such that*

$$l_a(\alpha_{ij}\beta_{ij}) = 4q\,.$$

Proof. Let $x_1$ in $R$ be so that there is a derivation of $x_1$ according to $G$ in which the production $A_i \to \alpha_{ij} A_i \beta_{ij}$ is applied:

$$A_1 \Rightarrow^* \alpha A_i \beta \Rightarrow \alpha\alpha_{ij} A_i \beta_{ij} \beta \Rightarrow^* \alpha\alpha_{ij}\gamma_i\beta_{ij}\beta = x_1\,.$$

Since for each $x$ in $R$ there is a nonnegative integer $k$ with $l_a(x) = 4k + 1$, there exists an $i_0$ such that

$$l_a(\gamma_i) = 4i_0 + 1 - l_a(\alpha\beta) - l_a(\alpha_{ij}\beta_{ij}) .$$

For $x_2 = \alpha\alpha_{ij}\alpha_{ij}\gamma_i\beta_{ij}\beta_{ij}\beta$ we have

$$l_a(x_2) = 4i_0 + 1 + l_a(\alpha_{ij}\beta_{ij}) .$$

Since $x_2$ is in $R$ there exists a $j_0$ such that $l_a(x_2) = 4j_0 + 1$. Thence,

$$j_0 = i_0 + \frac{l_a(\alpha_{ij}\beta_{ij})}{4} .$$

This proves Assertion 1.

In the sequel, we consider words of the form

$$x = (b^l a)^{4m}\, a .$$

Let $A \to \beta_1 A \beta_2$ be a production of $G$ with $l_a(\beta_1\beta_2) > 0$. Then $l_a(\beta_1\beta_2) \geqq 4$ by Assertion 1; so either $\beta_1$ or $\beta_2$ or both can be written in the form

$$uab^{n_1}av$$

where $u$ and $v$ are in $T^*$.

If $r = \max\{l(\beta) \mid A \to \beta$ is a production of $G\}$ then $n_1 < r$. Consequently, if a production $A \to \beta_1 A \beta_2$ with $l_a(\beta_1\beta_2) > 0$ is applied in a derivation of a word $x = (b^l a)^{4m}\, a$ according to $G$, then $l < r$ holds.

Let

$$\begin{aligned} P_1 &= \{A_i \to \bar{\alpha}_{ij} A_i \bar{\beta}_{ij} \text{ in } P \mid l_a(\bar{\alpha}_{ij}\bar{\beta}_{ij}) = 0\} , \\ P_2 &= \{A_i \to \xi_{ij} A_j \eta_{ij} \text{ in } P \mid 1 \leqq i \leqq j \leqq n\} , \\ P_3 &= \{A_i \to \gamma \text{ in } P \mid \gamma \text{ in } T^*\} \end{aligned}$$

and let

$$k = \sum_{P_2} l_a(\xi_{ij}\eta_{ij}) .$$

Consider $x = (b^r a)^{4(k+r)}\, a$. By the above remark, no production $A \to \beta_1 A \beta_2$ with $l_a(\beta_1\beta_2) > 0$ can be applied in a derivation of $x$. Since $G$ is assumed to be linear and sequential, each production of $P_2$ can only be applied once in a derivation according to $G$. Hence, any word generated by productions in $P_1 \cup P_2 \cup P_3$ contains at most $k + r$ occurrences of $a$.

Clearly, by the above construction

$$x = (b^r a)^{4(k+r)}\, a$$

is not in $L(G)$; but $x$ in $R$, a contradiction.

This proves Theorem 2.

**Corollary 3.** NLev *and* Ind *are incompatible.*

Proof. Let $R$ be as in the proof of Theorem 2. Clearly, $\mathrm{NLev}(R) = 0$. Since $\mathrm{Depth}_{\mathrm{lin}}(R) > 1$, also $\mathrm{NLev}(R) > 0$ holds.

**Theorem 4.** Prod *and* Ind *are incompatible.*

Proof. Let $L = \{a^i \mid 0 \leqq i \leqq 10\}$.

$L$ can be generated by the following grammar

$$G = (\{S, A\}, \{a\}, \{S \to A^{10},\ A \to a,\ A \to \varepsilon\}, S).$$

Thence, $\mathrm{Prod}(L) \leqq 3$.

We show that $\mathrm{Prod}_{\mathrm{lin}}(L) > 3$.

Assume $\mathrm{Prod}_{\mathrm{lin}}(L) = 3$ and let $G$ be a linear grammar with $\mathrm{Prod}(G) = 3$ generating $L$. For no nonterminal symbol $A$, $A \Rightarrow^* \alpha A \beta$ holds. Therefore, the set of productions of $G$ is of one of the following forms:

| | | | |
|---|---|---|---|
| (1) | $S \to \alpha_1 A \beta_1$, | $A \to \alpha_2 B \beta_2$, | $B \to \gamma$, |
| (2.1) | $S \to \alpha_1 A \beta_1$, | $A \to \gamma_1$, | $A \to \gamma_2$, |
| (2.2) | $S \to \alpha_1 A \beta_1$, | $S \to \gamma_1$, | $A \to \gamma_2$, |
| (2.3) | $S \to \alpha_1 A \beta_1$, | $S \to \alpha_2 A \beta_2$, | $A \to \gamma_1$, |
| (3) | $S \to \gamma_1$, | $S \to \gamma_2$, | $S \to \gamma_3$ |

where all $\alpha_i$, $\beta_i$, $\gamma_i$ are in $T^*$.

But no one of these production sets can generate $L$, a contradiction.

This proves Theorem 4.

(Received February 21, 1973.)

REFERENCES

[1] Brainerd, B.: An Analog of a Theorem about Context-Free Languages. Information and Control *11* (1968), 561—567.

[2] Gruska, J.: Some Classifications of Context-Free Languages. Information and Control *14* (1969), 152—173.

[3] Gruska, J.: A Few Remarks on the Index of Context-Free Grammars and Languages. Information and Control *19* (1971), 216—223.

[4] Jones, N. D.: A Note on the Index of a Context-Free Language. Information and Control *16* (1970), 201—202.

[5] Salomaa, A.: On the Index of Context-Free Grammars and Languages. Information and Control *14* (1969), 474—477.

*Dr. A. B. Cremers, University of Southern California, Computer Science Program; Los Angeles, Cal. 90007. U.S.A.*

*K. Weiss, Universität Karlsruhe, Institut für Informatik I; Postfach 6380, 75 Karlsruhe 1. BRD.*