Theoretical Computer Science 112 (1993) 277–289
Elsevier

Descriptional complexity of context-free grammar forms

Erzsébet Csuhaj-Varjú

Computer and Automation Institute, Hungarian Academy of Sciences, Victor Hugo u. 18–22, H-1132 Budapest, Hungary

Alica Kelemenová

Institute for Informatics, Slovak Academy of Sciences, Dúbravská cesta 9, 842 35 Bratislava, Slovakia

Communicated by A. Salomaa
Received August 1989
Revised November 1991

Abstract

Csuhaj-Varjú, E. and A. Kelemenová, Descriptional complexity of context-free grammar forms, Theoretical Computer Science 112 (1993) 277–289.

Descriptional complexity aspects of grammar forms are studied. It is shown that grammatical complexity measures $HEI_{\mathscr{G}}$, $LEV_{\mathscr{G}}$, $VAR_{\mathscr{G}}$, $PROD_{\mathscr{G}}$ and $DEP_{\mathscr{G}}$ related to any appropriate infinite class $\mathscr{G}$ of grammars are unbounded on the infinite class of languages determined by strict/general interpretations of any infinite grammar form.

1. Introduction

Descriptional (grammatical) complexity measures were introduced in [1, 4, 5] in order to classify context-free languages according to the size and/or structural properties of their grammars. For the size of grammars they are expressed by such complexity measures as the number of nonterminals (VAR) and the number of productions ($PROD$). The number of grammatical levels (LEV), the maximal number of elements of grammatical levels (DEP) and the height of the digraph of grammatical levels (HEI) are the complexity measures reflecting the structure of grammars.

One of the aspects of grammatical complexity theory is the study of the functional behaviour of the complexity measures on language classes. Complexity measures are

Correspondence to: E. Csuhaj-Varjú, Computer and Automation Institute, Hungarian Academy of Sciences, Victor Hugo u. 18–22, H-1132 Budapest, Hungary.

0304-3975/93/$06.00

functions defined on context-free languages, with values being natural numbers; thus, one can ask for the set of all values of complexity of languages or simply for the boundedness/unboundedness of the complexity measure (on a given class of languages); the latter leads to the finiteness/infinity of the corresponding language hierarchy. Obviously, this behaviour depends strictly on the grammar class that is used to specify languages. For a large variety of complexity measures it was proved that, related to appropriate grammar classes, for an arbitrary natural number n, there is a context-free language with the complexity equal to n (see e.g. [4, 5, 7]). Following this line, the study of the problem of boundedness/unboundedness for remarkable subclasses of the context-free language class is of interest. In this paper we concentrate on language families defined by grammar forms which present a natural generalization of the class of all context-free languages.

Context-free grammar forms define infinite families of structurally related grammars via special finite substitutions (interpretations) of terminals and nonterminals in the production set. (For details the reader is referred to [11].) The main result of this paper establishes unboundedness of complexity measures $VAR_{\mathscr{G}}$, $PROD_{\mathscr{G}}$, $LEV_{\mathscr{G}}$, $DEP_{\mathscr{G}}$, $HEI_{\mathscr{G}}$ on the classes of languages defined by grammar forms. This property holds for a rather large variety of grammar classes $\mathscr{G}$ describing these languages. The statements are presented with full technical details. They complete the earlier results given in [3, 8, 9].

The paper is organized as follows.

Section 2 lists some basic definitions from formal language theory.

In Section 3 we construct, for every fixed natural number k, $k \geq 1$, some context-free languages that are of complexity at least k for an arbitrary subclass of context-free grammars which enables one to generate these languages. The results are of auxiliary character and serve in proving the main statement in Section 5.

In Section 4 some special interpretations are presented to obtain the interpretation grammars generating languages of the previous section. These mappings are isolation, linear isolation, copy and renaming a single symbol.

In Section 5 we show that grammatical complexity measures $VAR_{\mathscr{G}}$, $PROD_{\mathscr{G}}$, $LEV_{\mathscr{G}}$, $HEI_{\mathscr{G}}$ and $DEP_{\mathscr{G}}$ are not bound on strict and on general grammatical families of self-embedding (infinite non-self-embedding linear) grammar forms for an arbitrary class of reduced (e.g. non-self-embedding linear) grammars, that is, for every natural number k and for each of the above complexity measures, there is a language of grammatical complexity at least k in the strict/general grammatical family of grammar forms. From these statements some results from previous papers can be derived as corollaries.

2. Basic definitions

We assume the reader to be familiar with the basics of formal language theory. For the details not explained here the reader is referred to [10].

We denote context-free grammars (shortly, grammars) by $G=(N, T, P, S)$, where N, T, P are the sets of nonterminals, terminals and productions, respectively, and S is the start symbol. The context-free language (the language) generated by G is denoted by $L(G)$.

By $SF(G)$ we mean the set of sentential forms derivable in a context-free grammar G from S.

A context-free grammar G is said to be reduced iff, for all $A \in N$, there is a derivation $S \Rightarrow^* uAv \Rightarrow^+ w$ in G, with $u, v, w \in T^*$. For a context-free grammar G, we denote by G^{red} a grammar obtained from G by elimination of all nonterminals A for which no derivation $S \Rightarrow^* uAv \Rightarrow^+ w$ can be found in G, where $u, v, w \in T^*$.

A nonterminal A of G is said to be recursive iff there is a derivation $A \Rightarrow^+ uAv$ in G, where $uv \in T^+$.

A reduced context-free grammar G is said to be self-embedding if there is a nonterminal A in N such that a derivation $A \Rightarrow^* uAv$, with $u, v \in T^+$ exists; otherwise, it is said to be non-self-embedding.

For a language L, we denote by $alph(L)$ the smallest alphabet T such that $L \subseteq T^*$.

For $w \in L$, we denote by $|w|$ the length of w and by $suf_l(w)$ the suffix of length l of w.

For a class $\mathcal{G}$ of context-free grammars, we denote by $\mathcal{L}(\mathcal{G})$ the class of languages generated by elements of $\mathcal{G}$.

In what follows, we review the notions of descriptional complexity measures (size and structural complexity measures) of context-free grammars (languages) introduced in [1, 4, 5].

The size measures for a context-free grammar G are the number of its nonterminals, denoted by $VAR(G)$, and the number of its productions, denoted by $PROD(G)$.

In order to define structural complexity measures, we have to introduce relation $\rhd$ on N for $G=(N, T, P, S)$. For two nonterminals A and B of G, we write $A \rhd B$ if there is a production $A \to uBv$ in G, with $u, v \in (N \cup T)^*$. $\rhd^+$ denotes the transitive closure of $\rhd$ and $\rhd^*$ the reflexive and transitive closure of $\rhd$.

An equivalence relation $\equiv$, defined as $A \equiv B$ iff $A \rhd^* B$ and $B \rhd^* A$, determines on N equivalence classes, called grammatical levels. For two grammatical levels Q_1 and Q_2 of G, where $Q_1 \neq Q_2$, we write $Q_1 \succ Q_2$ iff there are nonterminals $A \in Q_1$ and $B \in Q_2$, with $A \rhd B$.

Structural complexity measures for a context-free grammar G are defined as follows:

$LEV(G)$ denotes the number of grammatical levels of G^{red},

$DEP(G) = \max\{\text{card}(Q)\text{: } Q \text{ is a grammatical level of } G^{\text{red}}\}$,

$HEI(G) = \max\{HEI(Q)\text{: } Q \text{ is a grammatical level of } G^{\text{red}}\}$,

where $HEI(Q) = 1$ iff $S \in Q$ and $HEI(Q_i) = 1 + \max\{HEI(Q_t)\text{: } Q_t \succ Q_i\}$.

In what follows, we use for complexity measures VAR, LEV, HEI, DEP and $PROD$ the common denotation K.

The descriptional complexity measure of a language L with respect to a class of grammars $\mathcal{G}$ is defined as follows:

$$K_{\mathcal{G}}(L) = \begin{cases} \min\{K(G)\text{: } G \in \mathcal{G},\ L(G) = L\} & \text{if } L = L(G) \text{ for some } G \in \mathcal{G}, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Note that, by definition, for an arbitrary class $\mathscr{G}$ of grammars, with $L \in \mathscr{L}(\mathscr{G})$, $HEI_{\mathscr{G}}(L) \leqslant LEV_{\mathscr{G}}(L) \leqslant VAR_{\mathscr{G}}(L) \leqslant PROD_{\mathscr{G}}(L)$ holds.

In what follows, we review the notions of a grammar form and its strict and general interpretations. For further details, see [11].

Let $G_i = (N_i, T_i, P_i, S_i)$, where $i = 1, 2$ be context-free grammars. We say G_2 is obtained from grammar form G_1 by a general interpretation (shortly, a g-interpretation) μ, denoted by $G_2 \rhd_g G_1(\mu)$, if μ is a finite substitution on $(N_1 \cup T_1)^*$ and conditions (i)–(iv) hold:

(i) $\mu(A) \subseteq N_2$ for all $A \in N_1$ and $\mu(A) \cap \mu(B) = \emptyset$ for $A, B \in N_1$, with $A \neq B$;
(ii) $\mu(a) \subseteq T_2^*$ for all $a \in T_1$;
(iii) $P_2 \subseteq \mu(P_1) = \{u \to v\colon u \in \mu(\alpha), v \in \mu(\beta), \alpha \to \beta \in P_1\}$;
(iv) $S_2 \in \mu(S_1)$.

G_2 is said to be obtained from G_1 by a strict interpretation (shortly, an s-interpretation) μ, denoted by $G_2 \rhd_s G_1(\mu)$, if condition (ii) is modified as follows: $\mu(a) \subseteq T_2$ for every $a \in T_1$ and $\mu(a) \cap \mu(b) = \emptyset$ for all $a, b \in T_1$, where $a \neq b$.

The collection of grammars obtained by x-interpretations from a grammar G, where $x \in \{g, s\}$, is denoted by $\mathscr{G}_x(G)$.

The class of languages $\mathscr{L}_x(G) = \{L\colon L = L(G), G \in \mathscr{G}_x(G)\}$ is called the x-grammatical family of G.

The grammar G itself is often referred to as a grammar form.

A grammar form G is said to be infinite if $L(G)$ is infinite; otherwise, it is said to be finite.

3. On descriptional complexity of context-free languages

In this section we determine the complexity measures of some special context-free languages. The results obtained here will be used in Section 5 to prove the main results of the paper.

Definition 3.1. Let $G = (N, T, P, S)$ be a context-free grammar with a derivation tree t of $w = \alpha v \beta$ in G, with $\alpha, \beta \in T^*$, $v \in T^+$. We say t_v is a minimal subtree of t completely deriving v if t_v is a derivation tree of xvy, where $\alpha = x_0 x$, $\beta = y y_0$ and t_v has no subtree t'_v such that t'_v is a derivation tree of $x'vy'$, where $x = x_1 x'$, $y = y' y_1$ and $x_1 y_1 \in T^*$.

We shall use the pumping property of context-free grammars in the form specified by the following lemma.

Lemma 3.2. *Let $G = (N, T, P, S)$ be a context-free grammar. Let $w = xvy$ be in $L(G)$, where $|v| > d^m$ for $d = \max\{|\alpha|\colon A \to \alpha \in P\}$ and $m = \text{card}(N)$.*

Let t be a derivation tree of w with no subderivation $A \Rightarrow^+ A$ for any A in N and let t_v be a minimal subtree of derivation tree t completely deriving v. Then there is an $A_v \in N$ which occurs twice on the same branch of t_v. Moreover, the subderivation $A_v \Rightarrow^+ v_1 A_v v_2$

is determined in t_v by two consecutive occurrences of A_v on this branch, where v_1 and v_2 are subwords of w and $v_1 v_2 \neq \varepsilon$.

Proof. Suppose by contradiction that no nonterminal occurs twice on the same branch of t_v. Then the length of any branch of t_v is at most m, and $|v| \leqslant d^m$. This contradicts the assumption of the lemma.

Let A_v occur twice on the same branch of t_v and let $A_v \Rightarrow^+ v_1 A_v v_2$, with $v_1, v_2 \in T^*$, be a subderivation determined in t_v by two consecutive occurrences of A_v in this branch. Since $A_v \Rightarrow^+ A_v$ is not a subderivation in t, we have immediately $v_1 v_2 \neq \varepsilon$. □

The following theorem is about $LEV_{\mathscr{G}}$, $VAR_{\mathscr{G}}$ and $PROD_{\mathscr{G}}$ complexity of context-free languages being finite union of languages over pairwise disjoint alphabets.

Theorem 3.3. *Let $L = \bigcup_{i=1}^{k} L_i$, where L_i, $1 \leqslant i \leqslant k$, are infinite context-free languages over pairwise disjoint alphabets. Let $\mathscr{G}$ be a class of grammars such that L, $L_i \in \mathscr{L}(\mathscr{G})$ hold. Then $K_{\mathscr{G}}(L) \geqslant k$ for $K \in \{LEV, VAR, PROD\}$.*

Proof. Let $G = (N, T, P, S)$ be in $\mathscr{G}$ and $L(G) = L$. Let, for a given i, $1 \leqslant i \leqslant k$, $w_i = x_i v_i y_i \in L_i$ have a derivation tree t_i and $|v_i| > d^m$, where t_i, d and m are as in Lemma 3.2. Let t_i' be a minimal subtree of t_i completely deriving v_i. Then, by Lemma 3.2, there is a nonterminal A_i and a derivation $A_i \Rightarrow^+ u_i A_i v_i$ determined by A_i in t_i, with $u_i v_i \in (alph(L_i))^+$. Since $alph(L_i) \cap alph(L_j) = \emptyset$ for $i \neq j$, $1 \leqslant i, j \leqslant k$, neither $A_i \rhd^+ A_j$ nor $A_j \rhd^+ A_i$ holds for $i \neq j$, $1 \leqslant i, j \leqslant k$. This implies the statement for $K = LEV$ and, thus, also for $K = VAR$ or $K = PROD$.

Notation 3.4. Let u_i, v_i, $1 \leqslant i \leqslant k$, be nonempty words and $alph(u_i v_i) \cap alph(u_j v_j) = \emptyset$ for $i \neq j$.

Let $L_{k,0} = u_1^+ \dots u_k^+$.

For $x \in L_{k,0}$, let $mi(x)$ be defined as follows: for $x = u_1^{m_1} \dots u_k^{m_k} \in L_{k,0}$, let $mi(x) = v_k^{m_k} \dots v_1^{m_1}$, where $m_1, \dots, m_k \geqslant 1$. For $x = x_1 x_2$, where $x_1, x_2 \in L_{k,0}^+$, let $mi(x_1 x_2) = mi(x_2) mi(x_1)$.

Let u, v, w be arbitrary words with disjoint alphabets, also disjoint with those of u_i, v_i, $1 \leqslant i \leqslant k$. (In the case where u, v or w are nonempty.)

Let

$$M_k^{(1)} = \{uxw : x \in L_{k,0}\},$$

$$M_k^{(+)} = \{uxw : x \in L_{k,0}^+\},$$

$$L_k^{(1)} = \{uxwmi(x)v : x \in L_{k,0}\}$$

and

$$L_k^{(+)} = \{uxwmi(x)v : x \in L_{k,0}^+\}.$$

The structure of any context-free grammar generating any of the above language is determined in the following sense: all words with sufficiently many (say s) repetitions of the subword $u_i(v_i)$ are generated by pumping a subword of $u_i^s(v_i^s)$, with the length equal to some multiple of the length of u_i (v_i).

Lemma 3.5. *Let L_k be any of the languages $L_k^{(1)}$ and $L_k^{(+)}$. Let $L_k = L(G_k)$ for a context-free grammar G_k. Then, for every i, $1 \leqslant i \leqslant k$, there exists a nonterminal A_i in G_k and a number $n_i \geqslant 1$ such that $A_i \Rightarrow^+ \bar{u}_i^{n_i} A_i \bar{v}_i^{n_i}$ holds, where $\bar{u}_i = yx$ for some x, y, with $xy = u_i$, and $\bar{v}_i = tz$ for some z, t, with $zt = v_i$. Moreover, for $L_k = L_2^{(1)}$ and for L_k with $k \geqslant 3$, $A_i \neq A_j$ for $i \neq j$.*

Proof. Let d and m be as in Lemma 3.2. Consider $w_s = uu_1^s \ldots u_k^s w v_k^s \ldots v_1^s v$, where $s > d^m$. Obviously, $w_s \in L_k^{(1)} \subseteq L_k^{(+)}$. Let t be a derivation tree of w_s in G_k fulfilling the conditions of Lemma 3.2. According to Lemma 3.2, for every i, $1 \leqslant i \leqslant k$, and for every minimal subtree t_i of t completely deriving u_i^s, it holds that there is a nonterminal A_i and a subderivation $A_i \Rightarrow^+ u_i' A_i z_i'$, determined in t_i by A_i, where u_i' is a nonempty subword of u_i^s and z_i' is a terminal word. (The case where z_i' is a nonempty subword of u_i^s leads to a contradiction with the structure of L_k.) Let $u_i' = yu_i^{r_i}x$ for some x, y, where $0 \leqslant |x| < |u_i|$ and $0 < |y| \leqslant |u_i|$. We prove that $xy = u_i$, which leads to $A_i \Rightarrow^+ \bar{u}_i^{r_i+1} A_i z_i'$, where $\bar{u}_i = yx$.

Consider the derivation

$$S \Rightarrow^+ \bar{u} A_i \bar{v} \Rightarrow^+ \bar{u}(yu_i^{r_i}x)^j A_i (z_i')^j \bar{v} \Rightarrow^+ \bar{u}(yu_i^{r_i}x)^j w_i (z_i')^j \bar{v}$$

in G_k, where $\bar{u}$, $\bar{v}$, w_i are terminal words and w_i is derived with a minimal number of steps in G_k resulting from A_i a terminal word. By the structure of words in L_k, we have, for $j = 1$ and $j = 2$,

$$\bar{u} y u_i^{r_i} x w_i z_i' \bar{v} = \bar{x} u_i^{l_1} \bar{w} v_i^{l_1} \bar{y} \quad \text{and} \quad \bar{u} y u_i^{r_i} x y u_1^{r_i} x w_i z_i' z_i' \bar{v} = \bar{x} u_i^{l_2} \bar{w} v_i^{l_2} \bar{y}$$

for maximal numbers $l_2 > l_1 \geqslant 1$. This gives $u_i^{r_i} xy u_i^{r_i} = u_i^{s_i}$, i.e. $xy = u_i$ and $l_2 - l_1 = r_i + 1 = n_i$. Then $suf_l(w_i) z_i' \bar{v} = \bar{w} v_i^{l_1} \bar{y}$ and $suf_l(w_i) z_i' z_i' \bar{v} = \bar{w} v_i^{l_2} \bar{y}$ for some $l \leqslant |w_i|$. This implies $z_i' = t v_i^{r_i} z$ for some t, z, where $0 < |t| \leqslant |v_i|$, $0 \leqslant |z| < |v_i|$ and $zt = v_i$.

Let $A_i = A_j$ for some i, j, $1 \leqslant i, j \leqslant k$. Then $\bar{u}_i^{n_i} \bar{u}_j^{n_j} \bar{u}_i^{n_i}$, $n_i \geqslant 1$, $n_j \geqslant 1$, is a subword of some word in L_k. This implies $i = j$ for $L_k = L_k^{(1)}$, where $k \geqslant 2$, and for $L_k = L_k^{(+)}$, where $k \geqslant 3$.

In the case of regular languages $M_k^{(1)}$ and $M_k^{(+)}$ an analogous theorem holds for the non-self-embedding linear class of grammars. Using similar methods and arguments as in the proof of Lemma 3.5, we can prove the following lemma.

Lemma 3.6. *Let L_k be any of the languages $M_k^{(1)}$ and $M_k^{(+)}$. Let G_k be a non-self-embedding linear grammar generating L_k. Then, for every i, $1 \leqslant i \leqslant k$, there exists a nonterminal A_i in G_k and $n_i \geqslant 1$ such that $A_i \Rightarrow^+ \bar{u}_i^{n_i} A_i$ or $A_i \Rightarrow^+ A_i \bar{u}_i^{n_i}$ holds, where $\bar{u}_i = yx$ for some x, y, with $xy = u_i$. Moreover, $A_i \neq A_j$ for $i \neq j$, $1 \leqslant i, j \leqslant k$.*

Theorem 3.7. (i) *Let $\mathscr{G}$ be a class of context-free grammars such that, for every $k \geqslant 1$, $L_k^{(1)} \in \mathscr{L}(\mathscr{G})$. Then $HEI_{\mathscr{G}}(L_k^{(1)}) \geqslant k$.*

(ii) *Let $\mathscr{G}$ be a class of non-self-embedding linear grammars such that, for every $k \geqslant 1$, $M_k^{(1)} \in \mathscr{L}(\mathscr{G})$. Then $HEI_{\mathscr{G}}(M_k^{(1)}) \geqslant k$.*

Proof. Consider an arbitrary grammar G_k in $\mathscr{G}$ for which $L(G_k) = L_k^{(1)}$ ($L(G_k) = M_k^{(1)}$) holds. Let $A_1, \ldots, A_k$ be nonterminals of G_k determined in Lemma 3.5 (in Lemma 3.6), which are used in the derivation of $w_s = uu_1^s \ldots u_k^s wv_k^s \ldots v_1^s v$ ($w_s = uu_1^s \ldots u_k^s w$), where $s > d^m$, d and m being the numbers given in Lemma 3.2. Since no v_i can preceed u_j for $1 \leqslant i, j \leqslant k$ [since G_k is a linear grammar for (ii)], no sentential form xA_iyA_jz can be derived in G_k, where $x, y, z \in (N \cup T)^*$. This implies that either $A_i \rhd^* A_{i+1}$ or $A_{i+1} \rhd^* A_i$ for $1 \leqslant i \leqslant k-1$. Since u_{i+1} never preceeds u_i, in the case of $L_k^{(1)}$, $A_{i+1} \rhd^+ A_i$ does not hold and then $A_1 \rhd^+ A_2 \rhd^+ \cdots \rhd^+ A_k$. In the case of $M_k^{(1)}$, each A_i is either right-linear or left-linear but not both; so, a permutation $(p_1, \ldots, p_k)$ of $(1, \ldots, k)$ can be determined such that $A_{p_1} \rhd^+ A_{p_2} \rhd^+ \cdots \rhd^+ A_{p_k}$. Hence, $HEI(G_k) \geqslant k$.

Theorem 3.8. (i) *Let $\mathscr{G}$ be a class of context-free grammars such that, for every $k \geqslant 3$, $L_k^{(+)} \in \mathscr{L}(\mathscr{G})$. Then $DEP_{\mathscr{G}}(L_k^{(+)}) \geqslant k$.*

(ii) *Let $\mathscr{G}$ be a class of non-self-embedding linear grammars such that, for every $k \geqslant 1$, $M_k^{(+)} \in \mathscr{L}(\mathscr{G})$. Then $DEP_{\mathscr{G}}(M_k^{(+)}) \geqslant k$.*

Proof. Let G_k be an arbitrary element of $\mathscr{G}$ for which $L(G_k) = L_k^{(+)}$ ($L(G_k) = M_k^{(+)}$) holds. Let $A_1^{(t)}, \ldots, A_k^{(t)}$, for $1 \leqslant t \leqslant m+1$, be nonterminals of G_k, determined in Lemma 3.5 (in Lemma 3.6), which are used in the derivation of the word $\tilde{w}_s = u\tilde{w}^{m+1} wmi(\tilde{w}^{m+1})v$ ($\tilde{w}_s = u\tilde{w}^{m+1}w$), where $\tilde{w} = u_1^s \ldots u_k^s$ for some $s > d^m$, where d and m are defined in Lemma 3.2 (i.e. $A_i^{(t)}$ is a nonterminal producing $\bar{u}_i$'s in the tth position of u_i^s in $\tilde{w}_s$.) Note that $A_i^{(s_1)} \neq A_j^{(s_2)}$ for $i \neq j$ and for arbitrary s_1, s_2, $1 \leqslant s_1, s_2 \leqslant m+1$. As no v_j can preceed u_i for $0 \leqslant i, j \leqslant k$ in $L_k^{(+)}$, $A_i^{(s)} \rhd^+ A_{i+1}^{(s)}$ for $1 \leqslant i \leqslant k-1$ and $A_k^{(s)} \rhd^+ A_1^{(s+1)}$ for $1 \leqslant s \leqslant m$.

Since $t = 1, 2, \ldots, m+1$, there exist two different positions s_1 and s_2 such that $A_i^{(s_1)} = A_i^{(s_2)}$ for a fixed i. For $i = 1$ let us choose s_1, s_2, where $s_2 > s_1$, and $s_2 - s_1$ is minimal. Then $A_1^{(s_1)} \rhd^+ A_2^{(s_1)} \rhd^+ \cdots \rhd^+ A_k^{(s_1)} \rhd^+ A_1^{(s_1+1)} \rhd^* A_1^{(s_2)} = A_1^{(s_1)}$. (In the case of $M_k^{(+)}$, permutations $(p_{t1}, \ldots, p_{tk})$ of $(1, \ldots, k)$ for $1 \leqslant t \leqslant m+1$ exist such that, for $B_r^{(t)}$ equal to $A_{p_{tr}}^{(t)}$, $B_1^{s_1} \rhd^+ B_2^{s_1} \rhd^+ \cdots \rhd^+ B_k^{s_1} \rhd^+ B_1^{(s_1+1)} \rhd^* B_1^{s_2} = B_1^{s_1}$ holds.) Thus, $DEP(G_k) \geqslant k$.

4. Basic interpretations – auxiliary results

In this section we specify some strict interpretations used in the sequel. The basic idea behind them is a suitable renaming of nonterminals and arising rules that are not of interest.

By isolation we mean an interpretation which, roughly speaking, isolates a given derivation of a sentential form.

Lemma 4.1. *Let* $G=(N, T, P, S)$ *be a context-free grammar and let, for* A *in* N, $D: A=u_0 \Rightarrow u_1 \Rightarrow u_2 \Rightarrow \cdots \Rightarrow u_n$, $n \geqslant 1$, *be a derivation in* G, *where* $u_j \in (N \cup T)^*$, $1 \leqslant j \leqslant n$. *Then there is a strict interpretation* μ_D, *called an isolation of* D, and a grammar $G_D=(N_D, T_D, P_D, A)$ *such that* $G_D \rhd_s G(\mu_D)$ *and*

(i) $SF(G_D) \cap (N \cup T)^* = SF(G_{u_n})$, *where* $G_{u_n}=(N, T, \{A \rightarrow u_n\}, A)$,

(ii) $A=v_0 \Rightarrow v_1 \Rightarrow \cdots \Rightarrow v_{n-1} \Rightarrow v_n = u_n$, *where* $v_i \in \mu_D(u_i)$ *for* $i=1, 2, \ldots, n-1$, *is the only derivation in* G_D *of length* n *starting with* A, *and*

(iii) P_D *consists exactly of productions used in* $v_i \Rightarrow v_{i+1}$ *for* $i=0, 1, \ldots, n-1$.

Proof. Let us define μ_D as follows:

for $a \in T$,

$$\mu_D(a)=\{a\},$$

for $B \in N$,

$$\mu_D(B)=\{B^{[i,j]}\text{: for every } [i,j] \text{ such that } B \text{ occurs as the } j\text{th letter in } u_i,\ 1 \leqslant i \leqslant n-1\} \cup p(B),$$

where

$$p(B)=\begin{cases}\{B\} & \text{for } B=A \text{ or for } B \text{ being a letter of } u_n,\\ \emptyset & \text{otherwise.}\end{cases}$$

Let, for j, $1 \leqslant j \leqslant n-1$, $u_j = X_{j,1} \ldots X_{j,l_j}$, where $X_{j,k} \in (N \cup T)$, $1 \leqslant k \leqslant l_j$. We associate a word v_j with u_j, where $v_j = X'_{j,1} \ldots X'_{j,l_j}$, where

$$X'_{j,k}=\begin{cases}X_{j,k} & \text{if } X_{j,k} \in T,\\ X_{j,k}^{[j,k]} & \text{if } X_{j,k} \in N.\end{cases}$$

Let us consider the derivation $D': A \Rightarrow v_1 \Rightarrow v_2 \Rightarrow \cdots \Rightarrow v_{n-1} \Rightarrow u_n$ and let P_D be the set of productions used in this derivation. Then, obviously, $P_D \subseteq \mu_D(P)$ and, for the grammar G_D given implicitly by P_D, we have $G_D \rhd_s G(\mu_D)$ and $SF(G_D) \cap (N \cup T)^* = SF(G_{u_n})$.

Remark. $SF(G_{u_n})$ is infinite if A is a letter of u_n, where $u_n \neq A$; otherwise, $SF(G_{u_n}) = \{A, u_n\}$.

Linear isolations (constructed in the next lemma) cause fixed derivations to be isolated, and terminals derived left and right from a fixed branch of the derivation tree to be distinguished.

Lemma 4.2. *Let* $G=(N, T, P, S)$ *be a context-free grammar and let, for* A *in* N, $D: A=u_0 \Rightarrow u_1 \Rightarrow u_2 \Rightarrow \cdots \Rightarrow uAv$, $n \geqslant 1$, *be a derivation in* G, *where* $u, v \in T^+$. *Let* T' *be a primed version of* T. *Then there is a strict interpretation* μ'_D, *called a linear isolation of* D, *and a grammar* $G'_D=(N'_D, T \cup T', P'_D, A)$ *such that* $G'_D \rhd_s G(\mu'_D)$ *and*

(i) $SF(G'_D) \cap (N \cup T \cup T')^* = SF(G'_{u_n})$, *where* $G'_{u_n}=(N, T \cup T', \{A \rightarrow uAv'\}, S)$,

(ii) $A = v'_0 \Rightarrow v'_1 \Rightarrow \cdots \Rightarrow v'_{n-1} \Rightarrow v'_n = uAv'$, *where* $v'_i \in \mu'_D(u_i)$ *for* $i = 1, 2, \ldots, n-1$, *is the only derivation in* G'_D *of length n starting with A, and*

(iii) P'_D *consists exactly of productions used in* $v'_i \Rightarrow v'_{i+1}$ *for* $i = 0, 1, \ldots, n-1$.

Proof. Let $u_i = \alpha_i X_i \beta_i$, $1 \leqslant i \leqslant n-1$, where $\alpha_i, \beta_i \in (N \cup T)^*$ and X_i are nonterminals lying on the branch of the derivation tree of D beginning and ending with A. We define μ'_D as follows:

$$\mu'_D(a) = \{a, a'\} \text{ for } a \in T \quad \text{and} \quad \mu'_D(B) = \mu_D(B) \text{ for } B \in N,$$

where μ_D is the isolation defined in Lemma 4.1. Let, for j, $1 \leqslant j \leqslant n$, $u_j = X_{j,1} \ldots X_{j,n_j} \ldots X_{j,l_j}$, where $X_{j,k} \in (N \cup T)$, $1 \leqslant k \leqslant l_j$, $n_j \leqslant l_j$ and $X_{j,n_j} = X_j$.

We associate with u_j a word v'_j, where $v'_j = \bar{X}_{j,1} \ldots \bar{X}_{j,l_j}$, where

$$\bar{X}_{j,k} = \begin{cases} X_{j,k} & \text{for } X_{j,k} \in T,\ k < n_j, \\ X'_{j,k} & \text{for } X_{j,k} \in T,\ k > n_j, \\ X^{[j,k]}_{j,k} & \text{for } X_{j,k} \in N. \end{cases}$$

Let us consider the derivation $\bar{D}$: $A \Rightarrow v'_1 \Rightarrow v'_2 \Rightarrow \cdots \Rightarrow v'_{n-1} \Rightarrow uAv'$ and let P'_D be the set of productions used in this derivation. Then, obviously, $P'_D \subseteq \mu'_D(P)$ and, for the grammar G'_D given implicitly by P'_D, we have $G'_D \rhd_s G(\mu'_D)$ as well as $SF(G'_D) \cap (N \cup T \cup T')^* = SF(G'_{u_n})$.

Next we fix the notions of jth copy and renaming a single symbol by special isomorphic interpretations and define the corresponding grammars isomorphic to the core grammar.

Definition 4.3. Let $G = (N, T, P, S)$ be a grammar and j be a natural number. By μ_{cj} we denote an interpretation, called a jth copy, defined by $\mu_{cj}(X) = X^{(j)}$, for $X \in (N \cup T)$.

The jth copy $G^{(j)}$ of G is the grammar $G^{(j)} \lhd_s G(\mu_{cj})$, with $P^{(j)} = \mu_{cj}(P)$.

Interpretations can change some fixed occurrences of some symbol in the set of productions.

Definition 4.4. Let $G = (N, T, P, S)$ be a grammar and let $X \in N$, $Y \notin (N \cup T)$. By $\mu_{X \to Y}$, called renaming X (by Y) we denote an interpretation with $\mu_{X \to Y}(X) = \{X, Y\}$ and $\mu_{X \to Y}(Z) = \{Z\}$ for $Z \neq X, Z \in N \cup T$. By $G_{X \to Y}$ we denote the grammar $G_{X \to Y} = (N \cup \{Y\}, T, P_{X \to Y}, S)$, where $P_{X \to Y} = \{S \to \alpha_Y : S \to \alpha \in P\} \cup \{Y \to \alpha_Y : X \to \alpha \in P\} \cup \{A \to \alpha_Y : A \to \alpha \in P, A \neq X\}$, where α_Y denotes the word obtained from α by replacing all occurrences of X by Y.

Informally, $G_{X \to Y}$ is such an interpretation of G in which $X \in N$ is replaced by $Y \notin (N \cup T)$ in any position of X except where X is the start symbol of G and all other letters remain unchanged.

5. Complexity of grammar forms

In this section we show that grammatical complexity measures $VAR_{\mathscr{G}}$, $PROD_{\mathscr{G}}$, $LEV_{\mathscr{G}}$, $HEI_{\mathscr{G}}$ and $DEP_{\mathscr{G}}$ are unbounded on strict and on general grammatical families of self-embedding (infinite non-self-embedding linear) grammar forms.

Theorem 5.1. *Let G be a self-embedding context-free grammar form. Let $\mathscr{G}$ be a class of context-free grammars such that $\mathscr{L}_x(\mathscr{G})\subseteq\mathscr{L}(\mathscr{G})$, where $x\in\{g, s\}$. Let $K\in\{VAR, PROD, LEV, HEI, DEP\}$. Then $K_{\mathscr{G}}$ is unbounded on $\mathscr{L}_x(G)$.*

Proof. Let $x=g$. If G is a self-embedding grammar, then $\mathscr{L}_g(G)$ contains all linear languages (see [11, p. 43]). By [5] and Theorem 3.7, for each K and for every natural number k, there is a linear language L_k such that $K_{CF}(L_k)\geqslant k$. This results in $K_{\mathscr{G}}$ being unbounded on $\mathscr{L}_g(G)$.

Let $x=s$. First we prove the result for $K=HEI$. This gives the proof for VAR, $PROD$, LEV, too. Since G is self-embedding, there is a nonterminal A in G with derivations

$$D_S: S\Rightarrow^+ xAy,$$

$$D_I: A\Rightarrow^+ uAv,$$

$$D_F: A\Rightarrow^+ w,$$

with $x, y, w\in T^*$ and $u, v\in T^+$. Denote by P_S, P_I, P_F the sets of productions of P used in derivations D_S, D_I, D_F, respectively. According to Theorem 3.7, it is sufficient to give, for any $k\geqslant 1$, a grammar G_k such that $G_k\lhd_s G$ and $L(G_k)=L_k^{(1)}=\{xu_1^{m_1}\ldots u_k^{m_k}w_{k+1}v_k^{m_k}\ldots v_1^{m_1}y: m_i\geqslant 1, 1\leqslant i\leqslant k\}$. Let P'_S, P'_I, P'_F be the sets of productions obtained from P_S, P_I, P_F by isolations μ_{D_S}, μ'_{D_I}, μ_{D_F}, defined in Lemmas 4.1 and 4.2, respectively. We use abbreviation μ_{r0} for $\mu_{A\to A_1}$ and μ_{rj} for $\mu_{A_j\to A_{j+1}}$, where $1\leqslant j\leqslant k$.

Let $P_k=\mu_{r0}(P'_S)\cup\bigcup_{i=1}^{k}\mu_{ci}(P'_I)\cup\bigcup_{j=1}^{k}\mu_{rj}(\mu_{cj}(P'_I))\cup\mu_{ck+1}(P'_F)$ and let G_k be the grammar given implicitly by the productions of P_k. We shall prove that $L(G_k)=L_k^{(1)}$. Let $w=xu_1^{m_1}\ldots u_k^{m_k}w_{k+1}v_k^{m_k}\ldots v_1^{m_1}y$. Then w can be derived in G_k by using the following partial derivations:

$S\Rightarrow^+ xA_1y$, which uses the productions of $\mu_{r0}(P'_S)$,

$A_i\Rightarrow^+ u_iA_iv_i$, which uses the productions of $\mu_{ci}(P'_I)$ for $1\leqslant i\leqslant k$,

$A_j\Rightarrow^+ u_jA_{j+1}v_j$, which uses the productions of $\mu_{rj}(\mu_{cj}(P'_I))$ for $1\leqslant j\leqslant k$,

$A_{k+1}\Rightarrow^+ w_{k+1}$, which uses productions of $\mu_{ck+1}(P'_F)$.

Thus, $L_k^{(1)}\subseteq L(G_k)$.

We show that the opposite inclusion holds. Let $D: S\Rightarrow w_1\Rightarrow w_2\Rightarrow\cdots\Rightarrow w_r=w\in T^*$ be a derivation in G_k. Following P_k and Lemmas 4.1 and 4.2 any sentential form of G_k contains at most one recursive letter. The recursive nonterminal A_i, $2\leqslant i\leqslant k$, does not appear before A_{i-1} is rewritten. Moreover, every terminating derivation contains each A_i, $1\leqslant i\leqslant k$, at least once. Without loss of generality, we may assume that all nonrecursive nonterminals in D are rewritten before a recursive symbol is rewritten.

Then indices $2 \leqslant i_1 \leqslant i_2 \leqslant \cdots \leqslant i_{k+1} \leqslant r-1$ can be found such that A_t occurs in the sentential form w_{i_t} and it does not occur in any w_k for $k \leqslant i_t$. This gives

$$w_{i_1} = xA_1 y,$$

$$w_{i_2} = xu_1^{m_1} A_2 v_1^{m_1} y \quad \text{for some } m_1 \geqslant 1,$$

$$\vdots$$

$$w_{i_{k+1}} = xu_1^{m_1} \ldots u_k^{m_k} A_{k+1} v_k^{m_k} \ldots v_1^{m_1} y$$

for some $m_1, \ldots, m_k \geqslant 1$ and nonrecursive nonterminal A_{k+1}. Thus, $w_r = xu_1^{m_1} \ldots u_k^{m_k} w_{k+1} v_k^{m_k} \ldots v_1^{m_1} y$ and $L(G_k) \subseteq L_k^{(1)}$.

Let $K = DEP$. We show that there is a grammar $\bar{G}_k$, where $\bar{G}_k \lhd_s G$, such that $L(\bar{G}_k) = L_k^{(+)}$. Let P_k have the same meaning as above. Let $\bar{P}_k = P_k \cup \mu_{rk}(\mu_{ck}(P'_l))$, where μ_{rk} abbreviates $\mu_{A_k \to A_1}$. Let $\bar{G}_k$ be the grammar given implicitly by the elements of $\bar{P}_k$. It can be shown that $L(\bar{G}_k) = L_k^{(+)} = \{xzw_{k+1} mi(z) y : z \in L_{k,0}^{+}\}$. According to Theorem 3.8, $DEP_{\mathscr{G}}(L(\bar{G}_k)) \geqslant k$. □

Example. We illustrate the constructions of G_k and $\bar{G}_k$ from the previous proof.

Let G contain the productions

$$S \to aA,$$

$$A \to aA \mid Aa \mid a.$$

We give G_k and $\bar{G}_k$ corresponding to the derivations

$$D_S : S \Rightarrow aA,$$

$$D_I : A \Rightarrow aA \Rightarrow aAa,$$

$$D_F : A \Rightarrow a.$$

G_k is given by the productions

$$S \to aA_1,$$

$$A_i \to a_i A_i^{[1,2]},$$

$$A_i^{[1,2]} \to A_i a'_i \quad \text{for } i = 1, 2, \ldots, k,$$

$$A_i^{[1,2]} \to A_{i+1} a'_i \quad \text{for } i = 1, 2, \ldots, k,$$

$$A_{k+1} \to a_{k+1}.$$

$\bar{G}_k$ contains the same productions as G_k and, moreover, the production $A_k^{[1,2]} \to A_1 a'_k$.

To continue our study, we discuss the case where G is a non-self-embedding infinite grammar form. In this case $\mathscr{L}_x(G) \subseteq \mathscr{L}(\mathscr{REG})$. Now we have to distinguish between complexity measures in $\{HEI, DEP\}$ and in $\{VAR, PROD, LEV\}$ since, for any regular language R, $HEI_{CF}(R) \leqslant 2$ and $DEP_{CF}(R) = 1$, while VAR_{CF}, $PROD_{CF}$, and LEV_{CF} form infinite hierarchies on the class of regular languages.

Theorem 5.2. *Let G be a non-self-embedding infinite context-free grammar form and let $\mathcal{G}$ be a class of context-free grammars such that $\mathcal{L}_x(G) \subseteq \mathcal{L}(\mathcal{G})$, where $x \in \{g, s\}$. Then $LEV_{\mathcal{G}}$, $VAR_{\mathcal{G}}$ and $PROD_{\mathcal{G}}$ are unbounded on $\mathcal{L}_x(G)$.*

Proof. Let us discuss first $K = LEV$. Let $x = g$. $\mathcal{L}_g(G) = \mathcal{L}(R\mathcal{E}\mathcal{G})$; by [5], for every $k \geqslant 1$, there is a regular language R_k such that $LEV_{CF}(R_k) = k$ holds. Let $x = s$.

Without loss of generality, we may assume that G has a recursive nonterminal A with derivations

$$D_S : S \Rightarrow^* xAy,$$

$$D_I : A \Rightarrow^+ uA \text{ (or } D_I : A \Rightarrow^+ Au, \text{ but not both)}$$

$$D_F : A \Rightarrow^+ w,$$

with $x, y, w \in T^*$ and $u \in T^+$. Let P_S, P_I, P_F be the sets of productions used in derivations D_S, D_I, D_F, respectively. To prove the theorem, we construct, for any $k \geqslant 1$, a grammar G_k such that $L(G_k) = L$ satisfies the conditions of Theorem 3.1. Let μ_{D_S}, μ_{D_I}, μ_{D_F} be the isolations defined in Lemma 4.1 and denote by P'_S, P'_I, P'_F sets of productions obtained by them from P_S, P_I, P_F, respectively. Let $P_k = \bigcup_{i=1}^{k} (\{S \to \mu_{ci}(\alpha) : S \to \alpha \in P'_S\} \cup \mu_{ci}(P'_S \cup P'_I \cup P'_F))$.

Let G_k be the grammar given implicitly by productions of P_k. Then P_k determines grammar G_k with the following derivations:

$$S \Rightarrow^* x_i A_i y_i, \quad A_i \Rightarrow^+ u_i A_i \text{ (or } A_i \Rightarrow^+ A_i u_i), \quad A_i \Rightarrow^+ w_i, \quad 1 \leqslant i \leqslant k,$$

where $alph(x_i y_i u_i w_i)$ are pairwise disjoint for different i. Since $L(G_k) = \bigcup_{i=1}^{k} L_i$, where $L_i \subseteq alph(x_i y_i u_i w_i)^+$, $LEV_{\mathcal{G}} L(G_k) \geqslant k$, according to Theorem 3.1. Since $PROD_{\mathcal{G}} L(G_k) \geqslant VAR_{\mathcal{G}} L(G_k) \geqslant LEV_{\mathcal{G}} L(G_k)$, the proof is completed. □

If we restrict $\mathcal{G}$ to be a class of non-self-embedding linear grammars then for G a non-self-embedding linear infinite grammar form we obtain infinite hierarchy for $HEI_{\mathcal{G}}$ and $DEP_{\mathcal{G}}$ on $\mathcal{L}_x(G)$, too.

Theorem 5.3. *Let G be a non-self-embedding infinite linear grammar form. Let $\mathcal{G}$ be a class of non-self-embedding linear grammars such that $\mathcal{L}_x(G) \subseteq \mathcal{L}(\mathcal{G})$, where $x \in \{g, s\}$. Let $K \in \{HEI, DEP\}$. Then $K_{\mathcal{G}}$ is unbounded on $\mathcal{L}_x(G)$.*

Proof (*sketch*). The theorem can be proved by constructing languages $M_k^{(1)}$, $M_k^{(+)}$ using similar methods and arguments as in Theorem 5.1.

References

[1] W. Brauer, On grammatical complexity of context-free languages (extended abstract), in: *Proc. of MFCS'73*, 191–196.

[2] E. Csuhaj-Varjú, A connection between descriptional complexity of context-free grammars and grammar form theory, in: A. Kelemenová and J. Kelemen, eds., *Trends, Techniques, and Problems in Theoretical Computer Science*, Lecture Notes in Computer Science, Vol. 281 (Springer, Berlin, 1987) 60–74.

[3] E. Csuhaj-Varjú and J. Dassow, On bounded interpretations of grammar forms, *Theoret. Comput. Sci.* **87** (1991) 287–313.

[4] J. Gruska, On a classification of context-free grammars, *Kybernetika* **3** (1967) 22–29.

[5] J. Gruska, Some classifications of context-free languages, *Inform. and Control* **14** (1969) 152–179.

[6] A. Kelemenová, Grammatical levels of position restricted grammars, in: J. Gruska and M. Chytil, eds., *Proc. of MFCS '81*, Lectures Notes in Computer Science, Vol. 118 (Springer, Berlin, 1981) 347–359.

[7] A. Kelemenová, Grammatical complexity of context-free languages and normal forms of context-free grammars, in: P. Mikulecký, ed., *Proc. of IMYCS '82* (Bratislava, 1982) 239–258.

[8] A. Kelemenová, Complexity of normal form grammars, *Theoret. Comput. Sci.* **28** (1984) 288–314.

[9] A. Kelemenová, Structural complexity measures on grammar forms, in: *Proc. Conf. on Automata, Languages and Programming Systems*, Salgótarján DM 88-4 (Budapest, 1988) 73–76.

[10] A. Salomaa, *Formal Languages* (Academic Press, New York, 1973).

[11] D. Wood, *Grammar and L-forms: An Introduction*, Lecture Notes in Computer Science, Vol. 91 (Springer, Berlin, 1980).