DCFS 2007 (Draft)

Descriptional Complexity of Formal Systems
Deadline for submissions: March 20, 2007
Final versions: May 30, 2007

Descriptional Complexity of Generalized Forbidding Grammars

Tomáš Masopust Alexander Meduna

Faculty of Information Technology – Brno University of Technology –
Božetěchova 2 – Brno 61266 – Czech Republic
masopust@fit.vutbr.cz (T. Masopust) meduna@fit.vutbr.cz (A. Meduna)

Abstract. This paper discusses the descriptional complexity of generalized forbidding grammars in context of degrees, numbers of nonterminals and conditional productions, and a new descriptional complexity measure—an index—of generalized forbidding grammars.

Keywords: formal languages, descriptional complexity, generalized forbidding grammars

1 Introduction

Grammars whose derivations are regulated by various context conditions have always represented an important investigation area of formal language theory (see [7] for an overview, and [6] for the result that every recursively enumerable language can be generated by a generalized forbidding grammar of degree two with no more than thirteen conditional productions and fifteen nonterminals).

The present paper continues with this vivid topic of formal language theory by investigating their descriptional complexity. Specifically, it proves that every recursively enumerable language is generated (A) by a generalized forbidding grammar that has no more than nine nonterminals, ten conditional productions, six strings in the conditional set of any production, and any condition consists of two or fewer symbols; (B) by a generalized forbidding grammar that has no more than ten nonterminals, eleven conditional productions, four strings in the conditional set of any production, and any condition consists of two or fewer symbols; (C) by a generalized forbidding grammar that has no more than eight nonterminals, nine conditional productions, unlimited number of strings in the conditional set of any production, and any condition consists of two or fewer symbols.

This work was supported by the Czech Grant Agency projects 201/07/0005 and 102/05/H050, FRVŠ grant FR762/2007/G1, and the Czech Ministry of Education under the Research Plan MSM 0021630528.

2 Preliminaries and Definitions

We assume that the reader is familiar with formal language theory (see [1, 5]). For a set Q, $|Q|$ denotes the cardinality of Q. An alphabet is a finite nonempty set. For an alphabet V, V^* represents the free monoid generated by V. The unit of V^* is denoted by ε. Set $V^+ = V^* - \{\varepsilon\}$. For $w \in V^*$, $|w|$ and w^R denote the length and the mirror image of w, respectively. Set $sub(w) = \{u : u \text{ is a substring of } w\}$. RE denotes the family of all recursively enumerable languages.

Recall the result from [2].

Theorem 1. *Every recursively enumerable language is generated by a grammar in the* Geffert normal form $G_1 = (\{S, A, B, C, D\}, T, P \cup \{AB \to \varepsilon, CD \to \varepsilon\}, S)$, *where P contains context-free productions of the form*

$S \to uSa$, *where* $u \in \{A, C\}^*$, $a \in T$,
$S \to uSv$, *where* $u \in \{A, C\}^*$, $v \in \{B, D\}^*$,
$S \to uv$, *where* $u \in \{A, C\}^*$, $v \in \{B, D\}^*$.

In addition, any derivation generating a terminal string (or a terminal derivation, *for short) in G_1 is of the form $S \Rightarrow^* w_1w_2w$ by using productions from P, where $w_1 \in \{A, C\}^*$, $w_2 \in \{B, D\}^*$, $w \in T^*$, and $w_1w_2w \Rightarrow^* w$ by using $AB \to \varepsilon$ and $CD \to \varepsilon$.*

Definition 2. A *generalized forbidding grammar* (see [4]) is a quadruple $G = (N, T, P, S)$, where N is a nonterminal alphabet, T is a terminal alphabet such that $N \cap T = \emptyset$, $S \in N$ is the start symbol, and P is a finite set of productions of the form $(X \to \alpha, For)$ with $X \in N$, $\alpha \in (N \cup T)^*$, and $For \subseteq (N \cup T)^+$ being a finite set. If $For \neq \emptyset$, then the production $(X \to \alpha, For) \in P$ is said to be a *conditional* production; $cond(P)$ denotes the set of all conditional productions in P. For $x \in (N \cup T)^+$ and $y \in (N \cup T)^*$, x *directly derives* y according to the production $(X \to \alpha, For) \in P$, denoted by $x \Rightarrow y$, if $x = x_1Xx_2$, $y = x_1\alpha x_2$, for some $x_1, x_2 \in (N \cup T)^*$, and $For \cap sub(x) = \emptyset$. As usual, $\Rightarrow$ is extended to $\Rightarrow^i$, for $i \geq 0$, $\Rightarrow^+$, and $\Rightarrow^*$. The *language* generated by a generalized forbidding grammar, G, is defined as $\mathscr{L}(G) = \{w \in T^* : S \Rightarrow^* w\}$.

For $i, j, k, l \geq 0$, the language family $GF(i, j, k, l)$ is defined by this equivalence: $L \in GF(i, j, k, l)$ if and only if there is a generalized forbidding grammar $G = (N, T, P, S)$ that simultaneously satisfies:

(I) $L = \mathscr{L}(G)$,

(II) $(X \to \alpha, For) \in P$ and $x \in For$ implies $|x| \leq i$ (G's *degree*),

(III) $(X \to \alpha, For) \in P$ implies $|For| \leq j$ (G's *index*),

(IV) $|N| \leq k$,

(V) $|cond(P)| \leq l$.

3 Main Results

This section presents the main results of this paper.

Lemma 3. *Let $L \in RE$, $L = \mathscr{L}(G_1)$, G_1 is a grammar in the Geffert normal form. Then, there is a grammar $G = (\{S, 0, 1, \$\}, T, P \cup \{0\$0 \to \$, 1\$1 \to \$, \$ \to \varepsilon\}, S)$ with P containing only context-free productions of the form*

$S \to h(u)Sa$ *if* $S \to uSa$ *in* G_1,
$S \to h(u)Sh(v)$ *if* $S \to uSv$ *in* G_1,
$S \to h(u)\$h(v)$ *if* $S \to uv$ *in* G_1,

where $h : \{A, B, C, D\}^* \to \{0, 1\}^*$ is a homomorphism defined as $h(A) = h(B) = 0$ and $h(C) = h(D) = 1$, such that $\mathscr{L}(G) = \mathscr{L}(G_1)$.

Proof: Any terminal derivation in G_1 is, after the application of $S \to uv$, of the form $\{A, C\}^*\{B, D\}^*T^*$. From this, any terminal derivation in G is, after generating \$, of the form $h(\{A, C\}^*)\$h(\{B, D\}^*)T^*$. It is easy to see that if the production $AB \to \varepsilon$ or $CD \to \varepsilon$ is applied in G_1, then the production $0\$0 \to \$$ or $1\$1 \to \$$ is applied in G, respectively, and vice versa. Moreover, the last production applied in G in any terminal derivation is $\$ \to \varepsilon$. □

First, recall the result from [3].

Theorem 4. $RE = GF(2, 9, 10, 8)$.

We prove that the index and the number of nonterminals can be improved. However, the number of conditional productions increases.

Theorem 5. $RE = GF(2, 6, 9, 10)$.

The main idea of the proof is to simulate a terminal derivation of a grammar, G, in the form from Lemma 3. To do this, we first apply all context-free productions as applied in the G's derivation, and then we simulate the production $0\$0 \to \varepsilon$ so that we mark with $'$ two 0s and check that these marked symbols form a substring $0'\$0'$ of the current sentential form. If so, the marked symbols can be removed, which completes the simulation of the production $0\$0 \to \varepsilon$ in G; otherwise, the derivation must be blocked. Production $1\$1 \to \varepsilon$ is simulated analogously.

The formal proof follows.

Proof: Let L be a recursively enumerable language. Then, there is a grammar $G = (\{S, 0, 1, \$\}, T, P \cup \{0\$0 \to \$, 1\$1 \to \$, \$ \to \varepsilon\}, S)$ such that $L = \mathscr{L}(G)$ and P contains productions of the form shown in Lemma 3. Construct the grammar

$$G' = (\{S', Z, S, 0, 1, 0', 1', \$, \#\}, T, P' \cup P'', S'),$$

where P' contains productions of the form

$(S' \to ZSZ, \emptyset)$,
$(S \to uSZaZ, \emptyset)$ if $S \to uSa \in P$,
$(S \to uSv, \emptyset)$ if $S \to uSv \in P$,

$(S \to u\$v, \emptyset)$ if $S \to uv \in P$,

and P'' contains following ten conditional productions:

(I) $(0 \to 0', \{0', 1', \#\})$,
(II) $(1 \to 1', \{0', 1', \#\})$,
(III) $(0 \to 0'1', \{1', \#\})$,
(IV) $(1 \to 1'0', \{0', \#\})$,
(V) $(\$ \to \#, \{0\$, 1\$, Z\$, \$0, \$1, \$Z\})$,
(VI) $(0' \to \varepsilon, \{\$, S\})$,
(VII) $(1' \to \varepsilon, \{\$, S\})$,
(VIII) $(\# \to \$, \{0', 1'\})$,
(IX) $(Z \to \varepsilon, \{\$, \#, S\})$,
(X) $(\$ \to \varepsilon, \{0, 1, 0', 1'\})$,

To prove that $\mathscr{L}(G) \subseteq \mathscr{L}(G')$, consider a derivation, $S \Rightarrow^* w\$w^Rv$, in G using only productions from P, where $w \in \{0,1\}^*$ and $v \in T^*$. This can be derived in G' by productions from P' as $S' \Rightarrow^* Zw\$w^RZv'$, where $h(v') = v$ for a homomorphism $h : (T \cup \{Z\})^* \to T^*$ defined as $h(a) = a$, for $a \in T$, and $h(Z) = \varepsilon$. If $w = \varepsilon$, then $Z\$Zv' \Rightarrow ZZv' \Rightarrow^* v$, by productions (X) and (IX). If $w = w'0$, then

$$\begin{aligned} Zw'0\$0w'^RZv' &\Rightarrow Zw'0'\$0w'^RZv' \\ &\Rightarrow Zw'0'\$0'1'w'^RZv' \\ &\Rightarrow Zw'0'\#0'1'w'^RZv' \\ &\Rightarrow Zw'\#0'1'w'^RZv' \\ &\Rightarrow Zw'\#1'w'^RZv' \\ &\Rightarrow Zw'\#w'^RZv' \\ &\Rightarrow Zw'\$w'^RZv' \end{aligned}$$

by productions (I), (III), (V), (VI), (VI), (VII), and (VIII). The case of $w = w'1$ is analogous. The inclusion follows by induction.

To prove that $\mathscr{L}(G) \supseteq \mathscr{L}(G')$, consider a terminal derivation in G', $S' \Rightarrow^* Zw_1\$w_2Zw_3$, by productions from P', and $Zw_1\$w_2Zw_3 \Rightarrow^* w$, where $w_1, w_2 \in \{0,1\}^*$ and $w \in T^*$. We prove that $w_3 \in (T \cup \{Z\})^*$.

Assume that $Z0$ or $Z1$ is in $sub(Zw_3)$. Then, to eliminate this 0 or 1, production (VI) or (VII) must be applied. To apply production (VI) or (VII), production (V) must be applied before. Then, however, there is 0, 1, or Z next to \$; indeed, there cannot be more than two $0'$s or $1'$s in the derivation (there is no more than either $0'$ and $0'1'$, or $1'$ and $1'0'$). Thus, $w_3 \in (T \cup \{Z\})^*$ and $w = h(w_3)$. Then, $S \Rightarrow^* w_1\$w_2w$ in G by productions from P. We prove that $w_1\$w_2w \Rightarrow^* w$.

Assume that $w_1 = w_2 = \varepsilon$. Then, the only applicable production in G' is production (X). After production (X), only production (IX) is applicable. Thus, $Z\$Zw_3 \Rightarrow ZZw_3 \Rightarrow^* h(w_3)$.

Assume that $\varepsilon \in \{w_1, w_2\}$ and $w_1 \neq w_2$. Then,

$$Zw_1\$w_2Zw_3 \in \{Z\$w_2Zw_3, Zw_1\$Zw_3\}.$$

In both cases, neither 0 nor 1 can be eliminated (see production (v)).

By induction on the length of w_1, we prove that $w_1 = w_2^R$. The basic step has already been proved. Assume that $Zw_1\$w_2Zw_3 = Zw_1'0\$xw_2'Zw_3$, where $x \in \{0,1\}$. Then, only productions (i), (ii), (iii), (iv) can be applied. Notice that production (i) or (ii) is applied before production (iii) or (iv); otherwise, if production (iii) or (iv) is applied, then neither production (i) nor (ii) is applicable. Moreover, if production (i) is applied, then only production (iii) is applicable, and, similarly, if production (ii) is applied, then only production (iv) is applicable. According to production (v), 0\$ is rewritten by production (i) or (iii). Therefore, 0 is rewritten by production (i) and x is rewritten by production (iii), or vice versa. Thus, $x = 0$ and

$$Zw_1'0\$0w_2'Zw_3 \Rightarrow^2 Zw_1'0'\$0'1'w_2'Zw_3 \quad \text{or} \quad Zw_1'0'1'\$0'w_2'Zw_3.$$

Then, only production (v) is applicable;

$$\Rightarrow \;\; Zw_1'0'\#0'1'w_2'Zw_3 \quad \text{or} \quad Zw_1'0'1'\#0'w_2'Zw_3$$

and only productions (vi) and (vii) are applicable;

$$\Rightarrow^3 Zw_1'\#w_2'Zw_3$$

and only production (viii) is applicable;

$$\Rightarrow \;\; Zw_1'\$w_2'Zw_3.$$

The proof for $Zw_1\$w_2Zw_3 = Zw_1'1\$xw_2'Zw_3$, where $x \in \{0,1\}$, is analogous. By the induction hypothesis, $w_1 = w_2^R$.

Thus, if $S' \Rightarrow^* Zw_1\$w_1^RZw_3 \Rightarrow^* h(w_3)$ in G', where $w_1 \in \{0,1\}^*$ and $w_3 \in (T \cup \{Z\})^*$, then $S \Rightarrow^* w_1\$w_1^Rh(w_3) \Rightarrow^* h(w_3)$ in G. □

As a consequence of the previous theorem, we get the following corollary.

Corollary 6. *$RE = GF(2,4,10,11)$.*

Proof: Modify the set P'' from the proof of Theorem 5 in the following way.

(i) $(0 \to 0', \{0', 1', @\})$,
(ii) $(1 \to 1', \{0', 1', @\})$,
(iii) $(\$ \to \#, \{0\$, 1\$, Z\$\})$,
(iv) $(0 \to 0'1', \{1', @\})$,
(v) $(1 \to 1'0', \{0', @\})$,
(vi) $(\# \to @, \{\#0, \#1, \#Z\})$,
(vii) $(0' \to \varepsilon, \{\$, \#, S\})$,
(viii) $(1' \to \varepsilon, \{\$, \#, S\})$,
(ix) $(@ \to \$, \{0', 1'\})$,
(x) $(Z \to \varepsilon, \{\$, \#, @, S\})$,
(xi) $(\$ \to \varepsilon, \{0, 1\})$,

It is not hard to see that the only modification is that production (v) is split into two productions, (iii) and (vi). Thus, the proof is very similar to the previous one. We only demonstrate the main idea.

Assume the following sentential form, $Zw'0\$0w'^R Zv'$. Then,

$$\begin{aligned} Zw'0\$0w'^R Zv' &\Rightarrow Zw'0'\$0w'^R Zv' \\ &\Rightarrow Zw'0'\#0w'^R Zv' \\ &\Rightarrow Zw'0'\#0'1'w'^R Zv' \\ &\Rightarrow Zw'0'@0'1'w'^R Zv' \\ &\Rightarrow Zw'@0'1'w'^R Zv' \\ &\Rightarrow Zw'@1'w'^R Zv' \\ &\Rightarrow Zw'@w'^R Zv' \\ &\Rightarrow Zw'\$w'^R Zv' \end{aligned}$$

by productions (i), (iii), (iv), (vi), (vii), (vii), (viii), and (ix). □

If we allow the index to have no limitation, then the number of nonterminals and conditional productions can be decreased. To prove this, we first need to modify Lemma 3. More precisely, only the homomorphism h is modified.

Lemma 7. *Let $L \in RE$, $L = \mathscr{L}(G_1)$, G_1 is a grammar in the Geffert normal form. Then, there is a grammar $G = (\{S, 0, 1, \$\}, T, P \cup \{0\$0 \to \$, 1\$1 \to \$, \$ \to \varepsilon\}, S)$ with P containing only context-free productions of the form*

$S \to h(u)Sa$ *if* $S \to uSa$ *in* G_1,
$S \to h(u)Sh(v)$ *if* $S \to uSv$ *in* G_1,
$S \to h(u)\$h(v)$ *if* $S \to uv$ *in* G_1,

where $h : \{A, B, C, D\}^* \to \{0, 1\}^*$ is a homomorphism defined as $h(A) = h(B) = 00$, $h(C) = 01$, and $h(D) = 10$, such that $\mathscr{L}(G) = \mathscr{L}(G_1)$.

Now, we can prove the following theorem giving the best result with respect to the number of nonterminals.

Theorem 8. $RE = GF(2, \infty, 8, 9)$.

Proof: Let L be a recursively enumerable language. Then, there is a grammar $G = (\{S, 0, 1, \$\}, T, P \cup \{0\$0 \to \$, 1\$1 \to \$, \$ \to \varepsilon\}, S)$ such that $L = \mathscr{L}(G)$ and P contains productions of the form shown in Lemma 7. Construct the grammar

$$G' = (\{S', S, 0, 1, 0', 1', \$, \#\}, T, P' \cup P'', S'),$$

where P' contains productions of the form

$(S' \to 111S11, \emptyset)$,
$(S \to uS11a, \emptyset)$ if $S \to uSa \in P$,
$(S \to uSv, \emptyset)$ if $S \to uSv \in P$,
$(S \to u\$v, \emptyset)$ if $S \to uv \in P$,

and P'' contains following nine conditional productions:

(i) $(0 \to 0', \{0', 1', \#\})$,
(ii) $(1 \to 1', \{0', 1', \#\})$,
(iii) $(0 \to 0'1', \{1', \#\})$,
(iv) $(1 \to 1'0', \{0', \#\})$,
(v) $(\$ \to \#, \{0\$, 1\$, \$0, \$1\} \cup \{\$\}T)$,
(vi) $(0' \to \varepsilon, \{\$, S\})$,
(vii) $(1' \to \varepsilon, \{\$, S\})$,
(viii) $(\# \to \$, \{0', 1'\})$,
(ix) $(\$ \to \varepsilon, \{0, 0'\})$,

To prove that $\mathscr{L}(G) \subseteq \mathscr{L}(G')$, consider a derivation, $S \Rightarrow^* w\$w^R v$, in G using only productions from P, where $w \in \{00, 01\}^*$ and $v \in T^*$. This can be derived in G' by productions from P' as $S' \Rightarrow^* 111w\$w^R 11v'$, where $v' \in (T\{11\})^*$ and $h(v') = v$ for a homomorphism $h : (T \cup \{1\})^* \to T^*$ defined as $h(a) = a$, for $a \in T$, and $h(1) = \varepsilon$. If $w = \varepsilon$, then $111\$11v' \Rightarrow 11111v' \Rightarrow^* v$, by productions (ix), and repeating productions (ii) and (vii). If $w = w'0$, then

$$\begin{aligned} 111w'0\$0w'^R 11v' &\Rightarrow 111w'0'\$0w'^R 11v' \\ &\Rightarrow 111w'0'\$0'1'w'^R 11v' \\ &\Rightarrow 111w'0'\#0'1'w'^R 11v' \\ &\Rightarrow 111w'\#0'1'w'^R 11v' \\ &\Rightarrow 111w'\#1'w'^R 11v' \\ &\Rightarrow 111w'\#w'^R 11v' \\ &\Rightarrow 111w'\$w'^R 11v' \end{aligned}$$

by productions (i), (iii), (v), (vi), (vi), (vii), and (viii). The case of $w = w'1$ is analogous. The inclusion follows by induction.

To prove that $\mathscr{L}(G) \supseteq \mathscr{L}(G')$, consider a terminal derivation in G', $S' \Rightarrow^* 111w_1\$w_2 11w_3$, by productions from P', and $111w_1\$w_2 11w_3 \Rightarrow^* w$, where $w_1 \in \{00, 01\}^*$, $w_2 \in \{00, 10\}^*$, and $w \in T^*$.

Assume that $\varepsilon \in \{w_1, w_2\}$ and $w_1 \neq w_2$. Then,

$$111w_1\$w_2 11w_3 \in \{111\$w_2 11w_3, 111w_1\$11w_3\}.$$

First, assume that $111\$w_2 11w_3 = 111\$xw_2' 11w_3$, where $x \in \{00, 10\}$. As in the proof of Theorem 5, only productions (i), (ii), (iii), and (iv) can be applied. Moreover, production (i) (or (ii)) is applied before production (iii) (or (iv)). If production (i) is applied, then only production (iii) is applicable, and, similarly, if production (ii) is applied, then only production (iv) is applicable. According to production (v), $1\$$ is rewritten by production (ii) or (iv). Therefore, 1 is rewritten by production (ii) and x is rewritten by production (iv), or vice versa. Thus, $x = 10$ and $111\$10w_2' 11w_3 \Rightarrow^7 11\$0w_2' 11w_3$. Similarly, assume that $111w_1\$11w_3 = 111w_1' x\$11w_3$, $x \in \{00, 01\}$. Then, $x = 01$ and $111w_1' 01\$11w_3 \Rightarrow^* 111w_1' 0\$1w_3$. In both cases, the derivation is blocked.

Assume that $w_1 = w_2 = \varepsilon$, i.e. $S' \Rightarrow^* 111\$11w_3$, where $w_3 = aw_3'$, for some $a \in T$, or $w_3 = \varepsilon$. Then, $111\$11w_3 \Rightarrow^* \alpha$, where

$$\alpha \in \{111\$11w_3, 11\$1w_3, 1\$aw_3', 1\$\}.$$

In all cases, to remove \$, production (IX) is applied. However, production (IX) is applicable if and only if there is no 0 in w_3. Thus, $w_3 \in (T \cup \{1\})^*$, i.e., $h(w_3) = w$. Notice that if there is no \$ in the sentential form, then all 1s can be removed by productions (II) and (VII). Clearly, $\$w \Rightarrow w$ in G.

Analogously to the proof of Theorem 5, we can prove that $w_1 = w_2^R$.

Thus, we have proved that $0 \notin sub(w_3)$ and if $S' \Rightarrow^* 111w_1\$w_1^R11w_3 \Rightarrow^* h(w_3)$ in G', where $w_1 \in \{00, 01\}^*$, then $S \Rightarrow^* w_1\$w_1^R h(w_3) \Rightarrow^* h(w_3)$ in G. $\square$

References

[1] J. Dassow and Gh. Păun. *Regulated Rewriting in Formal Language Theory.* Springer-Verlag, Berlin, 1989.

[2] V. Geffert. Context-free-like forms for the phrase-structure grammars. In M. Chytil, L. Janiga, and V. Koubek, editors, *MFCS*, volume 324 of *Lecture Notes in Computer Science*, pages 309–317. Springer, 1988.

[3] T. Masopust and A. Meduna. Descriptional complexity of grammars regulated by context conditions. In *LATA 2007 Pre-proceedings*, pages 403–411, Tarragona, Spain, 2007.

[4] A. Meduna. Generalized forbidding grammars. *International Journal of Computer Mathematics*, 36:31–38, 1990.

[5] A. Meduna. *Automata and Languages: Theory and Applications.* Springer-Verlag, London, 2000.

[6] A. Meduna and M. Švec. Descriptional complexity of generalized forbidding grammars. *International Journal of Computer Mathematics*, 80:11–17, 2003.

[7] A. Meduna and M. Švec. *Grammars with Context Conditions and Their Applications.* John Wiley & Sons, New York, 2005.