Theoretical Computer Science 475 (2013) 66–72

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Necessary conditions for subclasses of random context languages

Sigrid Ewert [a,*], Andries van der Walt [b,1]

[a] School of Computer Science, University of the Witwatersrand, Johannesburg, Private Bag 3, Wits, 2050, South Africa

[b] Department of Mathematical Sciences, Computer Science Division, University of Stellenbosch, Private Bag X1, Matieland, 7602, South Africa

ARTICLE INFO

Article history:
Received 28 September 2011
Received in revised form 6 September 2012
Accepted 30 December 2012
Communicated by M. Ito

Keywords:
Formal language
Regulated rewriting
Random context language
Random forbidding context language
Random permitting context language
Context-free language
Necessary condition

ABSTRACT

Random context grammars belong to the class of context-free grammars with regulated rewriting. Their productions depend on context that may be randomly distributed in a sentential form. Context is classified as either permitting or forbidding, where permitting context enables the application of a production and forbidding context inhibits it. We have proven a pumping lemma for random permitting context languages and a shrinking lemma for random forbidding context languages. We now present new necessary conditions for both these classes of languages and illustrate them with examples. We also present and illustrate a new necessary condition for context-free languages.

1. Introduction

Random context grammars (rcgs)[2] [7] belong to the class of context-free grammars with regulated rewriting [2], i.e., the productions of a grammar are context-free, but are applied in a non-context-free manner.

In the case of random context grammars, the application of a production at any step in a derivation may depend on the set of symbols that appear in the sentential form of the derivation at that step. As opposed to context-sensitive grammars, the context may be distributed in a random manner in the sentential form. Context is classified as either permitting or forbidding: permitting context enables the application of a production, while forbidding context inhibits it. When a grammar uses permitting context only or forbidding context only, it is called a random permitting context grammar (rPcg) or random forbidding context grammar (rFcg), respectively. The corresponding languages are called random permitting context languages (rPcls) and random forbidding context languages (rFcls).

Dassow and Păun [2] showed that random context grammars without erasing productions lie strictly between the context-free and context-sensitive grammars. When erasing productions are allowed, random context grammars are as powerful as the recursively enumerable grammars. In the remainder of this paper, we will use the term *random context grammars* when referring to random context grammars without erasing productions.

* Corresponding author. Tel.: +27 11 717 6180; fax: +27 86 616 3839.
E-mail address: sigrid.ewert@wits.ac.za (S. Ewert).

[1] Andries van der Walt passed away in December 2008.

[2] Abbreviations: rcg - random context grammar; rcl - random context language; rPcg - random permitting context grammar; rPcl - random permitting context language; rFcg - random forbidding context grammar; rFcl - random forbidding context language; cfg - context-free grammar; cfl - context-free language; ET0L - extended table-driven context-free Lindenmayer

0304-3975/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2012.12.038

To our knowledge, an example of a context-sensitive language that cannot be generated by a random context grammar has not been found yet. In [2], Dassow and Păun conjectured that the following language is such an example: $\mathcal{L} = \{xcx \mid x \in \mathcal{D}\}$, where $\mathcal{D}$ is the language of balanced brackets over $\{[,]\}$. However, in [1], we proved the conjecture to be false.

We proved a pumping lemma for random permitting context languages in [3], and used it to show that random permitting context grammars are strictly weaker than random context grammars. A language that cannot be generated by any rPcg is $\mathcal{L} = \left\{a^{2^n} \mid n \geq 1\right\}$.

In [6] we proved a shrinking lemma for random forbidding context languages and showed that random forbidding context grammars are strictly weaker than random context grammars. A language that cannot be generated by any rFcg is $\mathcal{L} = \{z_1, z_2, \ldots\}$, where $z_1 = [a]$, $z_i = (\left[a^i\right])^{4|z_{i-1}|}$, $i > 1$, and a, [and] are terminal symbols.

Furthermore, Rabkin [5] developed analogues of Ogden's lemma [4] for random permitting and forbidding context languages.

We now present new necessary conditions for both rPcls and rFcls and illustrate them with examples. We also present a new necessary condition for context-free languages.

We formally introduce random context grammars in Section 2. In Section 3 we state two lemmas that are required for the work following. In Section 4 we concentrate on random permitting context languages and prove a necessary condition for a language to be generated by an rPcg. We then use this condition to show that a specific language is not an rPcl. In Section 5 we concentrate on random forbidding context languages and prove a necessary condition for these languages. We then illustrate this condition with some examples. In Section 6 we concentrate on context-free languages, which are strictly contained in both the random permitting context languages and random forbidding context languages. We prove a necessary condition for a language to be generated by a context-free grammar, and use it to show that a specific language is not context-free. In Section 7 we recommend future work.

2. Definitions

In this section we present the necessary notation and terminology.

Let $\mathbb{N}_+ = \{1, 2, \ldots\}$. Moreover, for $m \in \mathbb{N}_+$, let $[m] = \{1, 2, \ldots, m\}$.

Let ϵ denote the empty string.

A random context grammar $G = (V_N, V_T, P, S)$ has a finite alphabet V of symbols, consisting of the disjoint subsets V_N of variables and V_T of terminals. P is a finite set of productions of the form $A \to \alpha\ (\mathcal{P}; \mathcal{F})$, where $A \in V_N$, $\alpha \in V^+$ and $\mathcal{P}, \mathcal{F} \subseteq V_N$. Finally, there is a start symbol S, $S \in V_N$.

If there is a production $A \to \alpha\ (\mathcal{P}; \mathcal{F})$ in G and if γ_1, γ_2 are in V^*, then we may write $\gamma_1 A \gamma_2 \Longrightarrow \gamma_1 \alpha \gamma_2$ if every $B \in \mathcal{P}$ is in the string $\gamma_1\gamma_2$ and no $B \in \mathcal{F}$ is in the string $\gamma_1\gamma_2$. As usual, $\Longrightarrow^*$ denotes the reflexive transitive closure of $\Longrightarrow$. The random context language (rcl) $\mathcal{L}(G)$ generated by an rcg G is the set $\left\{z \in V_T^* \mid S \Longrightarrow^* z\right\}$.

For the sake of simplicity, we write a production of the form $A \to \alpha\ (\emptyset; \emptyset)$ as $A \to \alpha$.

If every production in an rcg G has $\mathcal{P} = \mathcal{F} = \emptyset$, G is a context-free grammar (cfg); if $\mathcal{F} = \emptyset$ for every production, we call G a random permitting context grammar, and if $\mathcal{P} = \emptyset$ for every production, we call G a random forbidding context grammar. We call the corresponding languages context-free languages (cfls), random permitting context languages and random forbidding context languages, respectively.

Let $G = (V_N, V_T, P, S)$ be an rcg. For $\alpha \in {V_N}^*$, let $l(\alpha) = \min\{B \subseteq V_N \mid \alpha \in B^*\}$. We refer to $l(\alpha)$ as the labels in α.

For $\alpha \in V^*$, we denote the length of α by $|\alpha|$. For $z, w \in {V_T}^*$, we write $w \sqsubseteq z$ if z can be written $z = z_1 w z_2$; we write $w \sqsubset z$ if $|z_1 z_2| \neq 0$. We call w a factor and a proper factor of z, respectively.

Suppose $S \Longrightarrow^* \alpha \Longrightarrow^* \beta$ is a derivation in G, where $\alpha = A_1 A_2 \ldots A_s$ and $\beta = \gamma_1\gamma_2 \ldots \gamma_s$, with $s \in \mathbb{N}_+$, $A_j \in V_N$ and $\gamma_j \in V^*$ for $j \in [s]$. We define the derivation tree corresponding to a derivation in the usual way [4]. Consider α and β as two cuts in the derivation tree. If the nodes in γ_j are all the descendants of A_j in cut β, then we write $A_j \Longrightarrow_c^* \gamma_j$.

Suppose $|V| = n$ and that V is ordered. Then we can represent a sentential form α as an n-vector of nonnegative integers, written as $\vec{\alpha}$, such that, if $\vec{\alpha} = (m_1, m_2, \ldots, m_n)$, then α contains exactly m_i occurrences of the ith symbol in V.

Let $n \in \mathbb{N}_+$. Let $\vec{\alpha} = (m_1, m_2, \ldots, m_n)$ and $\vec{\beta} = (p_1, p_2, \ldots, p_n)$ be n-vectors of integers. Then let $|\vec{\alpha}| = \sum_{i=1}^{n} m_i$ and $\zeta(\vec{\alpha}) = |\left\{j \in [n] \mid m_j = 0\right\}|$. Moreover, we write $\vec{\alpha} \leq \vec{\beta}$ if and only if for all $i \in [n]$, $m_i \leq p_i$. Similarly, we write $\vec{\alpha} < \vec{\beta}$ if and only if for all $i \in [n]$, $m_i < p_i$.

3. Useful results

In this section we present two results that are required for the work following.

Lemma 1. *Let $m_1, m_2, \ldots$ be an infinite sequence of nonnegative integers. Let n be any positive integer. Then, for any $h \geq 2$, there exists an integer b, which depends on h, such that if $\vec{\alpha_1}, \vec{\alpha_2}, \ldots$ is an infinite sequence of n-vectors of nonnegative integers with $|\vec{\alpha_i}| \leq m_i$, $i \geq 1$, then there are h indices $i_1, i_2, \ldots, i_h$, with $1 \leq i_1 < i_2 < \ldots < i_h \leq b$, such that $\vec{\alpha_{i_1}} \leq \vec{\alpha_{i_2}} \leq \ldots \leq \vec{\alpha_{i_h}}$.*

Proof. Given on Page 153 of [6]. □

Lemma 2. *Let n be any positive integer. Let $\vec{\alpha_1} < \vec{\alpha_2} < \ldots < \vec{\alpha_{n+1}}$ be a sequence of non-null n-vectors of nonnegative integers. Then there exist r and s, $1 \leq r < s \leq n+1$, such that $\zeta(\vec{\alpha_r}) = \zeta(\vec{\alpha_s})$.*

Proof. Assume the lemma is false. Then $n > \zeta(\vec{\alpha_1}) > \zeta(\vec{\alpha_2}) > \cdots > \zeta(\vec{\alpha_n}) > \zeta(\vec{\alpha_{n+1}}) \geq 0$. This is impossible. Therefore the assumption is false. □

4. A necessary condition for rPcls

In this section we concentrate on random permitting context languages and prove a necessary condition for a language to be generated by a grammar that uses permitting context only.

Necessary conditions for rPcls already exist. For example, in [3], we proved a pumping lemma for rPcls and in [5], Rabkin developed an analogue of Ogden's lemma [4] for these languages. An immediate consequence of the pumping property is that the length set of each infinite language generated by an rPcg contains an infinite arithmetic progression. This implies that the language $\mathcal{L} = \{a^{2^n} \mid n \geq 1\}$ cannot be generated by any rPcg. Since this language is an rcl [2], it follows that random permitting context grammars are strictly weaker than random context grammars.

For the necessary condition that we will prove in this section, Theorem 5, we need the following technical lemma. It states that in the permitting case, additional context cannot inhibit the application of productions.

Lemma 3. *Let $G = (V_N, V_T, P, S)$ be an rPcg. Let $s \in \mathbb{N}_+$. Suppose $S \Longrightarrow^* \alpha = A_1A_2 \ldots A_s$, where $A_j \in V_N \cup \{\epsilon\}$ for $j \in [s]$. Suppose $S \Longrightarrow^* \beta = A'_1A'_2 \ldots A'_s$, where $A'_j \in V_N$ and $A'_j = A_j$ if $A_j \neq \epsilon$ for $j \in [s]$.*

Suppose $\alpha \Longrightarrow^* \gamma_1\gamma_2 \ldots \gamma_s$, with $A_j \Longrightarrow_c^* \gamma_j$ for $j \in [s]$. Then a derivation for β is $\beta \Longrightarrow^* \gamma'_1\gamma'_2 \ldots \gamma'_s$, with $A'_j \Longrightarrow_c^* \gamma'_j$ for $j \in [s]$, where $\gamma'_j = \gamma_j$ if $A'_j = A_j$.

Proof. By induction on k, the length of the derivation.

1. Suppose $k = 1$. Then

$$\begin{aligned} \alpha &= A_1A_2 \ldots A_s \\ &= A_1A_2 \ldots A_{i-1}A_iA_{i+1} \ldots A_s \\ &\Longrightarrow A_1A_2 \ldots A_{i-1}\gamma_iA_{i+1} \ldots A_s \end{aligned}$$

using a production $A_i \to \gamma_i(\mathcal{P}; \emptyset)$.

Consider $\beta = A'_1A'_2 \ldots A'_{i-1}A'_iA'_{i+1} \ldots A'_s$. If $A'_i = A_i$, then

$$\begin{aligned} \beta &= A'_1A'_2 \ldots A'_{i-1}A_iA'_{i+1} \ldots A'_s \\ &\Longrightarrow A'_1A'_2 \ldots A'_{i-1}\gamma_iA'_{i+1} \ldots A'_s, \end{aligned}$$

using the production $A_i \to \gamma_i(\mathcal{P}; \emptyset)$, since $\{A_1, A_2, \ldots, A_{i-1}, A_{i+1}, \ldots, A_s\} \subseteq \{A'_1, A'_2, \ldots, A'_{i-1}, A'_{i+1}, \ldots, A'_s\}$.

2. Suppose the statement is true for k, i.e., if $\alpha \Longrightarrow^k \gamma_1\gamma_2 \ldots \gamma_s$, with $A_j \Longrightarrow_c^* \gamma_j$ for $j \in [s]$, then $\beta \Longrightarrow^* \gamma'_1\gamma'_2 \ldots \gamma'_s$, with $A'_j \Longrightarrow_c^* \gamma'_j$ for $j \in [s]$, where $\gamma'_j = \gamma_j$ if $A'_j = A_j$.
3. Consider $k+1$:

Suppose $\alpha = A_1A_2 \ldots A_s \Longrightarrow^{k+1} \gamma_1\gamma_2 \ldots \gamma_s$, with $A_j \Longrightarrow_c^* \gamma_j$ for $j \in [s]$.

Then, for some $B \in V_N$,

$$\begin{aligned} \alpha &\Longrightarrow^k \gamma_1\gamma_2 \ldots \gamma_{i-1}\delta_lB\delta_r\gamma_{i+1} \ldots \gamma_s \\ &\Longrightarrow \gamma_1\gamma_2 \ldots \gamma_{i-1}\delta_l\kappa\delta_r\gamma_{i+1} \ldots \gamma_s \\ &= \gamma_1\gamma_2 \ldots \gamma_{i-1}\gamma_i\gamma_{i+1} \ldots \gamma_s, \end{aligned}$$

using production $B \to \kappa(\mathcal{P}; \emptyset)$.

According to the hypothesis,

$$\begin{aligned} \beta &= A'_1A'_2 \ldots A'_s \\ &\Longrightarrow^* \gamma'_1\gamma'_2 \ldots \gamma'_{i-1}\delta'_lB'\delta'_r\gamma'_{i+1} \ldots \gamma'_s, \end{aligned}$$

with $A'_j \Longrightarrow_c^* \gamma'_j$ for $j \in [s]$, where

- for $j \neq i$, $\gamma'_j = \gamma_j$ if $A'_j = A_j$,
- for $j = i$, $\delta'_lB'\delta'_r = \delta_lB\delta_r = \gamma_i$ if $A'_i = A_i$.

Since $l(\gamma_1\gamma_2 \ldots \gamma_{i-1}\delta_l\delta_r\gamma_{i+1} \ldots \gamma_s) \subseteq l(\gamma'_1\gamma'_2 \ldots \gamma'_{i-1}\delta'_l\delta'_r\gamma'_{i+1} \ldots \gamma'_s)$, the production $B \to \kappa(\mathcal{P}; \emptyset)$ is enabled. Therefore $\beta \Longrightarrow^* \gamma'_1\gamma'_2 \ldots \gamma'_s$. □

In the case of random context, only the presence or absence of the context variables is important, and not the order in which the variables appear. Therefore we have the following:

Corollary 4. *Let $G = (V_N, V_T, P, S)$ be an rPcg. Suppose $S \Longrightarrow^* \alpha = A_1A_2 \dots A_s$, where $A_j \in V_N \cup \{\epsilon\}$ for $j \in [s]$. Suppose $S \Longrightarrow^* \beta = B_1B_2 \dots B_s$, where $B_j \in V_N$ for $j \in [s]$. Let $\{A_1, A_2, \dots, A_s\} \setminus \{\epsilon\} \subseteq \{B_1, B_2, \dots, B_s\}$.*

Let $\alpha \Longrightarrow^* \gamma_1\gamma_2 \dots \gamma_s$, with $A_j \Longrightarrow_c^* \gamma_j$, $j \in [s]$. Then a derivation for β is $\beta \Longrightarrow^* \gamma_{i_1}\gamma_{i_2} \dots \gamma_{i_s}$, with $B_j \Longrightarrow_c^* \gamma_{i_j}$, $j \in [s]$, where $\gamma_{i_j} = \gamma_j$ if $B_j = A_j$.

We now present the main result of this section, a necessary condition for random permitting context languages. In essence we prove that if a word is sufficiently long, then any derivation contains two sentential forms α and β such that α derives β, but they have the same labels. Starting from β, we can copy the derivation sequence that led from α to β, since in the case of an rPcg, any additional context in β cannot inhibit the application of productions.

Theorem 5. *Let $\mathcal{L}$ be an rPcl. Then there exists an n such that any word $z \in \mathcal{L}$ with $|z| \geq n$ has a factor v with $|v| \geq |z|/n$ that is a proper factor of a word $y \in \mathcal{L}$ with $|y| > |z|$.*

Proof. Let $\mathcal{L}$ be generated by an rPcg $G = (V_N, V_T, P, S)$. Let t be the length of the longest right-hand side of all productions in P. Let $m_j = 1 + (j-1)(t-1)$, $j \in \mathbb{N}_+$. Let $p = |V_N \cup V_T|$. Let b be the integer of Lemma 1 that depends on $p+1$.

Let $n = 1 + (b-1)(t-1)$. Let $z \in \mathcal{L}$ with $|z| \geq n$. Consider a derivation of z, i.e., $S \Longrightarrow^* z$. This derivation can be written as

$$S = \alpha_1 \Longrightarrow^* \alpha_2 \Longrightarrow^* \cdots \Longrightarrow^* \alpha_q \Longrightarrow^* z \ ,$$

where $|\alpha_j| < |\alpha_{j+1}|$ for $j \in [q-1]$, and q is as large as possible.

We note that, for all j, $j \in [q]$, $|\alpha_j| \leq m_j$. Then, according to Lemma 1, there are $p+1$ indices $i_1, i_2, \dots, i_{p+1}$, with $1 \leq i_1 < i_2 < \cdots < i_{p+1} \leq b$, such that $\alpha_{i_1} \leq \alpha_{i_2} \leq \cdots \leq \alpha_{i_{p+1}}$, where $\leq$ for strings means that their Parikh vectors have this relation. By construction, $\alpha_{i_1} < \alpha_{i_2} < \cdots < \alpha_{i_{p+1}}$.

For every j, $j \in [p+1]$, $|\alpha_{i_j}| \leq n$. Therefore α_{i_j} has maximally n variables and consequently at least one variable in α_{i_j} generates a string of length at least $|z|/n$.

Consider $\alpha_{i_1} < \alpha_{i_2} < \cdots < \alpha_{i_{p+1}}$. According to Lemma 2, there exist r and s, $1 \leq r < s \leq p+1$, such that $\zeta(\vec{\alpha}_r) = \zeta(\vec{\alpha}_s)$. Let B be a variable in α_r that derives a factor, say v, of length at least $|z|/n$. Starting from α_s and using Corollary 4, we can ensure that a copy of B in α_s derives v. Let y be the word derived in this way. Then y contains the factor v. Since $|\alpha_r| < |\alpha_s|$, v is a proper factor of y. □

With Theorem 5, it can easily be shown that the following language is not an rPcl.

Example 6. The language $\mathcal{L} = \left\{ \left(ga^k\right)^l \mid 0 \leq l \leq k \right\}$ cannot be generated by any rPcg.

Proof. Suppose $\mathcal{L}$ is generated by an rPcg. Let n be the integer of Theorem 5.

Now consider $z = \left(ga^{2n}\right)^{2n}$. Then $z \in \mathcal{L}$. Moreover, $|z| = 2n(2n+1) \geq n$. According to Theorem 5, z has a factor v with $|v| \geq |z|/n$ that is a proper factor of a word $y \in \mathcal{L}$ with $|y| > |z|$.

Consider a string v with $|v| \geq |z|/n = 2(2n+1)$. Then v contains the factor $ga^{2n}g$. Then there is a word $y \in \mathcal{L}$ with $|y| > |z|$ such that y contains $ga^{2n}g$. This contradicts the definition of $\mathcal{L}$. Therefore $\mathcal{L}$ cannot be generated by any rPcg. □

5. A necessary condition for rFcls

In this section we concentrate on random forbidding context languages and prove a necessary condition for a language to be generated by a grammar that uses forbidding context only.

Necessary conditions for rFcls already exist. For example, in [6], we proved a shrinking lemma for random forbidding context languages and in [5], Rabkin developed an analogue of Ogden's lemma [4] for these languages. As shown in [6], the language $\mathcal{L} = \{z_1, z_2, \dots\}$, where $z_1 = [a]$, $z_i = (\left[a^i\right])^{4|z_{i-1}|}$, $i > 1$, and a, [and] are terminal symbols, cannot be generated by any rFcg. Since this language is an rcl [6], it follows that random forbidding context grammars are strictly weaker than random context grammars.

For the necessary condition that we will prove in this section, Theorem 11, we need the following normal form for rFcgs:

Lemma 7. *Let $G = (V_N, V_T, P, S)$ be an rFcg. Then there exists an rFcg $G' = \left(V'_N, V_T, P', S\right)$ such that $\mathcal{L}\left(G'\right) = \mathcal{L}(G)$ and every production in P' has one of the following types:*

1. $A \to BC\ (\emptyset; \mathcal{F})$, $A, B, C \in V'_N$;
2. $A \to B\ (\emptyset; \mathcal{F})$, $A, B \in V'_N$;
3. $A \to a$, $A \in V'_N$, $a \in V_T$.

Proof. Given on Page 68 of [7]. □

Due to the normal form, every word has a derivation such that no variable is introduced into the derivation once a terminal appears in the derivation.

Lemma 8. *Let $\mathcal{L}$ be an rFcl. Let $\mathcal{L}$ be generated by an rFcg $G = (V_N, V_T, P, S)$ in normal form. Let $z \in \mathcal{L}$. Then there is a derivation of z in G of the form*

$$S = \alpha_1 \Longrightarrow^* \alpha_2 \Longrightarrow^* \cdots \Longrightarrow^* \alpha_{|z|} \Longrightarrow^* z \ ,$$

where, for $1 \leq i \leq |z|$, $|\alpha_i| = i$ and α_i consists of nonterminals only.

Proof. Let $V_X = \{X_a \mid a \in V_T\}$. Then let $G' = \left(V'_N, V_T, P', S\right)$, where

1. $V'_N = V_N \cup V_X$, and
2. P' is constructed by
   (a) adding the two productions $A \to X_a$ and $X_a \to a$ $(\emptyset; V_N)$ to P' for any production in P of the form $A \to a$, with $A \in V_N$ and $a \in V_T$,
   (b) adding the production $A \to BC$ $(\emptyset; \mathcal{F})$ to P' for any production in P of the form $A \to BC$ $(\emptyset; \mathcal{F})$, with $A, B, C \in V_N$, and
   (c) adding the production $A \to B$ $(\emptyset; \mathcal{F})$ to P' for any production in P of the form $A \to B$ $(\emptyset; \mathcal{F})$, with $A, B \in V_N$.

Then it should be clear that

- $\mathcal{L}\left(G'\right) = \mathcal{L}(G)$, and that
- no element of V_N can be introduced into the derivation once an element of V_T appears in the derivation. □

For Theorem 11 we also need the following technical lemma. It states that in the forbidding case, the lack of context cannot inhibit the application of productions.

Lemma 9. *Let $G = (V_N, V_T, P, S)$ be an rFcg. Let $s \in \mathbb{N}_+$. Suppose $S \Longrightarrow^* \alpha = A_1A_2 \ldots A_s$, where $A_j \in V_N$ for $j \in [s]$. Suppose $S \Longrightarrow^* \beta = A'_1A'_2 \ldots A'_s$, where $A'_j = A_j$ or $A'_j = \epsilon$ for $j \in [s]$.*

Suppose $\alpha \Longrightarrow^* \gamma_1\gamma_2 \ldots \gamma_s$, with $A_j \Longrightarrow_c^* \gamma_j$ for $j \in [s]$. Then a derivation for β is $\beta \Longrightarrow^* \gamma'_1\gamma'_2 \ldots \gamma'_s$, with $A'_j \Longrightarrow_c^* \gamma'_j$ for $j \in [s]$, where $\gamma'_j = \gamma_j$ if $A'_j = A_j$, and $\gamma'_j = \epsilon$ if $A'_j = \epsilon$.

Proof. By induction on k, the length of the derivation.

1. Suppose $k = 1$. Then

$$\begin{aligned} \alpha &= A_1A_2 \ldots A_s \\ &= A_1A_2 \ldots A_{i-1}A_iA_{i+1} \ldots A_s \\ &\Longrightarrow A_1A_2 \ldots A_{i-1}\gamma_iA_{i+1} \ldots A_s \end{aligned}$$

   using a production $A_i \to \gamma_i(\emptyset; \mathcal{F})$.
   Consider $\beta = A'_1A'_2 \ldots A'_{i-1}A'_iA'_{i+1} \ldots A'_s$. If $A'_i = A_i$, then

$$\begin{aligned} \beta &= A'_1A'_2 \ldots A'_{i-1}A_iA'_{i+1} \ldots A'_s \\ &\Longrightarrow A'_1A'_2 \ldots A'_{i-1}\gamma_iA'_{i+1} \ldots A'_s, \end{aligned}$$

   using the production $A_i \to \gamma_i(\emptyset; \mathcal{F})$, since $\left\{A'_1, A'_2, \ldots, A'_{i-1}, A'_{i+1}, \ldots, A'_s\right\} \subseteq \{A_1, A_2, \ldots, A_{i-1}, A_{i+1}, \ldots, A_s\}$.
2. Suppose the statement is true for k, i.e., if $\alpha \Longrightarrow^k \gamma_1\gamma_2 \ldots \gamma_s$, with $A_j \Longrightarrow_c^* \gamma_j$ for $j \in [s]$, then $\beta \Longrightarrow^* \gamma'_1\gamma'_2 \ldots \gamma'_s$, with $A'_j \Longrightarrow_c^* \gamma'_j$ for $j \in [s]$, where $\gamma'_j = \gamma_j$ if $A'_j = A_j$, and $\gamma'_j = \epsilon$ if $A'_j = \epsilon$.
3. Consider $k + 1$:
   Suppose $\alpha = A_1A_2 \ldots A_s \Longrightarrow^{k+1} \gamma_1\gamma_2 \ldots \gamma_s$, with $A_j \Longrightarrow_c^* \gamma_j$ for $j \in [s]$.
   Then, for some $B \in V_N$,

$$\begin{aligned} \alpha &\Longrightarrow^k \gamma_1\gamma_2 \ldots \gamma_{i-1}\delta_lB\delta_r\gamma_{i+1} \ldots \gamma_s \\ &\Longrightarrow \gamma_1\gamma_2 \ldots \gamma_{i-1}\delta_l\kappa\,\delta_r\gamma_{i+1} \ldots \gamma_s \\ &= \gamma_1\gamma_2 \ldots \gamma_{i-1}\gamma_i\gamma_{i+1} \ldots \gamma_s, \end{aligned}$$

   using production $B \to \kappa(\emptyset; \mathcal{F})$.
   According to the hypothesis,

$$\begin{aligned} \beta &= A'_1A'_2 \ldots A'_s \\ &\Longrightarrow^* \gamma'_1\gamma'_2 \ldots \gamma'_{i-1}\delta'_lB'\delta'_r\gamma'_{i+1} \ldots \gamma'_s, \end{aligned}$$

   with $A'_j \Longrightarrow_c^* \gamma'_j$ for $j \in [s]$, where
   - for $j \neq i$, $\gamma'_j = \gamma_j$ if $A'_j = A_j$, and $\gamma'_j = \epsilon$ if $A'_j = \epsilon$, and
   - for $j = i$, $\delta'_lB'\delta'_r = \delta_lB\delta_r$ if $A'_i = A_i$, and $\delta'_lB'\delta'_r = \epsilon$ if $A'_i = \epsilon$.

   Since $l(\gamma'_1\gamma'_2 \ldots \gamma'_{i-1}\delta'_l\delta'_r\gamma'_{i+1} \ldots \gamma'_s) \subseteq l(\gamma_1\gamma_2 \ldots \gamma_{i-1}\delta_l\delta_r\gamma_{i+1} \ldots \gamma_s)$, the production $B \to \kappa(\emptyset; \mathcal{F})$ is enabled. Therefore $\beta \Longrightarrow^* \gamma'_1\gamma'_2 \ldots \gamma'_s$. □

As already noted earlier, in the case of random context, only the presence or absence of the context variables is important, and not the order in which the variables appear. Therefore we have the following:

Corollary 10. *Let $G = (V_N, V_T, P, S)$ be an rFcg. Suppose $S \Longrightarrow^* \alpha = A_1A_2 \dots A_s$, where $A_j \in V_N$ for $j \in [s]$. Suppose $S \Longrightarrow^* \beta = B_1B_2 \dots B_s$, where $B_j \in V_N \cup \{\epsilon\}$ for $j \in [s]$. Let $\{B_1, B_2, \dots, B_s\} \subseteq \{A_1, A_2, \dots, A_s\}$.*

Let $\alpha \Longrightarrow^* \gamma_1\gamma_2 \dots \gamma_s$, with $A_j \Longrightarrow_c^* \gamma_j$, $j \in [s]$. Then a derivation for β is $\beta \Longrightarrow^* \gamma_{i_1}\gamma_{i_2} \dots \gamma_{i_s}$, with $B_j \Longrightarrow_c^* \gamma_{i_j}$, $j \in [s]$, where $\gamma_{i_j} = \gamma_j$ if $B_j = A_j$, and $\gamma_{i_j} = \epsilon$ if $B_j = \epsilon$.

We now present the main result of this section, a necessary condition for random forbidding context languages. In essence we prove that if a word is sufficiently long, then any derivation contains two sentential forms α and β such that α derives β, but they have the same labels. Starting from α, we can copy the derivation sequence that led from α to β, since in the case of a rFcg, the lack of context in α cannot inhibit the application of productions.

Theorem 11. *Let $\mathcal{L}$ be an rFcl. Then there exists an n such that any word $z \in \mathcal{L}$ with $|z| \geq n$ has a proper factor v with $|v| \geq |z|/n$ that is also a factor of a word $y \in \mathcal{L}$ with $|y| < |z|$.*

Proof. Let $\mathcal{L}$ be generated by an rFcg $G = (V_N, V_T, P, S)$ in normal form. Let $p = |V_N|$. Let n be the integer of Lemma 1 that depends on $p + 1$.

Let $z \in \mathcal{L}$ with $|z| \geq n$. Due to Lemma 8, there exists a derivation of z in the form

$$S = \alpha_1 \Longrightarrow^* \alpha_2 \Longrightarrow^* \cdots \Longrightarrow^* \alpha_{|z|} \Longrightarrow^* z \ ,$$

where $|\alpha_i| = i$ for $1 \leq i \leq |z|$, and $\alpha_i \in V_N^*$.

According to Lemma 1, there are $p + 1$ indices $i_1, i_2, \dots, i_{p+1}$, with $1 \leq i_1 < i_2 < \cdots < i_{p+1} \leq n$, such that $\alpha_{i_1} \leq \alpha_{i_2} \leq \cdots \leq \alpha_{i_{p+1}}$. By construction, $\alpha_{i_1} < \alpha_{i_2} < \cdots < \alpha_{i_{p+1}}$.

For every $j \in [p + 1]$, $|\alpha_{i_j}| \leq n$. Therefore each α_{i_j} has maximally n variables and consequently at least one variable in α_{i_j} generates a string of length at least $|z|/n$.

Consider $\alpha_{i_1} < \alpha_{i_2} < \cdots < \alpha_{i_{p+1}}$. According to Lemma 2 there exist r and s, $i_1 \leq r < s \leq i_{p+1}$, such that $\zeta(\vec{\alpha_r}) = \zeta(\vec{\alpha_s})$. Let B be a variable in α_s that derives a factor, say v, of length at least $|z|/n$. Starting from α_r and using Corollary 10, we can ensure that a copy of B in α_r derives v. Let y be the word derived in this way. Then y contains the factor v. Moreover, since $|\alpha_r| < |\alpha_s|$, v is a proper factor of z. □

With Theorem 11, it can easily be shown that many languages are not rFcls.

Example 12. The language $\mathcal{L} = \left\{ \left(ga^k\right)^l \mid 0 \leq k \leq l \right\}$ cannot be generated by any rFcg.

Proof. Suppose $\mathcal{L}$ is generated by an rFcg. Let n be the integer of Theorem 11.

Now consider $z = \left(ga^{2n}\right)^{2n}$. Then $z \in \mathcal{L}$. Moreover, $|z| = 2n(2n + 1) \geq n$. According to Theorem 11, z has a proper factor v with $|v| \geq |z|/n$ that is also a factor of a word $y \in \mathcal{L}$ with $|y| < |z|$.

Consider a string v with $|v| \geq |z|/n = 2(2n + 1)$. Then v contains the factor $ga^{2n}g$. Then there is a word $y \in \mathcal{L}$ with $|y| < |z|$ such that y contains the factor $ga^{2n}g$. This contradicts the definition of $\mathcal{L}$. Therefore $\mathcal{L}$ cannot be generated by any rFcg. □

The same proof can be used to show that the language $\mathcal{L} = \left\{ (ga^m)^m \mid m > 0 \right\}$ cannot be generated by any rFcg.

It was shown in Lemma 4 on Page 153 of [6] that the language in Example 13 is not an rFcl. However, that proof is more complicated than the following one, which uses Theorem 11.

Example 13. Consider the language $\mathcal{L} = \{z_1, z_2, \dots\}$, where $z_1 = [a]$, $z_2 = \left(\left[a^2\right]\right)^{4|z_1|}$, in general $z_i = \left(\left[a^i\right]\right)^{4|z_{i-1}|}$ for $i > 2$ and a, [and] are terminals. $\mathcal{L}$ cannot be generated by any rFcg.

Proof. Suppose $\mathcal{L}$ is generated by an rFcg. Let n be the integer of Theorem 11.

Now consider $z_n = ([a_n])^{4|z_{n-1}|}$. Then $z_n \in \mathcal{L}$. Moreover, $|z_n| = 4|z_{n-1}|(n + 2) \geq n$. According to Theorem 11, z has a proper factor v with $|v| \geq |z_n|/n$ that is a factor of a word $y \in \mathcal{L}$ with $|y| < |z_n|$.

Consider a string v with

$$\begin{aligned} |v| &\geq \frac{|z_n|}{n} \\ &= \frac{4|z_{n-1}|(n+2)}{n} \\ &= \frac{4n|z_{n-1}|}{n} + \frac{4 \times 2|z_{n-1}|}{n} \\ &= 4|z_{n-1}| + \frac{8|z_{n-1}|}{n} \\ &> 4|z_{n-1}|. \end{aligned}$$

Then $4|z_{n-1}| < |y| < |z_n|$. This contradicts the definition of $\mathcal{L}$. Therefore $\mathcal{L}$ cannot be generated by any rFcg. □

6. A necessary condition for cfls

In this section we concentrate on context-free languages, which are strictly contained in both the random permitting and the random forbidding context languages. We prove a necessary condition for a language to be generated by a context-free grammar, and use it to show that a particular language is not context-free.

Theorem 14. *Let $\mathcal{L}$ be a cfl. Then there exists an n such that any word $z \in \mathcal{L}$ with $|z| \geq n$ has a factor v with $|z|/n \leq |v| < |z|$ such that*

1. *v is a factor of a word $z_2 \in \mathcal{L}$ with $|z_2| > |z|$, and*
2. *v is a factor of a word $z_0 \in \mathcal{L}$ with $|z_0| < |z|$.*

Proof. Let $\mathcal{L}$ be generated by a cfg $G = (V_N, V_T, P, S)$ in Chomsky normal form. Let $p = |V_N|$. Let n be the integer of Lemma 1 that depends on $p + 1$.

Let $z \in \mathcal{L}$ with $|z| \geq n$. Consider a derivation of z, i.e., $S \Longrightarrow^* z$. Due to the normal form, this derivation can be written as

$$S = \alpha_1 \Longrightarrow^* \alpha_2 \Longrightarrow^* \cdots \Longrightarrow^* \alpha_{|z|} \Longrightarrow^* z \ ,$$

where $|\alpha_i| = i$ for $1 \leq i \leq |z|$, and $\alpha_i \in {V_N}^*$.

According to Lemma 1, there are $p + 1$ indices $i_1, i_2, \ldots, i_{p+1}$, with $1 \leq i_1 < i_2 < \cdots < i_{p+1} \leq n$, such that $\alpha_{i_1} \leq \alpha_{i_2} \leq \cdots \leq \alpha_{i_{p+1}}$. By construction, $\alpha_{i_1} < \alpha_{i_2} < \cdots < \alpha_{i_{p+1}}$.

For every $j \in [p + 1]$, $|\alpha_{i_j}| \leq n$. Therefore each α_{i_j} has maximally n variables and consequently at least one variable in α_{i_j} generates a string of length at least $|z|/n$.

Consider $\alpha_{i_1} < \alpha_{i_2} < \cdots < \alpha_{i_{p+1}}$. According to Lemma 2 there exist r and s, $i_1 \leq r < s \leq i_{p+1}$, such that $\zeta\ (\vec{\alpha_r}) = \zeta\ (\vec{\alpha_s})$.

Let B be a variable in α_s that derives a factor, say v, of length at least $|z|/n$. Since $|\alpha_s| \geq 2$, v is a factor of z.

Starting from α_s and using Corollary 4, we can ensure that a copy of B in α_s derives v. Let z_2 be the word derived in this way. Then z_2 contains the factor v. Since $|\alpha_r| < |\alpha_s|$, $|z_2| > |z|$.

Starting from α_r and using Corollary 10, we can ensure that a copy of B in α_r derives v. Let z_0 be the word derived in this way. Then z_0 contains the factor v. Since $|\alpha_r| < |\alpha_s|$, $|z_0| < |z|$. □

Consider the language $\mathcal{L} = \left\{(ga^m)^m \mid m > 0\right\} \cup \left\{a^i g^j \mid i, j \geq 1\right\}$. Due to the second term in its definition, it is not easy to prove with the pumping lemma for cfls [4] that it is not context-free. However, by using Condition 1 of Theorem 14 in the manner of the proof of Example 6, we can show that $\mathcal{L}$ is not context-free.

7. Future work

To the first author's knowledge, it is not known whether there exists an rPcl that cannot be generated by any extended table-driven context-free Lindenmayer (ET0L) system or any rFcg. Perhaps the necessary conditions for rPcls and rFcls presented above can aid in finding answers to these questions.

Moreover, since ET0L systems are strictly weaker than rFcgs [2], it may be possible to prove a stronger result than the above necessary condition for rFcls for the special case of ET0L languages.

Acknowledgements

We would like to thank the referees for their constructive comments.

References

[1] B. Atcheson, S. Ewert, D. Shell, A note on the generative capacity of random context, South African Computer Journal 36 (2006) 95–98.
[2] J. Dassow, G. Păun, Regulated Rewriting in Formal Language Theory, in: EATCS Monographs on Theoretical Computer Science, vol. 18, Springer-Verlag, 1989.
[3] S. Ewert, A. van der Walt, A pumping lemma for random permitting context languages, Theoretical Computer Science 270 (2002) 959–967.
[4] J.E. Hopcroft, J.D. Ullman, Introduction to Automata Theory, Languages, and Computation, in: Addison-Wesley Series in Computer Science, Addison-Wesley, Reading, Massachusetts, 1979.
[5] M. Rabkin, Ogden's lemma for random permitting- and forbidding-context and ET0L languages, Master of Science Dissertation (submitted May 2012), School of Computer Science, University of the Witwatersrand, Johannesburg, South Africa, 2012.
[6] A. van der Walt, S. Ewert, A shrinking lemma for random forbidding context languages, Theoretical Computer Science 237 (1–2) (2000) 149–158.
[7] A.P.J. van der Walt, Random context languages, Information Processing 71 (1972) 66–68.