NORMAL FORMS OF ONE-SIDED RANDOM CONTEXT GRAMMARS

Petr Zemek
Doctoral Degree Programme (2), FIT BUT
E-mail: xzemek02@stud.fit.vutbr.cz

Supervised by: Alexander Meduna
E-mail: meduna@fit.vutbr.cz

Abstract: The present paper discusses normal forms of one-sided random context grammars. More specifically, it (1) gives an overview of previously established normal forms and (2) establishes three new normal forms. All normal forms are established in terms of one-sided random context grammars with and without erasing rules. A discussion of an open problem concludes the paper.

Keywords: formal languages, one-sided random context grammars, normal forms

1 INTRODUCTION

A formal grammar is in a *normal form* (see [4]) if all its rules satisfy some prescribed uniform form. If the grammar posses more sets of rules, like in grammar systems (see [2]), conditions may also be placed upon the identity or disjointness of these sets. Indisputably, normal forms represent a crucially important part of formal language theory as a whole. Indeed, from a theoretical viewpoint, they are useful in simplifying proofs (see [3, 4]). From a practical viewpoint, they fulfill a useful role in parsing (see [1, 5]).

In the present paper, we discuss this important topic in terms of one-sided random context grammars. Recall that a *one-sided random context grammar* (see [6]) represents a variant of a random context grammar (see [3]). In this variant, a set of *permitting symbols* and a set of *forbidding symbols* are attached to every rule, and its set of rules is divided into the set of *left random context rules* and the set of *right random context rules*. A left random context rule can rewrite a nonterminal if each of its permitting symbols occurs to the left of the rewritten symbol in the current sentential form while each of its forbidding symbols is absent therein. A right random context rule is applied analogically except that the symbols are examined to the right of the rewritten symbol. Recall that one-sided random context grammars are computationally complete (see Theorem 3 in [6]), and if erasing rules are not permitted, they characterize the family of context-sensitive languages (see Theorem 2 in [6]).

The goal of this paper is twofold. It (1) gives an overview of previously established normal forms of one-sided random context grammars and (2) establishes three new normal forms. The first new normal form represents an analogy of the Chomsky normal form for context-free grammars (see [4]). The second new normal form requires that each rule has its permitting or forbidding context empty—that is, no rule can both permit and forbid symbols. In the third new normal form, the sets of left and right random context rules are disjoint. All normal forms are established in terms of one-sided random context grammars with and without erasing rules.

The paper is organized as follows. Section 2 gives all the necessary terminology. Then, Section 3 establishes the abovementioned normal forms of one-sided random context grammars. In the conclusion of this paper, Section 4 states an open problem.

2 PRELIMINARIES AND DEFINITIONS

We assume that the reader is familiar with formal language theory (see [3, 4]). For a set Q, $\mathrm{card}(Q)$ denotes the cardinality of Q, and 2^Q denotes the power set of Q. 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 $x \in V^*$, $|x|$ denotes the length of x, and $\mathrm{alph}(x)$ denotes the set of symbols occurring in x.

Definition 1 (see [6])**.** A *one-sided random context grammar* is a quintuple

$$G = \big(N, T, P_L, P_R, S\big)$$

where N and T are two disjoint alphabets, $S \in N$, and

$$P_L, P_R \subseteq N \times \big(N \cup T\big)^* \times 2^N \times 2^N$$

are two finite relations. Set $V = N \cup T$. The components V, N, T, P_L, P_R, and S are called the *total alphabet*, the alphabet of *nonterminals*, the alphabet of *terminals*, the set of *left random context rules*, the set of *right random context rules*, and the *start symbol*, respectively. Each $(A, x, U, W) \in P_L \cup P_R$ is written as $\lfloor A \rightarrow x,\, U,\, W\rfloor$ throughout this paper. For $\lfloor A \rightarrow x,\, U,\, W\rfloor \in P_L$, U and W are called the *left permitting context* and the *left forbidding context*, respectively. For $\lfloor A \rightarrow x,\, U,\, W\rfloor \in P_R$, U and W are called the *right permitting context* and the *right forbidding context*, respectively. If $\lfloor A \rightarrow x,\, U,\, W\rfloor \in P_L \cup P_R$ implies that $|x| \geq 1$, then G is said to be *propagating*. The *direct derivation relation* over V^* is denoted by $\Rightarrow$ and defined as follows. Let $u, v \in V^*$ and $\lfloor A \rightarrow x,\, U,\, W\rfloor \in P_L \cup P_R$. Then, $uAv \Rightarrow uxv$ in G if and only if

$$\lfloor A \rightarrow x, U, W\rfloor \in P_L, U \subseteq \mathrm{alph}(u), \text{ and } W \cap \mathrm{alph}(u) = \emptyset$$

or

$$\lfloor A \rightarrow x, U, W\rfloor \in P_R, U \subseteq \mathrm{alph}(v), \text{ and } W \cap \mathrm{alph}(v) = \emptyset$$

Let $\Rightarrow^n$ and $\Rightarrow^*$ denote the nth power of $\Rightarrow$, for some $n \geq 0$, and the reflexive-transitive closure of $\Rightarrow$, respectively. The *language of* G is denoted by $L(G)$ and defined as

$$L\big(G\big) = \big\{w \in T^* \mid S \Rightarrow^* w\big\} \qquad \square$$

Next, we illustrate the previous definition by an example.

Example 1. Consider the one-sided random context grammar

$$G = \big(\{S, A, B, \bar{A}, \bar{B}\}, \{a, b, c\}, P_L, P_R, S\big)$$

where P_L contains the following four rules:

$$\lfloor S \rightarrow AB, \emptyset, \emptyset\rfloor \qquad \lfloor \bar{B} \rightarrow B, \{A\}, \emptyset\rfloor$$
$$\lfloor B \rightarrow b\bar{B}c, \{\bar{A}\}, \emptyset\rfloor \qquad \lfloor B \rightarrow \varepsilon, \emptyset, \{A, \bar{A}\}\rfloor$$

and P_R contains the following three rules:

$$\lfloor A \rightarrow a\bar{A}, \{B\}, \emptyset\rfloor \qquad \lfloor \bar{A} \rightarrow A, \{\bar{B}\}, \emptyset\rfloor \qquad \lfloor A \rightarrow \varepsilon, \{B\}, \emptyset\rfloor$$

It is rather easy to see that every derivation that generates a non-empty string of $L(G)$ is of the form

$$\begin{aligned} S &\Rightarrow AB \\ &\Rightarrow a\bar{A}B \\ &\Rightarrow a\bar{A}b\bar{B}c \\ &\Rightarrow aAb\bar{B}c \\ &\Rightarrow aAbBc \\ &\Rightarrow^* a^nAb^nBc^n \\ &\Rightarrow a^nb^nBc^n \\ &\Rightarrow a^nb^nc^n \end{aligned}$$

where $n \geq 1$. The empty string is generated by $S \Rightarrow AB \Rightarrow B \Rightarrow \varepsilon$. Based on the previous observations, we see that G generates the non-context-free language $\{a^nb^nc^n \mid n \geq 0\}$. □

3 NORMAL FORMS OF ONE-SIDED RANDOM CONTEXT GRAMMARS

In this section, we give an overview of previously established normal forms of one-sided random context grammars, and establish the three new normal forms already mentioned in Section 1.

3.1 PREVIOUSLY ESTABLISHED NORMAL FORMS

To our knowledge, only a single normal form of one-sided random context grammars has been established. In this normal form, the set of left random context rules coincides with the set of right random context rules.

Theorem 1 (see [6]). *Let $G = (N, T, P_L, P_R, S)$ be a one-sided random context grammar. Then, there is a one-sided random context grammar, $H = (N', T, P'_L, P'_R, S)$, such that $L(H) = L(G)$ and $P'_L = P'_R$. Furthermore, if G is propagating, then so is H.* □

3.2 NEW NORMAL FORMS

Now, we establish three new normal forms of one-sided random context grammars. The first new normal form represents an analogy of the Chomsky normal form for context-free grammars (see [4]).

Theorem 2. *Let $G = (N, T, P_L, P_R, S)$ be a one-sided random context grammar. Then, there is a one-sided random context grammar, $H = (N', T, P'_L, P'_R, S)$, such that $L(H) = L(G)$ and $\lfloor A \rightarrow x, U, W \rfloor \in P'_L \cup P'_R$ implies that $x \in N'N' \cup T \cup \{\varepsilon\}$. Furthermore, if G is propagating, then so is H.*

Proof. This theorem can be established by analogy with the well-known conversion of context-free grammars into the Chomsky normal form (see Algorithm 5.1.4.1.1 in [4]). Due to space requirements, a rigorous proof of this simple theorem is left to the reader. □

In the second new normal form, each rule has its permitting or forbidding context empty.

Theorem 3. *Let $G = (N, T, P_L, P_R, S)$ be a one-sided random context grammar. Then, there is a one-sided random context grammar, $H = (N', T, P'_L, P'_R, S)$, such that $L(H) = L(G)$ and $\lfloor A \rightarrow x, U, W \rfloor \in P'_L \cup P'_R$ implies that $U = \emptyset$ or $W = \emptyset$. Furthermore, if G is propagating, then so is H.*

Proof. Let $G = (N, T, P_L, P_R, S)$ be a one-sided random context grammar. Set $V = N \cup T$ and

$$F = \{\langle r, i, d\rangle \mid r = \lfloor A \rightarrow x, U, W \rfloor \in P_d, d \in \{L, R\}, i \in \{1, 2\}\}$$

Without any loss of generality, we assume that $F \cap V = \emptyset$. Construct $H = (N', T, P'_L, P'_R, S)$ as follows. Initially, set $N' = N \cup F$, $P'_L = \emptyset$, and $P'_R = \emptyset$. Perform (1) and (2), given next.

(1) for each $r = \lfloor A \to x, U, W \rfloor \in P_L$,

(1.1) add $\lfloor A \to \langle r, L, 1 \rangle, \emptyset, F \rfloor$ to P'_R;
(1.2) add $\lfloor \langle r, L, 1 \rangle \to \langle r, L, 2 \rangle, \emptyset, W \cup F \rfloor$ to P'_L;
(1.3) add $\lfloor \langle r, L, 2 \rangle \to x, U, \emptyset \rfloor$ to P'_L;

(2) for each $r = \lfloor A \to x, U, W \rfloor \in P_R$,

(2.1) add $\lfloor A \to \langle r, R, 1 \rangle, \emptyset, F \rfloor$ to P'_L;
(2.2) add $\lfloor \langle r, R, 1 \rangle \to \langle r, R, 2 \rangle, \emptyset, W \cup F \rfloor$ to P'_R;
(2.3) add $\lfloor \langle r, R, 2 \rangle \to x, U, \emptyset \rfloor$ to P'_R.

A single rule from P_L and P_R is simulated in three steps by rules introduced in (1) and (2), respectively. As we cannot check both the presence and absence of symbols in a single step, we split this check into two consecutive steps. Clearly, $L(G) \subseteq L(H)$, so we only prove that $L(H) \subseteq L(G)$.

Observe that if we consecutively apply the three rules from (1) in H, then we can apply the original rule in G. Likewise for the rules introduced in (2). Therefore, it remains to be shown that H cannot generate false sentences by invalid intermixed simulations of more than one rule of G at a time. In what follows, we consider only simulations of rules from P_L; the situation for rules from P_R is analogical.

Let us consider a simulation of some $r = \lfloor A \to x, U, W \rfloor \in P_L$. Observe that the only situation where a false simulation may occur is that after a rule from (1.2) is applied, another simulation takes places which transforms a nonterminal to the left of $\langle r, L, 2 \rangle$ that is not in U into a nonterminal that is in U. To investigate this possibility, set $V' = N' \cup T$ and consider any successful derivation in H, $S \Rightarrow^* z$, where $z \in L(H)$. This derivation can be written in the form

$$S \Rightarrow^* w \Rightarrow y \Rightarrow^* z$$

where $w = w_1 \langle r, L, 1 \rangle w_2$, $y = w_1 \langle r, L, 2 \rangle w_2$, and $w_1, w_2 \in V'^*$. Since $w \Rightarrow y$ in H by $\lfloor \langle r, L, 1 \rangle \to \langle r, L, 2 \rangle, \emptyset, W \cup F \rfloor$, introduced to P'_L in (1.2) from r,

$$\mathrm{alph}(w_1) \cap (W \cup F) = \emptyset$$

From the presence of $\langle r, L, 2 \rangle$, no rule from (1) is now applicable to w_1. Let $w_1 = w'_1 B w''_1$ and $\lfloor B \to \langle s, R, 1 \rangle, \emptyset, F \rfloor \in P'_L$, introduced in (2.1) from some $s = \lfloor B \to v, X, Y \rfloor \in P_R$ such that $B \notin U$ and

$$\mathrm{alph}(v) \cap \big(U - \mathrm{alph}(w_1)\big) \neq \emptyset$$

(This last requirement implies that by successfully simulating s prior to r, we end up with an invalid simulation of r.) Then,

$$w'_1 B w''_1 \langle r, L, 2 \rangle w_2 \Rightarrow w'_1 \langle s, R, 1 \rangle w''_1 \langle r, L, 2 \rangle w_2$$

in H. Since $\langle s, R, 1 \rangle$ cannot be rewritten to $\langle s, R, 2 \rangle$ by a rule from (2.2) ($\langle r, L, 2 \rangle$ is present to the right of $\langle s, R, 1 \rangle$), we can either (a) correctly finish the simulation of r by rewriting $\langle r, L, 2 \rangle$ to x (recall that $B \notin U$) or (b) rewrite some nonterminal in w'_1 or w''_1. However, observe that in (b), we end up in the same situation as we are now.

Based on these observations, we see that no invalid intermixed simulations of more than one rule of G at a time are possible in H. Hence, $L(H) \subseteq L(G)$, so $L(H) = L(G)$. Clearly, $\lfloor A \to x, U, W \rfloor \in P'_L \cup P'_R$ implies that $U = \emptyset$ or $W = \emptyset$. Furthermore, observe that if G is propagating, then so is H. Thus, the theorem holds. $\square$

The third new normal form represents a dual normal form to the one in Theorem 1. Indeed, we next show that every one-sided random context grammar can be turned into an equivalent one-sided random context grammar with the sets of left and right random context rules being disjoint.

Theorem 4. *Let $G = (N, T, P_L, P_R, S)$ be a one-sided random context grammar. Then, there is a one-sided random context grammar, $H = (N', T, P'_L, P'_R, S)$, such that $L(H) = L(G)$ and $P'_L \cap P'_R = \emptyset$. Furthermore, if G is propagating, then so is H.*

Proof. Let $G = (N, T, P_L, P_R, S)$ be a one-sided random context grammar. Construct $H = (N', T, P'_L, P'_R, S)$, where

$$N' = N \cup \{L, R\};$$

$$P'_L = \{\lfloor A \to x, U, W \cup \{L\} \rfloor \mid \lfloor A \to x, U, W \rfloor \in P_L\};$$

$$P'_R = \{\lfloor A \to x, U, W \cup \{R\} \rfloor \mid \lfloor A \to x, U, W \rfloor \in P_R\}.$$

(Without any loss of generality, we assume that $\{L, R\} \cap (N \cup T) = \emptyset$.) Observe that the new nonterminals L and R cannot appear in any sentential form. Therefore, it is easy to see that $L(H) = L(G)$. Furthermore, observe that if G is propagating, then so is H. Since $P'_L \cap P'_R = \emptyset$, the theorem holds. □

4 CONCLUDING REMARKS

We conclude the paper by suggesting several open problem areas. Let $G = (N, T, P_L, P_R, S)$ be a one-sided random context grammar, and consider the following three normal forms:

I. either $P_L = \emptyset$ or $P_R = \emptyset$;

II. $\lfloor A \to x, U, W \rfloor \in P_L \cup P_R$ implies that $\text{card}(U) + \text{card}(W) \leq 1$;

III. $P_R = \emptyset$ and $\lfloor A \to x, U, W \rfloor \in P_L$ implies that $W = \emptyset$.

Can we turn G into an equivalent grammar in any of the abovementioned forms?

ACKNOWLEDGEMENTS

This work was supported by the following grants: BUT FIT FIT-S-11-2, EU CZ 1.05/1.1.00/02.0070, MŠMT FR271/2012/G1, and CEZ MŠMT MSM0021630528.

REFERENCES

[1] A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley, 2nd edition, 2006.

[2] E. Csuhaj-Varjú, J. Dassow, J. Kelemen, and G. Păun. *Grammar Systems: A Grammatical Approach to Distribution and Cooperation*. Gordon and Breach, 1994.

[3] J. Dassow and G. Păun. *Regulated Rewriting in Formal Language Theory*. Springer, 1989.

[4] A. Meduna. *Automata and Languages: Theory and Applications*. Springer, 2000.

[5] A. Meduna. *Elements of Compiler Design*. Auerbach Publications, 2007.

[6] A. Meduna and P. Zemek. One-sided random context grammars. *Acta Informatica*, 48(3):149–163, 2011.