HOW TO GENERATE RECURSIVELY ENUMERABLE LANGUAGES USING ONLY CONTEXT-FREE PRODUCTIONS AND EIGHT NONTERMINALS

Ing. Radek BIDLO, Doctoral Degree Programme (1)
Dept. of Information Systems, FIT, BUT
E-mail: bidlor@fit.vutbr.cz

Ing. Petr BLATNÝ, Doctoral Degree Programme (1)
Dept. of Information Systems, FIT, BUT
E-mail: blatny@fit.vutbr.cz

Supervised by: Dr. Alexander Meduna

ABSTRACT

The notion of a context-free grammar over a free group is introduced. The transformation of any type-0 grammar to an equivalent context-free grammar over a free group is demonstrated. This approach causes an undesirable increase of the number of nonterminal symbols. Hence we introduce a method for their reduction.

1 INTRODUCTION

The notion of a direct derivation in context free grammars has always been defined on letter monoids generated by total alphabets of these grammars (see [2]). In our paper, however, this notion is introduced on free letter groups generated by total alphabets of these grammars.

The resulting grammars are simple and natural modification of the classic definition of context-free grammars. Moreover, their generative capacity is remarkably increased.

By this approach, the number of nonterminal symbols is increased. Hence we will deal with their reduction using a suitable encoding.

2 PRELIMINARIES

We assume that the reader is familiar with formal language theory (see [1]).

For an alphabet, V, $\overline{V}$ denotes the letter (free) group generated by V under the operation of concatenation, ε is the unit of $\overline{V}$ and for every $Q \in V$, there is $\overline{Q} \in V$ called the *inverse* to Q with the property $Q\overline{Q} = \overline{Q}Q = \varepsilon$. For a word, $w = a_1a_2 \dots a_n \in \overline{V}$, $|w|$ denotes the length of w and $\overline{w} = \overline{a_n a_{n-1}} \dots \overline{a_1}$ denotes the inverse word to w with the property $w\overline{w} = \overline{w}w = \varepsilon$. Note that the equivalent notation is $inv(w)$. The word w is said to be *reduced*, if it contains no substrings of the form $a\bar{a}$, where $a \in V$. It is possible to prove that for every $w \in \overline{V}$, there exists exactly one reduced word without reference to the order of reductions (see [5]).

A context free grammar is a quadruple, $G = (N,T,P,S)$, where N and T are two disjoint alphabets of nonterminal and terminal symbols respectively, P is a finite set of productions of the form $A \rightarrow x$, where $A \in N$, $x \in (N \cup T)^*$, and $S \in N$ is the axiom (start symbol). The total alphabet of G is defined as $V = N \cup T$.

A phrase-structure grammar, $G = (N,T,P,S)$, is specified in Penttonen normal form, where N, T and S have the same meaning as for a context free grammar and every production in P is either of the form $AB \rightarrow AC$ or $A \rightarrow x$ or $A \rightarrow \varepsilon$, where $A,B,C \in N$, $x \in (T \cup N^2)$, see theorem 4 on page 391 in [3].

The left-hand side and right-hand side of $p \in P$ is denoted by $lhs(p)$ and $rhs(p)$ respectively. The set of all symbols occuring in $w \in (N \cup T)^*$ is denoted by $alph(w)$.

We define the direct derivation $\Rightarrow_G$ over V^* as follows. For every $x,y \in V^*$, $x \Rightarrow_G y$ if $x = uAv$, $y = uwv$, $u,v \in V^*$ and $A \rightarrow w \in P$. In the standard manner, we can introduce the relations $\Rightarrow_G^i$, $\Rightarrow_G^+$ and $\Rightarrow_G^*$, where the subscript G is usually omited when understood. If we want to express that $x \Rightarrow y$ using $p \in P$ in G, where $x,y \in (N \cup T)^*$, we write $x \Rightarrow_G y[p]$.

We now introduce a new structure for generating languages. Let $G_\Gamma = (N,T,P,S)$ be a context-free grammar, $V = N \cup T$. Then $\Gamma = (G_\Gamma, \overline{V})$ is the context-free grammar over a free group. It is clear that for every $Q \in V$, there must be $\overline{Q} \in V$ such that $Q\overline{Q} = \overline{Q}Q = \varepsilon$. In G_Γ, the *direct derivation over* $\overline{V}$ is defined as follows. For every $x,y \in \overline{V}$, $x \Rightarrow_\Gamma y$ if $x = uAv$, $y = uwv$, $u,v \in \overline{V}$ and $A \rightarrow w \in P$. The relations $\Rightarrow_\Gamma^i$, $\Rightarrow_\Gamma^+$ and $\Rightarrow_\Gamma^*$ have the usual meaning and are defined on $\overline{V}$.

We denote by **FG**(2) the family of languages generated by context-free (type-2 by the Chomsky hierarchy) grammars over free groups. By **FG**(2)**R**, we denote the family of languages generated by context-free grammars over free groups with the reduced number of nonterminals. The family of recursively enumerable languages is denoted by **RE**.

3 RESULTS

In this section, we examine the generative capacity of context-free grammars over free groups. We will outline the prove that for every phrase-structure grammar, G, there exists an equivalent context free grammar over a free group, Γ, such that $L(G) = L(\Gamma)$. Next, we introduce the reduction of nonterminal symbols. Note that the rigorous proofs exceed the page limit, so we only describe the constructions.

Lemma 3.1 For every phrase-structure grammar, $H = (N,T,P_H,S_H)$, there exists an equivalent phrase-structure grammar, $G = (N \cup \{X,Y\},T,P_G,S_G)$, such that $\{X,Y,S_G\} \cap \{N \cup T\} = \emptyset$ and the set P_G contains only productions of the forms

- $AB \rightarrow CD$
- $A \rightarrow BC$

- $A \rightarrow a$
- $XY \rightarrow X$
- $X \rightarrow \varepsilon$

where $A, B, C, D \in N$, $a \in T$.

Proof 3.1

Construction

Consider the grammar $H = (N', T, P', S_H)$. Without any loss of generality, assume that H satisfies the Penttonen normal form and assume that $\{S_G, X, Y\} \cap (N' \cup T) = \emptyset$. Define the grammar $G = (N, T, P, S_G)$, where $N = N' \cup \{S_G, X, Y\}$ and P is constructed as follows:

(1) add every $p \in P'$ with two nonterminals on its right-hand side to P

(2) for every $A \rightarrow \varepsilon \in P'$, add $A \rightarrow Y$ to P

(3) for every $A \in N'$, add $AY \rightarrow YA$ to P

(4) add $S_G \rightarrow XS_H$, $XY \rightarrow X$, and $X \rightarrow \varepsilon$ to P

The construction of G is completed. Now, we establish the following theorem.

Theorem 3.1 **FG**(2)=**RE**

Proof 3.2 Consider that $G = (N, T, P, S)$ is a phrase-structure grammar. Without any loss of generality, assume that G satisfies the properties described in Lemma 3.1.

Construction

We construct the context-free grammar over a free group, $G_\Gamma = (N_\Gamma, T_\Gamma, P_\Gamma, S)$, where we define $N_\Gamma = N \cup N_{CS} \cup \overline{N}$ and $T_\Gamma = T \cup \overline{T}$ such that

$$\begin{aligned}
\overline{N} &= \{\overline{A} | A \in N\} \\
N_{CS} &= \{\langle ABCD\rangle, \langle\overline{ABCD}\rangle | AB \rightarrow CD \in P\} \cup \{\langle XYX\rangle, \langle\overline{XYX}\rangle\} \\
\overline{T} &= \{\overline{a} | a \in T\}
\end{aligned}$$

Note that in general, Q is any symbol and $\overline{Q}$ is its coresponding inverse symbol.

The new set of productions P_Γ is constructed as follows:

I add every $A \rightarrow BC \in P$ and every $A \rightarrow a \in P$ to P_Γ, where $A, B, C \in N$, $a \in T$

II for every $AB \rightarrow CD \in P$, add $A \rightarrow C\langle ABCD\rangle$ and $B \rightarrow \langle\overline{ABCD}\rangle D$ to P_Γ

III for every $XY \rightarrow X \in P$, add $X \rightarrow X\langle XYX\rangle$ and $Y \rightarrow \langle\overline{XYX}\rangle$ to P_Γ

IV for every $X \to \varepsilon \in P$, add $X \to X\overline{X}$ to P_Γ

The construction of G_Γ is completed.

Observe that for every non-context-free production, there are two new nonterminals in the resulting grammar. Hence we will show that for every phrase-structure grammar, there exists an equivalent context-free grammar over a free group with only eight nonterminals, namely 0, $\overline{0}$, 1, $\overline{1}$, 2, $\overline{2}$, S and $\overline{S}$. Note that the existence of $\overline{S}$ follows from the definition of the free group. In fact, there is no production containing $\overline{S}$, so no sentential form contains it too.

Lemma 3.2 For every phrase-structure grammar, $H = (N, T, P, S)$, there exists an equivalent phrase-structure grammar, $G = (N_G, T, P_G, S)$, such that each production in P_G has one of these forms:

(1) $AB \to CD$, where $A \neq C$ and $A, B, C, D \in N_G$

(2) $A \to BC$, where $A \neq B$ and $A, B, C \in N_G$

(3) $A \to x$, where $A \in N_G$, $x \in T \cup \{\varepsilon\}$

Proof 3.3 Let $H = (N, T, P, S)$ be a grammar. Without any loss of generality, assume that H satisfies the Kuroda normal form; that is, every production in P has one of these forms:

(1) $AB \to CD$, where $A, B, C, D \in N$

(2) $A \to BC$, where $A, B, C \in N$

(3) $A \to x$, where $A \in N$, $x \in T \cup \{\varepsilon\}$

Define the grammar $G = (N_G, T, P_G, S)$, where P_G is constructed as follows:

(1) for every $AB \to AD \in P$, add $AB \to A'D'$, $A'D' \to AD$ to P_G and A', D' to N_G, where A' and D' are two new nonterminals

(2) for every $A \to AB \in P$, add $A \to A'B'$, $A'B' \to AB$ to P_G and A', B' to N_G, where A' and B' are two new nonterminals

(3) add all other productions from P to P_G

A formal proof that H and G are equivalent is simple and left to the reader.

Theorem 3.2 **FG**(2)**R**=**RE**

Proof 3.4

Consider that $G = (N, T, P, S)$ is a phrase-structure grammar, $N = \{A_1, \ldots, A_n\}$, $A_1 = S$. Without any loss of generality, assume that G satisfies the properties described in Lemma 3.2.

Construction

We construct the **FG**(2)**R** grammar $\Gamma = (N_\Gamma, T, P_\Gamma, S_\Gamma)$, where $N_\Gamma = \{0, \overline{0}, 1, \overline{1}, 2, \overline{2}, S_\Gamma, \overline{S_\Gamma}\}$. Define the injections, $h : N \to \{0,1\}^+$ and $\overline{h} : N \to \{\overline{0}, \overline{1}\}^+$, such that for every $A_i \in N$, $h(A_i) = (i)_2 rev((i)_2)$ and $\overline{h}(A_i) = inv(h(A_i))$, where $(i)_2$ is the binary representation of i on $\lceil \log_2 |N| \rceil$ bits, for $i = 1, 2, \ldots, n$. Note that the inverse symbol to $0, 1, 2$ and $S_\Gamma \in N_\Gamma$ is $\overline{0}, \overline{1}, \overline{2}$ and $\overline{S_\Gamma} \in N_\Gamma$ respectively.

The set of productions P_Γ is constructed as follows:

I add $S_\Gamma \to h(S)2$ to P_Γ

II for every $AB \to CD \in P$, add $2 \to \overline{h}(B)\overline{2}\overline{h}(A)\overline{2}2h(C)2h(D)2$ to P_Γ

III for every $A \to BC \in P$, add $2 \to \overline{h}(A)\overline{2}2h(B)2h(C)2$ to P_Γ

IV for every $A \to x \in P$, add $2 \to \overline{h}(A)x$ to P_Γ

where $A, B, C, D \in N$ and $x \in T \cup \{\varepsilon\}$.

The construction of Γ is completed.

Corollary 3.1 **FG**(2)=**FG**(2)**R**

4 CONCLUSIONS

The paper has presented a new type of common context-free grammar. By adding of the free group to the common context-grammar, we increase its generative capacity to the level of non-restricted grammar. The inverses from the free group allow to remove the non-context-free productions. Moreover, the encoding of nonterminal symbols reduces their number to exactly eight.

REFERENCES

[1] Salomaa, A.: Formal Languages, New York: Academic Press 1973

[2] Meduna, A.: Context Free Derivations on Word Monoids, Acta informatica 27: 781-786, Springer-Verlag, 1990

[3] Penttonen, M.: One-Sided and Two-Sided Context in Fromal Grammars, Information and Control 25, 371-392, 1974

[4] Meduna, A., Kolář, D.: Homogenous Grammars with a Reduced Number of Non-Context-Free Productions, In: Information Processing Letters 81, p. 253-257, Amsterdam, NL, 2002

[5] Milne, J.S.: Group Theory, Course Notes, 2003

[6] Geffert, V.: How to Generate Languages Using Only Two Pairs of Parentheses, J. Inform. Process. Cybernet. (EIK), 27, 305-15, 1991