Rule-Homogeneous CD Grammar Systems

Radim Kocman Zbyněk Křivka Alexander Meduna
Centre of Excellence IT4Innovations
Faculty of Information Technology
Brno University of Technology
Božetěchova 2, Brno 612 66, Czech Republic
{ikocman,krivka,meduna}@fit.vutbr.cz

A homogeneous rule has its left-hand side formed by a string of identical symbols. Consider two-component CD grammar systems that work under the $*$ mode or the t mode. This study presents two transformations that turn arbitrary grammars into equivalent two-component CD grammar systems with a context-free component and a homogeneous component. From one transformation, the homogeneous component results with two rules of the form $11 \to 00$ and $0000 \to 2222$, while the other transformation produces the homogeneous component with two rules of the form $11 \to 00$ and $0000 \to \varepsilon$. Apart from this significant restriction of non-context-free rules, the study describes several other useful properties concerning these systems and the way they work.

1 Introduction

The present study, which assumes a familiarity with formal language theory, concerns grammar systems. It concentrates its attention on two-component CD grammar systems working under the $*$ and t modes. Under the former mode, these systems obviously generate the family of context-free languages. More surprisingly, under the latter mode, they are also as powerful as ordinary context-free grammars, too. These results give rise to an idea of changing one of the context-free components with a simple non-context-free component so this change results into an increase of the generative power. An investigation of this idea represents the principal subject of the present study.

Before sketching this study and its achievement, we recall that a grammatical rule of the form $x \to y$, where x and y are strings, is homogeneous if x is formed by a string of identical symbols. A homogeneous rule $x \to y$ is evenly homogeneous if y is also formed by a string of identical symbols and $|x| = |y|$. In a CD grammar system, a component is homogeneous if all its rules are homogeneous, and it is evenly homogeneous if all its rules are evenly homogeneous. A CD grammar system is rule-homogeneous if all its components are homogeneous. Observe that any CD grammar system with context-free components is rule-homogeneous. As obvious, if a CD grammar system contains only evenly homogeneous rules, its language is a subset of the terminal alphabet in the system.

To give an insight into the present study, take any grammar G in Kuroda normal form—that is, a grammar in which every rule is of the form $AB \to CD$, $A \to BC$, $A \to a$, or $A \to \varepsilon$, where A, B, C, D are nonterminals, a is a terminal, and ε denotes the empty string. We demonstrate two transformations that turn G into a two-component rule-homogeneous CD grammar system with a context-free component H and a homogeneous non-context-free component I. One transformation produces $I = \{11 \to 00,\ 0000 \to 2222\}$, so I is evenly homogeneous. The other transformation produces $I = \{11 \to 00,\ 0000 \to \varepsilon\}$, which is thus homogeneous. The study proves that working under the $*$ and t modes, both systems resulting from these transformations are equivalent to G. Thus, more generally speaking, CD grammar systems of these two forms are computationally complete—that is, they characterize the family of recursively enumerable languages—because so are the Kuroda normal form grammars.

Apart from the computational completeness, we demonstrate two other properties:

(i) Consider the system with $I = \{11 \to 00,\ 0000 \to \varepsilon\}$. The study demonstrates that working under the t mode, during every generation of a sentence, it changes its components no more than once. Furthermore, if the system simulates at least one non-context-free rule, it changes its components precisely once.

(ii) From a general and intuitive viewpoint, taking a closer look at language-generating rewriting systems, we intuitively see that some of them generate the language in a more similar way than others. Formal language theory has formalized this generative phenomenon in terms of close derivation simulations. To give an insight into this formalization, consider grammatical models X and Y. If there is a constant k such that for every derivation of the form

$$x_0 \Rightarrow x_1 \Rightarrow \ldots \Rightarrow x_n$$

in X, where x_0 is its start symbol, there is a derivation of the form

$$x_0 \Rightarrow^{k_1} x_1 \Rightarrow^{k_2} \ldots \Rightarrow^{k_n} x_n$$

in Y, where $k_i \leq k$ for each $1 \leq i \leq n$, we tend to say that Y closely simulates X. In this sense, the study demonstrates that under the $*$ mode, the system with $I = \{11 \to 00,\ 0000 \to 2222\}$ closely simulates G, and so does the system with $I = \{11 \to 00,\ 0000 \to \varepsilon\}$ under this mode.

2 Results

This section gives a brief insight into the results. First, we start with a two-component rule-homogeneous CD grammar system that works in the $*$ mode and has the second component homogeneous.

Theorem 1. *Let $G = (N, T, P, S)$ be a grammar. Then, there exists a two-component rule-homogeneous CD grammar system, $\Gamma = (N', T, H, I, S)$, such that $I = \{11 \to 00,\ 0000 \to \varepsilon\}$ and $L_*(\Gamma) = L(G)$.*

Proof.
Construction. Let $G = (N, T, P, S)$ be a grammar. Without any loss of generality, assume that G satisfies the Kuroda normal form and $(N \cup T) \cap \{0, 1\} = \emptyset$. For some $m \geq 3$, define an injection, g, from $\mathrm{NonContextFree}(P)$ to $(\{01\}^+\{00\}\{01\}^+ \cap \{01, 00\}^m)$.

From G, we construct the two-component rule-homogeneous CD grammar system, $\Gamma = (N', T, H, I, S)$, where $N' = N \cup \{0, 1\}$, $I = \{11 \to 00,\ 0000 \to \varepsilon\}$, and H is defined as follows:

(I) For every $AB \to CD \in P$ where $A, B, C, D \in N$,
add $A \to Cg(AB \to CD)$ and $B \to \mathrm{rev}(g(AB \to CD))D$ to H.

(II) For every $A \to x \in P$ where $A \in N$ and $x \in (\{\varepsilon\} \cup T \cup N^2)$, add $A \to x$ to H.

The construction of Γ is completed.

Basic idea.

(a) The rules of (I) and I simulate the derivation steps made by $\mathrm{NonContextFree}(P)$ in G. That is, $xABy \Rightarrow xCDy$ according to $AB \to CD \in P$, where $x, y \in (N \cup T)^*$, in G is simulated in Γ as

$$\begin{aligned} xABy &\Rightarrow_H xCg(AB \to CD)By \\ &\Rightarrow_H xCg(AB \to CD)\,\mathrm{rev}(g(AB \to CD))Dy \\ &\Rightarrow_I^{2m-1} xCDy. \end{aligned}$$

Γ makes the $(2m-1)$-step derivation $xCg(AB \to CD)\,\mathrm{rev}(g(AB \to CD))Dy \Rightarrow_I^{2m-1} xCDy$ by using only rules from $I = \{11 \to 00,\ 0000 \to \varepsilon\}$. During this $(2m-1)$-step derivation, the string between C and D always contains exactly one occurrence of consecutive identical symbols that can be rewritten, so Γ actually verifies that the simulation of $xABy \Rightarrow xCDy$ is made properly.

(b) The rules of (II) simulate the use of $\mathrm{ContextFree}(P)$ in G.

The reader may notice that the simulation of non-context-free rules resembles similar techniques used in phrase-structure grammars. However, this is traditionally done by using several types of matching parentheses, which is not a suitable form for homogeneous rules, or it requires a significant non-local change in the generation flow of the grammar.

Corollary 2. *The resulting two-component rule-homogeneous CD grammar system Γ from the proof of Theorem 1 closely simulates the original grammar G in Kuroda normal form.*

Next, we consider a CD grammar system with the same structure but working in the t mode.

Theorem 3. *Let $G = (N,T,P,S)$ be a grammar. Then, there exists a two-component rule-homogeneous CD grammar system, $\Gamma = (N',T,H,I,S)$, such that $I = \{11 \to 00,\ 0000 \to \varepsilon\}$ and $L_t(\Gamma) = L(G)$.*

Proof.
Construction. The process of construction remains identical to Theorem 1.

Basic idea.
Recall that, during the generation of a sentence, a CD grammar system working in the t mode switches its components only if the process is not finished and there are no possible derivations with the previous component. Consider the general behavior of Γ. It starts the generation with S. For the first derivation, applicable rules can be found only in H, so this component has to be used. However, H also contains all rules simulating the original rules of G. Consequently, the first derivation in the t mode has to simulate all rules in G without completing the verification process for non-context-free rules. Nonetheless, we prove that the verification process can be done successfully afterwards for all simulated rules at once.

Corollary 4. *The resulting two-component rule-homogeneous CD grammar system Γ from the proof of Theorem 3 changes its components, during every generation of a sentence, no more than once.*

For the remaining results, we change the second component of the two-component rule-homogeneous CD grammar system so it is evenly homogeneous. We show that the previous ideas still hold, but the resulting behavior and properties are slightly modified.

Theorem 5. *Let $G = (N,T,P,S)$ be a grammar. Then, there exists a two-component rule-homogeneous CD grammar system, $\Gamma = (N',T,H,I,S)$, such that $I = \{11 \to 00,\ 0000 \to 2222\}$ and $L_*(\Gamma) = L(G)$.*

Corollary 6. *The resulting two-component rule-homogeneous CD grammar system Γ from the proof of Theorem 5 closely simulates the original grammar G in Kuroda normal form.*

Theorem 7. *Let $G = (N,T,P,S)$ be a grammar. Then, there exists a two-component rule-homogeneous CD grammar system, $\Gamma = (N',T,H,I,S)$, such that $I = \{11 \to 00,\ 0000 \to 2222\}$ and $L_t(\Gamma) = L(G)$.*

Acknowledgment

This work was supported by The Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability (NPU II); project IT4Innovations excellence in science - LQ1602; the TAČR grant TE01020415; and the BUT grant FIT-S-17-3964.