Leftmost Derivations of Propagating Scattered Context Grammars: A New Proof

Tomáš Masopust† and Jiří Techet‡

Faculty of Information Technology
Brno University of Technology
Božetěchova 2, Brno 61266
Czech Republic
masopust@fit.vutbr.cz,techet@fit.vutbr.cz

received September 17, 2007, revised April 8, 2008, accepted April 11, 2008.

In 1973, V. Virkkunen proved that propagating scattered context grammars which use leftmost derivations are as powerful as context-sensitive grammars. This paper brings a significantly simplified proof of this result.

Keywords: formal languages, propagating scattered context grammars, leftmost derivations, generative power

1 Introduction

Propagating scattered context grammars, introduced in [3], represent an important type of semi-parallel rewriting systems. Since their introduction, however, the exact relationship of the family of languages they generate to the family of context-sensitive languages is unknown. The language family generated by these grammars is included in the family of context-sensitive languages; on the other hand, the question of whether this inclusion is proper represents an open problem in formal language theory. There have been several attempts to modify the definition of propagating scattered context grammars to obtain the family of context-sensitive languages (see [1, 2, 4, 5, 6]). The approach discussed in [6] allows the productions to be applied only in a leftmost way and, thereby, obtain the family of context-sensitive languages generated by these grammars. This result is of some interest as the use of context-free, context-sensitive, and unrestricted productions in a leftmost way in the corresponding grammars of the Chomsky hierarchy does not have any impact on their generative power.

The proof in [6] consists of two parts; first, two preliminary lemmas (Lemma 2 and Lemma 3) are given and then the main result, stated in Theorem 2, is presented as a straightforward corollary of these
two lemmas. In Lemma 2 it is demonstrated how any sentence of some context-sensitive language can be derived by a propagating scattered context grammar which uses leftmost derivations. Every sentence generated in such a way contains, however, some additional symbols. Lemma 3 shows how these symbols can be removed. Together, the proof consists of six-page-long construction part and not even one-page-long basic idea of the construction which makes it extremely hard to follow. A more formal proof of the correctness of the construction is missing.

This paper aims to present the proof of this result in a much simpler and more readable way. The main difference of our proof lies (1) in the way how the symbols to be rewritten are selected and (2) the way context-sensitive productions are simulated. Furthermore, the proof is based on a single construction instead of two. All this leads to a significantly simpler and more transparent construction.

2 Preliminaries and definitions

We assume that the reader is familiar with formal language theory (see [10]). For an alphabet \( V \), \( |V| \) denotes the cardinality of \( V \). \( V^* \) represents the free monoid generated by \( V \). The unit of \( V^* \) is denoted by \( \varepsilon \). Set \( V^+ = V^* - \{ \varepsilon \} \). For \( w \in V^* \), \( |w| \) and \( \text{alph}(w) \) denote the length of \( w \) and the set of symbols occurring in \( w \), respectively.

A grammar is a quadruple \( G = (V, T, P, S) \), where \( V \) is the total alphabet, \( T \subset V \) is the set of terminals, \( P \) is a finite set of productions of the form \( x \rightarrow y \), where \( x \in V^*(V - T)V^* \), \( y \in V^* \), and \( S \in V - T \) is the start symbol of \( G \). If \( u = z_1x_2z_2, v = z_1y_2z_2, \) and \( x \rightarrow y \in P \), where \( z_1, z_2 \in V^* \), then \( G \) makes a derivation step from \( u \) to \( v \) according to \( x \rightarrow y \), symbolically written as \( u \Rightarrow_G v \) or, simply, \( u \Rightarrow v \). Let \( \Rightarrow^+_G \) and \( \Rightarrow^*_G \) denote the transitive closure of \( \Rightarrow_G \) and the reflexive and transitive closure of \( \Rightarrow_G \), respectively. If \( S \Rightarrow^*_G w \), where \( w \in T^* \), \( S \Rightarrow^*_G w \) is said to be a successful derivation of \( G \). The language of \( G \), denoted by \( L(G) \), is defined as \( L(G) = \{ w \in T^* : S \Rightarrow^*_G w \} \). If each production of \( G \) is of the form \( xAy \rightarrow xuy \), where \( x, y \in V^* \), \( A \in V - T \), \( u \in V^+ \), then \( G \) is a context-sensitive grammar. The family of context-sensitive languages is denoted by \( \mathcal{L}(CS) \). If each production of \( G \) is of one of the following forms: \( AB \rightarrow CD, A \rightarrow BC, A \rightarrow a \), where \( A, B, C, D \in V - T \), and \( a \in T \), then \( G \) is a grammar in the Kuroda normal form.

**Lemma 1 ([4])** For every context-sensitive grammar there exists an equivalent grammar in the Kuroda normal form.

A scattered context grammar (see [1] 2 3 5 6 7 8 9 11) is a quadruple \( G = (V, T, P, S) \), where \( V \) is the total alphabet, \( T \subset V \) is the set of terminals, \( S \in V - T \) is the start symbol of \( G \), and \( P \) is a finite set of productions such that each production has the form \( (A_1, A_2, \ldots, A_n) \rightarrow (x_1, x_2, \ldots, x_n) \), for some \( n \geq 1 \), where \( A_i \in V - T \), and \( x_i \in V^* \), for all \( 1 \leq i \leq n \). If each production \( (A_1, A_2, \ldots, A_n) \rightarrow (x_1, x_2, \ldots, x_n) \in P \) satisfies \( x_i \in V^+ \) for all \( 1 \leq i \leq n \), then \( G \) is a propagating scattered context grammar. If \( (A_1, A_2, \ldots, A_n) \rightarrow (x_1, x_2, \ldots, x_n) \in P, u = u_1A_1u_2A_2 \ldots u_nA_nu_{n+1}, \) and \( v = u_1x_1u_2x_2 \ldots u_nx_nu_{n+1} \), where \( u_i \in V^* \) for all \( 1 \leq i \leq n + 1 \), then \( G \) makes a derivation step from \( u \) to \( v \) according to \( p = (A_1, A_2, \ldots, A_n) \rightarrow (x_1, x_2, \ldots, x_n) \), symbolically written as \( u \Rightarrow_G v \) or, simply, \( u \Rightarrow v \). In addition, if \( A_i \notin \text{alph}(u_i) \) for all \( 1 \leq i \leq n \), then the direct derivation is leftmost, and we write \( u_{\text{lm}} \Rightarrow_G v \) or \( A_i \notin \text{alph}(u_{i+1}) \) for all \( 1 \leq i \leq n \), then the direct derivation is rightmost, and we write \( u_{\text{rm}} \Rightarrow_G v \). The language of \( G \), denoted by \( L(G) \), is defined as \( L(G) = \{ w \in T^* : S \Rightarrow^*_G w \} \). A propagating scattered context grammar \( G = (V, T, P, S) \) uses leftmost or rightmost derivations if its language is defined as \( L(G, \text{lm}) = \{ w \in T^* : S \Rightarrow^*_G w \} \) or \( L(G, \text{rm}) = \{ w \in T^* : S \Rightarrow^*_G w \} \).
$w\}$, respectively. The family of languages generated by propagating scattered context grammars which use leftmost or rightmost derivations is denoted by $L(PSC, \text{lm})$ or $L(PSC, \text{rm})$, respectively.

## 3 Main Results

The following theorem and its proof, which represent the main result of this paper, demonstrate that propagating scattered context grammars which use leftmost derivations are equivalent to context-sensitive grammars.

**Theorem 1** $L(PSC, \text{lm}) = L(\text{CS})$.

**Proof:** As propagating scattered context grammars do not contain erasing productions, their derivations can be simulated by linear bounded automata. As a result, $L(PSC, \text{lm}) \subseteq L(\text{CS})$. In what follows, we demonstrate that also $L(\text{CS}) \subseteq L(PSC, \text{lm})$ holds true by demonstrating that for every grammar in the Kuroda normal form there exists an equivalent propagating scattered context grammar which uses leftmost derivations.

Let $G = (V, T, P, S)$ be a grammar in the Kuroda normal form. Set $N_1 = (V - T) \cup \{\overline{a} : a \in T\}$ (suppose that $(V - T) \cap \{a : a \in T\} = \emptyset$), $N_1 = \{A : A \in N_1\}$. Let $n = |N_1|$; then, we denote the elements of $N_1$ as $\{A_1, A_2, \ldots, A_n\}$. Define the homomorphism $\alpha$ from $V^*$ to $N_1^*$ as $\alpha(A) = A$ for each $A \in V - T$, and $\alpha(a) = \overline{a}$ for each $a \in T$. Set $N'_2 = \{A' : A \in V - T\}$, $N'_3 = \{\langle ab \rangle : a, b \in V\}$, $N'_4 = \{\langle Aa \rangle' : A \in V - T, a \in V\}$, and

$$N_5 = \{\langle a, 0 \rangle, \langle ab, 0 \rangle : a, b \in V\}$$
$$\cup \{\langle a, i, j \rangle : a \in V - T, 1 \leq i \leq 3, 1 \leq j \leq n\}$$
$$\cup \{\langle ab, 4 \rangle : a, b \in T\}.$$  

Without loss of generality, assume that the sets $N_1, N'_1, N_2, N'_3, N'_4, N_5, \{S, X\}$, and $T$ are pairwise disjoint. Define the propagating scattered context grammar

$$\hat{G} = (N_1 \cup N'_1 \cup N_2 \cup N'_3 \cup N'_4 \cup N_5 \cup \{S, X\} \cup T, T, \hat{P}, \hat{S}),$$

where $\hat{P}$ is constructed as follows:

1. (a) For each $a \in L(G)$, where $a \in T$, add $$(S) \to (a) \to \hat{P};$$
   (b) For each $S \Rightarrow ab$, where $a, b \in V$, add $$(S) \to (\langle ab, 0 \rangle, X) \to \hat{P};$$

2. For each $a, b, c \in V$, add
   (a) $(\langle a, 0 \rangle, \alpha(b)) \to (\alpha(a), \langle b, 0 \rangle),$ (b) $(\alpha(a), \langle b, 0 \rangle)) \to (\langle a, 0 \rangle, \alpha(b)),$
   (c) $(\langle a, 0 \rangle, \langle bc \rangle) \to (\alpha(a), \langle bc, 0 \rangle),$ (d) $(\alpha(a), \langle bc, 0 \rangle) \to (\langle a, 0 \rangle, \langle bc \rangle) \to \hat{P};$

3. For each $A \to a \in P$ and $b \in V$, add
4. For each $A \rightarrow BC \in P$ and $a \in V$, add
   (a) $((A, 0)) \rightarrow ((a, 0))$,
   (b) $((Ab, 0)) \rightarrow ((ab, 0))$,
   (c) $((bA, 0)) \rightarrow ((ba, 0))$ to $\bar{P}$;

5. For each $AB \rightarrow CD \in P$, $a \in V$, $E \in N_3 \cup N_4$, $F' \in \{B', (Ba)'\}$, $1 \leq i \leq n$, and $1 \leq j \leq n - 1$, add
   (a) $((AB, 0)) \rightarrow ((CD, 0))$,
   (b) i. $((A, 0), B, X) \rightarrow ((A, 1, 1), B', A_1)$,
      ii. $((A, 0), (Ba), X) \rightarrow ((A, 1, 1), (Ba)', A_1)$,
   (c) i. $((A, 1, i), A_1) \rightarrow ((A, 2, i), \hat{A}_1)$,
      ii. $((A, 2, i), F', \hat{A}_1) \rightarrow ((A, 3, i), F', A_1)$,
      iii. $((A, 3, j), E, A_j) \rightarrow ((A, 1, j + 1), E, A_{j+1})$,
   (d) i. $((A, 3, n), B', E, A_n) \rightarrow ((C, 0), D, E, X)$,
      ii. $((A, 3, n), (Ba)', A_n) \rightarrow ((C, 0), (Da), X)$ to $\bar{P}$;

6. For each $a, b, c \in T$, add
   (a) $((ab, 0)) \rightarrow ((ab, 4))$,
   (b) $(\bar{\epsilon}, (ab, 4)) \rightarrow (c, (ab, 4))$,
   (c) $((ab, 4), X) \rightarrow (a, b)$ to $\bar{P}$.

In short, productions introduced in (1) initiate the derivation, productions from (2) are used to select the nonterminal to be rewritten, productions from (3), (4), and (5) simulate $G$’s productions of the form $A \rightarrow a$, $A \rightarrow BC$, and $AB \rightarrow CD$, respectively, and, finally, productions from (6) finish the derivation. In the following paragraphs, we describe the derivation of $\bar{G}$ in greater detail.

Every derivation starts either by a production introduced in (1a) to generate sentences $a \in L(G)$, where $a \in T$, or by a production introduced in (1b) to generate sentences $x \in L(G)$, where $|x| \geq 2$. As $\bar{S}$ does not occur on the right-hand side of any production, productions from (1) are not used during the rest of the derivation.

Consider $G$’s sentential form $a_1a_2\ldots a_k$, where $a_1, a_2, \ldots, a_k \in V$, for some $k \geq 2$. In $\bar{G}$, this sentential form corresponds to

$$b_1b_2\ldots b_{r-1}(a_r, 0)b_{r+1}b_{r+2}\ldots b_{k-2}(a_{k-1}a_k)X,$$

where $b_i = \alpha(a_i)$ for all $i \in \{1, 2, \ldots, r - 1, r + 1, r + 2, \ldots, k - 2\}$, for some $1 \leq r \leq k - 2$, or to

$$b_1b_2\ldots b_{k-2}(a_{k-1}a_k, 0)X,$$
where $b_i = \alpha(a_i)$ for all $1 \leq i \leq k - 2$ (observe that every right-hand side of a production from $\text{(1b)}$ represents a sentential form of this kind). To simulate a $G$’s production, the leftmost nonterminal from its left-hand side has to be selected in the sentential form of $\bar{G}$. This is done by appending 0 to the symbol to be selected by productions from $\text{(2)}$. Specifically, for a symbol $a \in V$, $\text{(2a)}$ selects the leftmost symbol $a$ immediately following the currently selected symbol and $\text{(2b)}$ selects the leftmost symbol $a$ preceding the currently selected symbol. Productions from $\text{(2c)}$ and $\text{(2d)}$ are used to select and unselect the penultimate nonterminal in $G$’s sentential form which is composed of two symbols from $V$. Observe that in this way, any symbol (except for the final $X$) in every sentential form of $\bar{G}$ can be selected. Further, observe that during a derivation, always one symbol is selected.

After the nonterminal is selected, the use of the $G$’s production can be simulated. Productions of the form $A \rightarrow a$ are simulated by $\text{(3a)}$ for every selected nonterminal $a_1, a_2, \ldots, a_{k-2}$ and by $\text{(5a)}, \text{(5c)}$ if the penultimate nonterminal (which contains $a_{k-1}, a_k$) of the $G$’s sentential form is selected. Analogously, productions of the form $A \rightarrow BC$ are simulated by productions from $\text{(4)}$.

Productions from $\text{(3a)}$ are used to simulate an application of productions of the form $AB \rightarrow CD$ within the penultimate nonterminal of $\bar{G}$’s sentential form. In what follows, we demonstrate how productions from $\text{(5b)}, \text{(5c)},$ and $\text{(5d)}$ are used if this production is simulated within $a_1 a_2 \ldots a_{k-2}$. Suppose that the sentential form in $\bar{G}$ is of the form

$$b_1 b_2 \ldots b_{r-1} (a_r, 0)b_{r+1}b_{r+2} \ldots b_{k-2} (a_{k-1} a_k) X$$

and we simulate the application of $a_ra_{r+1} \rightarrow c_rc_{r+1} \in P$. Recall that $N_1 = \{A_1, A_2, \ldots, A_n\}$ denotes the set of all symbols which may appear in $b_{r+1}b_{r+2} \ldots b_{k-2}$. First, to select $b_{r+1} = \alpha(a_{r+1})$, the production

$$((a_r, 0), b_{r+1}, X) \rightarrow ((a_r, 1, 1), b_{r+1}', A_1)$$

from $\text{(5b)}$ is applied in a successful derivation, so

$$b_1 b_2 \ldots b_{r-1} (a_r, 0)b_{r+1}b_{r+2} \ldots b_{k-2} (a_{k-1} a_k) X \\quad \text{lm} = \bar{G} \quad b_1 b_2 \ldots b_{r-1} (a_r, 1, 1)b_{r+1}'b_{r+2}b_{r+3} \ldots b_{k-2} (a_{k-1} a_k) X \quad A_1.$$  

Observe that if $b_{r+1}$ does not immediately follow $(a_r, 0)$, the leftmost $b \in \text{alph}(b_{r+2}b_{r+3} \ldots b_{k-2})$ satisfying $b = b_{r+1}$ is selected by the production from $\text{(5b)}$. The purpose of productions from $\text{(5c)}$ is to verify that the nonterminal immediately following $(a_r, 0)$ has been selected. First, the production

$$((a_r, 1, 1), A_1) \rightarrow ((a_r, 2, 1), \hat{A}_1)$$

from $\text{(5c)}$ is applied to tag the first $A_1$ following $(a_r, 1, 1)$, so

$$b_1 b_2 \ldots b_{r-1} (a_r, 1, 1)b_{r+1}'b_{r+2}b_{r+3} \ldots b_{k-2} (a_{k-1} a_k) A_1 \\quad \text{lm} = \bar{G} \quad b_1 b_2 \ldots b_{r-1} (a_r, 2, 1)b_{r+1}'y_1 (a_{k-1} a_k) d_1,$$

where either

$$y_1 = b_{r+2}b_{r+3} \ldots b_{m-1} \hat{A}_1 b_m b_{m+2} \ldots b_{k-2}, \quad d_1 = A_1,$$

satisfying $A_1 \notin \text{alph}(b_{r+2}b_{r+3} \ldots b_{m-1})$, for some $1 \leq m \leq k - 2$, or

$$y_1 = b_{r+2}b_{r+3} \ldots b_{k-2}, \quad d_1 = \hat{A}_1.$$
satisfying $A_1 \notin \text{alph}(y_1)$. Then, the production
\[
(\langle a_r, 2, 1 \rangle, b'_{r+1}, A_1) \rightarrow (\langle a_r, 3, 1 \rangle, b'_{r+1}, A_1)
\]
from (5(c)i) is applied to untag the first symbol $\hat{A}_1$ following $b'_{r+1}$, so
\[
b_1 b_2 \ldots b_{r-1} \langle a_r, 2, 1 \rangle b'_{r+1} y_1 \langle a_{k-1} a_k \rangle d_1,
\]
then, the production
\[
b_1 b_2 \ldots b_{r-1} \langle a_r, 3, 1 \rangle b'_{r+1} b_{r+2} b_{r+3} \ldots b_{k-2} \langle a_{k-1} a_k \rangle A_1.
\]
This means that if $A_1$ occurs between $\langle a_r, 2, 1 \rangle$ and $b'_{r+1}$, it is tagged by the production from (5(c)i) but it cannot be untagged by any production from (5(c)ii), so the derivation is blocked. Finally, the production
\[
(\langle a_r, 3, 1 \rangle, \langle a_{k-1} a_k \rangle, A_1) \rightarrow (\langle a_r, 1, 2 \rangle, \langle a_{k-1} a_k \rangle, A_2)
\]
from (5(c)ii) is applied, so
\[
b_1 b_2 \ldots b_{r-1} \langle a_r, 3, 1 \rangle b'_{r+1} b_{r+2} b_{r+3} \ldots b_{k-2} \langle a_{k-1} a_k \rangle A_1
\]
and the same verification continues for $A_2$. This verification proceeds for all symbols from $N_1$ so this part of the derivation can be expressed as
\[
\text{im} \Rightarrow G u_1 \ [p_{11}] \text{im} \Rightarrow G u_1 \ [p_{12}]
\]
\[
\text{im} \Rightarrow G u_2 \ [p_{13}] \text{im} \Rightarrow G u_2 \ [p_{21}]
\]
\[
\vdots
\]
\[
\text{im} \Rightarrow G u_n \ [p_{n-1}]
\]
with
\[
u_i = b_1 b_2 \ldots b_{r-1} \langle a_r, 1, i \rangle b'_{r+1} b_{r+2} b_{r+3} \ldots b_{k-2} \langle a_{k-1} a_k \rangle A_i,
\]
\[
v_i = b_1 b_2 \ldots b_{r-1} \langle a_r, 2, i \rangle b'_{r+1} y_1 \langle a_{k-1} a_k \rangle d_i,
\]
\[
u_i = b_1 b_2 \ldots b_{r-1} \langle a_r, 3, i \rangle b'_{r+1} b_{r+2} b_{r+3} \ldots b_{k-2} \langle a_{k-1} a_k \rangle A_i,
\]
where $p_{i1}$, $p_{i2}$, and $p_{i3}$ are productions from (5(c)i), (5(c)ii), and (5(c)iii), respectively, for all $1 \leq i \leq n$, $1 \leq j \leq n - 1$, and either
\[
y_i = b_{r+2} b_{r+3} \ldots b_{i_{m-1}} A_i b_{i_{m+1}} b_{i_{m+2}} \ldots b_{k-2},
\]
satisfying $A_i \notin \text{alph}(b_{r+2} b_{r+3} \ldots b_{i_{m-1}})$, for some $1 \leq i_{m} \leq k - 2$, or
\[
y_i = b_{r+2} b_{r+3} \ldots b_{k-2},
\]
satisfying $A_i \notin \text{alph}(y_i)$. After the verification is finished, the application of $a_r a_{r+1} \rightarrow c_r c_{r+1} \in P$ is simulated by
\[
(\langle a_r, 3, n \rangle, b'_{r+1}, \langle a_{k-1} a_k \rangle, A_n) \rightarrow (\langle c_r, 0 \rangle, c_{r+1}, \langle a_{k-1} a_k \rangle, X)
\]
from (5(d)ii), so
\[
b_1 b_2 \ldots b_{r-1} \langle a_r, 3, n \rangle b'_{r+1} b_{r+2} b_{r+3} \ldots b_{k-2} \langle a_{k-1} a_k \rangle A_n
\]
then, the production
\[
b_1 b_2 \ldots b_{r-1} \langle c_r, 0 \rangle c_{r+1} b_{r+2} b_{r+3} \ldots b_{k-2} \langle a_{k-1} a_k \rangle X.
\]
Observe that in order to simulate a production of the form $AB \rightarrow CD$ within $a_{k-2}a_{k-1}$, productions from $S(6b)$ and $S(6c)$ have to be used instead of productions from $S(6a)$ and $S(6d)$ in the simulation described above. The details are left to the reader.

Finally, consider a $G$’s sentence $a_1a_2\ldots a_k \in T^+$. This corresponds to

$$\bar{a}_1\bar{a}_2\ldots \bar{a}_{r-1}\langle a_r, 0\rangle \bar{a}_{r+1}\bar{a}_{r+2}\ldots \bar{a}_{k-2}\langle a_{k-1}a_k \rangle X,$$

or

$$\bar{a}_1\bar{a}_2\ldots \bar{a}_{k-2}\langle a_{k-1}a_k, 0 \rangle X$$

in $\tilde{G}$ after finishing the simulation. To enter the final phase in $\tilde{G}$, we need the sentential form to be in the second above described form. This can be achieved by applying a production from (2c) to the first sentential form. The rest of the derivation can be expressed as

$$\begin{align*}
\text{lm} & \Rightarrow_{\bar{G}} ^{k-2} \bar{a}_1\bar{a}_2\ldots \bar{a}_{k-2}\langle a_{k-1}a_k, 4 \rangle X \ [\Xi_{6b}] \\
\text{lm} & \Rightarrow_{\bar{G}} ^{k-2} a_1a_2\ldots a_{k-2}\langle a_{k-1}a_k, 4 \rangle X \ [\Xi_{6b}] \\
\text{lm} & \Rightarrow_{\bar{G}} ^{k-2} a_1a_2\ldots a_{k-2}a_{k-1}a_k \ [p_{6c}],
\end{align*}$$

where $p_{6a}$ and $p_{6c}$ are productions introduced in steps (6a) and (6c), respectively, and $\Xi_{6b}$ is a sequence of $k-2$ productions from (6b). As a result, $x \in L(G, \text{lm})$ if and only if $x \in L(G)$. Therefore, $\mathcal{L}(CS) \subseteq \mathcal{L}(PSC, \text{lm})$.

As $\mathcal{L}(PSC, \text{lm}) \subseteq \mathcal{L}(CS)$ and $\mathcal{L}(CS) \subseteq \mathcal{L}(PSC, \text{lm})$, we obtain $\mathcal{L}(PSC, \text{lm}) = \mathcal{L}(CS)$, so the theorem holds.

Next, we state the following corollary.

**Corollary 1** $\mathcal{L}(PSC, \text{rm}) = \mathcal{L}(CS)$.

**Proof:** This corollary can be proved by a straightforward modification of the proof of Theorem 1 and its proof is, therefore, left to the reader.
References


