On computational complexity of contextual languages

Lucian Ilie

Turku Centre for Computer Science

TUCS Technical Reports

No 11, May 1996

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi

May 1996

ISBN 951-650-753-0
ISSN 1239-1891

Abstract

We consider the following restriction of internal contextual grammars, called *local*: in any derivation in a grammar, after applying a context, further contexts can be added only inside of or at most adjacent to the previous ones. We further consider a natural restriction of this derivation mode by requiring that no superword of the word considered as selector can be used as selector. We investigate the relevance of the latter type of grammars for natural language study. In this aim, we show that all the three basic non-context-free constructions in natural languages, that is, multiple agreements, crossed agreements, and duplication, can be realized using this type of grammars. Our main result is that these languages are parsable in polynomial time. The problem of semilinearity remains open.

TUCS Research Group
"Mathematical Structures of Computer Science"

1 Introduction

The contextual grammars have been introduced in [M2], based on the basic phenomenon in descriptive linguistics, that of acceptance of a word by a context (or conversely); see [M1]. Thus, the generative process in a contextual grammar is based on two dual linguistic operations most important in both natural and artificial languages: insertion of a word in a given context (pair of words) and adding a context to a given word. Any derivation in a contextual grammar is a finite sequence of such operations, starting from an initial finite set of words, simple enough to be considered as primitive well formed words (axioms).

During the over 25 years since they have been introduced, many variants were already investigated: determinism, [PRS1], parallelism, [PRS2], normal forms, [EPR1], modularity, [PRS3], etc.; the reader is referred to the monograph [P1] for details (up to 1981) or to [EPR3] for a recent survey. Discussions about the motivations of contextual grammars can be found also in [M3].

In [MMMP], one considers internal contextual grammars with the following restriction: the words selecting the context must be maximal with respect to the subword order. The power of such grammars and their relevance for natural language study have been further investigated in [MMP1] and [MMP2]. The basic idea in restricting the derivation in such a way was to obtain classes of contextual languages more appropriate from natural languages point of view. In fact, the class of languages searched for should have the following basic properties, which define the so-called *mildly context-sensitive languages*:

- it contains the three basic non-context-free constructions in natural languages, that is *multiple agreements, crossed agreements*, and *duplication*,
- its languages are *semilinear*, and
- its languages are *polynomial time parsable.*

It is known that the basic variant of internal contextual grammars can generate non-semilinear languages. This was proved first in [EIPRS] for grammars with regular selection and the result was recently improved in [Il] where it is proved that grammars with finite selection only are powerful enough to generate non-semilinear languages.

Unfortunately, the restrictions imposed so far still allow non-semilinear languages, as proved recently in [MMP2], and it is not known whether or not the obtained languages are polynomial time parsable.

In this paper, we make an attempt towards finding a class of mildly context-sensitive contextual languages. We impose also the maximality of the selectors but, before this, we restrict the usual derivation in a contextual grammar in a *local* way, that is, in any derivation, after applying a context, further contexts can be added to the obtained word only inside of or at most adjacent to the previous ones. We prove that all the three non-context-free constructions in natural languages mentioned above can be realized using local contextual grammars with regular selection, working in the maximal mode of derivation. Our main result is that all languages generated by local contextual grammars with finite or regular selection, working or not in the maximal mode, are parsable in polynomial time. The problem of semilinearity of these languages remains open.

2 Basic definitions

A finite non-empty set Σ is called an alphabet. We denote by Σ^* the free monoid generated by Σ, by λ its identity, and $\Sigma^+ = \Sigma^* - \{\lambda\}$. The elements of Σ^* are called words. For any word $x \in \Sigma^*$, we denote by $|x|$ the length of x. The families of finite, regular, context-free, and context-sensitive languages are denoted by FIN, REG, CF, and CS, respectively. For further notions of formal language theory we refer to [Sa] or [HU].

An *(internal) contextual grammar* is a construct

$$G = (\Sigma, A, (S_1, C_1), \ldots, (S_n, C_n)),$$

$n \geq 1$, where

Σ is an alphabet,

$A \subseteq \Sigma^*$ is a finite set, called the set of *axioms*,

$S_i \subseteq \Sigma^*, 1 \leq i \leq n$, are the sets of *selectors*,

$C_i \subseteq \Sigma^* \times \Sigma^*, C_i$ finite, $1 \leq i \leq n$, are the sets of *contexts*.

There exists another type of contextual grammars called *external* but, since we work here only with internal ones, we will omit the word "internal" for the rest of the paper.

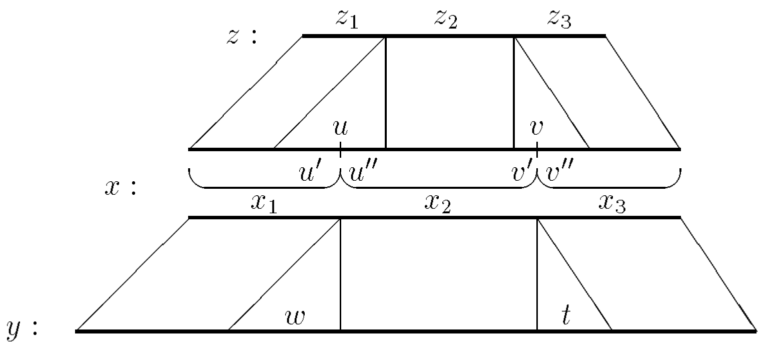
The usual derivation is defined, for $x, y \in \Sigma^*$, as follows:

$$\begin{aligned} x \Longrightarrow y \quad \text{iff} \quad & x = x_1x_2x_3, y = x_1ux_2vx_3, \\ & x_2 \in S_i, (u,v) \in C_i, \text{ for some } 1 \leq i \leq n. \end{aligned}$$

We consider the following modification of the basic mode of derivation: if we have the derivation $z \Longrightarrow x$ such that $z = z_1z_2z_3, z_1, z_2, z_3 \in \Sigma^*, z_2 \in S_i, (u,v) \in C_i$, for some $1 \leq i \leq n$, and $x = z_1uz_2vz_3$, then a derivation $x \Longrightarrow y$ is called *local* with respect to the derivation $z \Longrightarrow x$, denoted

$$z \Longrightarrow x \Longrightarrow_{loc} y,$$

iff we have $u = u'u'', v = v'v'', u', u'', v', v'' \in \Sigma^*$, $x = x_1x_2x_3$ for $x_1 = z_1u', x_2 = u''z_2v', x_3 = v''z_3$, $x_2 \in S_j, (w,t) \in C_j$, for some $1 \leq j \leq n$, and $y = x_1wx_2tx_3$. (The derivation is called local because the new places where contexts are introduced are inside or at most adjacent to the previous contexts introduced. From this definition follows also the term *local contextual grammars*.) For illustration, the picture below can be seen.



We further consider a natural restriction of the derivation newly defined, namely, using the same notations, the derivation $x \Longrightarrow y$ is called *maximal local* w.r.t. $z \Longrightarrow x$ iff it is local and there is no other local derivation w.r.t. $z \Longrightarrow x$, $x \Longrightarrow y'$, such that the decomposition used for x, say $x = x_1'x_2'x_3', x_2' \in S_j$, verifies $|x_1'| \leq |x_1|, |x_2'| > |x_2|, |x_3'| \leq |x_3|$. We denote this by $z \Longrightarrow x \Longrightarrow_{Mloc} y$.

For $\alpha \in \{loc, Mloc\}$ and $x, y \in \Sigma^*$ we put $x \Longrightarrow_\alpha^* y$ iff we have a finite sequence of derivations, each of them, excepting the first one, being performed in α mode w.r.t. the previous one, that is

$$x = x_0 \Longrightarrow x_1 \Longrightarrow_\alpha x_2 \Longrightarrow_\alpha \cdots \Longrightarrow_\alpha x_k = y$$

for some $k \geq 0$, and, moreover, in the case $\alpha = Mloc$, also the first derivation, $x \Longrightarrow x_1$, should be performed in a maximal way (in the above sense).

The language generated by a grammar G in the mode α is

$$L_\alpha(G) = \{w \in \Sigma^* \mid \text{ there is } x \in A \text{ such that } x \Longrightarrow_\alpha^* w\}.$$

If all the sets of selectors $S_1, S_2, \ldots, S_n$, are in a family F of languages, then we say that the grammar G is *with F selection*.

The family of languages $L_\alpha(G)$ for G with F selection will be denoted by $\mathcal{LC}_\alpha(F)$. (We will consider only the cases when F is FIN or REG.)

3 Non-context-free constructions in natural languages

We show in this section that all the three basic non-context-free constructions in natural languages, that is, multiple agreements, crossed agreements, and duplication, can be realized using local contextual grammars with regular selection, working in the maximal mode of derivation.

Lemma 3.1 *(multiple agreements)*
$L_1 = \{a^n b^n c^n \mid n \geq 1\} \in \mathcal{LC}_{Mloc}(REG)$.

Proof. Consider the following grammar

$$G_1 = (\{a, b, c\}, \{abc\}, (\{b^n \mid n \geq 1\}, \{(ab, c)\})).$$

Any derivation in the mode $Mloc$ in G is as follows:

$$abc \Longrightarrow a^2b^2c^2 \Longrightarrow_{Mloc} a^3b^3c^3 \Longrightarrow_{Mloc} \cdots \Longrightarrow_{Mloc} a^nb^nc^n,$$

for some $n \geq 1$, because the context (ab, c) must be applied all the time arround the longest subword $b^i, i \geq 1$, of the current sentential form. Remark further that the derivation above is indeed performed in the local mode. It follows that

$$L_{Mloc}(G_1) = L_1$$

and our lemma is proved.

□

Lemma 3.2 *(crossed agreement)*
$L_2 = \{a^n b^m c^n d^m \mid m, n \geq 1\} \in \mathcal{LC}_{Mloc}(REG)$.

Proof. Take the grammar

$$G_2 = (\{a, b, c, d\}, \{abcd\}, (\{b^m c^n \mid m, n \geq 1\}, \{(ab, cd)\})).$$

Arguing as in the proof of Lemma 3.1, we get that any derivation in the mode $Mloc$ in G_2 is as follows:

$$abcd \Longrightarrow a^2b^2c^2d^2 \Longrightarrow_{Mloc} a^3b^3c^3d^3 \Longrightarrow_{Mloc} \cdots \Longrightarrow_{Mloc} a^nb^nc^nd^n,$$

for some $n \geq 1$. It follows that

$$L_{Mloc}(G_2) = L_2$$

and the result is proved.

□

Lemma 3.3 *(duplication)*
$L_3 = \{wcw \mid w \in \{a,b\}^*\} \in \mathcal{LC}_{Mloc}(REG)$.

Proof. Consider the grammar

$$G_3 = (\{a,b,c\}, \{c\}, (\{cx \mid x \in \{a,b\}^*\}, \{(a,a),(b,b)\})).$$

Any word wcw in the language L_3 can be derived as below (we suppose that $w = w_1 w_2 \dots w_{|w|}, w_i \in \{a,b\}, 1 \le i \le |w|$):

$$\begin{gathered} c \Longrightarrow w_1 c w_1 = w_1(cw_1) \Longrightarrow_{Mloc} w_1 w_2 c w_1 w_2 = w_1 w_2 (c w_1 w_2) \Longrightarrow_{Mloc} \\ \Longrightarrow_{Mloc} \cdots \Longrightarrow_{Mloc} w_1 \dots w_i c w_1 \dots w_i = w_1 \dots w_i (c w_1 \dots w_i) \Longrightarrow_{Mloc} \\ \Longrightarrow_{Mloc} w_1 \dots w_{i+1} c w_1 \dots w_{i+1} = w_1 \dots w_{i+1} (c w_1 \dots w_{i+1}) \Longrightarrow_{Mloc} \\ \Longrightarrow_{Mloc} \cdots \Longrightarrow_{Mloc} w_1 w_2 \dots w_{|w|} c w_1 w_2 \dots w_{|w|} = wcw \end{gathered}$$

(we have put in parentheses the selector in each sentential form).

It should be clear that the maximal selector in a sentential form

$$w_1 w_2 \dots w_i c w_1 w_2 \dots w_i, 1 \le i \le |w|,$$

is the suffix beginning at the position where the letter c occurs and that each context, excepting the first one, is added adjacently to the previous one, hence the derivation is local.

Consequently,

$$L_{Mloc}(G_3) = L_3$$

and our lemma is proved.

□

From the three lemmas above we get the following theorem:

Theorem 3.4 *The three basic non-context-free constructions in natural languages can be realized using local contextual grammars with regular selection, working in the maximal mode of derivation.*

4 Computational complexity

We show now that all languages in the family $\mathcal{LC}_\alpha(F)$ for $\alpha \in \{loc, Mloc\}$, $F \in \{FIN, REG\}$, are parsable in polynomial time, that is, the computational complexity of the membership problem for these languages is polynomial-time.

We will use the following result from [Pa]:

Proposition 4.1 (see Theorem 7.4 in [Pa]) $\mathbf{NSPACE}(\log n) \subseteq \mathbf{P}$.

We intend to prove that $\mathcal{LC}_\alpha(F) \subseteq \mathbf{P}$, for $\alpha \in \{loc, Mloc\}$, $F \in \{FIN, REG\}$.

Lemma 4.2 *For any grammar G with F selection, $F \in \{FIN, REG\}$,*

$$G = (\Sigma, A, (S_1, C_1), \ldots, (S_n, C_n)),$$

$n \geq 1$, and $\alpha \in \{loc, Mloc\}$, there is an input preserving nondeterministic TM M which accepts $L_\alpha(G)$ and requires $\mathcal{O}(\log |x|)$ space on any input x.

Proof. We will present the proof for the case $\alpha = loc$, $F = REG$, the other cases being similar.

Let us first see more detailed how a word w in the language $L_{loc}(G)$ is obtained. According to the definition, there must be a derivation

$$x = x_0 \Longrightarrow x_1 \Longrightarrow_{loc} x_2 \Longrightarrow_{loc} \cdots \Longrightarrow_{loc} x_k = w,$$

for some $x \in A$, such that all the time the new contexts are added inside of or adjoint to the places where the previous contexts were added.

More formally, this means (and it follows directly from the definition of the local mode of derivation):

(i) for each $0 \leq i \leq k-1, x_i = x_i' x_i'' x_i'''$, with $x_i'' \in S_{j_i}$, for some $1 \leq j_i \leq n$,

(ii) for each $1 \leq i \leq k, x_i = x_{i-1}' u_i x_{i-1}'' v_i x_{i-1}'''$, for some context $(u_i, v_i) \in C_{j_{i-1}}$,

(iii) for each $1 \leq i \leq k-1, x_i' = x_{i-1}' u_i', x_i'' = u_i'' x_{i-1}'' v_i', x_i''' = v_i'' x_{i-1}'''$, where $u_i = u_i' u_i'', v_i = v_i' v_i''$.

Consider now the grammar G above and let us construct an input preserving nondeterministic TM M which accepts $L_{loc}(G)$ under the required conditions.

Our machine will not have an output tape since it is supposed to decide only whether or not the input is in our language. (In fact nothing is computed; M answers only "yes" or "no".)

M will work as follows (we use the notations concerning the derivation of the word w above):

(1) Start at the beginning with w on the input tape.

(2) Check whether or not $w \in A$. If it is, then M **accepts**, otherwise goes to step (3).

(3) All the time, M tries to reconstruct backward the derivation of w. Thus, M has to remember all the time both the current sentential form of G, say x_i, and the context added to it in order to obtain the next one, x_{i+1}.

As M is input preserving, so it is not allowed to alter the input, it will remember eight indices, $p_1, p_2, r_1, r_2, s_1, s_2, t_1, t_2$: there are two words for each context and each of them is split into two subwords; see (i)–(iii).

(4) Suppose that the decomposition of w is $w = w_1 w_2 \ldots w_{|w|}$, $w_1, w_2, \ldots, w_{|w|} \in \Sigma$. Then, at the beginning, M guesses the four numbers $p_1 = p_2, r_1 = r_2, s_1 = s_2, t_1 = t_2$ such that

$$\begin{aligned}
&u_k = w_{p_1} w_{p_1+1} \ldots w_{r_1-1},\\
&v_k = w_{s_1} w_{s_1+1} \ldots w_{t_1-1},\\
&w_{r_1} w_{r_1+1} \ldots w_{s_1-1} \in S_{j_k}, (u_k, v_k) \in C_{j_k}.
\end{aligned}$$

(Remeber that the context (u_k, v_k) was the last one added in the derivation of w; see (ii) above.)

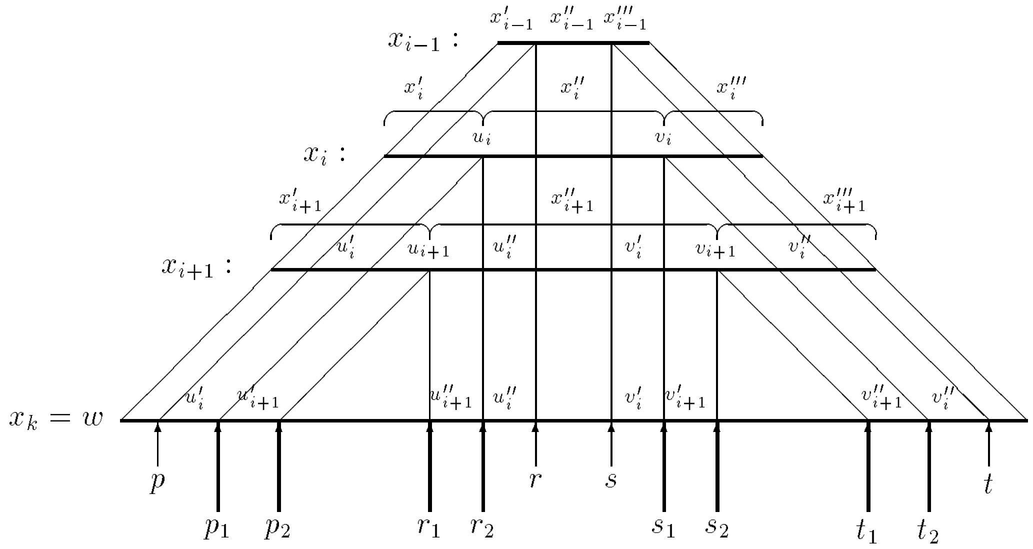
(5) If there are no numbers p_1, r_1, s_1, t_1 such that $(w_{p_1} \ldots w_{r_1-1}, w_{s_1} \ldots w_{t_1-1}) \in C_j, w_{r_1} \ldots w_{s_1-1} \in S_j$, for some $1 \leq j \leq n$, then M **rejects**.

(6) Suppose that, at some stage, M knows the eight indices $p_1, p_2, r_1, r_2, s_1, s_2, t_1, t_2$ with the properties mentioned at step (3). Knowing these indices, M can determine exactly the derivation of x_{i+1} from x_i because

$$x_{i+1} = x'_i u'_{i+1} u''_{i+1} x''_i v'_{i+1} v''_{i+1} x'''_i,$$

$$\begin{aligned}
&x'_i = w_1 \ldots w_{p_1-1}, && u'_{i+1} = w_{p_1} \ldots w_{p_2-1}, && v'_{i+1} = w_{s_1} \ldots w_{s_2-1},\\
&x''_i = w_{r_2} \ldots w_{s_1-1}, && u''_{i+1} = w_{r_1} \ldots w_{r_2-1}, && v''_{i+1} = w_{t_1} \ldots w_{t_2-1},\\
&x'''_i = w_{t_2+1} \ldots w_{|w|},
\end{aligned}$$

(see also (i)–(iii)). For more clarity, see the picture below.



(7) M guesses the numbers p, r, s, t such that the ends of the components of the context added in order to obtain the current sentential form of G, that is, x_i, was added in these positions. Because the derivation is supposed to be performed in *local* mode, there should exist numbers p, r, s, t such that

$$p \leq p_1, r_2 \leq r \leq s \leq s_2, t_2 \leq t.$$

For instance, in our case M could guess p, r, s, t such that

$$u'_i = w_p \dots w_{p_1-1}, u''_i = w_{r_2} \dots w_{r-1},$$
$$v'_i = w_s \dots w_{s_1-1}, v''_i = w_{t_2} \dots w_{t-1}.$$

(8) If there are no numbers p, r, s, t as above, then M **rejects**.

(9) Suppose that M has found the numbers p, r, s, t with the respective properties. Then the new current indices remembered by M will be

$$(p_1, p_2, r_1, r_2, s_1, s_2, t_1, t_2) \longleftarrow (p, p_1, r_2, r, s, s_1, t_2, t)$$

(where the replacements are made simultaneously).

Notice that the local mode of derivation plays an essential role at this point because M may remember only a set of eight numbers. If the derivation were of the usual type, then M should remember two pairs of numbers for each context added at each step and the space needed for this would not be any more logarithmic, as it is easy to see.

(10) If $z \in A$, then M **accepts**, otherwise it continues backward to reconstruct the derivation of w by guessing another context added. (So, M continues the work from step (7).)

Some remarks are in order:

- M never changes its input, so it is input preserving.

- If the input w belongs to the language $L_{loc}(G)$, then there must be some derivations in the local mode for w in G. Consequently, if the respective numbers are not found in steps (4) or (7), this means that there is no backward continuation hence M rejects correctly in steps (5) or (8), respectively.

- If the input w is an axiom, then it belongs to the language generated and M accepts correctly in step (2).

- If the word z is found as being an axiom, then, obviously, the tested word belongs to our language and M accepts correctly in step (10).

It is easy to see how much space M requires on the input w: it needs to store all the time the eight numbers $p_1, p_2, r_1, r_2, s_1, s_2, t_1, t_2$ and in step (6) to use the numbers p, r, s, t. As each of these twelve numbers is smaller that the length of the input, that is, $|w|$, it follows that each of them can be written in binary using at most $[\log_2 |w|] + 1$ memory cells. All the other operations performed by M (comparisons between indices or verifications of the membership for the selectors and contexts – the sets of contexts are finite and the sets of selectors are regular) require only a finite amount of space. It follows that M operates in space $\mathcal{O}(12 \log |x|) = \mathcal{O}(\log |x|)$ as claimed.

□

Directly from Lemma 4.2 follows the next result.

Corollary 4.3 $\mathcal{LC}_\alpha(F) \subseteq$ **NSPACE**$(\log n)$, *for* $\alpha \in \{loc, Mloc\}$, $F \in \{FIN, REG\}$.

From Corollary 4.3 and Proposition 4.1 we get the main result of this section.

Theorem 4.4 $\mathcal{LC}_\alpha(F) \subseteq$ **P**, *for* $\alpha \in \{loc, Mloc\}$, $F \in \{FIN, REG\}$.

5 Generative power

We are not concerned very much with the generative power of our devices because the main aim of this paper is not to investigate this but to show their relevance and suitability for natural languages.

Anyway, we prove that all families introduced above are strictly contained in the family of context-sensitive languages, the family of languages generated by grammars with regular choice contains non-context-free languages, and any regular language is the image by a weak coding (that is, a morphism which maps each letter into another or into λ) of a language in the family $\mathcal{LC}_{Mloc}(REG)$. However, it remains an open problem whether or not the family REG is included in any of the families $\mathcal{LC}_\alpha(F)$, $\alpha \in \{loc, Mloc\}$, $F \in \{FIN, REG\}$.

Theorem 5.1 *$\mathcal{LC}_\alpha(F) \subset CS$, for any $\alpha \in \{loc, Mloc\}, F \in \{FIN, REG\}$.*

Proof. The inclusion follows by Corollary 4.3 since

$$\mathcal{LC}_\alpha(F) \subseteq \mathbf{NSPACE}(\log n) \subseteq \mathbf{NSPACE}(n) = CS$$

for any $\alpha \in \{loc, Mloc\}, F \in \{FIN, REG\}$. Hence, it remains to prove its strictness. Consider the context-sensitive language

$$L = \{a^{2^n} \mid n \geq 1\}.$$

It is clear that all one-letter languages in the families $\mathcal{LC}_\alpha(F)$ are semilinear; this is due to the fact that in a derivation in a contextual grammar all sentential forms belong to the language generated. As L is not semilinear, it follows that $L \in CS - \mathcal{LC}_\alpha(F)$, for any $\alpha \in \{loc, Mloc\}, F \in \{FIN, REG\}$, and we are done.

□

Theorem 5.2 *$\mathcal{LC}_\alpha(REG) - CF \neq \emptyset$, for $\alpha \in \{loc, Mloc\}$.*

Proof. In what concers the case $\alpha = Mloc$, all the three languages in Lemmas 3.1 - 3.3 are not context-free. For the other case, consider the grammar in the proof of Lemma 3.1

$$G_1 = (\{a, b, c\}, \{abc\}, (\{b^n \mid n \geq 1\}, \{(ab, c)\})).$$

As it is easy to see, if we intersect the language generated by G_1 in the local mode with the regular language $a^+b^+c^+$ we get

$$L_{loc}(G_1) \cap a^+b^+c^+ = \{a^nb^nc^n \mid n \geq 1\}.$$

Indeed, by intersecting with $a^+b^+c^+$, we keep only those words from $L_{loc}(G_1)$ for which the local derivation is also maximal. Now, the equality follows from the proof of Lemma 3.1. As the right-hand member is a non-context-free language and the family CF is closed under intersection with regular sets, the result follows.

□

Theorem 5.3 *Any regular language is the image by a weak coding of a language in the family* $\mathcal{LC}_{Mloc}(REG)$.

Proof. Suppose that $L \in REG$ and consider a regular grammar

$$G_1 = (V_N, V_T, S, P)$$

which generates L. We construct the following contextual grammar

$$G_2 = (V_N \cup V_T, \{S\}, (S_X, C_X)_{X \in V_N}),$$

where, for each nonterminal symbol $X \in V_N$, we put

$$\begin{aligned} S_X &= V_N^*\{X\}, \\ C_X &= \ \{(a, Y) \mid X \longrightarrow aY \in P, Y \in V_N, a \in V_T\} \\ &\quad \cup \{(a, \lambda) \mid X \longrightarrow a \in P, a \in V_T\}. \end{aligned}$$

Consider also the weak coding

$$\begin{aligned} \varphi : \ & V_N \cup V_T \longrightarrow V_T \cup \{\lambda\}, \\ & \varphi(X) = \lambda, \text{ for any } X \in V_N, \\ & \varphi(a) = a, \text{ for any } a \in V_T. \end{aligned}$$

We claim that

$$L = \varphi(L_{Mloc}(G_2)).$$

Suppose that $w = a_1 a_2 \ldots a_k \in L, k \geq 0, a_1, \ldots, a_k \in V_T$. Then there must be a derivation in G_1 yielding w

$$S \overset{G_1}{\Longrightarrow} a_1 X_1 \overset{G_1}{\Longrightarrow} a_1 a_2 X_2 \overset{G_1}{\Longrightarrow} \cdots \overset{G_1}{\Longrightarrow} a_1 a_2 \ldots a_{k-1} X_{k-1} \overset{G_1}{\Longrightarrow} a_1 a_2 \ldots a_k.$$

The productions used are $S \longrightarrow a_1 X$, $X_i \longrightarrow a_{i+1} X_{i+1}, 1 \leq i \leq k-2$, $X_k \longrightarrow a_k$, and we can perform in G_2

$$\begin{aligned} S \overset{G_2}{\Longrightarrow} a_1 S X_1 = a_1(S X_1) \overset{G_2}{\Longrightarrow}_{Mloc} a_1 a_2 S X_1 X_2 = a_1 a_2 (S X_1 X_2) \overset{G_2}{\Longrightarrow}_{Mloc} \cdots \\ \overset{G_2}{\Longrightarrow}_{Mloc} a_1 a_2 \ldots a_{k-1} S X_1 X_2 \ldots X_{k-1} = \\ = a_1 a_2 \ldots a_{k-1} (S X_1 X_2 \ldots X_{k-1}) \overset{G_2}{\Longrightarrow}_{Mloc} a_1 a_2 \ldots a_k S X_1 X_2 \ldots X_{k-1}. \end{aligned}$$

Now, obviously,

$$\varphi(a_1 a_2 \ldots a_k S X_1 X_2 \ldots X_{k-1}) = a_1 a_2 \ldots a_k.$$

The other inclusion can be proved in a similar way and the proof is completed.

□

Further research concerning the generative power of these grammars could be done. We mention only a few problems left *open* here:

1. What is the relation between the families $\mathcal{LC}_\alpha(FIN), \alpha \in \{loc, Mloc\}$, and CF ?

2. Is the family REG included in any of the families $\mathcal{LC}_\alpha(F)$, $\alpha \in \{loc, Mloc\}$, $F \in \{FIN, REG\}$?

3. What about some homomorphic representations of recursively enumerable languages by local contextual languages, as the one in [EPR2] ?

6 Final remarks

We have tried to define a class of mildly context-sensitive contextual languages. We have imposed the following two restrictions on the basic model of internal contextual grammars: firstly, in any derivation, after applying a context, further contexts can be added only inside of or at most adjacent to the previous one and, secondly, the word selecting the context must be maximal with respect to the subword order.

We have proved that all the three basic non-context-free constructions in natural languages, multiple agreements, crossed agreements, and duplication, can be realized using this type of grammars. We have proved also that all languages considered are parsable in polynomial time.

The main *open problem* is that of semilinearity of the languages in our most important family, that of languages generated by local contextual grammars with regular choice, working in the maximal mode of derivation. We believe that the answer is affirmative. This conjecture is supported by the observation that the usual techniques used so far for finding non-semilinear contextual languages (see [EIPRS], [Il], and [MMP2]) do not work here.

If it turns out that the languages mentioned are semilinear, then the first class of mildly context-sensitive contextual languages has been found. Otherwise, one has to look for different types of contextual grammars or to impose further restrictions on the ones investigated here. We hope to return on this topic later.

Acknowledgement

I would like to thank my supervisor, Prof. Arto Salomaa, for his continuous encouragement and advice.

References

[EIPRS] A. Ehrenfeucht, L. Ilie, Gh. Păun, G. Rozenberg, A. Salomaa, On the generative capacity of certain classes of contextual grammars, in vol *Mathematical Linguistics and Related Topics* (Gh. Păun, ed.), Ed. Academiei, Bucharest, 1995, 105 – 118.

[EPR1] A. Ehrenfeucht, Gh. Păun, G. Rozenberg, Normal forms for contextual grammars, in vol *Mathematical Aspects of Natural and Formal Languages* (Gh. Păun, ed.), World Sci. Publ., Singapore, 1994, 79 – 96.

[EPR2] A. Ehrenfeucht, Gh. Păun, G. Rozenberg, On representing recursively enumerable languages by internal contextual languages, *Theoretical Computer Science*, to appear.

[EPR3] A. Ehrenfeucht, Gh. Păun, G. Rozenberg, Contextual grammars, in *Handbook of Formal Languages*, (G. Rozenberg, A. Salomaa, eds.), in preparation.

[HU] J. E. Hopcroft, J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, 1979.

[Il] L. Ilie, A non-semilinear language generated by an internal contextual grammar with finite selection, *Ann. Univ. Buc., Matem.-Inform. Series*, XIV, 1(1996), 63 – 70.

[M1] S. Marcus, *Algebraic Linguistics. Analytical Models*, Academic Press, New York, 1967.

[M2] S. Marcus, Contextual grammars, *Rev. Roum. Math. Pures Appl.*, 14, 10 (1969), 1525 – 1534.

[M3] S. Marcus, Deux types nouveaux de grammaires gnratives, *Cah. Ling. Th. Appl.* 6 (1969), 69 – 74.

[MMP1] S. Marcus, C. Martin-Vide, Gh. Păun, On the power of internal contextual grammars with maximal use of selectors, *Fourth Meeting on Mathematics of Language*, Philadelphia, 1995.

[MMP2] S. Marcus, C. Martin-Vide, Gh. Păun, On internal contextual grammars with maximal use of selectors, manuscript.

[MMMP] C. Martin-Vide, A. Mateescu, J. Miguel-Verges, Gh. Păun, Internal contextual grammars: minimal, maximal, and scattered use of selectors, *Bisfai '95 Conf. on Natural Languages and AI* (M. Kappel, E. Shamir, eds.), Jerusalem, 1995, 132 – 142.

[Pa] C. H. Papadimitriou, *Computational complexity*, Addison-Wesley, 1994.

[P1] Gh. Păun, *Contextual grammars*, Ed. Academiei, Bucharest, 1982.

[P2] Gh. Păun, Marcus contextual grammars. After 25 years, *Bulletin EATCS*, February 1994, 263 – 273.

[PN] Gh. Păun, X. M. Nguyen, On the inner contextual grammars, *Rev. Roum. Math. Pures Appl.*, 25, 4 (1980), 641 – 651.

[PRS1] Gh. Păun, G. Rozenberg, A. Salomaa, Contextual grammars: erasing, determinism, one-side contexts, *Proc. of Developments in Language Theory Symp.*, Turku, July 1993, 370 – 388.

[PRS2] Gh. Păun, G. Rozenberg, A. Salomaa, Contextual grammars: parallelism and blocking of derivations, *Fundamenta Inform.*, 25(1996), 381 – 397.

[PRS3] Gh. Păun, G. Rozenberg, A. Salomaa, Marcus contextual grammars: Modularity and leftmost derivation, in vol. *Mathematical Aspects of Natural and Formal Languages* (Gh. Păun, ed.), World Sci. Publ., Singapore, 1994, 375 – 392.

[Sa] A. Salomaa, *Formal Languages*, Academic Press, New York, 1973.

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi

University of Turku
- **Department of Mathematical Sciences**

Åbo Akademi University
- **Department of Computer Science**
- **Institute for Advanced Management Systems Research**

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**