# Some Recent Results on Contextual Languages

**Lucian Ilie**

Turku Centre for Computer Science
TUCS Technical Report No 96
February 1997
ISBN 951-650-966-5
ISSN 1239-1891

**Abstract**

We make a brief survey of some of the most important recent results in the area of contextual languages. We are mostly concerned with new classes of contextual languages more appropriate from natural languages point of view, the concept of mildly context-sensitivity playing a central role here, and also with some other topics, such as computational complexity and ambiguity. Some open problems are also mentioned.

# 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. For details, the reader is referred to the recent survey [EPR2]. For the most comprehensive study on the subject, the forthcoming book by Păun, [Pa1], is the best reference. Discussions about the motivations of contextual grammars can be found also in [M3].

In this paper, which can be viewed as a continuation of [Pa2], we intend to make a brief survey of some of the most important recent results in the area of contextual languages. We are mostly concerned with new classes of contextual languages more appropriate from natural languages point of view, the concept of mildly context-sensitivity playing a central role here, and also with some other topics, such as computational complexity and ambiguity. Some open problems are also mentioned.

# 2 Basic classes of contextual languages

This section contains the definitions of the basic classes of contextual languages.

For an alphabet $\Sigma$, we denote by $\Sigma^*$ the free monoid generated by $\Sigma$, by $\lambda$ its identity, and $\Sigma^+ = \Sigma^* - \{\lambda\}$. The families of finite, regular, linear, context-free, and context-sensitive languges are denoted by $FIN$, $REG$, $LIN$, $CF$, and $CS$, respectively. For further notions of formal language theory we refer to [HU] and [Sa].

A *contextual grammar* is a construct

$$G = (\Sigma, A, (S_1, C_1), (S_2, C_2), \ldots, (S_k, C_k)),$$

for some $k \geq 1$, where $\Sigma$ is an alphabet, $A \subset \Sigma^*$ is a finite set, called the set of *axioms*, $S_i \subseteq \Sigma^*, 1 \leq i \leq k$, are the sets of *selectors*, and $C_i \subset \Sigma^* \times \Sigma^*, C_i$

finite, $1 \leq i \leq k$, are the sets of *contexts*. There are two basic modes of derivation as follows. For two words $x, y \in \Sigma^*$, we have

- the *external* mode of derivation:

$$x \Longrightarrow_{ex} y \text{ iff } y = uxv, x \in S_i, (u, v) \in C_i, \text{ for some } 1 \leq i \leq k;$$

- the *internal* mode of derivation:

$$x \Longrightarrow_{in} y \text{ iff } x = x_1x_2x_3, y = x_1ux_2vx_3, x_2 \in S_i, (u, v) \in C_i, \text{ for some } 1 \leq i \leq k.$$

Notice that the external mode of derivation is essentially the one initially introduced by Marcus, while the internal mode, which is a generalization of the external one (take $x_1 = x_3 = \lambda$ in the definition) was introduced in [PN].

The language generated by $G$ with respect to each of the two modes of derivation is:

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

for $\alpha \in \{ex, in\}$, where $\Longrightarrow_\alpha^*$ denotes the reflexive-transitive closure of $\Longrightarrow_\alpha$.

If the sets $S_1, S_2, \ldots, S_k$ are languages in a given family $F$, then $G$ is said to be *with $F$ choice*. The family of languages generated by contextual grammars with $F$ choice in the mode $\alpha$ of derivation is denoted by $\mathcal{CL}_\alpha(F)$. Let us further remark that the languages generated in $\alpha = ex$ mode are called *external* contextual languages, while those generated using $\alpha = in$ are called *internal*.

A grammar $G$ as above such that $k = 1$ and $C_1 = \Sigma^*$ is called *without choice*, since any context in $C_1$ can be added to any string in $\Sigma^*$. The families of languages generated by contextual grammars without choice are denoted, for the two modes of derivation, by $\mathcal{CL}_{ex}$ and $\mathcal{CL}_{in}$, respectively.

**Examples**

1. For the grammar without choice

$$G_1 = (\{a, b\}, \{\lambda\}, \{(a, b)\}),$$

it is not too difficult to see that the languages generated in the two modes of derivation are:

$$L_{ex}(G_1) = \{a^nb^n \mid n \geq 0\},$$
$$L_{in}(G_2) = Dyck_{\{a,b\}},$$

where $Dyck_{\{a,b\}}$ is the Dyck language over the alphabet $\{a, b\}$.

2. An example a little bit more complicated is given below; the grammar is also without choice:

$$G_2 = (\{a, b, c\}, \{\lambda\}, \{(\alpha\beta, \gamma), (\alpha, \beta\gamma) \mid \{\alpha, \beta, \gamma\} = \{a, b, c\}\}).$$

The internal language generated by $G_2$ is

$$L_{in}(G_2) = \{x \in \{a, b, c\}^* \mid |x|_a = |x|_b = |x|_c\}.$$

3. Our last example is an interesting grammar with regular choice.

$$G_3 = (\{a, b, c, d\}, \{d\}, (b^*d, \{(ab, c)\})).$$

The internal language is now

$$L_{in}(G_3) = \{wdc^n \mid n \geq 0, w \in \{a, b\}^*, |w|_a = |w|_b = n, \text{ and } |x|_a \geq |x|_b, \text{ for any } x \in Pref(w)\}.$$

The relations among the basic families of contextual languages and the families in the Chomsky hierarchy are represented in Figure 1, where an arrow from a family $F_1$ to a family $F_2$ indicates a strict inclusion $F_1 \subset F_2$ and any two non-connected families are incomparable.

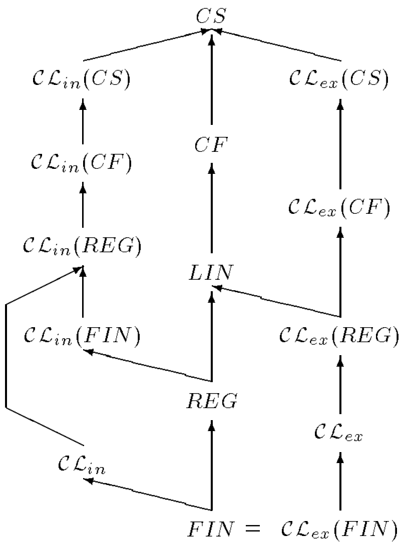


Figure 1: Basic families of contextual languages

## 3 Mildly context-sensitivity

After long debates started in sixties, the linguists seem to finally agree that the natural languages are not context-free; several languages (Bambara dialect, [Cu], Swiss German, [Sh], etc.) contain convincing non-context-free features and this is probably true for most natural languages. Thus, in order to obtain formal grammars aiming to model (specific constructions in) natural languages, we have to look for classes of grammars able to generate non-context-free languages. On the othe hand, they should not be too powerful, in the sense that they should not be able to generate languages without any linguistical relevance, such as, for instance, $\{a^{2^n} \mid n \geq 1\}$, as explicitely stated in [RRF]. More specifically, the idea of keeping the generative power under control has led to the notion of *mildly context-sensitive* families of languages, see [Jo]. The properties that characterize such families are the followings:

1. They contain the three basic non-context-free constructions in natural languages, that is,

- *multiple agreements*:

$$L_1 = \{a^n b^n c^n \mid n \geq 1\},$$

- *crossed agreements*:

$$L_2 = \{a^n b^m c^n d^m \mid n, m \geq 1\},$$

- *duplication*:

$$L_3 = \{wcw \mid w \in \{a, b\}^*\};$$

2. All their languages are *polynomial time parsable*,

3. They contain *semilinear* languages only.

Sometimes, the last property is considered to be too strong and it is replaced by the following weaker one:

3'. All languages have the *bounded growth property*, that is, there are not arbitrarily big gaps in the length set of any language in the respective family. More formally, this means that, for any infinite language $L$ in such a family, there is a constant depending on $L$ only, say $k_L$, such that, for any $n \geq 1$, if $L$ contains words of length $n$, then it contains also words of some length between $n + 1$ and $n + k_L$.

It is very important to notice that, due to the fact that in a derivation in a contextual grammar all sentential forms belong to the generated language, all contextual languages have the bounded growth property.

In what concerns the semilinearity, the basic classes of contextual grammars are able to generate non-semilinear languages, as we shall see in section 4. Also, the basic classes of contextual grammars cannot generate the basic non-context-free constructions above. Aiming to overcome these difficulties, some restrictions have been imposed on the basic types of derivations, as we shall see in sections 5 and 6.

# 4 Non-semilinearity in contextual languages

As we have mentioned already in the previous sections, there are basic classes of contextual grammars able to generate non-semilinear languages. We discuss in this section only the case of internal contextual languages, as being much more interesting from this point of view.

The first result about non-semilinearity in internal contextual languages concerns the family of languages generated by internal contextual grammars with regular choice and it was proved in [EIPRS].

**Theorem 4.1** *The family $\mathcal{CL}_{in}(REG)$ contains non-semilinear languages.*

The construction proving this result is given below:

$$G = (\Sigma, A, (S_1, C_1), (S_2, C_2), (S_3, C_3))$$

where

$$\begin{array}{ll}
\Sigma = \{a, b, c, d, e\}, & \\
A = \{dcbaabbabbcd\}, & \\
S_1 = \{dc(baa)^*(baab)^+\}, & C_1 = \{(\lambda, a)\}, \\
S_2 = \{(abb)^+cd\}, & C_1 = \{(b, \lambda)\}, \\
S_3 = \{cba(abb)^+abbc\}, & C_3 = \{(e, e)\}.
\end{array}$$

The basic idea behind this construction is to parse from left to right and conversely the current string doubling the occurrences of a certain symbol. After arbitrarily many such squarings, a new symbol which blocks the squaring is introduced. By intersecting the Parikh set of the language with a linear set which retains only those words containing the blocking symbol, we get exponentially many occurrences of the symbol which was pumped and so a non-semilinear set.

More specifically, in our example, when going from left to right, the context in $C_1$ is used in order to double the number of occurrences of $a$. From right to left, the context in $C_2$ doubles the number of $b$'s. The blocking symbol is here $e$; when the context in $C_3$ is used, no context in $C_1$ or $C_2$ can be applied further. Intersecting the Parikh set of the language $L_{in}(G)$ to keep those words having two occurrences of $e$, we get the set

$$\{2^n + 2, 2^{n+1} + 3, 2, 2, 2) \mid n \geq 0\}$$

which is not semilinear. For details, see [EIPRS].

The following problem was left open in [EIPRS]: is it possible to obtain non-semilinear languages by internal contextual grammars with the choice restricted to finite languages only, that is, does the smallest family of languages generated by internal contextual grammars with choice contain non-semilinear languages? The problem was solved in the affirmative in [Il1].

**Theorem 4.2** *The family $\mathcal{CL}_{in}(FIN)$ contains non-semilinear languages.*

The example in [Il1] is much more complicated than the one presented above and it is beyond the goal of this paper to present it. The idea is also to parse from one end to the other the current sentential form of a grammar and to pump exponentially many occurrences of a certain symbol until another symbol is introduced and blocks the exponentiation. But it is more difficult in the finite choice case since we cannot encompass arbitrarily many symbols, thus being forced to work only locally. The reader is referred to [Il1] for details.

# 5 Maximal use of selectors

As mentioned in the previos sections, the basic classes of contextual grammars are not powerful enough to generate the three basic non-context-free constructions in the definition of the concept of mildly context-sensitivity. In order to achieve such constructions, some attempts have been done lately.

One natural restriction that has been imposed on the use of selectors in [MMMP], is to require some length conditions on the selector to be used, such as minimality or maximality. This means that we can impose the restriction that any time when a context is added around a selector, no factor of the selector (in the minimal case) can be used as a selector, or no word containing the current selector as a factor can be used as a selector (in the maximal case). This restriction can be imposed with respect to the specified pair selectors-contexts or to the whole grammar. We discuss here in some details only the maximal case, as being more interesting.

More formally, for a contextual grammar as in the definition,

$$G = (\Sigma, A, (S_1, C_1), (S_2, C_2), \ldots, (S_k, C_k)),$$

we define, for two words $x, y \in \Sigma^*$, the *maximal* mode of derivation in $G$ as follows. When the maximum is considered with respect to the component currently in use, we have a *local maximum*, that is:

$$x \Longrightarrow_{lM} y \text{ iff } x = x_1x_2x_3, y = x_1ux_2vx_3, \text{ for } x_2 \in S_i, (u,v) \in C_i, 1 \le i \le k$$
$$\text{and for no } x_1', x_2', x_3' \in S_i, x = x_1'x_2'x_3', x_2' \in S_i, x_2 \text{ a factor of } x_2'.$$

The second case is when the maximum is considered with respect to the whole grammar $G$, that is, in the above definition, instead of $x_2' \in S_i$, we have $x_2' \in S_j$, for some $1 \le j \le k$. In this case we say that it is a *global maximum* and we denote the derivation relation by

$$x \Longrightarrow_{gM} y.$$

The reader should be careful not to make a confusion between the local maximum here and the local mode of derivation, which is another type of restriction imposed on the basic derivation in contextual grammars, and which will be discussed in section 6. Also the notion of maximal local is present there. In order to reduce the danger of confusion, we have changed here the notion of maximal local, as appeared originally in [MMMP], into local maximum.

The relations in Figure 2 among the new families of contextual languages and the basic ones, all of them with the choice restricted to finite or regular languages, have been proved in [MMMP] and [MMP1]. The notations for the new families are clear: $\mathcal{CL}_\alpha(F)$, $\alpha \in \{lM, gM\}$, $F \in \{FIN, REG\}$. In Figure 2, an arrow means an inclusion, not necessarily strict. Two non-connected families are not necessarily incomparable. For the open cases, see [MMMP] and [MMP1].

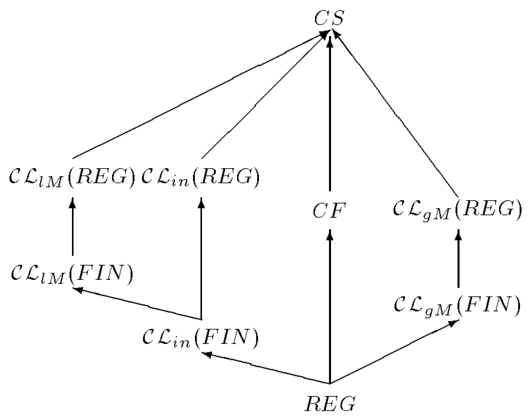


Figure 2: Maximal use of selectors

In what concerns the three basic non-context-free constructions, the following result was proved in [MMP3].

**Theorem 5.1** *Both families $\mathcal{CL}_{lM}(REG)$ and $\mathcal{CL}_{gM}(REG)$ contain all the three basic non-context-free constructions in natural languages.*

We only give the corresponding grammars. The reader may verify that they work as intended.

$$\begin{aligned}G_1 &= (\{a,b,c\},\{abc\},(b^+,\{(a,bc)\})),\\ G_2 &= (\{a,b,c,d\},\{abcd\},(ab^+c,\{(a,c)\}),(bc^+d,\{(b,d)\})),\\ G_3 &= (\{a,b,c\},\{c\},(\{c\}\{a,b\}^*,\{(a,a),(b,b)\})).\end{aligned}$$

While the polynomial time parsability problem is open, the semilinearity problem is completely solved. For the local maximum case, the families $\mathcal{CL}_{lM}(FIN)$ and $\mathcal{CL}_{lM}(REG)$, include the family $\mathcal{CL}_{in}(FIN)$ which contains non-semilinear languages, see Theorem 4.2.

About the global case, it has been proved in [MMP2] that we can still find non-semilinear languages.

**Theorem 5.2** *The family $\mathcal{CL}_{gM}(FIN)$ contains non-semilinear languages.*

Finally, even if the restrictions on the use of the selectors presented in this section are able to generate the three non-context-free constructions, some other attempts have been made in order to find families of contextual languages more appropriate from natural languages point of view, this fact being mainly due to the open problem about the computational complexity. We shall present another restriction in the next section.

# 6 Local contextual languages

In this section we describe another restriction which was imposed on the basic type of derivation in internal contextual grammars in order to obtain contextual languages more appropriate from natural languages point of view. This restriction was introduced and studied in [Il2].

The idea is here to maintain the restriction of the maximality of the selectors to be used but after imposing another restriction, namely, the derivation is restricted in a local way. This means that, 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 previuos ones. The maximality of selectors is imposed only after the local restriction above is fulfilled. Notice that it will follow from the formal definition that the maximum imposed here is global but all results proved in [Il2] hold for the local maximum as well.

More formally, for a contextual grammar as before,

$$G = (\Sigma, A, (S_1, C_1), (S_2, C_2), \ldots, (S_k, C_k)),$$

if we have an usual derivation in $G$, say $z \Longrightarrow x$, for some $z, x \in \Sigma^*$, such that $z = z_1z_2z_3, z_2 \in S_i, (u,v) \in C_i, 1 \leq i \leq k$, and $x = z_1uz_2vz_3$, then a derivation $x \Longrightarrow y$, for $x, y \in \Sigma^*$, is called *local* with respect to the derivation $z \Longrightarrow x$, denoted

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

if and only if we have $u = u'u'', v = v'v'', u', u'', v', v'' \in \Sigma^*, x = x_1x_2x_3$, for $x_1 = z_1u', x_2 = u''z_2v', x_3 = v''z_3, x_2 \in S_j, (w,t) \in C_j, 1 \leq j \leq k$, and $y = x_1wx_2tx_3$.

The derivation is called local because the new places where contexts can be introduced are locally positioned with respect to the old ones.

For illustration, see Figure 3 below.

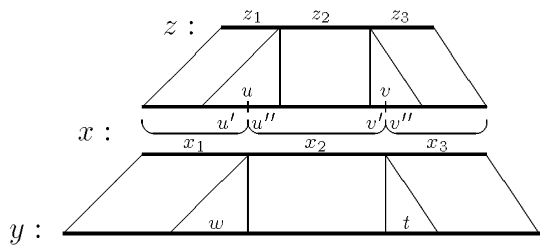


Figure 3: The local derivation

As mentioned, the derivation is further restricted in the maximal way. Thus, if $z \Longrightarrow x$ in $G$, then $x \Longrightarrow y$ is called *maximal local* with respect to $z \Longrightarrow x$ if and only if it is local with respect to $z \Longrightarrow x$ and there is no other local derivation with respect to $z \Longrightarrow x$, say $x \Longrightarrow y'$, such that the decomposition used for $x$, say $x = x_1'x_2'x_3', x_2' \in S_r, 1 \leq r \leq k$, verifies that $x_2$ is a factor of $x_2'$.

This is denoted by

$$z \Longrightarrow x \Longrightarrow_{Mloc} y.$$

As usual, the family of languages generated by contextual grammars working in the maximal local mode of derivation and with $F$ choice is denoted by $\mathcal{CL}_{Mloc}(F)$.

It turns out that these restrictions still allow the new contextual grammars to generate the three non-context-free constructions above.

**Theorem 6.1** *The family $\mathcal{CL}_{Mloc}(REG)$ contains all the three non-context-free constructions in natural languages.*

Three contextual grammars which realize the three languages are rather similar to the ones in the previous section for maximal use of selectors.

$$G_1 = (\{a,b,c\}, \{abc\}, (b^+, \{(ab,c)\})),$$
$$G_2 = (\{a,b,c,d\}, \{abcd\}, (b^+c^+, \{(a,c),(b,d)\})),$$
$$G_3 = (\{a,b,c\}, \{c\}, (\{c\}\{a,b\}^*, \{(a,a),(b,b)\})).$$

While the semilinearity problem for the languages in $\mathcal{CL}_{Mloc}(REG)$ is an open problem, this family has the property that all its languages are parsable in polynomial time. (In fact, as proved in [Il2], they are recognizable in logarithmic space.)

**Theorem 6.2** $\mathcal{CL}_{Mloc}(REG) \subseteq \mathbf{P}$.

Actually, this is the only family of contextual languages for which this fact is known, among all families able to generate the three non-context-free constructions mentioned above.

If we consider the definition of mildly context-sensitive families with the requirements 1,2, and 3' (see section 3), then the family $\mathcal{CL}_{Mloc}(REG)$ is the first family of mildly context-sensitive contextual languages.

The problem here is that the exact generative power of these grammars is not known (see the open problems in the last section). It might be that they are not strong enough. Therefore, we still have to look for even better contextual grammars from linguistics point of view.

# 7 Computational complexity

In this section, we are concerned with the computational complexity of contextual languages. Since the contextual grammars have been introduced with linguistical motivation, it is of crucial importance to study the computational complexity of their generated languages; they cannot be used effectively without having feasible parsing algorithms. Therefore, it is rather surprising that very little is known about this topic. In fact, the only results about the computational complexity of contextual languages are those in [Il2] about local contextual languages and some results about external contextual languages from [Il3], which we briefly present here.

More precisely, the main result proved in [Il3] is that the languages generated by external contextual grammars with context-free choice are polynomial time parsable.

In has been proved first that, for regular choice, the respective languages are parsable in logarithmic space.

**Theorem 7.1** $\mathcal{CL}_{ex}(REG) \subseteq \mathbf{NSPACE}(log)$.

This result is clear since $\mathcal{CL}_{ex}(REG) \subseteq LIN$ (see section 2), but the proof of the context-free case uses this proof.

**Theorem 7.2** $\mathcal{CL}_{ex}(CF) \subseteq \mathbf{P}$.

The idea of the construction of an input preserving Turing machine which recognizes a language generated by an external contextual grammar with regular choice in logarithmic space is by guessing the positions where contexts have been added. The machine has to remember only a finite number of such positions and each of them can be stored in logarithmic space. Since it can be tested whether or not the guessed selector belongs to the guessed regular set of selectors in constant space, the machine works within the required space.

For the context-free case, the above idea does not work any more, since it is not known how to recognize context-free languages in logarithmic space. Roughly, the idea in this case is as follows. The Turing machine will compute deterministically at the very beginning the tables in the Cocke-Younger-Kasami algorithm for the membership problem for context-free languages for each (context-free) set of selectors wih respect to the input. Then, it will guess nondeterministically the places where the contexts have been added, as in the regular case. Finally, it is shown that it is possible to simulate in polynomial time the nondeterministic part by a deterministic one. All in all, the final algorithm works in polynomial time. For details, see [Il3]. We remark again that most of the problems are open in this area. We shall point out some of the most important ones in section 9.

# 8 Ambiguity

For every class of grammars, a natural and important problem is that of syntactic ambiguity, that is, whether or not there are words in the language generated by the given grammar which have at least two syntactic descriptions. Intuitively, by syntactic description is meant distinct derivation. For Chomsky grammars for instance, the ambiguity problem is very clearly defined: a grammar is ambiguous if and only if there is a word in the generated language which has two distinct derivation trees.

For contextual grammars it is not clear what to take into account when defining ambiguity. In the simplest case, that of external contextual grammars without choice, the ambiguity is defined by going to linear languages. A derivation is described by the axiom and the sequence of contexts used. This is called a *control sequence*. This can be extended to all external contextual grammars, so an external contextual grammar is *ambiguous* if there are words which can be generated using two different control sequences. As usual, if all grammars generating a given language are ambiguous, then the respective language is called *inherently ambiguous*.

As proved in [MMPS], where the ambiguity of contextual grammars has been introduced, the external contextual grammars can specify more precisely some contextual languages than the external contextual grammars without choice, in the sense of the following theorem. Also, the ambiguity problem is undecidable.

**Theorem 8.1** *(i) There are inherently ambiguous languages with respect to external contextual grammars without choice, but all languages in the family $\mathcal{CL}_{ex}$ are unambiguous with respect to external contextual grammars with choice.*

*(ii) The ambiguity of external contextual grammars, with or without choice, is undecidable.*

In what concerns the internal case, it is more difficult to define the ambiguity since there are several things that can be taken into account in the definition. This is why five distinct levels of ambiguity have been introduced in [MMPS]. In order to define them, we need some other definitions.

Consider a contextual grammar

$$G = (\Sigma, A, (S_1, C_1), (S_2, C_2), \ldots, (S_k, C_k))$$

and a derivation from an axiom in $G$

$$\delta: \quad w_1 \Longrightarrow_{in} w_2 \Longrightarrow_{in} \cdots \Longrightarrow_{in} w_m,$$

where $m \geq 1, w_1 \in A, w_i = x_{1i}x_{2i}x_{3i}, x_{1i}, x_{2i}, x_{3i} \in \Sigma^*, w_{i+1} = x_{1i}u_ix_{2i}v_ix_{3i}$, for $x_{2i} \in S_j, (u_i, v_i) \in C_j, 1 \leq j \leq k$, for any $1 \leq i \leq m-1$.

The sequence $w_1, (u_1, v_1), \ldots, (u_{m-1}, v_{m-1})$ is called the *control sequence* of $\delta$. Clearly, the control sequence does not identify the derivation uniquely since the contexts might be adjoined at different positions. If we take into consideration also the positions, that is, the sequence

$$w_1, x_{11}(u_1)x_{21}(v_1)x_{31}, \ldots, x_{1,m-1}(u_{m-1})x_{2,m-1}(v_{m-1})x_{3,m-1},$$

which is called the *description* of $\delta$, then $\delta$ is uniquely identified. An intermediate case is to consider only the contexts and their selectors, without the positions. This is a *complete control sequence.* If we do not consider the order of application but just the multisets of the contexts (and the selectors) used, then the terms *unordered control sequence* and *unordered complete control sequence* should be clear.

The five types of ambiguity, as introduced in [MMPS], are, concisely:

*1-ambiguity* – two unordered control sequences,

*2-ambiguity* – two unordered complete control sequences,

*3-ambiguity* – two control sequences,

*4-ambiguity* – two complete control sequences,

*5-ambiguity* – two descriptions.

One more type of ambiguity, the strongest one, has been introduced in [Il4], namely:

*0-ambiguity* – two axioms.

The relations among the six types of ambiguity in internal contextual grammars defined above are represented in Figure 4. An arrow $i \longrightarrow j$ means that, for any internal contextual grammar $G$, if $G$ is $i$-ambiguous, then it is also $j$-ambiguous.

$$\begin{tikzcd}
 & 2 \arrow[dr] & & \\
0 \arrow[r] & 1 \arrow[u] \arrow[d] & 4 \arrow[r] & 5 \\
 & 3 \arrow[ur] & &
\end{tikzcd}$$

Figure 4: Levels of ambiguity

Several results in [MMPS] are resumed in the following theorem.

**Theorem 8.2** *For any pair* $(i, j) \in \{(5,4),(4,3),(4,2),(3,2),(3,1)\}$*, there are inherently* $i$*-ambiguous languages with respect to contextual grammars with arbitrary choice which are* $j$*-unambiguous with respect to grammars with finite choice.*

The same problem for the pairs $(2,3),(2,1)$, and $(1,0)$ is open. For $(1,0)$, we have only the following result from [Il4] .

**Theorem 8.3** *There are inherently* 1*-ambiguous languages with respect to internal contextual grammars without choice which are* 0*-unambiguous with respect to internal contextual grammars with finite choice as well as without choice.*

About the decidability of the ambiguity problem, we have from [MMPS]

**Theorem 8.4** *The ambiguity problem of any type* $t \in \{1,2,3,4,5\}$ *of internal contextual grammars with finite choice is undecidable.*

Notice that the ambiguity problem has been investigated in [Il4] for deterministic contextual grammars, as well as for grammars with one-letter alphabets.

## 9 Open problems

In the last section of this survey, we point out some of the most important open problems concerning the topics discussed in the previous sections.

**Open problem 9.1** *What is the computational complexity of the languages in the families* $\mathcal{CL}_\alpha(F), \alpha \in \{lM, gM\}, F \in \{FIN, REG\}$*?*

**Open problem 9.2** *Are the languages in the family* $\mathcal{CL}_\alpha(F), \alpha \in \{loc, Mloc\}$, $F \in \{FIN, REG\}$*, semilinear?*

**Open problem 9.3** *Find the places of the families* $\mathcal{CL}_\alpha(F), \alpha \in \{loc, Mloc\}$, $F \in \{FIN, REG\}$*, with respect to the families in Chomsky hierarchy as well as with respect to the other families of contextual languages.*

A very important basic problem is the following one:

**Open problem 9.4** *What is the computational complexity of the languages in the family* $\mathcal{CL}_{in}(FIN)$*? The same for* $\mathcal{CL}_{in}(REG), \mathcal{CL}_{in}(CF)$*.*

This is a particularly interesting problem since it is known that the family $\mathcal{CL}_{in}(FIN)$ contains languages which are not in $ET0L$ or $MAT^\lambda$. The place of $\mathcal{CL}_{in}(FIN)$ in Chomsky hierarchy is depicted in Figure 5.

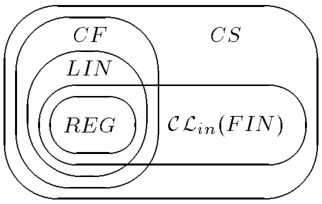


Figure 5: The family $\mathcal{CL}_{in}(FIN)$

About ambiguity, we mentione only the following problem. For further open problems in this area, see [MMPS] and [Il4].

**Open problem 9.5** *For each* $(i, j) \in \{(2, 3), (2, 1), (1, 0)\}$*, are there languages which are inherently i-ambiguous with respect to grammars with arbitrary choice but j-unambiguous with respect to grammars with finite choice?*

## References


[Cu] C. Culy, The complexity of the vocabulary of Bambara, *Linguistics and Philosophy*, 8 (1985) 345–351.

[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, Contextual grammars, in *Handbook of Formal Languages*, (G. Rozenberg, A. Salomaa, eds.), to appear.

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

[Il1] 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.

[Il2] L. Ilie, On computational complexity of contextual languages, *Theoretical Computer Science*, to appear.

[Il3] L. Ilie, The computational complexity of Marcus contextual grammars, submitted, 1996.

[Il4] L. Ilie, On ambiguity in internal contextual languages, *II Intern. Conf. Mathematical Linguistics*, Tarragona, 1996.

[Jo] A. K. Joshi, An introduction to tree adjoining grammars, in *Mathematics of Languages* (A. Manaster Ramer, ed.), John Benjamins, Amsterdam, Philadelphia, 1987, 87–114.

[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.

[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.

[MMP1] S. Marcus, C. Martin-Vide, Gh. Păun, Contextual grammars as generative models of natural languages, *Fourth Meeting on Mathematics of Languages*, MOL4, Philadelphia, 1995.

[MMP2] S. Marcus, C. Martin-Vide, Gh. Păun, On the internal contextual grammars with maximal use of selectors, *8th Conf. Automata and Formal Languages*, Salotarjan, 1996.

[MMP3] S. Marcus, C. Martin-Vide, Gh. Păun, Contextual grammars versus natural languages, *Speach and Computer '96 Conference*, St. Petersburg, 28–33.

[MMPS] C. Martin-Vide, J. Miguel-Verges, Gh. Păun, A. Salomaa, Attempting to define the ambiguity in internal contextual languages, *II Intern. Conf. Mathematical Linguistics*, Tarragona, 1996.

[Pa1] Gh. Păun, *Contextual grammars. From natural languages to contextual languages and back*, manuscript.

[Pa2] Gh. Păun, Marcus contextual grammars. After 25 years, *Bulletin EATCS*, 52(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.

[RRF] W. C. Rounds, A. Manaster Ramer, J. Friedman, Finding natural languages a home in formal language theory, *Mathematics of Languages* (A. Manaster Ramer, ed.), John Benjamins, Amsterdam, Philadelphia, 1987, 87–114.

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

[Sh] S. M. Shieber, Evidence against the context-freeness of natural languages, *Linguistics and Philosophy*, 8 (1985), 333–343.

**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**