# On Internal Contextual Grammars with Subregular Selection Languages

Florin Manea[1]* and Bianca Truthe[2]

[1] Christian-Albrechts-Universität zu Kiel, Institut für Informatik
Christian-Albrechts-Platz 4, D-24098 Kiel, Germany
flm@informatik.uni-kiel.de

[2] Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik
PSF 4120, D-39016 Magdeburg, Germany
truthe@iws.cs.uni-magdeburg.de

**Abstract.** In this paper, we study the power of internal contextual grammars with selection languages from subfamilies of the family of regular languages. If we consider families $\mathcal{F}_n$ which are obtained by restriction to $n$ states or nonterminals or productions or symbols to accept or to generate regular languages, we obtain four infinite hierarchies of the corresponding families of languages generated by internal contextual grammars with selection languages in $\mathcal{F}_n$.

## 1 Introduction

Contextual grammars were introduced by S. Marcus in [4] as a formal model that might be used in the generation of natural languages. The derivation steps consist in adding contexts to given well formed sentences, starting from an initial finite basis. Formally, a context is given by a pair $(u, v)$ of words and the external adding to a word $x$ gives the word $uxv$ whereas the internal adding gives all words $x_1ux_2vx_3$ when $x = x_1x_2x_3$. Obviously, by linguistic motivation, a context can only be added if the words $x$ or $x_2$ satisfy some given conditions. Thus, it is natural to define contextual grammars with selection in a certain family $\mathcal{F}$ of languages, where it is required that $x$ or $x_2$ have to belong to a language of the family $\mathcal{F}$ which is associated with the context. Mostly, the family $\mathcal{F}$ is taken from the families of the Chomsky hierarchy (see [3, 6, 5], and the references therein).

In [1], the study of external contextual grammars with selection in special regular sets was started. Finite, combinational, definite, nilpotent, regular suffix-closed, regular commutative languages and languages of the form $V^*$ for some alphabet $V$ were considered. In [2], the research was continued and new results on the effect of regular commutative, regular circular, definite, regular suffix-closed, ordered, combinational, nilpotent, and union-free selection languages on the generative power of external contextual grammars were obtained. Furthermore, families of regular languages which are defined by restrictions on the resources used

* F. Manea's work is supported by the DFG grant 582014. On leave of absence from: *Faculty of Mathematics and Computer Science, University of Bucharest, Str. Academiei 14, RO-010014 Bucharest, Romania (*flmanea@fmi.unibuc.ro*).*

to generate or to accept them were investigated. As measures, the number of states necessary to accept the regular languages and the number of nonterminals, production rules or symbols needed to generate the regular languages have been considered. In all these cases, infinite hierarchies were obtained.

In the present paper, we continue this line of research and investigate the effect of the number of resources (states, nonterminals, production rules, and symbols) on the generative power of internal contextual grammars. This case seems more complicated than the case of external contextual grammars, as there are two important differences between the way a derivation is conducted in internal grammars and in an external one. First, in the case of internal contextual grammars, the insertion of a context in a sentential form can be done in more than one place, so the derivation becomes, in a sense, non-deterministic; in the case of external grammars, once a context was selected there is at most one way to insert it: wrapped around the sentential form, when this word was in the selection language of the context. Second, if a context can be added internally, then it can be added arbitrarily often (because the subword where the context is wrapped around does not change) which does not necessarily hold for external grammars. However, we are able to obtain infinite hierarchies with respect to the descriptional complexity measures we use, but with different proof techniques.

## 2 Definitions

Throughout the paper, we assume that the reader is familiar with the basic concepts of the theory of automata and formal languages. For details, we refer to [6]. Here we only recall some notation and the definition of contextual grammars with selection which form the central notion of the paper.

Given an alphabet $V$, we denote by $V^*$ and $V^+$ the set of all words and the set of all non-empty words over $V$, respectively. The empty word is denoted by $\lambda$. For a word $w \in V^*$ and a letter $a \in V$, by $|w|$ and $\#_a(w)$ we denote the length of $w$ and the number of occurrences of $a$ in $w$, respectively. The cardinality of a set $A$ is denoted by $\#(A)$.

Let $G = (N, T, P, S)$ be a regular grammar (specified by finite sets $N$ and $T$ of nonterminals and terminals, respectively, a finite set of productions of the form $A \to wB$ or $A \to w$ with $A, B \in N$ and $w \in T^*$ as well as $S \in N$). Further, let $A = (X, Z, z_0, F, \delta)$ be a deterministic finite automaton (specified by sets $X$ and $Z$ of input symbols and states, respectively, an initial state $z_0$, a set $F$ of accepting states, and a transition function $\delta$) and $L$ be a regular language. Then we define

$$\begin{aligned}
\mathit{State}(A) &= \#(Z),\\
\mathit{Var}(G) = \#(N),\ \mathit{Prod}(G) &= \#(P),\ \mathit{Symb}(G) = \sum_{A\to w\in P} (|w|+2),\\
\mathit{State}(L) &= \min\left\{\, \mathit{State}(A) \mid A \text{ is a det. finite automaton accepting } L \,\right\},\\
K(L) &= \min\left\{\, K(G) \mid G \text{ is a reg. grammar for } L \,\right\} (K \in \{\, \mathit{Var}, \mathit{Prod}, \mathit{Symb}\}),
\end{aligned}$$

and, for $K \in \{\, \mathit{State}, \mathit{Var}, \mathit{Prod}, \mathit{Symb} \,\}$, we set

$$REG_n^K = \{\, L \mid L \text{ is a regular language with } K(L) \leq n \,\}\,.$$

**Remark**. We note that if we restricted ourselves to rules of the form $A \to aB$ and $A \to \lambda$ with $A, B \in N$ and $a \in T$, then we would have $\mathit{State}(L) = \mathit{Var}(L)$.

We now introduce the central notion of this paper.

Let $\mathcal{F}$ be a family of languages. A *contextual grammar with selection in* $\mathcal{F}$ is a triple

$$G = (V, \{\, (S_1, C_1), (S_2, C_2), \ldots, (S_n, C_n) \,\}\,, B)$$

where

- $V$ is an alphabet,
- for $1 \leq i \leq n$, $S_i$ is a language over $V$ in $\mathcal{F}$ and $C_i$ is a finite set of pairs $(u, v)$ with $u \in V^*$, $v \in V^*$,
- $B$ is a finite subset of $V^*$.

The set $V$ is called the basic alphabet; the languages $S_i$ and the sets $C_i$, $1 \leq i \leq n$, are called the *selection languages* and the sets of *contexts* of $G$, respectively; the elements of $B$ are called *axioms*.

We now define the internal derivation for contextual grammars with selection.

Let $G = (V, \{\, (S_1, C_1), (S_2, C_2), \ldots, (S_n, C_n) \,\}\,, B)$ be a contextual grammar with selection. A direct *internal derivation step* in $G$ is defined as follows: a word $x$ derives a word $y$ (written as $x \Longrightarrow y$) if and only if there are words $x_1$, $x_2$, $x_3$ with $x_1x_2x_3 = x$ and there is an integer $i$, $1 \leq i \leq n$, such that $x_2 \in S_i$ and $y = x_1ux_2vx_3$ for some pair $(u, v) \in C_i$. Intuitively, we can only wrap a context $(u, v) \in C_i$ around a subword $x_2$ of $x$ if $x_2$ belongs to the corresponding language $S_i$. We call a word of a selection language useful, if it is a subword of a word of the generated language – if it is really selected from wrapping a context around it.

By $\Longrightarrow^*$ we denote the reflexive and transitive closure of $\Longrightarrow$. The *internal language generated by* $G$ is defined as

$$L(G) = \{\, z \mid x \Longrightarrow^* z \text{ for some } x \in B \,\}\,.$$

By $\mathcal{L}(IC, \mathcal{F})$ we denote the family of all internal languages generated by contextual grammars with selection in $\mathcal{F}$. When we speak about contextual grammars in this paper, we mean contextual grammars with internal derivation (also called internal contextual grammars).

*Example 1.* Let $n \geq 1$ and $V = \{a\}$ be a unary alphabet. We set

$$B_n = \left\{\, a^i \mid 1 \leq i \leq n \,\right\}\,,\ U_n = \{\, a^n \,\}^+\,,\ \text{and } L_n = B_n \cup U_n.$$

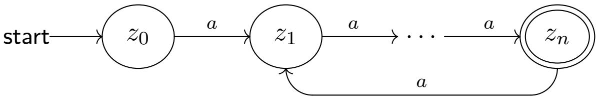
The contextual grammar $G_n = (V, \{\, (U_n, \{(\lambda, a^n)\}) \,\}\,, B_n)$ generates the language $L_n$. This can be seen as follows. The context $a^n$ can be added to a word $w$ if and only if $w$ contains at least $n$ letters. The only axiom to which a context can be added is $a^n$. From this, we get the unique derivation

$$a^n \Longrightarrow a^{2n} \Longrightarrow a^{3n} \Longrightarrow \cdots\,.$$

It is easy to see that the set $U_n$ is accepted by the automaton

$$(V, \{z_0, z_1, \ldots, z_n\}, z_0, \{z_n\}, \delta_n)$$

where the graph



represents the transition function $\delta_n$.

Hence, we have $L_n \in \mathcal{L}(IC, REG^{State}_{n+1})$. ♢

## 3 Selection with Bounded Resources

First we prove that we obtain an infinite hierarchy with respect to the number of states.

**Theorem 2.** *For any natural number $n \geq 1$, we have the proper inclusion*

$$\mathcal{L}(IC, REG^{State}_{n}) \subset \mathcal{L}(IC, REG^{State}_{n+1}).$$

*Proof.* For $n \geq 2$, let

$$B_n = \left\{ a^i \mid 1 \leq i \leq n \right\}, \; U_n = \left\{ a^n \right\}^+, \text{ and } L_n = B_n \cup U_n$$

be the languages from Example 1, where we have shown the relation

$$L_n \in \mathcal{L}(IC, REG^{State}_{n+1}).$$

We now prove that $L_n \notin \mathcal{L}(IC, REG^{State}_{n})$ for $n \geq 2$.

Let $G = (V, \{ (S_1, C_1), (S_2, C_2), \ldots, (S_m, C_m) \}, B)$ be a contextual grammar with $L(G) = L_n$ and where every selection language is an arbitrary regular language.

Let $k' = \max \{ |uv| \mid (u,v) \in C_i, 1 \leq i \leq m \}$, $k'' = \max \{ |z| \mid z \in B \}$, and $k = k' + k''$. Let us consider the word $w = a^{(k+1)n} \in L_n$. Obviously, $w \notin B$ because the length $(k+1)n$ is greater than $k''$. Thus, the word $w$ is obtained from some word $w' = w'_1 w'_2 w'_3 \in L_n$ by adding a context $(u,v) \in C_i$ for some index $i$ with $1 \leq i \leq m$ around the subword $w'_2 \in S_i$. Then $w = w'_1 u w'_2 v w'_3$. For the length of the word $w'$, we obtain

$$|w'| = |w| - |uv| = (k' + k'' + 1)n - |uv| \geq k'(n-1) + k''n + n > n.$$

The word $w'$ belongs to the language $L_n$ and has a length greater than $n$ which implies $w' \in U_n$. Hence, $w' = a^{jn}$ for some $j$ with $2 \leq j \leq k$ and $uv = a^{(k+1-j)n}$. If $S_i$ contains the empty word $\lambda$, then also the word $auv$ belongs to the language $L_n$ (since $a \in L_n$). However, the length $|auv| = 1 + (k+1-j)n$ is greater than $n$ but not a multiple of $n$ which is a contradiction. If $S_i$ contains a non-empty word $z$ that has a length smaller than $n$, then $z$ belongs to the language $L_n$ and also $uzv \in L_n$. But then the length $|uzv| = |z| + (k+1-j)n$ is greater than $n$ but not a multiple of $n$ which is again a contradiction. Hence, the set $S_i$ does not contain a word with a length smaller than $n$. Let $r = \min\left\{\, l \mid a^l \in S_i \,\right\}$. Then $r \geq n$. We set $z_i = a^{r-i}$ for $0 \leq i \leq r$. Then we have the relations $a^i z_i \in S_i$ and $a^j z_i \notin S_i$ for $0 \leq j < i \leq r$ because $a^i z_i = a^r$ and $|a^j z_i| < r$ for $0 \leq j < i \leq r$.

Therefore the words $\lambda, a, a^2, \ldots, a^r$ are pairwise not in the Myhill-Nerode relation. Thus, the minimal deterministic finite automaton accepting $S_i$ has at least $r+1 \geq n+1$ states.

For $n=1$, consider the language

$$L_1 = \{\,\lambda\,\} \cup \{\, w \mid w \in \{a,b\}^+,\ \#_b(w) = 1 \,\}.$$

This language can be generated by the contextual grammar

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

Since the language $\{a,b\}^+$ can be accepted by an automaton with two states, we have the relation

$$L_1 \in \mathcal{L}(IC, REG_2^{State}).$$

For every contextual grammar $G = (V, \{\, (S_1,C_1),(S_2,C_2),\ldots,(S_m,C_m) \,\}, B)$ with $S_i \in REG_1^{State}$ for $1 \leq i \leq m$, we have $S_i = \emptyset$ or $S_i = V^*$ for $1 \leq i \leq m$. If all selection languages are empty, the generated language is the set $B$ of axioms and hence finite. In order to obtain the infinite language $L_1$, one set $S_i$ is equal to $V^*$. Then any context of $C_i$ can be applied to any word of the language $L_1$. If such a context consists of letters $a$ only, we obtain from the empty word a word which does not belong to the language $L_1$ (which has only letters $a$). If such a context contains a letter $b$, we obtain from a word with a $b$ another word with two letters $b$ and hence does not belong to the language $L_1$. Thus, the language $L_1$ does not belong to the family $\mathcal{L}(IC, REG_1^{State})$. □

We now consider the measure *Var*.

**Theorem 3.** *For any natural number $n \geq 0$, we have the proper inclusion*

$$\mathcal{L}(IC, REG_n^{Var}) \subset \mathcal{L}(IC, REG_{n+1}^{Var}).$$

*Proof.* Let $V = \{\, a,b \,\}$. For each natural number $n \geq 1$, we consider a language $L_n$ which consists of words formed by $2n$ blocks of letters $a$, ended by the letter $b$ each, where the block lengths coincide in a crossed agreement manner:

$$L_n = \{\, a^{p_1}ba^{p_2}b\ldots a^{p_n}ba^{p_1}ba^{p_2}b\ldots a^{p_n}b \mid p_i \geq 1,\ 1 \leq i \leq n \,\}.$$

A contextual grammar generating the language $L_n$ is

$$G_n = (V, \{\, (S, \{\, (a,a) \,\}) \,\}, \{\, w \mid w = abab\ldots ab,\ |w| = 4n \,\})$$

with

$$S = (\{b\}\{a\}^+)^{n-1}\{b\}\{a\}.$$

The selection language $S$ can be generated by the following regular grammar

$$G = (\{\, N_1, N_2, \ldots, N_n \,\}, V, P, N_1)$$

with the rules

$$\begin{aligned} &N_1 \to baN_2,\\ &N_k \to aN_k \mid baN_{k+1} \text{ for } 2 \le k \le n-1,\\ &N_n \to aN_n \mid ba. \end{aligned}$$

Hence, we obtain $L_n \in \mathcal{L}(IC, REG_n^{Var})$ for all numbers $n \geq 1$.

It remains to show that, for any $n \geq 1$, the language $L_n$ cannot be generated by a contextual grammar where, for generating the selection languages, less than $n$ nonterminal symbols are sufficient.

Let $n \geq 1$ and $G'_n = (V, \{\, (S_1, C_1), (S_2, C_2), \ldots, (S_m, C_m) \,\}, B)$ be a contextual grammar with $L(G'_n) = L_n$ and where every selection language is an arbitrary regular language. Since the language $L_n$ is infinite, there are words that do not belong to the set $B$ of axioms and which are therefore generated by adding a context. Whenever a context $(u, v) \in C_i$ for some $i$ with $1 \leq i \leq m$ can be wrapped around a subword $z$ of a word $z_1 z z_2 \in L_n$, it can be added infinitely often because the subword $z$ remains unchanged. Hence, the letter $b$ cannot occur in any context because otherwise words with an unbounded number of the letter $b$ could be generated. Thus, every context $(u, v)$ of the system consists of letters $a$ only and, moreover, $u = v = a^{n_c}$ for some number $n_c$. Otherwise, a word would be generated which does not belong to the language $L_n$. Any useful word of a selection language of the system belongs to the set

$$F = \{a\}^*(\{b\}\{a\}^+)^n,$$

otherwise the application of a context would yield a word which does not belong to the language $L_n$. We consider only the useful words of the selection languages. All these words contain exactly $n$ letters $b$ and at least a letter $a$ after each $b$. The maximal word which is between two letters $b$ and consists of letters $a$ only is called an inner $a$-block.

There is a selection language $S_i$ $(1 \leq i \leq m)$ where all inner $a$-blocks are unbounded. Suppose that this is not the case. Then in every selection language, at least one inner $a$-block is bounded. Let $l$ be the maximal length of bounded inner $a$-blocks over all selection languages $S_i$ $(1 \leq i \leq m)$ and the set $B$ of axioms. Let $k = l + 1$. We consider the word $w = a^k b a^k b \ldots a^k b \in L_n$. By the choice of $k$, the word $w$ is not an axiom. Hence, it is derived from a word $w' \in L_n$ by wrapping a context $(a^{n_c}, a^{n_c})$ around a subword $w'' = a^{q_1}(ba^k)^{n-1}ba^{q_2}$ where $q_1 \geq 0$ and $q_2 \geq 1$. This word $w''$, however, does not belong to any selection language (because in every word of any selection language, at least one of the inner $a$-blocks has a length less than $k$). This contradiction implies that there is a selection language $S_i$ $(1 \leq i \leq m)$ where all inner $a$-blocks are unbounded.

When generating such a language $S_i$, a second nonterminal must appear before the second letter $b$ is generated because otherwise the first inner $a$-block would be bounded or the number of letters $b$ would be unbounded and a word with $n+1$ letters $b$ would be generated (because a derivation of a word $w \in S_i$ would exist with the form $A \Longrightarrow a^q ba^r A \Longrightarrow^* w \in S_i$ and hence, also the derivation $A \Longrightarrow^* a^q ba^{r+q} ba^r A \Longrightarrow^* w'$ would exist which yields a word $w'$ with $\#_b(w') = \#_b(w) + 1$). Such a word should not be in the language $S_i$ because otherwise the lengths of two $a$-blocks not corresponding to each other would be increased but not the corresponding ones. By induction, one obtains that an $n$-th nonterminal must appear before the $n$-th letter $b$ is generated.

From this follows, that $n-1$ nonterminals are not sufficient.

As a result, we have $L_n \in \mathcal{L}(IC, REG_n^{Var}) \setminus \mathcal{L}(IC, REG_{n-1}^{Var})$ for any natural number $n \geq 1$. □

As consequences from the previous theorem, we obtain also infinite hierarchies with respect to the number of production rules and the number of symbols. However, to show that the inclusions $\mathcal{L}(IC, REG_n^K) \subseteq \mathcal{L}(IC, REG_{n+1}^K)$ with $K \in \{\, Prod, Symb \,\}$ are proper for each number $n \geq n_0$ for some start number $n_0$, we need further investigations.

We start with the complexity measure *Prod*. Let us consider a generalization of the languages $L_n$ for $n \geq 1$ used in the proof of the previous theorem.

Let $m \geq 1$, $A_m = \{\, a_1, \ldots, a_m \,\}$, and $V = A_m \cup \{b\}$. The languages $L_n$ consist again of words formed by $2n$ $a$-blocks ended by the letter $b$ each and where the block lengths coincide in a crossed agreement manner. However, an $a$-block now consists of letters from the set $A_m$ instead of the single letter $a$ only.

Formally, we define for $n \geq 1$ the languages

$$L_n^{(m)} = \{\, w_1 b w_2 b \ldots w_n b w_{n+1} b w_{n+2} b \ldots w_{2n} b \mid \\ w_i, w_{n+i} \in A_m^+,\ |w_i| = |w_{n+i}|,\ 1 \leq i \leq n \,\}.$$

*Example 4.* Let us consider the languages $L_n^{(m)}$ for $n = 1$, $n = 2$, and $n = 3$. Then

$$\begin{aligned} L_1^{(m)} &= \left\{\, w_1 b w_2 b \mid \{w_1, w_2\} \subset A_m^+,\ |w_1| = |w_2| \,\right\}, \\ L_2^{(m)} &= \left\{\, w_1 b w_2 b w_3 b w_4 b \mid \{w_1, w_2, w_3, w_4\} \subset A_m^+, |w_1| = |w_3|, |w_2| = |w_4| \,\right\}, \\ L_3^{(m)} &= \left\{\, w_1 b w_2 b w_3 b w_4 b w_5 b w_6 b \mid \{w_i, w_{3+i}\} \subset A_m^+, |w_i| = |w_{3+i}|, 1 \leq i \leq 3 \,\right\}. \end{aligned}$$

The following contextual grammar generates the language $L_1^{(m)}$:

$$G_1^{(m)} = (V, \mathcal{S}, B)$$

with

$$B = \{\, xbyb \mid \{x, y\} \subseteq A_m \,\} \text{ and } \mathcal{S} = \{\, (S_x, C) \mid x \in A_m \,\}$$

where

$$S_x = \{\, bx \,\} \text{ and } C = \{\, (x, y) \mid \{x, y\} \subseteq A_m \,\}.$$

For generating such a selection language $S_x$, only one rule is necessary. Hence, $L_1^{(m)} \in \mathcal{L}(IC, REG_1^{Prod})$ for each number $m \geq 1$.

For generating the language $L_2^{(m)}$ by a contextual grammar, we first increase the lengths of the first and third $a$-blocks and then the lengths of the remaining blocks.

The language $L_2^{(m)}$ is generated by the contextual grammar

$$G_2^{(m)} = (V, \mathcal{S}, B)$$

with

$$\begin{aligned} B &= \{\, x_1bx_2bx_3bx_4b \mid x_i \in A_m, 1 \le i \le 4 \,\}\,, \\ \mathcal{S} &= \{\, (S_{i,j}, C) \mid 1 \le i, j \le m \,\} \end{aligned}$$

where

$$\begin{aligned} S_{i,j} &= \{\, ba_iwba_j \mid w \in A_m^* \,\}\,, \\ C &= \{\, (x,y) \mid \{x,y\} \subseteq A_m \,\}\,. \end{aligned}$$

Each selection language $S_{i,j}$ can be generated by $m+2$ rules ($N_1 \to ba_iN_2$, $N_2 \to xN_2$ for any $x \in A_m$ and $N_2 \to ba_j$). Hence, $L_2^{(m)} \in \mathcal{L}(IC, REG_{m+2}^{Prod})$ for each number $m \ge 1$.

A contextual grammar for the language $L_3^{(m)}$ for an arbitrary number $m \ge 1$ is

$$G_3^{(m)} = (V, \mathcal{S}, B)$$

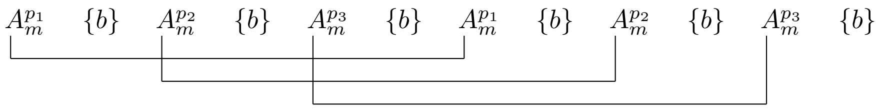
with

$$\begin{aligned} B &= \{\, x_1bx_2bx_3bx_4bx_5bx_6b \mid x_i \in A_m, 1 \le i \le 6 \,\}\,, \\ \mathcal{S} &= \{\, (S_{i,j,k}, C) \mid 1 \le i, j, k \le m \,\} \end{aligned}$$

where

$$\begin{aligned} S_{i,j,k} &= \{\, ba_iw_1ba_jw_2ba_k \mid \{w_1, w_2\} \subset A_m^* \,\}\,, \\ C &= \{\, (x,y) \mid \{x,y\} \subseteq A_m \,\}\,. \end{aligned}$$

The following picture shows where contexts can be added:



A grammar generating a selection language $S_{i,j,k}$ with $1 \le i, j, k \le m$ is $G_{i,j,k} = (\{N_1, N_2, N_3\}, V, P, N_1)$ with the rules

$$\begin{aligned} &N_1 \to ba_iN_2, \\ &N_2 \to xN_2 \mid ba_jN_3, \\ &N_3 \to xN_3 \mid ba_k \end{aligned}$$

for $x \in A_m$. Hence, $L_3^{(m)} \in \mathcal{L}(IC, REG_{2(m+1)+1}^{Prod})$ for each number $m \ge 1$. $\diamondsuit$

If the number of rules of a regular grammar is equal to zero, the language generated is empty. Hence, the class $\mathcal{L}(IC, REG_0^{Prod})$ contains only finite languages. Since $L_1^{(m)} \in \mathcal{L}(IC, REG_1^{Prod})$ (as shown in the previous example) and $L_1^{(m)} \notin \mathcal{L}(IC, REG_0^{Prod})$ (because the language is infinite), we have the first proper inclusion.

**Corollary 5.** *The proper inclusion*

$$\mathcal{L}(IC, REG_0^{Prod}) \subset \mathcal{L}(IC, REG_1^{Prod})$$

*holds.*

Let us now have a closer look at the relation between the $L_n^{(m)}$ languages and the *Prod*-hierarchy.

For $n = 3$, we have found in the Example 4 that $(m+1)(n-1)+1$ rules is an upper bound for generating the selection languages. This bound also holds for $n = 1$ and $n = 2$, as one can see from the example.

We now investigate the languages $L_n^{(m)}$ for $n \geq 4$ and prove the tightness of the bound for $n \geq 1$.

**Lemma 6.** *Let $m \geq 1$, $n \geq 1$, and*

$$L_n^{(m)} = \{\ w_1bw_2b \dots w_nbw_{n+1}bw_{n+2}b \dots w_{2n}b \mid \\ w_i, w_{n+i} \in A_m^+,\ |w_i| = |w_{n+i}|,\ 1 \leq i \leq n\ \}.$$

*Then*

$$L_n^{(m)} = \mathcal{L}(IC, REG_{(m+1)(n-1)+1}^{Prod}) \setminus \mathcal{L}(IC, REG_{(m+1)(n-1)}^{Prod})$$

*holds.*

*Proof.* A contextual grammar for the language $L_n^{(m)}$ for arbitrary numbers $m \geq 1$ and $n \geq 1$ is

$$G_n^{(m)} = (V, \mathcal{S}, B)$$

with

$$B = \{\ x_1bx_2b \dots x_{2n}b \mid x_i \in A_m, 1 \leq i \leq 2n\ \},$$
$$\mathcal{S} = \{\ (S_{i_1,i_2,\dots,i_n}, C) \mid 1 \leq i_j \leq m, 1 \leq j \leq n\ \}$$

where

$$S_{i_1,i_2,\dots,i_n} = \left\{\ ba_{i_1}w_1ba_{i_2}w_2 \dots ba_{i_{n-1}}w_{n-1}ba_{i_n} \mid w_i \in A_m^*, 1 \leq i \leq n_1\ \right\},$$
$$C = \{\ (x, y) \mid \{x, y\} \subseteq A_m\ \}.$$

A grammar generating a selection language $S_{i_1,i_2,\dots,i_n}$ with $1 \leq i_j \leq m$ for $1 \leq j \leq n$ is $G_{i_1,i_2,\dots,i_n} = (\{N_1, N_2, \dots, N_n\}, V, P, N_1)$ with the rules

$$N_1 \to ba_{i_1}N_2,$$
$$N_k \to xN_k \mid ba_{i_k}N_{k+1} \text{ for } 2 \leq k \leq n-1,$$
$$N_n \to xN_n \mid ba_{i_n}$$

for $x \in A_m$. Hence, $L_n^{(m)} \in \mathcal{L}(IC, REG^{Prod}_{(m+1)(n-1)+1})$ for each number $m \geq 1$ and $n \geq 2$ and, together with the considerations in the beginning, also for $n \geq 1$.

We now show that, for any $m \geq 1$ and $n \geq 1$, the language $L_n^{(m)}$ cannot be generated by a contextual grammar where, for generating every selection language, at most $(m+1)(n-1)$ nonterminal symbols are sufficient.

Let $n \geq 1$, $m \geq 1$, and $H_n^{(m)} = (V, \{\,(S_1, C_1), (S_2, C_2), \ldots, (S_m, C_m)\,\}\,, B)$ be a contextual grammar with $L(H_n^{(m)}) = L_n^{(m)}$ and where every selection language is an arbitrary regular language. Since the language $L_n^{(m)}$ is infinite, there are words that do not belong to the set $B$ of axioms and which are therefore generated by adding a context. Whenever a context $(u,v) \in C_i$ for some $i$ with $1 \leq i \leq m$ can be wrapped around a subword $z$ of a word $z_1 z z_2 \in L_n^{(m)}$, it can be added infinitely often because the subword $z$ remains unchanged. Hence, the letter $b$ cannot occur in any context because otherwise words with an unbounded number of the letter $b$ could be generated. Thus, every context $(u,v)$ of the system consists of letters from the alphabet $A_m$ only and, moreover, $|u| = |v|$. Otherwise, a word would be generated which does not belong to the language $L_n^{(m)}$. Any useful word of a selection language of the system belongs to the set

$$F = A_m^*(\{b\}A_m^+)^n,$$

otherwise the application of a context would yield a word which does not belong to the language $L_n^{(m)}$. We consider only the useful words of the selection languages. All these words contain exactly $n$ letters $b$ and at least a letter of the set $A_m$ after each $b$. The maximal word which is between two letters $b$ and consists of letters from the set $A_m$ only is called an inner $a$-block.

There is a selection language $S_i$ $(1 \leq i \leq m)$ where in every inner $a$-block the number of occurrences of each letter of the alphabet $A_m$ is unbounded. Suppose that this is not the case. Then for every selection language, there is a letter $a_j \in A_m$ such that at least one inner $a$-block contains a bounded number of the letter $a_j$. Let $l$ be the maximal number of each letter in an inner $a$-block (if the number of occurrences of this letter is bounded in this block) over all selection languages $S_i$ $(1 \leq i \leq m)$ and the set $B$ of axioms. Let $k = l+1$ and $w_a = a_1^k a_2^k \ldots a_m^k$. We consider the word $w = w_a b w_a b \ldots w_a b \in L_n$. By the choice of $k$, the word $w$ is not an axiom. Hence, it is derived from a word $w' \in L_n$ by wrapping a context $(u,v) \in A_m^+ \times A_m^+$ around a subword $w'' = u_a(bw_a)^{n-1}bv_a$ where $u_a$ is a suffix of the word $w_a$ and $v_a$ is a prefix of the word $w_a$. This word $w''$, however, does not belong to any selection language (because in every word of any language, at least one of the inner $a$-blocks contains less than $k$ occurrences of a certain letter of $A_m$). This contradiction implies that there is a selection language $S_i$ $(1 \leq i \leq m)$ where in every inner $a$-block the number of occurrences of each letter of the alphabet $A_m$ is unbounded.

When generating such a language $S_i$, a second nonterminal must appear before the second letter $b$ is generated because otherwise the first inner $a$-block would be bounded or the number of letters $b$ would be unbounded and a word with $n+1$ letters $b$ would be generated. Such a word should not be in the language $S_i$ because otherwise the lengths of two $a$-blocks not corresponding to each other would be increased but not the corresponding ones. By induction, one obtains that an $n$-th nonterminal must appear before the $n$-th letter $b$ is generated. Since, in every inner $a$-block, the number of each letter from $A_m$ is unbounded and the appearance is in arbitrary order, one needs a 'loop' for every letter of the set $A_m$. Hence, for each inner $a$-block, one needs a rule for generating every letter of $A_m$ and the letter $b$. Since, we have to remember in which block we are, we need those $m+1$ rules for every inner $a$-block. Furthermore, we need a rule for the first nonterminal symbol. Thus, $(n-1)(m+1)+1$ rules are necessary for at least one selection language. From this follows, that $(n-1)(m+1)$ rules are not sufficient.

Hence, $(m+1)(n-1)+1$ rules are sufficient for generating every selection language and necessary for at least one selection language in a contextual grammar generating the language $L_n^{(m)}$. □

From the previous result, we obtain the strictness of the following inclusions.

**Lemma 7.** *Let $m \geq 1$. For $n \geq 1$, the proper inclusion*

$$\mathcal{L}(IC, REG^{Prod}_{(m+1)(n-1)}) \subset \mathcal{L}(IC, REG^{Prod}_{(m+1)(n-1)+1})$$

*holds.*

Let us now consider the inclusion

$$\mathcal{L}(IC, REG^{Prod}_{k-1}) \subseteq \mathcal{L}(IC, REG^{Prod}_{k})$$

for some number $k \geq 1$.

For $k = 1$, we know from Corollary 5 that the inclusion is proper.

For $k \geq 3$, we obtain the strictness of the inclusion from Lemma 7 when we set $n = 2$ and $m = k - 2$.

We now prove that the inclusion is also proper for $k = 2$.

**Lemma 8.** *Let $V = \{\, a, b, c, d, e \,\}$ and*

$$L = \left\{\, w_{ab}cd^ne^m \mid m \geq 0, n \geq 0, w_{ab} \in \{\, a, b \,\}^*, \#_a(w_{ab}) = n, \#_b(w_{ab}) = m \,\right\}.$$

*Then $L \in \mathcal{L}(IC, REG^{Prod}_2)$ but $L \notin \mathcal{L}(IC, REG^{Prod}_1)$.*

*Proof.* The contextual grammar

$$G = (V, (\{ce\}, \{\, (b,e) \,\}), (\{b\}^*\{c\}, \{\, (a,d) \,\}), \{\, c, bce \,\})$$

generates the language $L$ as can be seen as follows.

Let $m \geq 0$, $n \geq 0$, and $w_{ab} \in \{\, a, b \,\}^*$ with $\#_a(w_{ab}) = n$ and $\#_b(w_{ab}) = m$. Then there exist numbers $m_1, m_2, \ldots m_{n+1}$ with $m_i \geq 0$ for $1 \leq i \leq n+1$ and $m_1 + m_2 + \cdots + m_{n+1} = m$ such that $w_{ab} = b^{m_1}ab^{m_2}a \ldots b^{m_n}ab^{m_{n+1}}$. How is the word $w = w_{ab}cd^ne^m \in L$ generated? If $n = m = 0$, then $w = c \in B$. If $n = 0$ and $m = 1$, then $w = bce \in B$. Otherwise, the word $w$ is derived from the axiom $bce$

by first wrapping the letters $b$ and $e$ around the subword $ce$ until the word $b^m ce^m$ is obtained and then wrapping the letters $a$ and $d$ around the subwords $b^{m-m_1}c$, $b^{m-m_1-m_2}c$, and so on until $b^{m_{n+1}}c$. This yields the derivation

$$\begin{aligned} bce \Longrightarrow^* b^m ce^m &\Longrightarrow b^{m_1}ab^{m-m_1}cde^m \\ &\Longrightarrow b^{m_1}ab^{m_2}ab^{m-m_1-m_2}cd^2e^m \\ &\Longrightarrow^* b^{m_1}ab^{m_2}ab^{m-m_1-m_2}a\ldots b^{m_n}ab^{m_{n+1}}cd^ne^m = w. \end{aligned}$$

Hence, every word of the language $L$ is generated by the contextual grammar $G$.

On the other hand, from a word of the language $L$, only words of the language $L$ are obtained. Let $w = w_{ab}cd^ne^m \in L$. If $n = 0$, then $w = b^m ce^m$. From the word $w$, the word $b^{m+1}ce^{m+1}$ can be obtained in one derivation step as well as any word $b^{m_1}ab^{m_2}cde^m$ with $m_1 \geq 0$, $m_2 \geq 2$, and $m_1 + m_2 = m$. These words belong to the language $L$, too; other words cannot be derived in one step. If $n > 0$, then the number of the letters $b$ and $e$ cannot be increased further. Hence, only words $w'_{ab}ab^kcd^{n+1}e^m$ with $k \geq 0$ and $w'_{ab}b^k = w_{ab}$ can be derived in one step. All these words also belong to the language $L$.

Thus, we obtain $L = L(G)$. The selection language $\{ce\}$ can be generated by a grammar with only one rule. The selection language $\{b\}^*\{c\}$ can be generated by a grammar with exactly two rules ($S \to bS$ and $S \to c$).

Hence, $L \in \mathcal{L}(IC, REG_2^{Prod})$.

We now show that the language $L$ cannot be generated by a contextual grammar where every selection language is generated by a grammar with one rule only.

Suppose, $L \in \mathcal{L}(IC, REG_1^{Prod})$. Then there is a contextual grammar

$$G' = (V, \{\, (S_1, C_1), (S_2, C_2), \ldots, (S_p, C_p)\, \}, B')$$

with $L(G') = L$ and every language $S_i$ is a singleton set for $1 \leq i \leq p$. Let $k$ be the length of the longest word in the union of these sets and $B'$ as well as among the contexts:

$$\begin{aligned} k_1 &= \max \{\, |w| \mid w \in B' \,\}, \\ k_2 &= \max \bigcup_{i=1}^{p} \{\, |w| \mid w \in S_i \,\}, \\ k_3 &= \max \bigcup_{i=1}^{p} \{\, |w| \mid (u, w) \in C_i \text{ or } (w, u) \in C_i \,\}, \\ k &= \max \{\, k_1, k_2, k_3 \,\}. \end{aligned}$$

Let us consider the word $w = a^{2k}b^{2k}cd^{2k}e^{2k} \in L$. Due to its length, it is not an axiom of the set $B'$. Hence, it is derived from another word of the language $L$. The number of letters $a$ and $d$ as well as $b$ and $e$ must be increased by the same number in each derivation step. For any context $(u, v)$, we know that $v \in \{d\}^+$ or $v \in \{e\}^+$, otherwise also a word which does not belong to the language $L$

could be generated. Let $(u,v)$ be a context that was added to a word to obtain the word $w$. If $v \in \{d\}^+$, then the word $u$ contains an occurrence of the letter $a$ and the subword where the context is wrapped around contains at least $k+1$ letters. If $v \in \{e\}^+$, then the word $u$ contains an occurrence of the letter $b$ and the subword also contains $k+1$ letters. However, no such word exists in any selection language. Because of this contradiction, the assumption is not true. This yields that $L \notin \mathcal{L}(IC, REG_1^{Prod})$. □

Together, we obtain the following result.

**Theorem 9.** *The relation*

$$\mathcal{L}(IC, REG_n^{Prod}) \subset \mathcal{L}(IC, REG_{n+1}^{Prod})$$

*holds for every natural number $n \geq 0$.*

We now consider the complexity measure *Symb*. Both the language families $\mathcal{L}(IC, REG_0^{Symb})$ and $\mathcal{L}(IC, REG_1^{Symb})$ are equal to the class of finite languages (every selection language is the empty set; the language generated coincides with the set of axioms). The class $\mathcal{L}(IC, REG_2^{Symb})$ contains infinite languages; for instance, the language $L = \{a\}^*$ is generated by the contextual grammar $G = (\{\,a\,\}, \{\,(\{\lambda\}, \{(\lambda, a)\})\,\}, \{\,\lambda\,\})$. This yields the proper inclusion

$$\mathcal{L}(IC, REG_1^{Symb}) \subset \mathcal{L}(IC, REG_2^{Symb}).$$

Also the further inclusions are proper.

**Lemma 10.** *We have $\mathcal{L}(EC, REG_n^{Symb}) \subset \mathcal{L}(EC, REG_{n+1}^{Symb})$ for $n \geq 2$.*

*Proof.* Let $n \geq 2$, $V_n = \{\,a, b, c_1, c_2, \ldots, c_{n-1}\,\}$, and

$$L_n = \{\,a^m c_1 c_2 \ldots c_{n-1} b^m \mid m \geq 0\,\}\,.$$

This language is generated by the contextual grammar

$$G_n = (V_n, \{\,(\{c_1 c_2 \ldots c_{n-1}\}, \{(a,b)\})\,\}, \{\,c_1 c_2 \ldots c_{n-1}\,\}).$$

Hence, $L_n \in \mathcal{L}(IC, REG_{n+1}^{Symb})$. Let $G'_n$ be a contextual grammar generating the language $L_n$. Every applicable context is of the form $(a^k, b^k)$ (with another context, a word would be generated which does not belong to the language). When such a context is inserted, the word that selects it contains the word $c_1 c_2 \ldots c_{n-1}$. Hence, any selection language contains a word which contains $n-1$ different letters. Thus, at least $n+1$ symbols are necessary for generating each selection language. Hence, $L_n \notin \mathcal{L}(IC, REG_n^{Symb})$. □

Together, we obtain the following result.

**Theorem 11.** *We have the relations*

$$\mathcal{L}(IC, REG_0^{Symb}) = \mathcal{L}(IC, REG_1^{Symb}) = FIN$$

*and*

$$\mathcal{L}(IC, REG_n^{Symb}) \subset \mathcal{L}(IC, REG_{n+1}^{Symb})$$

*for every natural number $n \geq 1$.*

## 4 Conclusions and Further Work

The main contribution of our paper consists in exhibiting a series of languages that highlight the difference between the power of internal contextual grammars that have the selection languages from different families of subregular languages. Here we only considered families defined by imposing restriction on the number of states, nonterminals, productions, or symbols needed to accept or generate them. It seems a natural continuation to consider in this context other families of subregular languages, defined by restrictions of combinatorial nature, like the ones considered in [1, 2] for external contextual grammars.

Also, it may be interesting to show that there are languages over a constant alphabet that separate the classes on consecutive levels of every hierarchy we defined. While such a result was obtained in the case of the *State*-hierarchy, by Theorem 2, as well as in the case of the *Var*-hierarchy, by Theorem 3, showing the existence of such languages remains an open problem in the other two cases.

**Acknowledgements.** We thank Jürgen Dassow for fruitful discussions and the anonymous referees for their remarks which helped to improve the paper.

## References

1. Dassow, J.: Contextual grammars with subregular choice. Fundamenta Informaticae **64**(1–4) (2005) 109–118
2. Dassow, J., Manea, F., Truthe, B.: On Contextual Grammars with Subregular Selection Languages. In Holzer, M., Kutrib, M., Pighizzini, G., eds.: Descriptional Complexity of Formal Systems – 13th International Workshop, DCFS 2011, Gießen/Limburg, Germany, July 25 – 27, 2011. Proceedings. Volume 6808 of LNCS, Springer-Verlag (2011) 135–146
3. Istrail, S.: Gramatici contextuale cu selectiva regulata. Stud. Cerc. Mat. **30** (1978) 287–294
4. Marcus, S.: Contextual grammars. Revue Roum. Math. Pures Appl. **14** (1969) 1525–1534
5. Păun, G.: Marcus Contextual Grammars. Kluwer Publ. House, Doordrecht (1998)
6. Rozenberg, G., Salomaa, A., eds.: Handbook of Formal Languages. Springer-Verlag, Berlin (1997)