On External Contextual Grammars with Subregular Selection Languages

Jürgen Dassow[a], Florin Manea[b,1], Bianca Truthe[a]

[a] *Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik*
PSF 4120, D-39016 Magdeburg, Germany
[b] *Institut für Informatik, Christian-Albrechts-Universität zu Kiel,*
Christian-Albrechts-Platz 4, D-24098 Kiel, Germany

Abstract

In this paper, we study the power of external 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 external contextual grammars with selection languages in $\mathcal{F}_n$. Moreover, we give some results on the power of external contextual grammars with regular commutative, regular circular, definite, suffix-free, ordered, combinational, nilpotent, and union-free selection languages.

Keywords: External contextual grammars, subregular selection languages.

1. Introduction

Contextual grammars were introduced by S. Marcus in [15] as a formal model that might be used in the generation of natural languages. The derivation steps consist of 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 an 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

Email addresses: `dassow@iws.cs.uni-magdeburg.de` (Jürgen Dassow), `flmanea@informatik.uni-kiel.de` (Florin Manea), `truthe@iws.cs.uni-magdeburg.de` (Bianca Truthe)

[1] Florin Manea is on leave of absence from the University of Bucharest, Faculty of Mathematics and Computer Science. His work is supported by the *DFG* grant 582014. Part of this paper was writen when Florin Manea was a Research Fellow of the *Alexander von Humboldt Foundation* at the Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, whose support is also acknowledged.

set of $\mathcal{F}$ which is associated with the context. Mostly, the family $\mathcal{F}$ is taken from the families of the Chomsky hierarchy (see [12, 17, 16] and the references therein).

In [4], 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 this paper, we continue this line of research. More precisely, we obtain some 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.

Moreover, we consider families of regular languages which are defined by restrictions on the resources used to generate or to accept them. As measures we consider the number of states necessary to accept the regular languages and the number of nonterminals, productions, or symbols needed to generate the regular languages. We prove that in all cases infinite hierarchies are obtained.

Our research is part of the study of problems and processes connected with regular sets. In the last years, many papers were published in which the effect of going from arbitrary regular sets to special regular sets was studied. Some of these topics are the following:

- Any nondeterministic finite automaton with n states can be transformed into a deterministic one with 2^n states which accepts the same language. This exponential blow-up with respect to the number of states is necessary in the worst cases. In [1], this problem is studied if one restricts to the case that the automata accept special regular languages only. It is shown, that the situation does not change for suffix-closed and star-free regular languages; however, for some classes of definite languages, the size of the deterministic automaton is bounded by $2^{n-1}+1$.
- A number α between some number n and 2^n is called magic with respect to n if there is no nondeterministic finite automaton with n states such that the equivalent minimal deterministic finite automaton has α states. It is known that no magic numbers exist if $n \geq 3$. This situation changes if one considers subregular families of languages. For instance, only the values α with $n+1 \leq \alpha \leq 2^{n-1}+1$ are possible for prefix-free regular languages (see [11]).
- For languages obtained by certain operations from other languages, upper bounds for the state complexity (the number of states that are needed for a minimal finite automaton to accept the language) have been determined depending on the state complexities of the other languages. Those upper bounds are tight. It has been shown that the upper bounds are smaller if one restricts to special regular languages (see [9], [10], [2], and [13]).
- In order to enlarge the generative power, some mechanisms connected with regular languages were introduced, which control the derivations in context-free grammars. For instance, the sequence of applied rules in a regularly controlled grammar, the current sentential form in a conditional grammar, and the levels of the derivation tree in a tree controlled grammar

have to belong to given regular languages. In the papers [3], [6], [5], and [8], the change in the generative power, if one restricts to special regular sets, is investigated.

- Networks of evolutionary processors with regular filters are computationally complete in that sense that they can generate or accept every recursively enumerable language. In [14] and [7], it was shown that such networks are still computational complete if the filters are taken from several subclasses of regular languages but, for other subclasses, they are not computational complete.

The present paper follows this direction with results on the influence of subregularity in the selection of contextual grammars.

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 [17]. 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 λ. 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 $\#(A)$.

For a language L over V, we set

$$\begin{aligned}
\mathit{Comm}(L) &= \{\, a_{\pi(1)}a_{\pi(2)}\dots a_{\pi(n)} \mid a_i \in V \text{ for } 1 \leq i \leq n,\ a_1a_2\dots a_n \in L, \\
&\qquad\qquad\qquad\qquad \pi \text{ is a permutation of } \{1,2,\dots,n\} \,\}, \\
\mathit{Circ}(L) &= \{\, yx \mid xy \in L \text{ for some } x, y \in V^* \,\}, \\
\mathit{Suf}(L) &= \{\, y \mid xy \in L \text{ for some } x \in V^* \,\}.
\end{aligned}$$

It is known that $\mathit{Suf}(L)$ and $\mathit{Circ}(L)$ are regular for a regular language L, whereas *Comm* does not preserve regularity.

Let V be an alphabet. We say that a language L over V is

- *combinational* iff it can be represented in the form $L = V^*A$ for some finite non-empty subset $A \subseteq V \cup \{\lambda\}$,
- *definite* iff it can be represented in the form $L = A \cup V^*B$ where A and B are finite subsets of V^*,
- *nilpotent* iff L is finite or $V^* \setminus L$ is finite,
- *commutative* iff $\mathit{Comm}(L) = L$,
- *circular* iff $\mathit{Circ}(L) = L$,
- *suffix-closed* (or *fully initial* or a *multiple-entry* language) iff $\mathit{Suf}(L) = L$,
- *ordered* iff L is accepted by some finite automaton $\mathcal{A} = (V, Z, z_0, F, \delta)$ where the set Z of states is totally ordered with respect to a relation $\preceq$ and, for any $a \in V$, the relation $z \preceq z'$ implies the relation $\delta(z,a) \preceq \delta(z',a)$,

- *union-free* iff L can be described by a regular expression which is only built by product and star.

It is obvious that combinational, definite, nilpotent, ordered, and union-free languages are regular, whereas non-regular languages of the other three types exist.

We denote by REG the class of regular languages. By FIN, COMB, DEF, NIL, COMM, CIRC, SUF, ORD, and UF we denote the families of all finite, combinational, definite, nilpotent, regular commutative, regular circular, regular suffix-closed, ordered, and union-free languages, respectively.

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 δ) and L be a regular language. Then we define

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

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

$$\mathit{REG}_n^K = \{\, L \mid L \text{ is a regular language with } K(L) \le n \,\}.$$

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 construct

$$G = (V, (P_1, C_1), (P_2, C_2), \ldots, (P_n, C_n), B)$$

where

- V is an alphabet,
- for $1 \le i \le n$, P_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 P_i and the sets C_i, $1 \le i \le n$, are called the selection languages and the sets of contexts of G, respectively; the elements of B are called axioms.

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

Let $G = (V, (P_1, C_1), (P_2, C_2), \ldots, (P_n, C_n), B)$ be a contextual grammar with selection. A direct *external 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 is an integer i, $1 \leq i \leq n$, such that $x \in P_i$ and $y = uxv$ for some pair $(u, v) \in C_i$. Intuitively, we can only wrap a context $(u, v) \in C_i$ around a word x if x belongs to the corresponding language P_i.

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

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

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

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

$$B_n = \left\{ \, a^i \mid 0 \leq i \leq n \, \right\}, \ U_n = \left\{ \, a^{n+1} \, \right\}^*, \text{ and } L_n = B_n \cup U_n.$$

The contextual grammar

$$G_n = (V, (U_n, \{(\lambda, a^{n+1})\}), B_n)$$

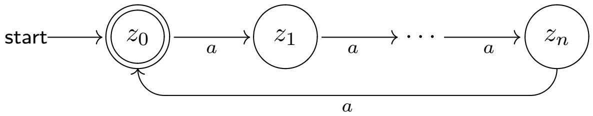
generates the language L_n. This can be seen as follows. The only axiom to which a context can be added is λ. From this, we get the unique derivation

$$\lambda \Longrightarrow a^{n+1} \Longrightarrow a^{2(n+1)} \Longrightarrow a^{3(n+1)} \Longrightarrow \cdots .$$

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

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

where the graph



represents the transition function δ_n. Hence, we have $L_n \in \mathcal{L}(EC, REG^{State}_{n+1})$.

Furthermore, the language U_n is generated by the regular grammar

$$(\{S\}, \{a\}, \{S \to a^{n+1}S, \ S \to \lambda\}, S).$$

Hence, we have also $L_n \in \mathcal{L}(EC, REG^{Symb}_{n+6})$. ◇

Obviously, the following lemma holds (see [4], Lemma 4.1).

Lemma 2 *For any two language classes X and Y with $X \subseteq Y$, we have $\mathcal{L}(EC, X) \subseteq \mathcal{L}(EC, Y)$.* □

3. Selection with Bounded Resources

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

Theorem 3 *For any natural number $n \geq 1$, we have the proper inclusion*

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

Proof. For $n \geq 1$, let

$$B_n = \left\{ a^i \mid 0 \leq i \leq n \right\}, \; U_n = \left\{ a^{n+1} \right\}^*, \text{ and } L_n = B_n \cup U_n$$

be the languages from Example 1, where we have shown $L_n \in \mathcal{L}(EC, REG_{n+1}^{State})$.

We now prove that $L_n \notin \mathcal{L}(EC, REG_n^{State})$.

Let $G = (V, (P_1, C_1), (P_2, C_2), \dots, (P_m, C_m), B)$ be a contextual grammar with selection in REG such that $L(G) = L_n$.

Let

$$\begin{aligned} k' &= \max \{ |uv| \mid (u,v) \in C_i, 1 \leq i \leq m \}, \\ k'' &= \max \{ |z| \mid z \in B \}, \\ k &= k' + k''. \end{aligned}$$

We consider the word $w = a^{k(n+1)} \in L_n$. Obviously, $w \notin B$. Thus, the word w is obtained from some word $w' \in L_n$ by adding a context $(u,v) \in C_i$ for some index i with $1 \leq i \leq m$ and $w' \in P_i$. Then $w = uw'v$. For the length of the word w', we obtain

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

Hence $w' \notin B_n$, so $w' \in U_n$, which implies $w' = a^{j(n+1)}$ for some j with $1 \leq j < k$ and $uv = a^{(k-j)(n+1)}$. Therefore, P_i contains some element of $\{a\}^+$. Further, if P_i contains a word z of B_n, then also $uzv \in L_n$. But $z \in B_n$ implies $|z| = s$ with $1 \leq s \leq n$, and thus $|uzv| = s + (k-j)(n+1)$ is greater than $n+1$ but not a multiple of $n+1$. This is impossible for words in L_n. Hence, the set P_i does not contain a word of B_n. Let $r = \min \left\{ l \mid a^l \in P_i, a^l \neq \lambda \right\}$. Then $r \geq n+1$. We set $z_i = a^{r-i}$ for $0 \leq i \leq r$. Then we have the relations $a^i z_i \in P_i$ and $a^j z_i \notin P_i$ for $1 \leq j < i \leq r$ because $a^i z_i = a^r$ and $|a^j z_i| < r$ for $1 \leq j < i \leq r$.

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

We now consider the measures *Var* and *Prod*. We start with the following result.

Theorem 4 *The equality $\mathcal{L}(EC, REG_1^{Prod}) = FIN$ holds.*

Proof. All languages from the class REG_1^{Prod} are finite. According to Lemma 2, we have the inclusion $\mathcal{L}(EC, REG_1^{Prod}) \subseteq \mathcal{L}(EC, FIN)$. Since the equality $\mathcal{L}(EC, FIN) = FIN$ holds ([4]), we get the inclusion

$$\mathcal{L}(EC, REG_1^{Prod}) \subseteq FIN.$$

Let L be a finite language. This language can be generated by the grammar $G = (V, (\{\lambda\}, \{(\lambda, \lambda)\}), L)$. The regular language $\{\lambda\}$ can be generated with one production $S \to \lambda$ only. Hence, $L \in \mathcal{L}(EC, REG_1^{Prod})$ which gives us the other inclusion

$$FIN \subseteq \mathcal{L}(EC, REG_1^{Prod})$$

and thus the equality claimed. □

For certain languages over a unary alphabet, we obtain the following.

Theorem 5 *Any language $L \in \mathcal{L}(EC, REG)$ over a unary alphabet belongs to the set*

$$\mathcal{L}(EC, REG_1^{Var}) \cap \mathcal{L}(EC, REG_2^{Prod}).$$

Proof. By Corollary 8.2 of [16], any language in $\mathcal{L}(EC, REG)$ is linear. Thus, any language in $\mathcal{L}(EC, REG)$ over a unary alphabet is regular (since all context-free languages over a unary alphabet are regular, see [17], Volume 1, Chapter 3, Theorem 2.6). Therefore, any language $L \in \mathcal{L}(EC, REG)$ can be represented in the form

$$L = \{a^{i_1}, a^{i_2}, \dots, a^{i_r}\} \cup \{a^p\}^* \{a^{j_1}, a^{j_2}, \dots, a^{j_s}\}$$

for some numbers $r \geq 0$, $s \geq 0$, $p \geq 1$, $i_1, i_2, \dots, i_r$, $j_1, j_2, \dots, j_s$ with

$$0 \leq i_1 < i_2 < \cdots < i_r < p \leq j_1 < j_2 < \cdots < j_s < 2p.$$

This representation follows immediately from an analysis of the form that a deterministic finite automaton accepting L may have. Such a deterministic automaton can be uniquely depicted as a chain of transitions (labelled with a) that ends with a cycle of length p; the values $i_1, \dots, i_r$ and $j_1, \dots, j_s$ are obtained by counting how many symbols a should be read in order to get from the initial state to the final states.

Thus, L can be generated by the contextual grammar

$$(\{a\}, (\{a^p\}^* \{a^{j_1}\}, \{(\lambda, a^p)\}), \dots, (\{a^p\}^* \{a^{j_s}\}, \{(\lambda, a^p)\}), B)$$

with $B = \{a^{i_1}, a^{i_2}, \dots, a^{i_r}, a^{j_1}, a^{j_2}, \dots, a^{j_s}\}$. Moreover, each selection language $\{a^p\}^* \{a^{j_\ell}\}$, with $1 \leq \ell \leq s$, is generated by the regular grammar

$$(\{S\}, \{a\}, \{S \to a^p S, S \to a^{j_\ell}\}, S).$$

Hence, $L \in \mathcal{L}(EC, REG_1^{Var}) \cap \mathcal{L}(EC, REG_2^{Prod})$. □

The above proof shows that the class of all unary languages from $\mathcal{L}(EC, REG)$ is equal to the class of all unary regular languages.

By Theorem 5, there are bounds (1 for the number of variables and 2 for the number of production rules) for the necessary size of selection languages if the language to be generated by the respective contextual grammar is unary. We now show that this does not hold if the underlying alphabet contains at least two letters.

Let p be a positive natural number. For each natural number $n > 1$, we define the following language over the alphabet $\{a, b\}$:

$$\begin{aligned} K_n = {} & \{a(ab^{k_1})\dots(ab^{k_{n-1}})ab^p \mid k_i > p \text{ for } 1 \leq i \leq n-1\} \\ & \cup \{a(ab^{k_1})\dots(ab^{k_\ell}) \mid 1 \leq \ell < n, k_i > p \text{ for } 1 \leq i \leq \ell\}. \end{aligned}$$

The parentheses are used above only to highlight the repetitive structure of the language.

Lemma 6 *For $n > 1$, we have $K_n \in \mathcal{L}(EC, REG_{n-1}^{Var})$.*

Proof. First, we assume that $n \geq 3$. Consider the following regular grammars:

- G_1 having $n-1$ nonterminals $S, A_3, A_4, \dots, A_n$ where S is the axiom and having the rules $S \to aabA_3$, $S \to bS$, $S \to \lambda$, and $A_i \to bA_i$, $A_i \to abA_{i+1}$ for $3 \leq i \leq n-1$ and $A_n \to bA_n$, $A_n \to abS$.
- G_2 having $n-1$ nonterminals $S, A_3, A_4, \dots, A_n$ where S is the axiom and having the rules $S \to aabA_3$ and $A_i \to bA_i$, $A_i \to \lambda$, $A_i \to abA_{i+1}$ for $3 \leq i \leq n-1$ and $A_n \to bA_n$, $A_n \to \lambda$.

The languages generated by the above grammars are:

- $L_1 = L(G_1) = \{a\}(\{a\}\{b\}^+)^{n-1} \cup L_1' \cup L_1'' \cup \{\lambda\}$ where L_1' is a regular language that contains only words with at least $n+2$ a-symbols and at least two distinct factors aa and L_1'' is a regular language that contains only words that start with b.
- $L_2 = L(G_2) = \bigcup\limits_{1 \leq \ell < n-1} \{a\}(\{a\}\{b\}^+)^\ell$.

In what follows, we show that K_n is generated by the contextual grammar

$$G_n = (V, (L_1, \{(\lambda, b)\}), (L_1, \{(\lambda, ab^p)\}), (L_2, \{(\lambda, b)\}), (L_2, \{(\lambda, ab^{p+1})\}), B)$$

with $V = \{a, b\}$ and $B = \{aab^{p+1}\}$.

First, consider a word $a(ab^{k_1})\dots(ab^{k_\ell})$ with $1 \leq \ell < n$ and $k_i > p$ for $1 \leq i \leq \ell$. This word is generated starting from the axiom aab^{p+1} by wrapping the context (λ, b) selected by L_2 around the current word until aab^{k_1} is obtained. Then, if $\ell > 1$, the context (λ, ab^{p+1}) selected by L_2 is added and we get $aab^{k_1}ab^{p+1}$. The process continues in the same way: First we add several times (λ, b) selected by L_2, then we add (λ, ab^{p+1}) selected by L_2. At some point, the word $a(ab^{k_1})\dots(ab^{k_{\ell-1}})(ab^{p+1})$ will be generated. Then, the context (λ, b) selected by L_1 (if $\ell = n-1$) or by L_2 (if $\ell < n-1$) is added until the word $a(ab^{k_1})\dots(ab^{k_\ell})$ is obtained.

Now consider the word $a(ab^{k_1})\dots(ab^{k_{n-1}})ab^p$. First, we generate the word $a(ab^{k_1})\dots(ab^{k_{n-2}})ab^{p+1}$ exactly as above. This word is in $L_1 \setminus L_2$. We continue by wrapping the context (λ, b) selected by L_1 around the current word until we obtain $a(ab^{k_1})\dots(ab^{k_{n-1}})$; to end the derivation, we add the context (λ, ab^p) selected by L_1 and obtain $a(ab^{k_1})\dots(ab^{k_{n-1}})ab^p$.

It is not hard to see that only words of these forms can be obtained. Using the contextual rules that have L_2 as selector, we only obtain words of the form $a(ab^{k_1})\dots(ab^{k_\ell})$ or $a(ab^{k_1})\dots(ab^{k_\ell})(ab^{p+1})$ with $1 \le \ell < n-1$ and $k_i > p$ for $1 \le i \le \ell$. Using additionally the contextual rules that have L_1 as selector, we obtain further the words of the form $a(ab^{k_1})\dots(ab^{k_\ell})$ or $a(ab^{k_1})\dots(ab^{k_\ell})(ab^p)$ with $1 \le \ell < n$ and $k_i > p$ for $1 \le i \le \ell$. The words of the first type can be further derived, while the words of the second type cannot be derived anymore as they are not in $L_1 \cup L_2$. Thus, $L_i \in REG_{n-1}^{Var}$ for $i \in \{1, 2\}$.

Now consider that $n = 2$. We define the grammar G_1 with one nonterminal S and the rules $S \to aabS$, $S \to bS$, and $S \to \lambda$. The language generated by this grammar is $L_1 = L(G_1) = \{aa\}\{b\}^+ \cup L_1' \cup L_1'' \cup \{\lambda\}$ where L_1' is a regular language that contains at least two distinct factors aa and L_1'' is a regular language that contains only words that start with b. The language K_n is generated by the grammar $G_n = (V, (L_1, \{(\lambda, b)\}), (L_1, \{(\lambda, ab^p)\}), B)$ where, as in the previous case, $B = \{aab^{p+1}\}$. Hence, we have $K_n \in \mathcal{L}(EC, REG_{n-1}^{Var})$ in this case, too. □

Lemma 7 *For $n > 1$, we have that $K_n \notin \mathcal{L}(EC, REG_{n-2}^{Var})$.*

Proof. Assume that $G = (V, (P_1, \{(u_1, v_1)\}), \dots, (P_\ell, \{(u_\ell, v_\ell)\}), B)$ is a contextual grammar that generates K_n and that $G_1, \dots, G_\ell$ are regular grammars that generate $P_1, \dots, P_\ell$, respectively.

We first note that $u_i = \lambda$ for $1 \le i \le \ell$. All the words derived by G have the form aaw with $w \in \{a, b\}^*$. Further, if a context has the form (u_i, v_i) with $u_i \ne \lambda$, then wrapping this context around a certain word derived by G yields a word of the language that contains the factor aa not as a prefix or at least twice. But this is a contradiction.

Now, denote by M the maximum length of the right hand sides of the productions of the grammars $G_1, \dots, G_\ell$ and denote by M' the maximum length of the words v_i for $1 \le i \le \ell$ and of the words from B. Then $M' \ge p$ because every word of the language K_n (and especially of B) has more than p letters.

Let $N = M + 2M' + 1$ and consider now the word $w = a(ab^N)\dots(ab^N)ab^p$ where $|w|_a = n + 1$. It does not belong to the set B since it is longer than any word of B. Thus, it is obtained by adding a context, denoted (λ, v_i), to another word w' of K_n. We note that this context is of the form $(\lambda, b^x ab^p)$ with $x \ge 0$ (by the choice of N sufficiently large and because $w' \in K_n$). Hence, $w' = a(ab^N)^{n-2}ab^m$ with $M + M' + 1 \le m \le N$.

In what follows, we show that the grammar G_i generating the selection language P_i of the contextual rule $(P_i, \{(\lambda, v_i)\})$ has at least $n-1$ nonterminals. Let us assume the opposite: The grammar G_i has at most $n-2$ nonterminals.

A possible derivation of the word $w' \in P_i$ has the form

$$S \Longrightarrow^* aab^{k_1}A_1 \Longrightarrow^* aab^N ab^{k_2}A_2 \Longrightarrow^* \cdots \Longrightarrow^* a(ab^N)^{i-1}ab^{k_i}A_i \Longrightarrow^* \cdots \\ \Longrightarrow^* a(ab^N)^{n-2}ab^{k_{n-1}}A_{n-1} \Longrightarrow^* a(ab^N)^{n-2}ab^m$$

where $A_1, \ldots, A_{n-1}$ are nonterminals and $k_i \leq M$ for $1 \leq i \leq n-1$. Since G_i has at most $n-2$ nonterminals by our assumption, there exist numbers j and k such that $1 \leq j < k \leq n-1$ and $A_j = A_k$. Hence, we can also derive by the grammar G_i a word that has at most $t \leq n-1$ symbols a and has the form $a(ab^{l_1}) \ldots (ab^{l_{t-1}})$ with $l_s > p$ for $1 \leq s \leq t-1$ (we derive $a(ab^N)^{j-1}ab^{k_j}A_j$ and then rewrite A_j as we rewrote A_k in the derivation above). In a word obtained in this way, every group of b symbols contains more than p symbols, because we basically decreased the number of symbols in such a group by at most M. Hence, $a(ab^{l_1}) \ldots (ab^{l_{t-1}}) \in P_i$ and the context $(\lambda, b^x ab^p)$ can be added which yields the word $a(ab^{l_1}) \ldots (ab^{l_{t-1}})ab^p$ with $t \leq n-1$ and $l_s > p$ for $1 \leq s \leq t-1$. This is a contradiction to the definition of K_n. Therefore, G_i has at least $n-1$ nonterminals. This concludes our proof. □

From the previous two lemmas we obtain the following result:

Theorem 8 *For any natural number $n \geq 1$, we have the proper inclusion*

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

The generation of the language K_n for a natural number $n > 1$ requires at least $2(n-2)+1$ productions in each component (each nonterminal different from the axiom has at least two productions; otherwise, we could substitute each occurrence of a nonterminal A which appears in only one rule $A \to w$ by w in all right hand sides and ignore the nonterminal A and its rule, which would give a regular grammar with a smaller number of nonterminals). From the proof of Lemma 6, we have $K_n \in \mathcal{L}(EC, REG_{3n-6}^{Prod})$ for $n \geq 5$. Hence, together, we obtain $K_n \in \mathcal{L}(EC, REG_{3n-6}^{Prod}) \setminus \mathcal{L}(EC, REG_{2n-4}^{Prod})$ for $n \geq 5$, from which immediately follows that the families $\mathcal{L}(EC, REG_n^{Prod})$ for $n \geq 1$ form an infinite chain with respect to inclusion. If we do not restrict the size of the alphabet we get a stronger result for which we first state the following lemma.

Lemma 9 *Let $G = (N, V, P, S)$ be a regular grammar, with $V = \{a_1, \ldots, a_n\}$. Assume that P contains at most $n-1$ rules whose right hand side contains a nonterminal symbol. Then there exist a number $k_G \in \mathbb{N}$ and a word $w_G \in V^*$ such that w_G is not a factor of the prefix of length $m - k_G$ of any word w of length m generated by G for all $m \in \mathbb{N}$.*

Proof. Let P' denote the subset of all the rules of P whose right hand side contains a nonterminal. Let k_G be the maximum length of the right hand side of the rules from $P \setminus P'$. Furthermore, denote by P'' the set of all the rules of P' that are not of the form $A \to B$ with $A, B \in N$. We can write $P'' = \{A_i \to x_i A'_i \mid A_i, A'_i \in N, x_i \in V^+, i \in \{1, \ldots, t\}\}$ for some $t \leq n-1$.

Let ℓ be a natural number. There are n^ℓ words of length ℓ over V. We will compute an upper bound on the number of factors of length ℓ that can appear in the prefixes of length $m - k_G$ of the words of length m generated by G, for all $m \in \mathbb{N}$.

If x is such a factor it is a factor of one of the words of the set

$$\{x_{j_1} \dots x_{j_\ell} \mid 1 \leq j_1, \dots, j_\ell \leq t\}.$$

But the number of factors of length ℓ of the words from that set is less or equal to $t^\ell(M\ell - (\ell - 1))$ where M is the maximum length of one of the words x_i with $i \in \{1, \dots, t\}$. Since $t \leq n - 1$ it follows that $(n-1)^\ell(M\ell - \ell + 1)$ upper bounds the number of factors of length ℓ that can appear in the prefixes of length $m - k_G$ of the words of length m generated by the grammar G for all $m \in \mathbb{N}$.

But for ℓ sufficiently large we have that $n^\ell > (n-1)^\ell(M\ell - \ell + 1)$, so there exists a word of length ℓ that does not appear as a factor in the prefix of length $m - k_G$ of any word w of length m generated by G for all $m \in \mathbb{N}$. Let such a word be w_G. This concludes our proof. □

We can now prove the infinite hierarchy without 'gaps' between the number of production rules.

Theorem 10 *For any natural number $n \geq 1$, we have the proper inclusion $\mathcal{L}(EC, REG_n^{Prod}) \subset \mathcal{L}(EC, REG_{n+1}^{Prod})$.*

Proof. Let $V = \{a_1, \dots, a_n\}$. The language V^* is generated by the regular grammar

$$G = (\{S\}, V, \{S \to a_i S \mid 1 \leq i \leq n\} \cup \{S \to \lambda\}, S).$$

The language V^* is also generated by the contextual grammar

$$G' = (V, (V^*, \{(\lambda, a_i) \mid 1 \leq i \leq n\}), \{\lambda\}).$$

Thus, $V^* \in \mathcal{L}(EC, REG_{n+1}^{Prod})$. Let us now prove that $V^* \notin \mathcal{L}(EC, REG_n^{Prod})$.

We assume the opposite: there exists a contextual grammar with regular selection $G' = (V, (L_1, C_1), (L_2, C_2) \dots, (L_p, C_p), B)$ that generates V^* such that the languages L_i can be generated by regular grammars with at most n productions for $1 \leq i \leq p$. For all $i \in \{1, \dots, p\}$, the grammar generating L_i must have at least a rule whose right hand side contains only terminals; consequently, by Lemma 9, for all $i \in \{1, \dots, p\}$, there exists $k_i \in \mathbb{N}$ and $w_i \in V^*$ such that w_i is not a factor of the prefix of length $m - k_i$ of any word w of length m of L_i.

Now let M_1 be the maximum length of an axiom of G', let M_2 be the maximum length of one side of the contexts of G', and let M_3 be the maximum of the set $\{k_i \mid 1 \leq i \leq p\}$. We denote by $M = M_1 + M_2 + M_3$.

Let $w = a_1^M w_1 w_2 \dots w_p a_1^M$. Due to its length, $w \notin B$, so it must be obtained by adding one context c to a word x generated by G'. But the word x would contain the factor w_i in its prefix of length $|x| - k_i$ for all $i \in \{1, \dots, p\}$. So x does not belong to any of the selection languages, and cannot select c. We have reached a contradiction, and, consequently, $V^* \notin \mathcal{L}(EC, REG_n^{Prod})$. □

We now consider the measure *Symb*. For any language L, one needs at least two symbols for generating it.

Lemma 11 *Let L be an infinite regular language. Let k be the minimal length of non-empty words in L. Then $Symb(L) \geq k+5$.*

Proof. Let G be a regular grammar generating L. Let us consider a derivation

$$S \Longrightarrow x_1A_1 \Longrightarrow x_1x_2A_2 \Longrightarrow \cdots \Longrightarrow x_1x_2\ldots x_{n-1}A_{n-1} \Longrightarrow x_1x_2\ldots x_n = w$$

of a word $w \in L$ of length k.

If $n=1$, then $S \Longrightarrow w$. The rule $S \to w$ contributes $k+2$ symbols. There is also a rule $A \to z$ with $z \notin T^*$ since L is infinite. Hence, $Symb(G) \geq k+5$.

If $n \geq 2$, we have rules $S \to x_1A_1$, $A_i \to x_{i+1}A_{i+1}$ for $1 \leq i \leq n-2$ and $A_{n-1} \to x_n$. Then we have $Symb(G) = k+5+3(n-2) \geq k+5$. □

By this rsult, we know especially that, for generating an infinite language, one needs at least six symbols. With this result, we can show the following theorem.

Theorem 12 *We have $\mathcal{L}(EC, REG_k^{Symb}) = FIN$ for $2 \leq k \leq 5$.*

Proof. According to Lemma 11, all languages from the set REG_k^{Symb} where k satisfies $2 \leq k \leq 5$ are finite. Thus, $\mathcal{L}(EC, REG_5^{Symb}) \subseteq \mathcal{L}(EC, FIN)$. Since $\mathcal{L}(EC, FIN) = FIN$ ([4]), we get $\mathcal{L}(EC, REG_5^{Symb}) \subseteq FIN$.

Let L be a finite language. This language can be generated by the grammar $G = (V, (\{\lambda\}, \{(\lambda,\lambda)\}), L)$. Hence, $L \in \mathcal{L}(EC, REG_2^{Symb})$. □

Finally, we also obtain the following result.

Theorem 13 *For any number $n \geq 6$, we have*

$$\mathcal{L}(EC, REG_{n-1}^{Symb}) \subset \mathcal{L}(EC, REG_n^{Symb}).$$

Proof. Let $L_n = \{\, a^i \mid 0 \leq i \leq n \,\} \cup \{\, a^{n+1} \,\}^*$ be the languages from Example 1, where we have shown that L_n belongs to the class $\mathcal{L}(EC, REG_{n+6}^{Symb})$.

We now prove that $L_n \notin \mathcal{L}(EC, REG_{n+5}^{Symb})$. Assume the contrary, i. e., there is a contextual grammar $G = (V, (P_1, C_1), (P_2, C_2), \ldots, (P_m, C_m), B)$ with selection in REG_{n+5}^{Symb} such that $L(G) = L_n$. We define k', k'', and k as in the proof of Theorem 3,

$$\begin{aligned} k' &= \max\{\, |uv| \mid (u,v) \in C_i, 1 \leq i \leq m \,\}, \\ k'' &= \max\{\, |z| \mid z \in B \,\}, \\ k &= k' + k'', \end{aligned}$$

and consider the words $w_p = a^{pk(n+1)}$ for $p \geq 1$. Again, there are words w'_p such that $w_p = u_p w'_p v_p$ for some context (u_p, v_p) associated with some regular set P_{i_p}, $1 \leq i_p \leq m$, and

$$\begin{aligned} pk(n+1) = |w_p| > |w'_p| &= p(k'+k'')(n+1) - |u_p v_p| \\ &\geq p(k'+k'')(n+1) - k' > (p-1)k(n+1). \end{aligned}$$

Thus, all these words w'_p are different and each of them belongs to at least one of the sets P_j, $1 \leq j \leq m$. Therefore at least one of the sets P_j, $1 \leq j \leq m$, say P_i, is infinite. As in the proof of Theorem 3, we can show that the shortest non-empty word of P_i has a length $r \geq n+1$. Now we get from Lemma 11 that $Symb(P_i) \geq n+6$. □

We make some final remarks on the results of this section. We have shown that, for $K \in \{State, Var, Prod\}$, we obtain infinite dense hierarchies, i.e., for any $n \geq 1$,

$$\mathcal{L}(EC, REG^K_{n+1}) \setminus \mathcal{L}(EC, REG^K_n) \neq \emptyset.$$

However, these results hold for all alphabets with respect to *State*; in the case of *Var*, it is valid for all alphabets with at least two letters (if the alphabet is unary, we have only one level in the hierarchy); and for *Prod* we need alphabets of arbitrary size. It is an open question whether the results hold for *Prod* and a fixed alphabet size. For *Symb*, we get a dense hierarchy over all alphabets, but starting with $n = 6$.

4. Circular, Ordered, and Union-Free Selection

We start with some results on circular selection languages.

Theorem 14 *We have the proper inclusion $\mathcal{L}(EC, COMM) \subset \mathcal{L}(EC, CIRC)$.*

Proof. By Lemma 2, we obtain $\mathcal{L}(EC, COMM) \subseteq \mathcal{L}(EC, CIRC)$. Let

$$P_1 = \{\, abc^n \mid n \geq 1 \,\}\,,\ P_2 = \{\, c^n ba \mid n \geq 1 \,\}\,,\ \text{and } L = P_1 \cup P_2.$$

The language L is generated by the contextual grammar

$$G = (\{\, a, b, c \,\}, (Circ(P_1), \{\, (\lambda, c) \,\}), (Circ(P_2), \{\, (c, \lambda) \,\}), \{\, abc, cba \,\})$$

because, for $1 \leq i \leq 2$, a word from the set P_i always derives a word from the set P_i since the circular closures of the sets P_1 and P_2 are disjoint; furthermore, a word abc^n derives the word abc^{n+1} and a word $c^n ba$ derives the word $c^{n+1}ba$ for $n \geq 1$.

Let us assume that the language L is also generated by a contextual grammar $G' = (\{\, a, b, c \,\}, (Q_1, C_1), (Q_2, C_2), \ldots, (Q_m, C_m), B)$ where all sets Q_i with $1 \leq i \leq m$ are commutative. We consider a word $abc^k \in L$ where k is greater than the maximal length of the words in the set B. Then this word is generated by deriving another word $z \in L$. Hence, there is an index i with

$1 \leq i \leq m$ such that $abc^k = uzv$ for some context $(u,v) \in C_i$ and $z \in Q_i$. Since $z \in L$ the word z has the form abc^l with $1 \leq l < k$. Consequently, the context (u,v) is (λ, c^s) for a number $s \geq 1$. Since $z \in Q_i$ and the set Q_i is commutative, it also contains the word $c^l ba$ which belongs to the language L. According to our assumption, this word $c^l ba$ belongs to the language $L(G')$ and therefore derives the word $c^l bac^s \in L(G')$ which does not belong to the language L. Hence, the two languages L and $L(G')$ are not equal which contradicts the assumption. Thus, the language L cannot be generated by a contextual grammar with commutative selection which yields the proper inclusion $\mathcal{L}(EC, COMM) \subset \mathcal{L}(EC, CIRC)$. □

Theorem 15 *The class $\mathcal{L}(EC, CIRC)$ is incomparable with each of the classes $\mathcal{L}(EC, DEF)$ and $\mathcal{L}(EC, SUF)$.*

Proof. Let $V = \{a,b,c\}$ and $L_1 = \{\, a^n b^n \mid n \geq 1 \,\} \cup \{\, cb^n \mid n \geq 1 \,\}$. The language L_1 can be generated by the contextual grammar with circular selection

$$G = (V, (P_1, \{\, (a,b) \,\}), (P_2, \{\, (\lambda, b) \,\}), \{\, ab, cb \,\})$$

with $P_1 = \{\, w \mid w \in V^* \text{ and } |w|_c = 0 \,\}$ and $P_2 = V^* \backslash P_1$. However, the language does not belong to the class $\mathcal{L}(EC, DEF)$ according to [4].

Let $L_2 = \{\, ab^n \mid n \geq 1 \,\} \cup \{\, b \,\}$. The language L_2 can be generated by

$$G = (\{\, a, b \,\}, (Circ(\{\, ab^n \mid n \geq 1 \,\}), \{\, (\lambda, b) \,\}), \{\, ab, b \,\})$$

with circular selection. However, $L_2 \notin \mathcal{L}(EC, SUF)$ according to [4].

Let $L_3 = \{\, ab^n \mid n \geq 1 \,\} \cup \{\, b^n a \mid n \geq 1 \,\}$. According to [4], the language L_3 belongs to both the classes $\mathcal{L}(EC, DEF)$ and $\mathcal{L}(EC, SUF)$. Let us assume that the language L_3 is also generated by a contextual grammar

$$G = (\{\, a, b \,\}, (P_1, C_1), (P_2, C_2), \ldots, (P_m, C_m), B)$$

where all sets P_i $(1 \leq i \leq m)$ are circular. We consider a word $ab^k \in L$ where k is greater than the maximal length of the words in the set B. Then this word is generated by deriving another word $z \in L_3$. Hence, there is an index i with $1 \leq i \leq m$ such that $ab^k = uzv$ for some context $(u,v) \in C_i$ and $z \in P_i$. Since $z \in L_3$ the word z has the form ab^l with $1 \leq l < k$. Consequently, the context (u,v) is (λ, b^s) for a number $s \geq 1$. Since $z \in P_i$ and the set P_i is circular, it also contains the word $b^l a$ which belongs to the language L_3. According to our assumption, this word $b^l a$ belongs to the language $L(G)$ and therefore derives the word $b^l ab^s \in L(G)$ which does not belong to the language L_3. Hence, the two languages L_3 and $L(G)$ are not equal which contradicts the assumption. Thus, the language L_3 does not belong to the class $\mathcal{L}(EC, CIRC)$.

The languages L_1, L_2, and L_3 are witnesses for the claimed incomparability. □

The language $L_3 = \{\, ab^n \mid n \geq 1 \,\} \cup \{\, b^n a \mid n \geq 1 \,\}$ from the proof of the previous theorem also belongs to the class $\mathcal{L}(EC, REG)$. This yields the following statement.

Corollary 16 *We have the proper inclusion $\mathcal{L}(EC, CIRC) \subset \mathcal{L}(EC, REG)$.* □

We now give some results on the class of languages generated by contextual grammars with ordered selection.

Let $P_1 = \{\, a^n b^n \mid n \geq 1 \,\}$, $P_2 = \{\, b^n \mid n \geq 1 \,\}$, and $L = P_1 \cup P_2$. The language L is generated by the contextual grammar

$$G = (\{\, a, b \,\}, (P_1', \{\, (a, b) \,\}), (P_2, \{\, (\lambda, b) \,\}), \{\, ab, b \,\})$$

with $P_1' = \{b\}^*\{a\}\{a, b\}^*$. The set P_1' can be accepted by the finite automaton

$$A = (\{\, a, b \,\}, \{\, z_0, z_1 \,\}, z_0, \{\, z_1 \,\}, \delta)$$

with the transition function δ defined by $\delta(z_0, a) = \delta(z_1, a) = z_1$ and $\delta(z_i, b) = z_i$ for $0 \leq i \leq 1$. Both the sets P_1' and P_2 are ordered. This leads us to the following conclusion.

Corollary 17 *The language $L = \{\, a^n b^n \mid n \geq 1 \,\} \cup \{\, b^n \mid n \geq 1 \,\}$ belongs to the class $\mathcal{L}(EC, ORD)$.* □

In [4], it was shown that the language $L_1 = \{\, a^n b^n \mid n \geq 1 \,\} \cup \{\, cb^n \mid n \geq 1 \,\}$ does not belong to the class $\mathcal{L}(EC, DEF)$. Using similar techniques, one can also show that the language $L = \{\, a^n b^n \mid n \geq 1 \,\} \cup \{\, b^n \mid n \geq 1 \,\}$ does not belong to the class $\mathcal{L}(EC, DEF)$. Together with Corollary 17, we obtain the following statement.

Theorem 18 *There exists a language in the set $\mathcal{L}(EC, ORD) \setminus \mathcal{L}(EC, DEF)$.*□

By Lemma 2, the class $\mathcal{L}(EC, COMB)$ is a subset of the class $\mathcal{L}(EC, DEF)$ (in [4], even its properness was shown) and of the class $\mathcal{L}(EC, ORD)$. This yields the following result.

Corollary 19 *The proper inclusion $\mathcal{L}(EC, COMB) \subset \mathcal{L}(EC, ORD)$ holds.* □

We use the language $L = \{\, a^n b^n \mid n \geq 1 \,\} \cup \{\, b^n \mid n \geq 1 \,\}$ given above to show the next statement.

Theorem 20 *The proper inclusion $\mathcal{L}(EC, NIL) \subset \mathcal{L}(EC, ORD)$ holds.*

Proof. By Lemma 2, we know the inclusion $\mathcal{L}(EC, NIL) \subseteq \mathcal{L}(EC, ORD)$. Let us consider the language $L = \{\, a^n b^n \mid n \geq 1 \,\} \cup \{\, b^n \mid n \geq 1 \,\}$ given above. By Corollary 17, we have $L \in \mathcal{L}(EC, ORD)$.

Assume, $L \in \mathcal{L}(EC, NIL)$. Then there is a contextual grammar

$$G = (\{\, a, b \,\}, (P_1, C_1), (P_2, C_2), \ldots, (P_m, C_m), B)$$

where all sets P_i $(1 \leq i \leq m)$ are nilpotent.

Let μ be the maximal length of all words in the base set B and the finite sets P_i:

$$\mu = \max \{\, |w| \mid w \in B \text{ or } w \in P_i \text{ and } \#(P_i) < \infty, 1 \leq i \leq m \,\}\,.$$

Let κ be the maximal length of all contexts:

$$\kappa = \max \{\, |u| + |v| \mid (u, v) \in C_i, 1 \leq i \leq m \,\}\,.$$

We now consider a word $a^n b^n \in L$ with $n > \mu + \kappa$. Then there exists a word $z \in L(G)$, an index i with $1 \leq i \leq m$, and a context $(u, v) \in C_i$ such that $z \in P_i$ and $a^n b^n = uzv$. Since $n > \kappa \geq |u|$ and $n > |v|$ there are numbers $s \geq 0$ and $t \geq 0$ such that $u = a^s$, $v = b^t$, and $s + t \geq 1$.

Because of the choice of n, we further have $|z| > \mu$. Hence, P_i is infinite (in the finite selection languages, there is no word with a length greater than μ). Since it is nilpotent, the set $C = \{\, a, b \,\}^* \setminus P_i$ is finite. Hence, the set C contains only finitely many words of the form $a^k b^k$ or b^k. Thus, the set P_i contains infinitely many words of the form $a^k b^k$ and infinitely many words of the form b^k. If $s \neq t$, we obtain from a word $a^k b^k \in P_i$ a word which is not in the language L. Hence, $s = t$. But then from a word $b^k \in P_i$ we can derive the word $ub^k v = a^s b^{k+s}$ which is due to $s \geq 1$ not a word of the language L. Thus, $L \notin \mathcal{L}(EC, NIL)$. □

We now discuss union-free languages. By the classical theorem by Kleene we know that every regular language over an alphabet V can be constructed from the sets $\{x\}$ for $x \in V$ and the empty set by finitely many operations of union, concatenation and Kleene-star. We now present a version of this statement using union-free sets.

Lemma 21 *Every regular language can be represented by a finite union of union-free languages.*

Proof. Let V be an alphabet. The atomic regular languages empty set and $\{x\}$ for $x \in V$ are union-free. Let U and W be two regular languages over V with representations as finite unions

$$U = U_1 \cup U_2 \cup \cdots \cup U_n \text{ and } W = W_1 \cup W_2 \cup \cdots \cup W_m$$

of union-free languages U_i for $1 \leq i \leq n$ and W_i for $1 \leq i \leq m$. Then also the languages $U \cup W$, $U \cdot W$, and U^* can be represented by a finite union of union-free languages:

$$\begin{aligned} U \cup W &= U_1 \cup U_2 \cup \cdots \cup U_n \cup W_1 \cup W_2 \cup \cdots \cup W_m, \\ U \cdot W &= (U_1 \cup U_2 \cup \cdots \cup U_n) \cdot (W_1 \cup W_2 \cup \cdots \cup W_m) = \bigcup_{i=1}^{n} \bigcup_{j=1}^{m} U_i \cdot W_j, \\ U^* &= (U_1 \cup U_2 \cup \cdots \cup U_n)^* = (U_1^* U_2^* \ldots U_n^*)^*. \end{aligned}$$

Hence, any regular language can be represented by a finite union of union-free languages. □

Obviously, there exist non-union-free regular languages. However, the previous statement helps us to show that the restriction to union-free selection languages does not imply a restriction of the generative capacity of contextual grammars.

Theorem 22 *We have the equality $\mathcal{L}(EC, UF) = \mathcal{L}(EC, REG)$.*

Proof. According to Lemma 2, we have $\mathcal{L}(EC, UF) \subseteq \mathcal{L}(EC, REG)$. Hence, we only have to show that the inclusion $\mathcal{L}(EC, REG) \subseteq \mathcal{L}(EC, UF)$ also holds.

Let $L \in \mathcal{L}(EC, REG)$ and $G = (V, (P_1, C_1), (P_2, C_2), \ldots, (P_n, C_n), B)$ be a contextual grammar generating L where the sets $P_1, P_2, \ldots, P_n$ are arbitrary regular languages. By Lemma 21, every language P_i $(1 \leq i \leq n)$ can be represented as a union $P_i = Q_{i,1} \cup Q_{i,2} \cup \cdots \cup Q_{i,n_i}$ of union-free languages $Q_{i,j}$ with $1 \leq j \leq n_i$ for a natural number $n_i \geq 1$. The contextual grammar

$$G' = (V, (Q_{1,1}, C_1), \ldots, (Q_{1,n_1}, C_1), \ldots, (Q_{n,1}, C_n), \ldots, (Q_{n,n_n}, C_1), B)$$

generates the same language as G because any context can by added by G if and only if it can be added by G'. Thus, for any language $L \in \mathcal{L}(EC, REG)$, there is also a contextual grammar that generates the language L and has only union-free selection languages. This proves the inclusion $\mathcal{L}(EC, REG) \subseteq \mathcal{L}(EC, UF)$. □

Note that the results of this section come as an extension of the hierarchy of classes of languages generated by contextual grammars with subregular choice reported in [4]. Once again, it is worth mentioning that all these results are valid for alphabets that contain at least two letters, except for Theorem 14, where three letters are needed; it remains to be settled whether these bounds are optimal. It also seems interesting how the hierarchies defined in the previous section can be compared with the hierarchy developed in [4] and here. It seems interesting to see if results similar to the ones shown in this and the previous section can be derived for internal contextual grammars with subregular selection.

References

[1] H. Bordihn, M. Holzer, M. Kutrib, Determinization of finite automata accepting subregular languages, Theoretical Computer Science 410 (2009) 3209–3222.

[2] J. Brzozowski, G. Jirásková, B. Li, Quotient complexity of ideal languages, in: LATIN 2010, volume 6034 of *LNCS*, Springer-Verlag Berlin, 2010, pp. 208–221.

[3] J. Dassow, Subregularly controlled derivations: the context-free case, Rostocker Mathematisches Kolloquium 34 (1988) 61–70.

[4] J. Dassow, Contextual grammars with subregular choice, Fundamenta Informaticae 64 (2005) 109–118.

[5] J. Dassow, Grammars with commutative, circular, and locally testable conditions, in: Automata, Formal Languages, and Related Topics – Dedicated to Ferenc Gécseg on the occasion of his 70th birthday, University of Szeged, 2009, pp. 27–37.

[6] J. Dassow, H. Hornig, Conditional grammars with subregular conditions, in: Proc. Internat. Conf. Words, Languages and Combinatorics II, World Scientific Singapore, 1994, pp. 71–86.

[7] J. Dassow, F. Manea, B. Truthe, Networks of Evolutionary Processors with Subregular Filters, in: A.H. Dediu, S. Inenaga, C. Martín-Vide (Eds.), Language and Automata Theory and Applications. Fifth International Conference, LATA 2011, Tarragona, Spain, May 26 – 31, 2011, volume 6638 of *LNCS*, Springer-Verlag, 2011, pp. 262–273.

[8] J. Dassow, R. Stiebe, B. Truthe, Generative capacity of subregularly tree controlled grammars, International Journal of Foundations of Computer Science 21 (2010) 723–740.

[9] Y.S. Han, K. Salomaa, State complexity of basic operations on suffix-free regular languages, Theoretical Computer Science 410 (2009) 2537–2548.

[10] Y.S. Han, K. Salomaa, D. Wood, Nondeterministic state complexity of basic operations for prefix-suffix-free regular languages, Fundamenta Informaticae 90 (2009) 93–106.

[11] M. Holzer, S. Jakobi, M. Kutrib, The magic number problem for subregular language families, in: Proceedings of 12th Internat. Workshop Descriptional Complexity of Formal Systems, University of Saskatchewan, Saskatoon, 2010, pp. 135–146.

[12] S. Istrail, Gramatici contextuale cu selectiva regulata, Stud. Cerc. Mat. 30 (1978) 287–294.

[13] G. Jirásková, T. Masopust, Complexity in union-free languages, in: Developments in Language Theory, volume 6224 of *LNCS*, Springer-Verlag Berlin, 2010, pp. 255–266.

[14] F. Manea, B. Truthe, Accepting Networks of Evolutionary Processors with Subregular Filters, in: P. Dömösi, S. Iván (Eds.), Automata and Formal Languages – 13th International Conference AFL 2011, Debrecen, Hungary, August 17–22, 2011, Proceedings, College of Nyíregyháza, 2011, pp. 300–314.

[15] S. Marcus, Contextual grammars, Revue Roum. Math. Pures Appl. 14 (1969) 1525–1534.

[16] G. Păun, Marcus Contextual Grammars, Kluwer Publ. House, Doordrecht, 1998.

[17] G. Rozenberg, A. Salomaa (Eds.), Handbook of Formal Languages, Springer-Verlag, Berlin, 1997.