Nonterminal Complexity of Some Operations on Context-Free Languages

Jürgen Dassow and Ralf Stiebe
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
PSF 4120, D-39016 Magdeburg, Germany

Abstract: We investigate context-free languages with respect to the measure *Var* of descriptional complexity, which gives the minimal number of nonterminals which is necessary to generate the language. Especially, we consider the behaviour of this measure with respect to operations. For given numbers $c_1, c_2, \ldots, c_n$ and an n-ary operation τ on languages we discuss the range of $Var(\tau(L_1, L_2, \ldots, L_n))$ where, for $1 \leq i \leq n$, L_i is a context-free language with $Var(L_i) = c_i$. The operation under discussion are the six AFL-operations union, concatenation, Kleene-closure, homomorphisms, inverse homomorphisms and intersections by regular sets.

1 Introduction

With respect to finite automata the number of states is the most natural and most investigated measure of descriptional complexity. For a given regular language L, its state complexity $c(L)$ can be defined as the number of states of a minimal automaton $\mathcal{A}$ which accepts L. Early papers concerning $c(L)$ are e.g. [9, 7]. A very natural question is the following one: Given n numbers $c_1, c_2, \ldots, c_n$ and an n-ary operation τ on languages, which values are possible for $c(\tau(L_1, L_2, \ldots, L_n))$ where, for $1 \leq i \leq n$, L_i is a regular language with $c(L_i) = c_i$. In the last years there appeared a lot of papers which have discussed the following special version: For $c_1, c_2, \ldots, c_n$ and τ, let $f'_\tau(c_1, c_2, \ldots, c_n)$ be the maximum of $c(\tau(L_1, L_2, \ldots, L_n))$ where the maximum is taken over all regular languages L_i with $c(L_i) = c_i$, $1 \leq i \leq n$. This problem has been solved for some operations, e.g., $f'_\cup(m, n) = mn$ and $f'_\cdot(m, n) = (2m - 1)2^{n-1}$. We refer to [1, 12, 6, 5] and the summarizing articles [10, 11]. In [4] this question is considered with respect to nondeterministic automata.

However, the question can be asked a little bit more general: For $c_1, c_2, \ldots, c_n$ and τ, let $r'_\tau(c_1, c_2, \ldots, c_n)$ be the set of all numbers $c(\tau(L_1, L_2, \ldots, L_n))$ where L_i is a regular with $c(L_i) = c_i$, $1 \leq i \leq n$. In [5] $r'_C(n)$, where C denotes the complementation, is partially determined.

Surprisingly, there are almost no results in this direction with respect to the descriptional complexity of context-free languages. The measure which corresponds to the state complexity is the number of nonterminals (if one restricts to regular grammars with rules of the form $A \to aB$ or $A \to \lambda$, where A and B are nonterminals and a is a terminal, then the number of nonterminals equals the state complexity with respect to nondeterministic finite automata). Formally, for a context-free grammar $G = (N, T, P, S)$ (with the sets N, T and P of nonterminals, terminals and productions, respectively, and the axiom S) we define $Var(G)$ as the cardinality of N. For a context-free language L, we set

$$Var(L) = \min\{ Var(G) \mid G \text{ is a context-free grammar and } L(G) = L\}.$$

This complexity measure was originally studied by J. GRUSKA (see [3]). As above now we can define the set $r_\tau(c_1, c_2, \ldots, c_n)$ of all numbers $Var(\tau(L_1, L_2, \ldots, L_n))$ where, for $1 \leq i \leq n$, L_i is a context-free language with $Var(L_i) = c_i$. In [8] GH. PĂUN has partially determined $r_\cup(m, n)$ and $r_*(n)$, more precisely, he has shown that

$$\{1, 3, 4, 5, \ldots, n\} \subseteq r_\cup(n, n) \quad \text{and} \quad \{1, 2, \ldots, n\} \subseteq r_*(n).$$

Moreover, he also discussed $r_\cap(n, n)$; however, this is not of such general interest since the class of context-free languages is not closed under intersection in general.

In this paper we discuss the general case for the operations defining an abstract family of languages (under which the family of context-free languages is closed). Thus we study $r_\tau(n, m)$ for τ being union and concatenation, and $r_\tau(n)$ for τ being Kleene-closure, homomorphisms, inverse homomorphisms and intersection with regular sets. For union, Kleene-closure, homomorphisms, inverse homomorphisms, and intersections with regular sets we determine the sets completely; for concatenation we only present a partial solution. For union, concatenation, Kleene-closure, and homomorphisms, we get especially the maximal value of $r_\tau(n, m)$ and $r_\tau(n)$, respectively, and we prove that such a maximal value does not exist for inverse homomorphisms and intersections with regular sets.

Throughout the paper we assume that the reader is familiar with basic concepts of the theory of (context-free) languages.

2 Nonterminal Complexity of Some Context-Free Languages

We start with the determination of the complexity of some languages which are needed later.

Lemma 2.1 *Let $i_1, i_2, \ldots, i_{2n}$ be $2n$ pairwise different positive integers and*

$$L = \{ab^{i_1}\}^*\{ab^{i_2}\}^* \ldots \{ab^{i_{2n}}\}^*.$$

Then $Var(L) = n$.

Proof. Let $m = \max\{i_1, i_2, \ldots, i_{2n}\}$. Let $G = (N, T, P, S)$ be a context-free grammar such that $L(G) = L$ and $Var(G) = Var(L)$. First we show that, for any nonterminal A different from S, there is a rule $A \to xAy$ with $xy \neq \lambda$. Let us assume the contrary. If there is no rule $A \to w$ in P where A occurs in w, we can construct a grammar G' by replacing any occurrence of A on a right hand side of a production by all right hand sides of productions with left hand side A and omitting all rules with left hand side A. Obviously, $L(G') = L$ and $Var(L) \leq Var(G') = Var(G) - 1 < Var(G) = Var(L)$ which is a contradiction. Thus there is a rule $A \to xAy$. If $xy = \lambda$, we can omit this rule without changing the language. Thus $xy \neq \lambda$.

We only discuss the case $x \neq \lambda$; the case $y \neq \lambda$ can be handled analogously. Obviously, G has to be reduced, i.e., there is a derivation

$$S \Longrightarrow^* uAv \Longrightarrow^* uwv \in L(G).$$

Moreover, let $x \Longrightarrow x' \in T^*$ and $y \Longrightarrow^* y' \in T^*$ two terminating derivations. Then, for any $n \geq 0$, we have a derivation

$$S \Longrightarrow^* uAv \Longrightarrow^* u(x')^n A(y')^n v \Longrightarrow^* u(x')^n w(y')^n v \in L(G) = L$$

If $n \geq 2m+1$, then x^n contains a subword $ab^{i_j}a$ for some j. Assume that there are i_k, $k \neq j$, and a derivation $A \Longrightarrow^* x''Ay''$ where x'' contains the subword $ab^{i_k}a$. Then we have the derivation

$$\begin{aligned} S \Longrightarrow^* uAv \quad &\Longrightarrow^* \quad ux''Ay''v \Longrightarrow^* ux''(x')^n A(y')^n y''v \Longrightarrow^* ux''(x')^n x''Ay''(y')^n y''v \\ &\Longrightarrow^* \quad ux''(x')^n x''wy''(y')^n y''v = p \in L(G) \end{aligned}$$

which generates a word containing a subword $ab^{i_k}azab^{i_j}az'ab^{i_j}a$ which is not in L. Thus a letter A can only contribute to one ab^{i_j} to the left. Analogously, A can only contribute to one $ab^{i_{j'}}$ to the right.

If there is a derivation $S \Longrightarrow^* xSy$ with $xy \neq \lambda$, the same argumentation holds for S. Since we have $2n$ numbers $i_1, i_2, \ldots, i_n$, we need at least n nonterminals for the generation of L, i.e., $Var(G) \geq n$. If there is no derivation $S \Longrightarrow^* xSy$ with $xy \neq \lambda$, then we need n additional letters to generate all sets $\{ab^{i_j}\}^*$, i.e., $Var(G) \geq n+1$. Hence, $Var(L) \geq n$.

On the other hand, since

$$(\{A_1, A_2, \ldots, A_n\}, \{a, b\}, P, A_1)$$

with

$$\begin{aligned} P \quad = \quad & \{A_n \to ab^{i_n} A_n, A_n \to A_n ab^{i_{n+1}}, A_n \to \lambda\} \\ & \cup \bigcup_{j=1}^{n-1} \{A_j \to ab^{i_j} A_j, A_j \to A_j ab^{i_{2n-j+1}}, A_j \to A_{j+1}\} \end{aligned}$$

generates L, we have $Var(L) \leq n$.

Thus $Var(L) = n$. □

Lemma 2.2 *Let $i_1, i_2, \ldots, i_{2n}$ be $2n$ pairwise different positive integers and*

$$\begin{aligned} L \quad = \quad & \{(ab^{i_1})^{k_1}(ab^{i_2})^{k_2} \ldots (ab^{i_n})^{k_n}(ab^{i_{n+1}})^{k_n}(ab^{i_{n+2}})^{k_{n-1}} \ldots (ab^{i_{2n}})^{k_1} \\ & \mid k_1, k_2, \ldots, k_n \geq 0\}. \end{aligned}$$

Then $Var(L) = n$.

Proof. The proof can be given analogously to Lemma 2.1. □

The following lemma is essentially shown in [3].

Lemma 2.3 *Let $i_1, i_2, \ldots, i_n$ be $n \geq 2$ pairwise different positive integers and*

$$L = \bigcup_{j=1}^{n} \{ab^{i_j}\}^*.$$

Then $Var(L) = n + 1$.

Lemma 2.4 *Let $i_1, i_2, \ldots, i_n$ and $j_1, j_2, \ldots, j_m$ be $n \geq 1$ and $m \geq 1$ pairwise different integers such that $i_l \geq 2$ and $j_k \geq 2$ for $1 \leq l \leq n$ and $1 \leq k \leq m$, respectively, and*

$$L = \{ba^{j_1}, ba^{j_2}, \ldots, ba^{j_m}\}^* \cup \bigcup_{j=1}^{n} \{ab^{i_j}\}^*.$$

Then $Var(L) = n + 2$.

Proof. Let $G = (N, T, P, S)$ be a context-free grammar with $L(G) = L$ and $Var(G) = Var(L)$. As above, we can show that, for any nonterminal A different from S, there is a derivation $A \Longrightarrow^* xAy$ such that x contains a subword $ab^{i_j}a$ or y contains a subword $ab^{i_j}a$ for some j, $1 \leq j \leq n$, or x contains a subword $ba^{j_k}b$ or y contains a subword $ba^{j_k}b$ for some k, $1 \leq k \leq m$. We say that A belongs to ab^{i_j} or to ba^{j_k}, respectively. It is easy to see that A cannot belong to two different words w and w' such that both words are in $M = \{ab^{i_1}, ab^{i_2}, \ldots, ab^{i_n}\}$ or one word is in M and the other is in $M' = \{ba^{j_1}, ba^{j_2}, \ldots, ba^{j_m}\}$. For example, let $w = ab^{i_j} \in M$ and $w' = ba^{j_k} \in M$. Then there are derivations $A \Longrightarrow^* xAy$ and $A \Longrightarrow^* x'Ay'$ where x' contains the subword $ab^{i_j}a$ and y' contains the subword $ba^{j_k}b$ (the other possibilities for the containments in x, x', y, y' can be handled analogously). Then we have a derivation

$$S \Longrightarrow^* uAv \Longrightarrow^* uxAyv \Longrightarrow uxx'Ay'yv \Longrightarrow^* uxx'wy'yv = z \in L(G).$$

However, z contains both subwords $ab^{i_j}a$ and $ba^{j_k}b$ and therefore z contains the subwords a^2 and b^2 which is impossible for words in L. Thus we have a contradiction to $L(G) = L$. (If w and w' belong to ab^{i_j} and ab^{i_k}, $j \neq k$; then z contains the subwords $ab^{i_j}a$ and $ab^{i_k}a$ which is impossible, too.)

Thus any nonterminal different from S belongs to only one word of M or to (possibly some) words of M'.

If there is a derivation $S \Longrightarrow^* xSy$ for some x and y with $xy \neq \lambda$, then with respects to containment of subwords we have the same situations as above. Assume that x contains wa for some $w \in M$. Then we have a derivation $S \Longrightarrow^* xSy \Longrightarrow^* xba^{j_1}y \in L(G)$ but $xba^{j_1}y$ contains the subwords a^2 and b^2 which contradicts $L(G) = L$. Hence there is no derivation $S \Longrightarrow^* xSy$. Therefore the generation of $\{w\}^*$ with $w \in M$ or $(M')^*$ needs a certain nonterminal $A \neq S$ which belongs to w or to some words of M', respectively. Since any nonterminal $A \neq S$ cannot belong to two words of M or to one word in M and one word in M' simultaneously, we need at least $n + 1$ nonterminals which are different from the axiom and the axiom, i.e., $Var(L) = Var(G) \geq n + 2$.

On the other hand, $H = (\{S, A_1, A_2, \ldots, A_n, B\}, \{a, b, c\}, P, S)$ with

$$P = \{S \to B\} \cup \bigcup_{k=1}^{m} \{B \to ba^{j_k}B, B \to \lambda\} \cup \bigcup_{j=1}^{n} \{S \to A_j, A_j \to ab^{i_j}A_j, A_j \to \lambda\}$$

generates L which implies $Var(L) \leq Var(H) = n+2$.

Thus $Var(L) = n+2$. $\square$

Lemma 2.5 *Let $i_1, i_2, \ldots, i_n$ be $n \geq 1$ pairwise different positive natural numbers and*

$$L = \{b\}^* \cup \bigcup_{j=1}^{n} \{ab^{i_j}\}^*.$$

Then $Var(L) = n+2$.

Proof. Again, let $G = (N, T, P, S)$ be a context-free grammar such that $L = L(G)$ and $Var(L) = Var(G)$. For $1 \leq j \leq n$, any derivation of $(ab^{i_j})^m$ for sufficiently large m contains a subderivation $A_j \Longrightarrow^* xA_jy$ such that x or y contains the subword $ab^{i_j}a$. Analogously, any derivation of b^m for sufficiently large m contains a subderivation $B \Longrightarrow xBy$ such that x or y contains a subword b^r with $r > i_j$ for $1 \leq j \leq n$. As in the proof of Lemma 2.4 we can show that all the letters $A_1, A_2, \ldots, A_n, B$ have to be different and different from the axiom. Thus $Var(L) \geq n+2$.

It is easy to prove that there is a grammar with $n+2$ nonterminals which generates L. Thus $Var(L) = n+2$. $\square$

Lemma 2.6 *For $L = a\{a,b\}^*a\{a,b\}^*$, $Var(L) = 2$.*

Proof. Clearly $Var(L) \leq 2$, since L is generated by

$$G = (\{S, B\}, \{a, b\}, \{S \to aBaB, B \to aB, B \to bB, B \to \lambda\}, S).$$

On the other hand, let H be some grammar with the single nonterminal S and let k be the greatest length of a right hand side in the rules of H. If there is a terminating rule $S \to w$ with $w \notin L$, then $L(H)$ contains $w \notin L$. Otherwise, all words in $L(H)$ contain a subword of length $\leq k$ which is in L; however, $ab^ka \in L$ does not contain a subword of length $\leq k$ which is in L, too. In both cases, we obtain $L(H) \neq L$, which means that $Var(L) \geq 2$. $\square$

Lemma 2.7 *Let $i_1, i_2, \ldots, i_n$ be $n \geq 1$ pairwise different positive natural numbers,*

$$L = \{b\}\{a,b\}^* \cup \bigcup_{j=1}^{n} \{ab^{i_j}\}^* \text{ and } L' = \{b\}\{a,b\}^*\{b\}\{a,b\}^* \cup \bigcup_{j=1}^{n} \{ab^{i_j}\}^*.$$

Then $Var(L) = Var(L') = n+2$.

Proof. The proof can be given analogously to that of Lemma 2.5. $\square$

Lemma 2.8 *Let $i_1, i_2, \ldots, i_n$ be $n \geq 1$ pairwise different positive integers, $i \geq 2$ and*

$$L = \{a^i\} \cup \bigcup_{j=1}^{n} \{ab^{i_j}\}^*.$$

Then $Var(L) = n+1$.

Proof. Let $n \geq 2$. As in the proof of Lemma 2.4 we can show that, for any number i_j, there is at most one nonterminal A_j which belongs to ab^{i_j} and that we need in addition to these n nonterminals an axiom. Thus $Var(L) \geq n+1$.

Now let $n = 1$. Let L be generated by a context-free grammar $G = (\{S\}, \{a,b\}, P, S)$ (with only one nonterminal). Since L is infinite, there is a derivation $S \Longrightarrow^* xSy$ with $xy \in \{a,b\}^+$. By iterating this derivation we get $S \Longrightarrow^* x^4Sy^4$ where at least one of the words x^4 or y^4 has length 4 and therefore it contains b. Thus we also have a derivation $S \Longrightarrow^* x^4a^iy^4 \in L(G)$. However, this contradicts $L = L(G)$ since $x^4a^iy^4$ contains the subwords b and a^2 (since $i \geq 2$) which is impossible for words in L. Therefore we need at least two nonterminals, i.e. $Var(L) \geq 2 = n+1$.

On the other hand

$$\Big(\{S, A_1, A_2, \ldots, A_n\}, \{a,b\}, \{S \to a^i\} \cup \bigcup_{j=1}^{n} \{S \to A_j, A_j \to ab^{i_j}A_j, A_j \to \lambda\}, S\Big)$$

generates L which proves $Var(L) \leq n+1$. □

Lemma 2.9 *For any context-free language L over a unary alphabet, $Var(L) \leq 2$.*

Proof. It is well-known that any context-free language over a unary alphabet consisting of the letter a can be represented as $L = U \cup \{a^p\}^*U'$ where U and U' are finite sets. Thus L can be generated by

$$(\{S, A\}, \{a\}, \{S \to w \mid w \in U\} \cup \{S \to A, B \to a^pB\} \cup \{B \to v \mid v \in U'\}, S)$$

which proves the statement. □

3 Nonterminal Complexity of Union

In this section we study the behaviour of nonterminal complexity with respect to union.

Theorem 3.1 *i) For any two context-free languages L_1 and L_2,*

$$Var(L_1 \cup L_2) \leq Var(L_1) + Var(L_2) + 1.$$

ii) For any three numbers $n \geq 1$, $m \geq 1$ and k such that $k \leq n+m+1$ and any alphabet T with at least two letters, there are context-free languages $L_n \subseteq T^*$ and $K_m \subseteq T^*$ such that

$$Var(L_n) = n, \quad Var(K_m) = m \quad \text{and} \quad Var(L_n \cup K_m) = k.$$

Proof. i) The statement follows by the standard construction to prove the closure of the family of context-free languages under union (one adds $S \to S_1$ and $S \to S_2$ where S is the new axiom).

ii) Without loss of generality we assume that $n \geq m$.

Let $n \geq 1$, $m \geq 1$ and $k = n+m+1$. We choose

$$L_n = \{ab\}^*\{ab^2\}^* \ldots \{ab^{2n}\}^* \quad \text{and} \quad K_m = \{ab^{2n+1}\}^*\{ab^{2n+2}\}^* \ldots \{ab^{2n+2m}\}^*.$$

By Lemma 2.1, we have $Var(L_n) = n$ and $Var(K_m) = m$. We now prove that $Var(L_n \cup K_m) = n + m + 1$.

Let $G = (N, \{a, b\}, P, S)$ be a context-free grammar with $L(G) = L_n \cup K_m$. As in the proof of Lemma 2.1 we can show that we need at least $n + m$ nonterminals in order to generate words with ab^i, $1 \le i \le 2n + 2m$.

Let us assume that one of these symbols, say A which generates ab^j with $1 \le j \le 2n$ (the case $2n + 1 \le j \le 2n + 2m$ can be handled analogously), is the axiom. Then there is a derivation

$$A \Longrightarrow^* uab^j au' Av \Longrightarrow^* uab^j au' ab^{2n+1} ab^{2n+2m} v \notin L_n \cup K_m$$

or

$$A \Longrightarrow^* uAvab^j av' \Longrightarrow^* uab^{2n+1} ab^{2n+2m} vab^j av' \notin L_n \cup K_m.$$

Thus we need in addition to the $n + m$ nonterminals a further nonterminal as the axiom. Hence $Var(L_n \cup K_m) \ge n + m + 1$. By the part i), we get $Var(L_n \cup K_m) = n + m + 1$.

Let $n \ge 2$, $m \ge 1$ and $k = n + m$. Then we consider

$$L_n = \{ab^1\}^*\{ab^2\}^* \dots \{ab^{2n}\}^*$$

and

$$\begin{aligned} K_m &= \{ab^{2n+1}\}^*\{ab^{2n+2}\}^* \dots \{ab^{2n+m-1}\}^*\{ab^n\}^* \\ &\quad \cdot \{ab^{n+1}\}^*\{ab^{2n+m+2}\}^*\{ab^{2n+m+3}\}^* \dots \{ab^{2n+2m}\}^*. \end{aligned}$$

By Lemma 2.1, $Var(L_n) = n$ and $Var(K_m) = m$.

Let $G = (N, T, P, S)$ be a context-free grammar with $L(G) = L_n \cup K_m$. As in the case $k = m + n + 1$ we can show that we need $n + m - 1$ nonterminals to generate words containing ab^j, $1 \le j \le 2n + 2m$, $j \ne 2n + m$ and $j \ne 2n + m + 1$ and in addition an axiom. Thus $Var(L_n \cup K_m) \ge n + m$.

The context-free grammar

$$H = (\{S, A_1, A_2, \dots, A_n, B_1, B_2, \dots, B_{m-1}\}, \{a, b\}, P, S)$$

with

$$\begin{aligned} P &= \{S \to A_1, S \to B_1\} \cup \bigcup_{i=1}^{n-1} \{A_i \to ab^i A_i, A_i \to A_i ab^{2n-i+1}, A_i \to A_{i+1}\} \\ &\quad \cup \{A_n \to ab^n A_n, A_n \to A_n ab^{n+1}, A_n \to \lambda\} \\ &\quad \cup \bigcup_{i=1}^{m-2} \{B_i \to ab^{2n+i} B_i, B_i \to B_i ab^{2n+2m-i+1}, B_i \to B_{i+1}\} \\ &\quad \cup \{B_{m-1} \to ab^{2n+m-1} B_{m-1}, B_{m-1} \to B_{m-1} ab^{2n+m+2}, B_{m-1} \to A_n\}. \end{aligned}$$

It is easy to see that $L(H) = L_n \cup K_m$ and $Var(H) = n + m$. Hence $Var(L_n \cup K_m) = n + m$.

Let $k = n + m$ and $n = 1$. Then $m = 1$ and $k = 2$. It is easy to see that $Var(\{ab^2\}^*) = 1$ and $Var(\{a^3\}) = 1$. By Lemma 2.8, we have $Var(\{a^3\} \cup \{ab^2\}^*) = 2$.

Let $n \geq m \geq 3$ and $n > k \geq 3$. We consider the languages

$$L_n = \bigcup_{j=2}^{n-1} \{ab^j\}^* \cup b\{a,b\}^* \text{ and } K_m = \bigcup_{j=2}^{m-1} \{ba^j\}^* \cup \{ab^{k-1}, ab^k, \ldots, ab^{n-1}\}^*.$$

We obtain $L_n \cup K_m = \bigcup_{j=2}^{k-2}\{ab^j\}^* \cup \{ab^{k-1}, ab^k, \ldots, ab^{n-1}\}^* \cup b\{a,b\}^*$. By Lemmas 2.5 and 2.7, $Var(L_n) = n$, $Var(K_m) = m$. Analogously to the proof in Section 2 it can be shown that $Var(L_n \cup K_m) = k$.

Let $n \geq m \geq 3$ and $n \leq k < n+m$. We consider the languages

$$L_n = \bigcup_{j=1}^{n-1} \{ab^j\}^* \text{ and } K_m = \bigcup_{j=k-m}^{k-1} \{ab^j\}^*.$$

We obtain $L_n \cup K_m = \bigcup_{j=1}^{k-1}\{ab^j\}^*$. By Lemma 2.3, $Var(L_n) = n$, $Var(K_m) = m$, and $Var(L_n \cup K_m) = k$.

Let $n \geq m \geq 3$ and $k = 2$. We consider the languages

$$L_n = \bigcup_{j=1}^{n-2} \{ab^j\}^* \cup b\{a,b\}^*b\{a,b\}^* \text{ and } K_m = \bigcup_{j=1}^{m-2} \{ba^j\}^* \cup a\{a,b\}^*a\{a,b\}^*.$$

By Lemma 2.7 and symmetry $Var(L_1) = n$ and $Var(L_2) = m$. The union of L_n and K_m is $L = a\{a,b\}^*a\{a,b\}^* \cup b\{a,b\}^*b\{a,b\}^*$. Analogous to Lemma 2.6 it can be shown that $Var(L) = 2$.

Let $n \geq m \geq 3$ and $k = 1$. We consider the languages

$$L_n = \bigcup_{j=1}^{n-2} \{ab^j\}^* \cup b\{a,b\}^* \text{ and } K_m = \bigcup_{j=1}^{m-2} \{ba^j\}^* \cup a\{a,b\}^*.$$

By Lemma 2.7 and symmetry $Var(L_n) = n$ and $Var(K_m) = m$. Finally, the union of L_1 and L_2 is $\{a,b\}^+$ which can be generated by a grammar with one nonterminal symbol.

We omit the complete proof for the remaining cases and only give the languages such that the requirements of the statement are satisfied.

n	m	k	L_n	K_m
≥ 3	2	$n+1$	$\{a^2\} \cup \bigcup_{i=1}^{n-1}\{ab^i\}^*$	$\{ab^n\}^* \cup \{a^2\}$
≥ 3	2	$n \geq k \geq 3$	$\{a^2\} \cup \bigcup_{i=1}^{n-1}\{ab^i\}^*$	$\{ab^{k-1}, ab^k, \ldots, ab^{n-1}\}^* \cup \{a^2\}$
≥ 3	2	2	$\bigcup_{i=1}^{n-1}\{ab^i\}^*$	$\{a^n\} \cup \{ab{\cdot}ab^2, \ldots, ab^{n-1}\}^*$
≥ 3	2	1	$\{b\}^* \cup \bigcup_{i=1}^{n-2}\{ab^i\}^*$	$\{a,b\}^*\{a\}\{a,b\}^*$
≥ 3	1	$n \geq k \leq 3$	$\{a^2\} \cup \bigcup_{i=1}^{n-1}\{ab^i\}^*$	$\{ab, ab^2, \ldots, ab^{n-1}\}^*$
≥ 3	1	2	$\bigcup_{i=1}^{n-1}\{ab^i\}^* \cup \{a^2\}$	$\{ab, ab^2, \ldots, ab^{n-1}\}^*$
≥ 3	1	1	$\bigcup_{i=1}^{n-1}\{ab^i\}^*$	$\{ab, ab^2, \ldots, ab^{n-1}\}^*$
2	2	3	$\{a^2\}^* \cup \{ab\}^*$	$\{a^2\}^* \cup \{ab^2\}^*$
2	2	2	$\{a^2\}^* \cup \{ab\}^*$	$\{a^2\}^* \cup \{ab\}^*$
2	2	1	$\{a, a^3\} \cup \{a^n \mid n \geq 5\}$	$\{a^2\} \cup \{a^n \mid n \geq 4\}$
1	1	1	$\{a^2\}$	$\{a^2\}$

□

4 Nonterminal Complexity of Further Operations

In this section we study the behaviour of the complexity with respect to concatenation, Kleene-closure, homomorphisms, inverse homomorphisms and intersection with regular sets.

Theorem 4.1 *i) For any two context-free languages L_1 and L_2,*

$$Var(L_1L_2) \le Var(L_1) + Var(L_2) + 1.$$

ii) For any three numbers $n \ge 1$, $m \ge 1$ and k such that $\max\{n,m\} < k \le n+m+1$ and any alphabet T with at least two letters, there are context-free languages $L_n \subseteq T^*$ and $K_m \subseteq T^*$ such that

$$Var(L_n) = n, \quad Var(K_m) = m \quad \text{and} \quad Var(L_nK_m) = k.$$

Proof. i) The statement follows by the standard construction to prove the closure of the family of context-free languages under concatenation (one adds $S \to S_1S_2$ where S is the new axiom).

ii) Let $n \ge m$. Let $k = n+1+t$. Then $0 \le t \le m$.

We consider the languages

$$\begin{aligned} L_n \;=\; & \{(ab^{2m+1})^{r_1}(ab^{2m+2})^{r_2}\dots(ab^{m+n+t})^{r_{n+t-m}}(ab^{t+1})^{r_{n+t-m+1}}(ab^{t+2})^{r_{n+t-m+2}} \\ & \dots(ab^{m})^{r_n}(ab^{m+1})^{r_n}(ab^{m+2})^{r_{n-1}}\dots(ab^{2m-t})^{r_{n+t-m+1}} \\ & (ab^{m+n+t+1})^{r_{n+t-m}}(ab^{m+n+t+2})^{r_{n+t-m-1}}\dots(ab^{2n+2t})^{r_1} \\ & \mid r_1, r_2, \dots, r_n \ge 0\}. \end{aligned}$$

and

$$\begin{aligned} K_m \;=\; & \{(ab)^{k_1}(ab^2)^{k_2}\dots(ab^m)^{k_m}(ab^{m+1})^{k_m}(ab^{m+2})^{k_{m-1}}\dots(ab^{2m})^{k_1} \\ & \mid k_1, k_2, \dots, k_m \ge 0\}. \end{aligned}$$

(note that $2n+2t = 2m+2(m+n+t-2m)$ and $(m+n+t-2m)+(m-t) = n$). By Lemma 2.2, $Var(K_m) = m$ and $Var(L_n) = n$. Moreover, the number of different exponents of b in L_nK_m is $2m+2(m+n+t-2m) = 2(n+t)$.

Let $G = (N, \{a,b\}, P, S)$ be a grammar with $L(G) = L_nK_m)$ and $Var(G) = Var(L_nK_m)$. Assume there is a derivation $S \Longrightarrow^* xSy$ with $xy \in \{a,b\}^+$. Since $ab^{2m+1}ab^{2n+2t}abab^{2m} \in L_nK_m$, for $s \ge 2(n+t)$, we also have a derivation

$$S \Longrightarrow x^sSy^s \Longrightarrow^* x^sab^{2m+1}ab^{2n+2t}abab^{2m}y^s.$$

By the structure of the words in L_nK_m we get $x^s \in \{ab^{2m+1}\}^*$ and $y^s = \{ab^{2m}\}^*$. Moreover, in order to ensure that the powers of ab^{2m+1} and ab^{2n+2t} have to be equal and the powers of ab and ab^{2m} have to be equal for words in L_nK_m, we get $x^s = y^s = \lambda$. This contradicts our assumption. Therefore there are no sentential forms different from the axiom which contain S.

On the other hand, as in the proof of Lemma 2.1 one can show that any letter A different from the axiom has a derivation $A \Longrightarrow^* xAy$ with $xy \in \{a,b\}^+$ and can contribute to at most two subwords ab^k, $1 \leq k \leq 2n+2t$. Therefore we need an axiom and at least $n+t$ additional nonterminals. Thus $Var(L_nK_m) \geq n+t+1 = k$.

Furthermore, the grammar

$$(\{S, A_1, A_2, \ldots, A_{n+t-m}, B_1, B_2, \ldots, B_m\}, \{a,b\}, Q, S)$$

with

$$\begin{aligned} Q &= \{S \to A_1B_1, A_{n+t-m} \to ab^{m+n+t}A_{n+t-m}ab^{m+n+t+1}, A_{n+t-m} \to B_1, \\ &\quad B_m \to ab^m B_m ab^{m+1}, B_m \to \lambda\} \\ &\quad \cup \bigcup_{i=1}^{m-1} \{B_i \to ab^i B_i ab^{2m-i+1}, B_i \to B_{i+1}\} \\ &\quad \cup \bigcup_{j=1}^{n+t-m-1} \{A_j \to ab^{2m+j}A_j ab^{2n+2t-j+1}, A_j \to A_{j+1}\} \end{aligned}$$

(note $2m + (n+t-m) = n+t+m$ and $2m + 2(n+t-m) = 2n+2t$) generates L_nK_m with $1 + (n+t-m) + m = n+t+1 = k$ nonterminals. Hence $Var(L_nK_m) \leq k$.

We conclude $Var(L_nK_m) = k$.

It is easy to give the modifications for the case $m \geq n$. □

Theorem 4.2 *i) For any context-free language L, $Var(L^*) \leq Var(L)+1$.*

ii) For any two natural numbers $n \geq 1$ and k with $1 \leq k \leq n+1$, there is a context-free language L_n such that

$$Var(L_n) = n \quad \text{and} \quad Var(L_n^*) = k.$$

Proof. i) can be shown by the standard construction (use an additional nonterminal S' and additional rules $S' \to SS'$ and $S' \to \lambda$).

ii) Let $k = n+1$. We choose

$$\begin{aligned} L_n &= \{(ab)^{k_1}(ab^2)^{k_2}\ldots(ab^n)^{k_n}(ab^{n+1})^{k_n}(ab^{n+2})^{k_{n-1}}\ldots(ab^{2n})^{k_1} \\ &\quad \mid k_1, k_2, \ldots, k_n \geq 0\}. \end{aligned}$$

By Lemma 2.2, $Var(L_n) = n$, and $Var(L_n^*) = n+1$ can be proved analogously to case $k = n+m+1$ in the proof of Theorem 4.1.

The statement for $k \leq n$ was shown in [8]. □

Theorem 4.3 *i) For any context-free language L and any homomorphism h, we have $Var(h(L)) \leq Var(L)$.*

ii) For any natural numbers $n \geq 1$ and k with $1 \leq k \leq n$ and any alphabet T which consists of at least 3 letters, there are a regular language $L_n \subseteq T^*$ and a homomorphism $h_{n,k} : T^* \to T^*$ such that $Var(L_n) = n$ and $Var(h_{n,k}(L_n)) = k$.

Proof. i) The standard construction to prove that, for any context-free language L and any homomorphism h, $h(L)$ is a context-free language, too, consists in the replacement of each rule $A \to w$ by $A \to h(w)$, where $h(B) = B$ for any nonterminal B. Thus we have immediately, that $Var(h(L)) \le Var(L)$.

ii) Let $k \ge 3$. We choose

$$L_n = \bigcup_{i=0}^{n-k+1} \{ab^{3i+2}\}^+ \cup \bigcup_{j=1}^{k-2} \{ac^j\}^+$$

and define $h_{n,k}$ by

$$h_{n,k}(a) = h_{n,k}(b) = a \text{ and } h_{n,k}(c) = c.$$

Obviously,

$$h_{n,k}(L_n) = \{a^3\}^+ \cup \bigcup_{j=1}^{k-2} \{ac^j\}^+$$

It is easy to prove by methods analogous to that in Section 2 that

$$Var(L_n) = 1 + (n-k+1) + (k-2) = n \text{ and } Var(h_{n,k}(L_n)) = 2 + (k-2) = k.$$

Let $k = 2$. Let $L_n = \{a^2\} \cup \bigcup_{i=0}^{n-2} \{ab^{3i+2}\}^+$. We define $h_{n,2}$ by $h_{n,2}(a) = h_{n,2}(b) = a$ and get $h_{n,2}(L_n) = \{a^2\} \cup \{a^{3i} \mid i \ge 1\}$. By Lemma 2.5, $Var(L_n) = n$ Moreover, it is easy to see that $Var(h_{n,2}(L_n)) = 2$.

Let $k = 1$ and $n \ge 2$. We consider $L_n = \{a^3\} \cup \bigcup_{i=0}^{n-2} \{ab^{3i+2}\}^+$ and $h_{n,1}$ given by $h_{n,1}(a) = h_{n,1}(b) = a$. Then $h_{n,1}(L_n) = \{a^{3i} \mid i \ge 1\}$. By Lemma 2.8, $Var(L_n) = n$ and $Var(h_{n,1}(L_n)) = 1$ holds obviously.

For $k = n = 1$, we choose $L_1 = \{a\}^+$ and $h_{1,1}$ as the identical mapping. Then $Var(L_1) = Var(h_{1,1}(L_1)) = 1$. □

For inverse homomorphisms, in general, there is no relation between $Var(L)$ and $Var(h^{-1}(L))$ where L is a context-free language and h is a homomorphisms. More precisely, we have the following statement.

Theorem 4.4 *i) For any two natural numbers $n \ge 1$ and k with $1 \le k \le n$ and any alphabet T with at least two letters, there are a regular language $L_n \subseteq T^*$ and a homomorphism $h_{n,k} : T^* \to T^*$ such that $Var(L_n) = n$ and $Var(h_{n,k}^{-1}(L_n)) = k$.*

ii) For any three natural numbers $n \ge 1$, $m \ge 3$ and k such that $n \le k \le (m-1)(n-1)+1$, there is an alphabet T_m with at least $m+1$ letters, a regular language $L_n \subseteq T_m^*$ and a homomorphism $h_{n,k} : T_m^* \to T_m^*$ such that $Var(L_n) = n$ and $Var(h_{n,k}^{-1}(L_n)) = k$.

Proof. i) If $k \ge 2$, we choose

$$L_n = \{a^2\} \cup \bigcup_{i=1}^{k-1} \{ab^{2i}\}^+ \cup \bigcup_{i=1}^{n-k} \{ab^{2i+1}\}^+$$

and define $h_{n,k}$ by $h_{n,k}(a) = a$ and $h_{n,k}(b) = b^2$. By Lemma 2.8 we have $Var(L_n) = 1 + (k-1) + (n-k) = n$. Moreover, $h_{n,k}^{-1}(L_n) = \{a^2\} \cup \bigcup_{i=1}^{k-1} \{ab^i\}^+$. Again, by Lemma 2.8, we get $Var(h_{n,k}-1(L_n)) = 1 + (k-1) = k$.

If $n \ge 2$ and $k = 1$, we choose L_n as above and give $h_{n,k}$ by $h_{n,k}(a) = a$ and $h_{n,k}(b) = a^2b$. Then $Var(L_n) = n$ and $h_{n,k}^{-1}(L_n) = \{a^2\}$ which can obviously be generated by one nonterminal.

The modifications for the case $n = k = 1$ are left to the reader.

ii) Let $T = \{a_1, a_2, \ldots, a_{m-1}, b, c\}$. Since $(m-1)(n-1)+1 = n+(m-2)(n-1)$ any number k with $n \leq k \leq (m-1)(n-1)+1$ can be represented as $k = n+n_2+n_3+\ldots n_{m-1}$ for some n_l with $0 \leq n_l \leq n-1$, where $2 \leq l \leq m-1$. We consider the language

$$L_n = \bigcup_{i=1}^{n-1} \{b\}\{a_1 b^{mi+m}\}^* \{b^{m-2}c\} \cup \bigcup_{l=2}^{m-1} \bigcup_{j=1}^{n_l} \{b^j\}\{a_1 b^{mi+m}\}^* \{b^{m-j-1}c\}.$$

It is easy to prove by arguments analogous to those given in Section 2 that $Var(L_n) \geq n$. On the other hand,

$$G = (\{S, A_1, A_2, \ldots, A_{n-1}\}, \{a_1, b, c\}, P, S)$$

with

$$P = (\bigcup_{l=1}^{m-1} \bigcup_{j=1}^{n_l} \{S \to b^l A_j b^{l-1} c\}) \cup \bigcup_{i=1}^{n-1} \{S \to bA_i b^{m-2} d, A_i \to ab^{mi+m} A_i, A_i \to \lambda\}$$

generates L_n which proves $Var(L_n) \leq n$. Moreover, let $h_{n,k}$ be the homomorphism given by

$$h_{n,k}(a_l) = b^l a_1 b^{m-l} \text{ for } 1 \leq l \leq m-1, \; h_{n,k}(b) = b^m, \text{ and } h_{n,k}(c) = b^{m-1}c.$$

It is easy to see that $h_{n,k}(a_l b^i c) = b^l ab^{mi+m} b^{m-l-1} c$ for $1 \leq l \leq m-1$ and thus

$$h_{n,k}^{-1}(L_n) = \bigcup_{i=1}^{n-1} \{a_1 b^i\}^+ \{c\} \cup \bigcup_{l=2}^{m-1} \bigcup_{j=1}^{n_l} \{a_l b^i\}^* \{c\}.$$

Again, it is easy to prove that $Var(h_{n,k}^{-1}(L_n)) = n-1+n_2+n_3 \ldots + n_{m-1}+1 = k$. □

For the intersection by regular sets, in general, there is also no relation between $Var(L)$ and $Var(L \cap R)$.

Theorem 4.5 *For any two natural numbers $n \geq 1$ and $k \geq 1$ and any alphabet T consisting of at least two symbols, there are a context-free language $L_n \subseteq T^*$ and a regular language $R_{n,k} \subseteq T^*$ such that $Var(L_n) = n$ and $Var(L_n \cap R_{n,k}) = k$.*

Proof. If $n \geq k \geq 1$, we choose

$$L_n = \{ab\}^* \{ab^2\}^* \ldots \{ab^{2n}\}^*$$

and

$$R_{n,k} = \{ab\}^* \{ab^2\}^* \ldots \{ab^k\}^* \{ab^{2n-k+1}\}^* \{ab^{2n-k+2}\}^* \ldots \{ab^{2n}\}^*.$$

By Lemma 2.1, $Var(L_n) = n$ and $Var(L_n \cap R_{n,k}) = Var(R_{n,k}) = k$.

If $k \geq n \geq 2$, we choose

$$L_n = \{b\}\{a,b\}^* \cup \bigcup_{i=2}^{n-1} \{ab^i\}^+ \text{ and } R_{n,k} = \{a\}\{a,b\}^* \cup \bigcup_{i=2}^{k-n+2} \{ba^i\}^+.$$

By Lemma 2.7, $Var(L_n) = n$, and it is easy to see that

$$Var(L_n \cap R_{n,k}) = Var\Big(\bigcup_{i=2}^{n-1} \{ab^i\}^+ \cup \bigcup_{i=2}^{k-n+2} \{ba^i\}^+\Big) = 1 + (n-2) + (k-n+1) = k.$$

The modification for the cases $1 = n \leq k$ are left to the reader. □

5 Summary

The results given in the two preceding sections can be summarized in the following theorem.

Theorem 5.1

$$
\begin{array}{rl}
i) & r_{\cup}(n,m) = \{1,2,\ldots,n+m+1\} \textit{ for } n \geq 1, m \geq 1,\\
ii) & r_{\cdot}(n,m) \supseteq \{\max\{n,m\}, \max\{n,m\}+1, \ldots n+m+1\} \textit{ for } n \geq 1, m \geq 1,\\
iii) & r_{*}(n) = \{1,2,\ldots,n+1\} \textit{ for } n \geq 1,\\
iv) & r_{h}(n) = \{1,2,\ldots,n\} \textit{ for } n \geq 1,\\
v) & r_{h^{-1}}(n) = \{1,2,3,\ldots\} \textit{ for } n \geq 1,\\
vi) & r_{\cap R}(n) = \{1,2,3,\ldots\} \textit{ for } n \geq 1.
\end{array}
$$

We left open the complete determination of $r_{\cdot}(n,m)$.

If we are only interested in the maximal value of $r_{\tau}(n,m)$ or $r_{\tau}(n)$, i.e., if we consider the function

$$f_{\tau}(c_1, c_2, \ldots, c_n) = \max\{r_{\tau}(c_1, c_2, \ldots, c_n)\}$$

for some n-ary operation τ, we obtain the following statement.

Theorem 5.2 *For $n \geq 1$ and $m \geq 1$,*

$$f_{\cup}(n,m) = n+m+1, \; f_{\cdot}(n,m) = n+m+1, f_{*}(n) = n+1, \textit{ and } f_{h}(n) = n.$$

We note that the values $f_{h^{-1}}(n)$ and $f_{\cap R}(n)$ are undefined for any $n \geq 1$ since the corresponding sets $r_{h^{-1}}(n)$ and $r_{\cap R}(n)$ coincide with the set of all positive integers.

Except for the cases of homomorphisms and inverse homomorphisms all our results are already valid for languages over alphabets with two letters. For homomorphisms we need three letters whereas for inverse homomorphisms we cannot bound the size of the alphabets. If one restricts to unary alphabets, the situation changes drastically by Lemma 2.9.

References

[1] C. Campeanu, K. Culik II, K. Salomaa and Sh. Yu, State complexity of basic operations on finite languages. In: *Proc. Workshop on Implementing Automata 1994*, LNCS 2214, Springer-Verlag, Berlin, 2001, 60–70.

[2] M. Domaratzki and K. Salomaa Transition complexity of language operations. In: *Proc. Intern. Workshop Descriptional Complexity of Formal Systems 2006*, New Mexico State Univ. Las Cruces, 2006, 141–152.

[3] J. Gruska, On a classification of context-free languages. *Kybernetika* **1** (1967) 22–29.

[4] M. Holzer and M. Kutrib, Nondeterministic descriptional complexity of regular languages. *Intern. J. Found. Comp. Sci.* **14** (2003) 1087–1102.

[5] J. Jirasek, G. Jiraskova and A. Szabari, State complexity of concatenation and complement of regular languages. In: *Proc. Conf. Implementation and Application of Automata 2004*, LNCS 3317, Springer-Verlag, Berlin, 2004, 178–189.

[6] G. Jiraskova and A. Okhotin, State Complexity of cyclic shift. In: *Proc. Descriptional Complexity of Formal Systems 2005*, Univ. Milano, 2005, 182–193.

[7] F. R. Moore, On the bounds of state-set size in the proofs of equivalence between deterministic, nondeterministic, and two-way finite automata. *IEEE Trans. Computers* **20** (1971) 1211–1214.

[8] Gh. Păun, On the smallest number of nonterminals required to generate a context-free language. *Mathematica – Revue d'Analyse Numerique et de la Theorie de L'Approximation* **18** (1976) 203–208.

[9] A. Salomaa, On the reducibility of events represented in automata. *Annales Academiai Scientarum Fennicae, Series A, I. Mathematica* **353**, 1964.

[10] Sh. Yu, State complexity of regular languages. *J. Automata, Languages and Combinatorics* **6** (2001) 221–234.

[11] Sh. Yu, State complexity of finite and infinite regular languages. *Bulletin of the EATCS* **76** (2002) 142–152.

[12] Sh. Yu, Q. Zhuang and K. Salomaa, The state complexity of some basic operations on regular languages. *Theor. Comp. Sci.* **125** (1994) 315–328.