INFORMATIQUE THÉORIQUE ET APPLICATIONS

VICTOR MITRANA

On languages satisfying "interchange lemma"

Informatique théorique et applications, tome 27, nº 1 (1993), p. 71-79
<http://www.numdam.org/item?id=ITA_1993__27_1_71_0>

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 27, n° 1, 1993, p. 71 à 79)

ON LANGUAGES SATISFYING "INTERCHANGE LEMMA" (*)

by Victor MITRANA [1]

Communicated by J. BERSTEL

Abstract. – *This paper deals with the closure properties of the family of languages satisfying the "interchange lemma" conditions and with a short comparison between these conditions and other iteration conditions on formal languages.*

Résumé. – *Dans cet article, on étudie les propriétés de clôture pour la famille des langages qui satisfont les conditions du « lemme de l'échange » et on compare ces conditions et d'autres conditions d'itération dans les langages formels.*

1. INTRODUCTION

To substitute a subword of a word from a given language with an other string, such that the new word remains in language is a topic with many significances.

W. Ogden, R. Ross, K. Winklmann [8] have shown, in a rather wide sense, that if L is a context-free language and $Q \subseteq L_n = \{x \in L : \lg(x) = n\}$ is an enough large set then there are k strings in Q, z_i, $1 \leqq i \leqq k$, so that:

(i) $z_i = x_i w_i y_i$, $1 \leqq i \leqq k$;

(ii) $\lg(x_i) = \lg(x_j)$, $\lg(y_i) = \lg(y_j)$, $1 \leqq i, j \leqq k$;

(iii) $x_i w_j y_i \in L$ for any $i, j \in \{1, 2, \ldots, k\}$.

Using this result, they prove that $C_2(V) = \{z \in V^+ : z = xy^2 w, y \neq \varepsilon\}$ is not context-free for any alphabet V such that $|V| \geqq 3$.

In analytical linguistics we find other applications [6].

(*) Received September 1991, accepted December 1991.
[1] University of Bucharest, Faculty of Mathematics, Academiei Street 14, 70109 Bucharest, Romania.

0988-3754/93/01 71 9/$2.90/© AFCET-Gauthier-Villars

We used the following notations:
$\lg(x)$ is the length of the string x,
ε is the empty world; $\lg(\varepsilon)=0$,
CF is the family of all context-free languages,
$|A|$ is the cardinality of the finite set A,
$N=\{0, 1, 2, \ldots\}$.

2. PRELIMINARIES

Let $p \geqq 2$, $q \geqq 1$ be two integers. A language $L \subseteq V^*$ is in $IL(p, q)$ iff for any integers n, m so that $p \leqq m \leqq n$ and any subset $Q \subseteq L_n$, there are $k \geqq |Q|/n^q$ strings $z_i \in Q$, $1 \leqq i \leqq k$, such that the following conditions are satisfied:

(i) $\quad z_i = x_i w_i y_i, \; x_i, w_i, y_i \in V^*, \qquad 1 \leqq i \leqq k;$

(ii) $\lg(x_i)=\lg(x_j)$, $\lg(y_i)=\lg(y_j)$, for any $1 \leqq i, j \leqq k$;

(iii) $\quad m/p < \lg(w_i) \leqq m, \qquad 1 \leqq i \leqq k;$

(iv) $\quad x_i w_j y_i \in L_n$, for any $i, j \in \{1, 2, \ldots, k\}$.

It is clear that $IL(p, q) \subseteq IL(p', q')$ for any $p' \geqq p$, $q' \geqq q$ and $L \in IL(p, q)$ iff $L - \{\varepsilon\} \in IL(p, q)$. Therefore, we consider only the languages which have not the empty word. We denote $IL = \bigcup_{p \geqq 2,\, q \geqq 1} IL(p, q)$.

The inclusion $CF \subseteq IL$ is immediate as a consequence of [8]. If L is a context-free language then $L \in IL(\max(2, q), 3)$ where q is the constant from the "interchange lemma" with respect to L. The language L is *polynomially bounded* iff exist the integers p, $q \geqq 1$ such that for any $n \geqq p$, $|L_n| \leqq n^q$.

It is obvious that if L is polynomially bounded then $L \in IL$.

A language L is *bounded* iff exist $w_1, w_2, \ldots, w_k \in V^*$, $w_i \neq \varepsilon$, $1 \leqq i \leqq k$, such that $L \subseteq \{w_1\}^* \{w_2\}^* \ldots \{w_k\}^*$.

PROPOSITION 1: *Any bounded language is polynomially bounded but not vice versa.*

Proof: Let $L \subseteq \{w_1\}^* \ldots \{w_s\}^*$ be a bounded language. Then $|L_n| \leqq (n+1)^s \leqq n^{s+1}$ for any $n \geqq 2^s$, hence L is polynomially bounded.

Conversely, we consider a 2-Thue language over $\{a, b, c\}$ obtained by the iteration of a morphism [2]. This language is polynomially bounded but it is not bounded.

PROPOSITION 2: *In IL there are strict context sensitive, recursive recursively enumerable, non recursively enumerable languages.*

Proof: It is known that there are subsets $P \subseteq N$ so that $L_P = \{ a^n : n \in P \}$ is context sensitive, recursive, recursively enumerable or non recursively enumerable language. Now, $L_P \in IL$ holds because L_P is polynomially bounded.

Examples: $L_1 = \{ z \in V^+ : z = xy^2 w,\ x,\ y,\ w \in V^*,\ y \neq \varepsilon \}$, $|V| \geqq 6$ is not in IL as a consequence of [8].

$L = \{ \alpha \# \alpha : \alpha \in \{ a, b \}^* \}$ is not in IL.

Proof: We assume that $L \in IL(p, q)$. Then let n be an odd integer such that $2^{(n-1)/2}/n^q > 2^{(n(p-1)-p)/(2p)}$ and $Q = L_n$. Obviously, $|Q| = 2^{(n-1)/2}$.

CLAIM 1: *If* $\{ x_1 w_1 y_1, x_2 w_2 y_2 \} \subseteq L_n$ *with* $\lg(x_1) = \lg(x_2)$, $w_1 \neq w_2$ *and*

$$0 < \lg(w_1) = \lg(w_2) \leqq n/2 \quad \textit{then } x_1 w_2 y_1 \notin L \textit{ and } x_2 w_1 y_2 \notin L.$$

We consider two cases:

Case 1: w_1 and w_2 are subwords of α; in this case by the interchange of w_1 with w_2 the obtained words are not of the form $\beta \# \beta$.

Case 2: $w_1 = \delta_1 \# \delta_2$, $w_2 = \gamma_1 \# \gamma_2$ where $\lg(\delta_1) = \lg(\gamma_1)$, $\lg(\delta_2) = \lg(\gamma_2)$. If $x_1 w_2 y_1 \in L$ then, because $\lg(w_2) \leqq n/2$, we have $\gamma_1 = \delta_1$ and $\gamma_2 = \delta_2$, hence, $w_1 = w_2$ which is contradictory.

CLAIM 2: *Let* n_1, n_2 *be two nonnegative integers such that* $n/(2p) < n - n_1 - n_2$ *and* $w \in \{ a, b, c \}^*$, $\lg(w) = n - n_1 - n_2$. *If* T_w *is defined as follows:*

$$T_w = \{ z \in Q : z = xwy,\ \lg(x) = n_1,\ \lg(y) = n_2 \}$$

then

$$|T_w| \leqq 2^{(n(p-1)-p)/(2p)}.$$

Again, we consider two cases:

Case 1: w is over $\{ a, b \}$. If $z \in T_w$ then $z = \beta w \gamma \# \beta w \gamma$, β, $\gamma \in \{ a, b \}^*$ so, $|T_w| \leqq 2^{(n-1)/2 - \lg(w)}$ and because $\lg(w) > n/(2p)$ it follows that

$$|T_w| < 2^{(n-1)/2 - n/(2p)} = 2^{(n(p-1)-p))/(2p)}.$$

Case 2: w does contain a letter $\#$. In this case if $z \in T_w$ then $z = w_2 \beta w_1 \# w_2 \beta w_1$, where $w = w_1 \# w_2$ and $\beta \in \{ a, b \}^*$. Hence

$$|T_w| \leqq 2^{(n-1)/2 - \lg(w_1 w_2)} \leqq 2^{(n-1)/2 - n/(2p)}.$$

Now, we choose $m=n/2$. It must exist $k>2^{(n(p-1)-p)/(2p)}$ strings in Q satisfying the conditions $(*)$. But, among these strings there are two strings $z_1 \in T_{w_1}$ and $z_2 \in T_{w_2}$ such that $w_1 \neq w_2$. Thus the claim 2 and the point (iv) from $(*)$ are contradictory. Hence, $L \notin IL$.

3. CLOSURE PROPERTIES

A faithfull rational transduction [1] is said to be bi-faithfull iff both of the morphisms h and g are strictly alphabetical. A class of languages closed under bi-faithfull rational transductions is a bi-faithfull rational cone.

We will prove that IL is a bi-faithfull cone closed under restricted morphisms and as a consequence we will show that the programing language ALGOL 60 is not in IL. An other result it will be that the problem "Is L in IL" for a given context sensitive language L is an undecidable problem.

THEOREM 1: *IL is closed under intersection with regular languages.*

Proof: Let L be in $IL(p, q)$, $L \subseteqq V^*$, and L'' be a regular language accepted by a deterministic finite automata $M=(V, K, s_0, f, F)$. Then, $L'=L \cap L'' \in IL(p', q')$ where $p'=\max(p, |K^2|)$, $q'=q+1$. For any $p' \leqq m \leqq n$ and any $Q \subseteqq L'_n \subseteqq L_n$ there exist k strings in Q satisfying the conditions $(*)$ with respect to L. Let R be the set of all these strings $z_i=x_i w_i y_i$, $1 \leqq i \leqq k$. We set

$$T(s_i, s_j)=\{z \in R : f(s_0, x)=s_i \text{ and } f(s_i, w)=s_j\},$$

for all $(s_i, s_j) \in K^2$. Then $|R|=\sum_{(s_i, s_j) \in K^2} |T(s_i, s_j)|$ therefore exists a pair $(s_i, s_j) \in K^2$ such that $k/n \leqq |R|/|K^2| \leqq |T(s_i, s_j)|$. All elements of $T(s_i, s_j)$ satisfy $(*)$ and the proof is complete.

LEMMA 1: *IL is closed under substitution with finite, ε-free languages.*

Proof: Let $L' \subseteqq \{a_1, a_2, \ldots, a_s\}^*$ be a language in $IL(p', q')$ and $f(a_i)=L^{(i)}$ $1 \leqq i \leqq s$ be a substitution. Let us denote

$$t_1=\max(|L^{(i)}|),\ 1 \leqq i \leqq s,$$
$$t_2=\max\{\lg(x) : x \in L^{(i)},\ 1 \leqq i \leqq s\} \text{ and } L=f(L').$$

Furthermore, we take $p=\max(p' t_2, t_1)$ and $q=q'+5$. If $p \leqq m \leqq n$ and $Q \subseteqq L_n$ there exists $Q_j \subseteqq L_j$, $p' \leqq j \leqq n$ such that $Q=\bigcup_{j=p}^{n}$, $(\mathrm{f}(\mathrm{Q_j}) \cap Q)$ and

$f(x) \cap Q \neq \varnothing$ for any $x \in Q_j$, $p' \leqq j \leqq n$. Hence, exists $j \in \{p', p'+1, \ldots, n\}$ such that $|f(Q_j) \cap Q| \geqq |Q|/n$. Moreover, $|Q_j| \geqq |f(Q_j) \cap Q|/t_1 . j$ and if $[a]_*$ means an integer so that $[a]_* \leqq a < [a]_* + 1$ then there are k' strings $z_i = x_i w_i y_i$ in Q_j with respect to (∗) for $m = [m/t_2]_*$. Let A' be the set of all these strings and $A = \{f(z') \cap Q : z' \in A'\}$.

For any integers n_1, n_2 such that $m/p < n - n_1 - n_2 \leqq m$ we construct the sets $T(n_1, n_2) = \{z \in A : z = xwy, \lg(x) = n_1, \lg(y) = n_2$ and exists $z' = x' w' y' \in A'$ so that $z \in f(z')$ and $x \in f(x'), y \in f(y')\}$. Since $m/p' . t_2 < \lg(w') \leqq m/t_2$ it follows that $m/p < \lg(w) \leqq m$ hence $A = \bigcup_{n_1, n_2} T(n_1, n_2)$. Therefore, exist $n_1\, n_2$ such that $|T(n_1, n_2)| \geqq |A|/n^2$. Because all words of $T(n_1, n_2)$ do satisfy the conditions (∗) with respect to L, $L \in IL(p, q)$ follows.

THEOREM 2: *IL is a bi-faithfull rational cone. However, IL is not closed under arbitrary morphisms.*

Proof: IL is a bi-faithfull cone follows from Theorem 1 and Lemma 1. If $L = \{xy\#xz : x \in \{a, b\}^*, y, z \in \{c, d\}^*, \lg(x) = \lg(y) = \lg(z)\}$ is in $IL(6, 1)$ and h is a morphism from $\{a, b, c, d, \#\}^*$ into $\{a, b, \#\}^*$, $h(a) = a$, $h(b) = b$, $h(\#) = \#$, $h(c) = h(d) = \varepsilon$ then we have $h(L) = \{x\#x : x \in \{a, b\}^*\}$ which is not in IL. However, IL is closed under an other kind of morphism, the restricted morphism.

If $L \subseteqq (V\{\varepsilon, \#, \#^2, \ldots, \#^{k-1}\})^*$ for some $\# \notin V$ and some constant k and if h is a morphism defined on $(V \cup \{\#\})^*$ by $h(a) = a$ for $a \in V$ and $h(\#) = \varepsilon$ then h is said *k-restricted* on L. This implies that $\lg(x) \leqq k . \lg(h(x))$ for all $x \in L$; moreover, given any subword x of a word in L, such that $\lg(x) \geqq k$, we know that $\lg(h(x)) \geqq 1$.

Now, we prove:

THEOREM 3: *IL is closed under restricted morphisms.*

Proof: Let h be a t-restricted morphism on

$$L \subseteqq (V\{\varepsilon, \#, \#^2, \ldots, \#^{t-1}\})^*, \qquad L \in IL(p, q).$$

Then $L' = h(L) \in IL(p . t, q+3)$. Let $p . t \leqq m \leqq n$ and $Q \subseteqq L_n$, $Q = \bigcup_{i=n}^{t . n} h(Q_i)$, $Q_i \subseteqq L_i$, $n \leqq i \leqq t . n$. Furthermore, if $x_1, x_2 \in Q_i$ then $h(x_1) \neq h(x_2)$ for any $n \leqq i \leqq t . n$ (*i.e.* $|h(Q_i)| = |Q_i|$). Since $|Q| \leqq \sum_{i=n}^{t . n} |Q_i|$ there is i between n and $t . n$ such that $|Q_i| \geqq |Q|/t . n$. There are k strings $z_j = x_j w_j y_j$, $1 \leqq j \leqq k$, in Q_i in keeping with the requirements of (∗), $|Q|/n^{q+2} \leqq |Q_i|/n^q \leqq k$. Now, exists a set $F \subseteq \{z_1, z_2, \ldots, z_k\}$ such that:

(i) if $z' = x'w'y'$, $z = xwy$ are in F then $N_{\#}(w') = N_{\#}(w)$ where $N_{\#}(v)$ is the number of occurrences of letter $\#$ in v;

(ii) $|F| \geqq k/n$.

All strings of $h(F)$ respect the conditions $(*)$ involving that the claim holds.

Now, following [4, pp. 20] we consider the ALGOL programs of the form:

```
begin integer x;
    y:=1
end
```

These programs are correct if and only if $x=y$. Let us denote R the set of all programs as above with arbitrary x, y over $\{a, b\}$, and let ALGOL be the set of all correct ALGOL programs. If h is a restricted morphism which erases all symbols different from; and those in x and y, $h(a)=a$, $h(b)=b$, $h(;)=\#$, then we obtain $h(\text{ALGOL} \cap R) = \{x\#x : x \in \{a, b\}^*\}$ which is not in IL. Since IL is closed under intersection with regular sets and under restricted morphisms it follows that ALGOL is not in IL.

We are going on with other closure properties of IL.

THEOREM 4: *IL is closed under union and concatenation but it is not closed under intersection and complementation.*

Proof: Let $E \in IL(p_1, q_1)$, $F \in IL(p_2, q_2)$ be two languages, E over V^*, F over U^*.

It is immediate that $E \cup F \in IL\,(\max(p_1, p_2), \max(q_1, q_2)+1)$.

If $L = E \circ F$ and $p = \max(2.\max(p_1, p_2), |V|, |U|)$, $q = \max(q_1, q_2)+3$ then $L \in IL(p, q)$. Let $p \leqq m' \leqq n$ and $Q \subseteq L_n$, then $Q = \bigcup_{i+j=n} (A_i B_j \cap Q)$, $A_i \subseteq E_i$, $B_j \subseteq F_j$.

Hence, there are i and j so that $i+j=n$ and $|A_i B_j \cap Q| \geqq |Q|/n$. If $i+j=n$ then we have $i \geqq n/2$ or $j \geqq n/2$. Let us assume that $i \geqq n/2$. If $\alpha \in A_i$ then $|M(\alpha) = \{\beta \in B_j : \alpha\beta \in Q\}| \leqq |U|^j < n^n$.

For each $1 \leqq h \leqq n$ we set M_h the union of all sets $\alpha M(\alpha)$ with $\alpha \in A_i$ and $n^{h-1} \leqq |M(\alpha)| < \mathrm{n}^{\mathrm{n}}$. So, $A_i B_j \cap Q = \bigcup_{h=1}^{n} M_h$ and exists h between 1 and n such that $|M_h| \geqq |A_i B_j \cap Q|/n$. If $M = \{\alpha \in A_i : \alpha\beta \in M_h\}$ then there are $k_1 \geqq |M|/n^{q_1}$ strings in M satisfying the conditions $(*)$ with respect to E for $m=m'$ if $i \geqq m'$ or $m=i$ if $m'>i$. Since $k_1 n^{h-1} \geqq |M_h|/n^{q_1+1}$ so $\left|\bigcup_{i=1}^{k_1} z_i M(z_i)\right| \geqq |\mathrm{M}_h|/n^{q_1+1}$.

Now, it is easy to observe that all strings from $\bigcup_{i=1}^{k_1} z_i M(z_i)$ verify (*), so, $E \circ F \in IL$.

For the last part of this theorem let us consider

$L_1 = \{ xy\#xz : x, y, z \in \{ a, b \}^*, \lg(y) = \lg(z) \}$,

$L_2 = \{ xy\#zy : \mathrm{x}, \mathrm{y}, \mathrm{z} \in \{ \mathrm{a}, \mathrm{b} \}^*, \lg(x) = \lg(z) \}$

which are in $IL(5, 1)$ (the details are left to the reader) and $L_1 \cap L_2 = \{ x\#x : x \in \{ a, b \}^* \}$ is not in IL. Using De Morgan relations the proof is complete.

For a language L and a word x, let us denote $\partial_x(L) = \{ y : xy \in L \}$ for the *left derivative* with respect to x.

LEMMA 2: *IL is closed under the left derivative with respect to a given word.*

Proof is immediate.

At the end, we prove that the problem "Is L in IL?" for an arbitrary context sensitive language L, is undecidable, using the following:

THEOREM 5 [5]: *Let $\mathscr{L}$ be a family of languages effectively closed under union, concatenation with regular sets and the problems "Is the complementary of L empty or no?" is undecidable for L.*

If P is a non trivial property on $\mathscr{L}$ such that:

(*a*) P is true on the regular languages,

(*b*) if L has the property P, R is a regular language and x is an arbitrary word, then $L \cap R$ and $\partial_x(L)$ have the property P, then P is undecidable on $\mathscr{L}$.

If $\mathscr{L}$ is the family of the context sensitive languages and P is the property of a language that to be in IL then it satisfies all conditions of Theorem 5.

4. A SHORT COMPARISON WITH OTHER ITERATION CONDITION ON FORMAL LANGUAGES

We denote

OG – the family of languages satisfying Ogden's lemma, [7]

BH – the family of languages satisfying Bar Hillel's lemma,

SO – the family of languages satisfying Sokolowski's lemma [9].

THEOREM 6:

1. *OG and IL are incomparable.*

2. *BH and IL are incomparable.*
3. *$IL \subset SO$ and the inclusion is proper.*
4. *$CF \subset OG \cap BH \cap IL$ is proper.*

Proof: We consider the languages:

$$L_1 = \{ a^n b^n c^n : n \geqq 1 \}$$
$$L_2 = \{ xy^2 z : x, y, z \in V^*, y \neq \quad \}, |V| = 6.$$

Now, $L_1 \in IL$ but $L_1 \notin BH \cup OG$, $L_2 \notin IL$ but $L_2 \in BH \cap OG$.

Let L be in $IL(p, q)$, $L \subseteqq V^*$ and $|V| \geqq 3$. Let U be a subset of V, $|U| \geqq 2$ $u_1, u_3 \in V^*$, $u_2 \in V^* - U^*$ (if $u_2 \in U^*$ then the Sokolowski's condition is trivial), so that $A = \{ u_1 x u_2 x u_3 : x \in U^+ \} \subseteqq L$. Let n be an integer such that

$$p \,.\, \max(\lg(u_1), \lg(u_2), \lg(u_3)) \leqq \mathrm{n}/2 < n - (\lg(u_1) + \lg(u_2) + \lg(u_3))$$

and verifying a similar condition as in Examples for the language L. We take $Q = A_n$, since $L \in IL$ there are k strings in Q satisfying the requirements ($*$), for $m = n/2$. Therefore (*see* Examples) there are $z_1 = x_1 w_1 y_1$, $z_2 = x_2 w_2 y_2$ in Q so that $w_1 \neq w_2$ and $x_1 w_2 y_1 \in L$. Hence, we get the desired x', x'' from U^+, $x' \neq x''$, such that $u_1 x' u_2 x'' u_3 \in L$. Because $L_3 \in SO$ the inclusion is proper.

Finally, $B_P = \{ (ab)^n : n \in P \} \cup (\{ a, b \}^* \{ aa, bb \} \{ a, b \}^*)$, where $P \subseteqq N$ is the set of all prime numbers, is not context-free; from [3] B_P is in $OG \cap BH$ and the relation $B_P \in IL\ (2, 2)$ is immediate.

We mention here some *open problems*: Is IL closed under substitutions, inverse morphisms or under the operations $*$?

ACKNOWLEDGEMENTS

I wish to thanks for the suggestions on parts of the first version of the text given to us by the referee.

REFERENCES

1. J. M. Autebert and L. Boasson, Generators of Cones and Cylinders, Formal Languages Theory: Perspectives and Open Problems, R. V. Book Ed., *Acad. Press*, 1980, pp. 49-88.
2. J. Berstel, Sur les mots sans carré définis par un morphisme, *Lecture Notes in Comput. Sci.*, 1979, *71*, pp. 16-25.

3. L. Boasson and S. Horvath, On language satisfying Ogden's lemma, *R.A.I.R.O. Inform. Theor. Appl.*, 1978, *12*, pp. 193-199.
4. J. Dassow and Gh. Paun, Regulated rewriting in formal language theory, *Akademie-Verlag*, Berlin, 1989.
5. S. A. Greibach, A note on undecidable properties of formal languages, *Math. Syst. Theory*, 1968, *2*, 1, pp. 1-6.
6. S. Marcus, Algebraic linguistics. Analytical models, New York, London, *Academic Press*, 1967.
7. W. Ogden, A helpful result for proving inherent ambiquity, *Math. Syst. Theory*, 1968, *2*, pp. 191-197.
8. W. Ogden, R. Ross and K. Winklemann, An "interchange lemma" for context-free languages, *S.I.A.M. J. Comput.*, 1985, *14*, pp. 410-415.
9. S. Sokolowski, A method for proving programming language non context-free, *Inf. Proc. Lett.*, 1978, *7*, pp. 151-153.