RAIRO. INFORMATIQUE THÉORIQUE

J. ALBERT

A note on the undecidability of contextfreeness

RAIRO. Informatique théorique, tome 16, nº 1 (1982), p. 3-11

<http://www.numdam.org/item?id=ITA_1982__16_1_3_0>

R.A.I.R.O. Informatique théorique/Theoretical Informatics
(vol. 16, n° 1, 1982, p. 3 à 11)

A NOTE ON THE UNDECIDABILITY OF CONTEXTFREENESS (*)

by J. ALBERT ([1])

Communicated by J. BERSTEL

Abstract. — *For subcontext-free language families K' and noncontext-free families K we consider the problem, if there is an algorithm to decide for all $L \in K$ whether or not $L \in K'$.*

Under the assumption that K and K' contain special languages related to the Post Correspondence Problem, this question is shown undecidable. This gives a simple and direct proof for a number of well-known results and yields also several new ones.

Résumé. — *Nous considérons, pour des familles de langages K' contenues dans les langages context-free et des familles non context-free de langages K, le problème de savoir s'il existe un algorithme pour décider, pour tout $L \in K$, si $L \in K'$ ou non.*

Sous l'hypothèse que K et K' contiennent des langages spéciaux liés au problème de correspondance de Post, on montre que cette question est indécidable. Ceci donne une preuve simple et directe pour un bon nombre de résultats bien connus et en établit aussi quelques uns qui sont nouveaux.

0. INTRODUCTION

One of the basic questions arising when two language families, say K and K', are compared, is whether there is an algorithm to decide for all $L \in K$ if $L \in K'$ or not. In the terminology of [24] this is the K'-ness problem for K, one of the so-called "comparative decision problems" concerning language families.

In this area the following results are perhaps the best known and the oldest ones. The finiteness problem is decidable, the regularity problem undecidable for the familiy of context-free languages [7].

Some more recent results are the decidability of the regularity and contextfreeness problem for HDOL languages [24] as well as the decidability of the DOL-ness for context-free languages [19].

(*) Received April 1980.

([1]) Institut für Angewandte Informatik und Formale Beschreibungsverfahren, Universität Karlsruhe (TH), D 7500 Karlsruhe, W. Germany.

R.A.I.R.O. Informatique théorique/Theoretical Informatics, 0399-0540/1981/3 /$ 5.00

The problem, whether a language contains an infinite regular set – is an IRS language – is decidable for the deterministic context-free languages and undecidable for the linear context-free languages, [5], to mention only a few results of this comparative decision type.

The main theorems presented in this paper will yield undecidability results for the K'-ness problem for K, where K' is subcontext-free (or CF itself) and K ranges over all supersets of a specific small family.

1. PRELIMINARIES

One of the common methods to show undecidability of certain questions arising in string manipulating systems, is to use the undecidability of the Post Correspondence Problem (PCP) [20] and to encode the instances of this problem as languages in an appropriate manner. This is done e. g. in the standard proof showing that the regularity problem is undecidable for the context-free languages, where it is even proven that the question "if $L=\Sigma^*$" is undecidable for linear context-free languages L.

In [12] a rather general undecidability result for predicates P on language families F is based on the undecidability of this question "$L \overset{?}{=} \Sigma^*$" and some closure properties of P and F. This theorem yields e. g. the undecidability of the regularity and contextfreeness problem for the one-way stack languages.

The latter result will appear here too in corollary 1. But in our main theorems we will not require closure properties or undecidability of "$L \overset{?}{=} \Sigma^*$" for the considered language families. In place of that we can prove undecidability of the K'-ness problem for K, whenever K and K' contain special languages related to the Post Correspondence Problem.

Throughout the following chapters we will assume that the language families F occurring are "effective" [12], i. e., there is a specification of F (grammar, automaton, ...) such that one can effectively enumerate the languages in F, each language in F is recursively enumerable and there is a partial recursive function which assigns "yes" to the pair consisting of the "name" of the language L in F and the word w if and only if $w \in L$ (*see* also [6]).

All notions not defined explicitely in the following chapters can be found in standard literature as [11, 16, 23], a. o.

2. RESULTS

As mentioned above we start with encoding the instances of the Post Correspondence Problem (PCP) as languages.

By this construction we generate a noncontext-free language if and only if the given instance of the PCP is solvable.

DEFINITION 1: For a given instance:

$$I = \langle (x_1, \ldots, x_t), (y_1, \ldots, y_t) \rangle$$

of the Post Correspondence Problem (PCP), where $x_i, y_i \in \Sigma_I^*$, $(x_i, y_i) \neq (\varepsilon, \varepsilon)$, we define:

$$\text{3-COPY}(I) := \{ u^3 \mid u = x_{i_1} \ldots x_{i_n} \cent y_{i_n}^R \ldots y_{i_1}^R \cent, \quad n \geqq 1, i_k \in \{1, \ldots, t\} \text{ for } 1 \leqq k \leqq n \},$$

where $\cent \notin \Sigma_I$,

$$H(\Sigma_I) := \{ u_1 \cent u_2 \cent u_3 \cent v_3 \cent v_2 \cent v_1 \cent \mid u_j, v_j \in \Sigma_I^*, \ u_j \neq v_j^R \text{ for } j = 1, 2, 3 \}$$

and $L(I) := \text{3-COPY}(I) \cup H(\Sigma_I)$.

LEMMA 1: *Let I be an instance of the PCP and $L(I)$ be defined as above.*
(a) If I is not solvable then $L(I) = H(\Sigma_I)$.
(b) If I is solvable then $L(I)$ is not context-free.

Proof: The proof of the (a)-part is straightforward by the definition of $L(I)$.

For the (b)-part we use the fact that for a solvable instance I we have infinitely many solutions, i. e., there are infinitely many words of the form:

$$x \cent x^R \cent x \cent x^R \cent x \cent x^R \cent \quad \text{in } \text{3-COPY}(I) \subseteqq L(I).$$

If we assume now $L(I)$ to be context-free, we can apply the context-free pumping lemma to such a word

$$z = x \cent x^R \cent x \cent x^R \cent x \cent x^R \cent \in L(I),$$

which is long enough.

Thus,

$$z = w_1 w_2 w_3 w_4 w_5, \qquad w_2 w_4 \neq \varepsilon$$

and:

$$z_i = w_1 w_2^i w_3 w_4^i w_5 \in L(I) \quad \text{for all } i \geqq 0.$$

Clearly, $w_2 w_4$ cannot contain any $¢$.

Otherwise, $z_2 = w_1 w_2 w_2 w_3 w_4 w_4 w_5$ would have more than 6 $¢$'s.

Since w_2 and w_4 are substrings of x or x^R and z contains three pairs of x, x^R the word z_2 is not of the form:

$$u_1 ¢ u_2 ¢ u_3 ¢ v_3 ¢ v_2 ¢ v_1 ¢, \qquad u_j \neq v_j^R,$$

i. e., $z_2 \notin H(\Sigma_I)$.

But z_2 is not in 3-COPY (I) either, because at least one and at most two x's, x^R's resp., have been changed by the insertion of w_2 and w_4.

With $z_2 \notin$ 3-COPY$(I) \cup H(\Sigma_I) = L(I)$ we get the desired contradiction, proving lemma 1. □

We can now formulate our first main theorem for families containing the languages $L(I)$, $H(\Sigma_I)$, resp.

THEOREM 1: *Let $L(I)$ and $H(\Sigma_I)$ be as in definition* 1 *and:*

$$K := \{ L(I) \mid I \text{ instance of the PCP} \},$$
$$K' := \{ H(\Sigma) \mid \Sigma \text{ alphabet} \}.$$

If $K \subseteq M$ and $K' \subseteq N \subseteq CF$ for some language families M and N then N-ness is undecidable for M.

Proof: Clear by the undecidability of the Post Correspondence Problem and lemma 1.

In the following corollary we will just list some of the language families satisfying the conditions of theorem 1.

COROLLARY 1: *Let K, K' be as in theorem* 1.

(*a*) *$K' \subseteq N \subseteq CF$ holds for the following classes N:*

- *linear context-free languages;*
- *metalinear context-free languages* (*of width k for each $k \geqq 1$*) [16]*;*
- *ultralinear context-free languages;*
- *deterministic context-free languages;*
- *sequential context-free languages* [10]*;*
- (*n, m*)*-bounded languages for all* $n \in \{3, 4, \ldots\} \cup \{\infty, \omega\}$, $m \in \{1, 2, \ldots\} \cup \{\infty, \omega\}$ [13]*;*
- *$LR(k)$ languages for all $k \geqq 0$* [16, 17]*;*
- *the intersection and union of any of the above families.*

(*b*) *$K \subseteq M$ holds for the following classes M:*

- *checking stack languages* (*of finite return*) [14, 15];
- *two-way finite state transducer languages* (*of finite return*) [15];
- *EDTOL languages* (*of finite index*) [21, 22];
- *Indian parallel languages* (*of finite index*) [22, 25, 26];
- *homomorphic replacement languages* (*finite axiom set, linear replacements*) [3].

Supersets of above families are e. g.:

- Russian parallel languages [18, 26];
- (nonerasing) stack languages [8, 14, 15];
- iterated deterministic context-free substitution languages [4];
- inside-out-macro languages [8, 9];
- outside-in-macro languages = indexed languages [1, 8, 9];
- ETOL languages (of finite index) [21, 22].

By a result in [22] ETOL of finite index equals about 15 classes also under the finite index restriction; among those the families of:

- scattered languages;
- context-free programmed languages;
- unconditional transfer context-free programmed languages;
- matrix languages;
- ordered languages;
- context-free languages with regular control;
- state languages;
- random context languages;
- forbidding languages;
- permitting languages;
- ETIL languages.

These lists surely can be extended by the interested reader.

Some of the results presented here are either well-known or can be concluded from theorems given by [6, 12].

The virtue of theorem 1 is due to the fact that all these undecidability results are obtained in one sweep by rather simple means. Another advantage is the applicability of theorem 1 to grammatical families generated by grammar forms or *L* forms, where closure-properties are rare.

By a slight variation in the definition of *K* and *K'* we get a similar result as in theorem 1 but now for some language families with less copying power than, say EDTOL or Indian Parallel.

DEFINITION 2: For a given instance $I=\langle (x_1, \ldots, x_t), (y_1, \ldots, y_t) \rangle$ of the PCP, where $x_i, y_i \in \Sigma_I^*$, $(x_i, y_i) \neq (\varepsilon, \varepsilon)$ we define:

$$\begin{aligned} \text{3-PAR}(I) := \{ x_{i_1} \ldots x_{i_n} ¢ y_{i_n}^R \ldots y_{i_1}^R ¢ x_{j_1} \ldots x_{j_n} \\ ¢ y_{j_n}^R \ldots y_{j_1}^R ¢ x_{k_1} \ldots x_{k_n} ¢ y_{k_n}^R \ldots y_{k_1}^R ¢ \mid n \geqq 1, \\ i_r, j_r, k_r \in \{1, \ldots, t\} \text{ for } 1 \leqq r \leqq n \}, \end{aligned}$$

where $¢ \notin \Sigma_I$,

$$\begin{aligned} J(\Sigma_I) := \{ u_1 ¢ v_1 ¢ u_2 ¢ v_2 ¢ u_3 ¢ v_3 ¢ \mid u_j, v_j \in \Sigma_I^*, \\ u_j \neq v_j^R \text{ for } j=1, 2, 3 \} \quad \text{and} \quad P(I) := \text{3-PAR}(I) \cup J(\Sigma_I). \end{aligned}$$

THEOREM 2: *Let $P(I)$ and $J(\Sigma_I)$ be defined as above and:*

$$Q := \{ P(I) \mid I \text{ instance of the PCP} \},$$
$$Q' := \{ J(\Sigma) \mid \Sigma \text{ alphabet} \}.$$

If $Q \subseteqq R$ and $Q' \subseteqq S \subseteqq CF$ for some language classes R and S then S-ness is undecidable for R.

Proof: Only some minor modifications are necessary to carry over the proof of lemma 1 and the conclusion in theorem 1 to this variation of encoding the PCP:

The equality $P(I)=J(\Sigma_I)$ in case instance I is not solvable, can be verified immediately.

Now assume that I has solutions and $P(I)$ be context-free. Again we can apply the context-free pumping lemma to a word:

$$z = x ¢ x^R ¢ x ¢ x^R ¢ x ¢ x^R \in \text{3-PAR}(I) \subseteqq P(I),$$

i. e.,

$$z = w_1 w_2 w_3 w_4 w_5, \qquad w_2 w_4 \neq \varepsilon$$

and:

$$z_i := w_1 w_2^i w_3 w_4^i w_5 \in P(I) \quad \text{for all } i \geqq 0.$$

As before $w_2 w_4$ cannot contain any $¢$, so w_2, w_4 are subwords of x or x^R. Thus, there is at least one subword $x ¢ x^R ¢$ of z which is kept unchanged in all:

$$z_i = w_1 w_2^i w_3 w_4^i w_5, \qquad i \geqq 2,$$

i. e.,

$$z_i \notin J(\Sigma_I) \quad \text{for all } i \geqq 2.$$

It remains the case $z_i \in$ 3-PAR(I) for all $i \geqq 2$, which means each z_i is of the form:

$$z_i = y_1^{(i)} \mathbin{¢} y_2^{(i)} \mathbin{¢} y_3^{(i)} \mathbin{¢},$$

where:

$$y_1^{(i)}, y_2^{(i)}, y_3^{(i)} \in \{ x_{r_1} \ldots x_{r_{n_i}} \mathbin{¢} y_{r_{n_i}}^R \ldots y_{r_1}^R \mid r_j \in \{1, \ldots, t\}, 1 \leqq j \leqq n_i \}$$

for some n_i.

The set of n_i's occurring here must be infinite because of $|z_{i+1}| > |z_i|$ for all $i \geqq 0$. But, as mentioned above, at least one of the $y_k^{(i)}$ equals $x \mathbin{¢} x^R$ for all $i \geqq 2$, contradicting the fact $|x_m y_m| \geqq 1$ for all $m \in \{1, \ldots, t\}$.

Thus, $P(I) \notin CF$, proving theorem 2. □

In the following corollary we will again list some of those language families R and S which fulfil the assumptions of theorem 2.

COROLLARY 2: *Let Q, Q' be as in theorem* 2.

(*a*) *$Q' \subseteqq S \subseteqq CF$ holds for the following classes S:*

- *reversal-bounded one-counter languages* [15];
- *metalinear context-free languages of width k for each* $k \geqq 3$ [16];
- *(n, m)-bounded languages for* $n, m \in \{1, 2, 3, \ldots\} \cup \{\infty, \omega\}$ [13].

(*b*) *$Q \subseteqq R$ holds for the following classes R:*

- *EOL languages* (*of index* 3) [22];
- *RMOL languages* [21];
- $R_{\text{iter}}^{(1)}$ [21];
- *m-block-indexed languages for all* $m \geqq 1$ [2].

Note that the families listed in (*b*) are incomparable to EDTOL, even of index 2 [2]. Since these families are subsets of ETOL (of finite index) or indexed languages, resp. the corresponding families from corollary 1. (*b*) can be added here too.

3. CONCLUSIONS

We hope to have demonstrated how a number of well-known undecidability results and quite a lot of new ones can be obtained in a direct and simple manner. The proofs are just standard and so short, that they could well serve as classroom-notes or exercises in a formal language course.

Not covered by these results are e. g. the regularity – and the contextfreeness problem for OL languages which seem to be extremely difficult because of the absence of nonterminals in these parallel rewriting systems.

REFERENCES

1. A. V. AHO, *Indexed Grammars – an Extension of Context – Free Grammars*, J. Assoc. Comput. Mach., Vol. 15, 1968, pp. 647-671.
2. J. ALBERT, *Über indizierte und m-Block-indizierte Grammatiken*, Ph. D. Thesis, Institut für Angewandte Informatik und Formale Beschreibungsverfahren, Univ. Karlsruhe, 1976.
3. J. ALBERT and L. WEGNER, *Languages with Homomorphic Replacements*, Theor. Comput. Sc., Vol. 16, 1981, pp. 291-305.
4. P. ASVELD and J. ENGELFRIET, *Iterated Deterministic Substitution*, Acta Informatica, Vol. 8, 1977, pp. 285-302.
5. J.-M. AUTEBERT, J. BEAUQUIER, L. BOASSON and M. LATTEUX, *Indécidabilité de la condition IRS*, Université de Lille I, publication n° 18.79, 1979.
6. B. S. BAKER and R. V. BOOK, *Reversal-Bounded Multipush-down Machines*, J. Comput. System Sci., Vol. 8, 1974, pp. 315-332.
7. Y. BAR-HILLEL, M. PERLES and E. SHAMIR, *On Formal Properties of Simple Phrase-Structure Grammars*, Z. f. Phonetik, Sprachwissenschaft, Kommunikationsforschung, Vol. 14, 1961, pp. 143-177.
8. J. ENGELFRIET, E. MEINECHE SCHMIDT and J. VAN LEEUWEN, *Stack Machines and Classes of Nonnested Macro Languages*, J. Assoc. Comput. Mach., Vol. 27, 1980, pp. 96-117.
9. M. J. FISCHER, *Grammars with Macro-like Productions*, Ph. D. Thesis, Haryard Univ., Cambridge, Mass., 1968.
10. S. GINSBURG and H. G. RICE, *Two Families of Languages Related to ALGOL*, J. Assoc. Comput. Mach., Vol. 9, 1962, pp. 350-371.
11. S. GINSBURG, *Formal Languages*, North-Holland, Amsterdam, 1975.
12. S. GREIBACH, *A Note on Undecidable Properties of Formal Languages*, Math. Systems Theory, Vol. 2, 1968, pp. 1-6.
13. S. GREIBACH, *An Infinite Hierarchy of Context-Free Languages*, J. Assoc. Comput. Mach., Vol. 16, 1969, pp. 9-106.
14. S. GREIBACH, *Checking Automata and one-way Stack Languages*, J. Comput. System Sc., Vol. 3, 1969, pp. 196-217.
15. S. GREIBACH, *One Way Finite Visit Automata*, Theor. Comput. Sc., Vol. 6, 1978, pp. 175-221.
16. M. A. HARRISON, *Introduction to Formal Language Theory*, Addison-Wesley, Reading, Massachusetts, 1978.
17. D. E. KNUTH, *On the Translation of Languages from Left to Right*, Information and Control, Vol. 8, 1965, pp. 607-635.
18. M. LEVITINA, *O nekotoryh grammatikah s pravilami globalnoi podstanovki*, Akad. Nauk S.S.S.R. Nauchno-Tekhn. Inform., Ser. 2, 1973, pp. 32-36.
19. M. LINNA, *The DOL-Ness for Context-Free Languages is Decidable*, Inf. Process. Letters, Vol. 5, 1976, pp. 149-151.

20. E. L. POST, *A Variant of a Recursively Unsolvable Problem*, Bulletin A.M.S., Vol. 52, 1946, pp. 264-268.
21. G. ROZENBERG and A. SALOMAA, *L Systems*, Lecture Notes in Computer Science, Vol. 15, 1974, Springer-Verlag.
22. G. ROZENBERG and D. VERMEIR, *L. Systems of Finite Index* in Automata, Languages and Programming, A. SALOMAA and M. STEINBY, Eds., Vol. 52, Lecture Notes in Computer Science, Springer-Verlag, 1977, pp. 430-439.
23. A. SALOMAA, *Formal Languages*, Academic Press, New York, 1973.
24. S. SALOMAA, *Comparative Decision Problems between Sequential and Parallel Rewriting*, Proc. Intern. Symp. Uniformly Structured Automata and Logic, Tokyo, 1975.
25. R. SIROMONEY and K. KRITHIVASAN, *Parallel Context-Free Languages*, Information and Control, Vol. 24, 1974, pp. 155-161.
21. S. SKYUM, *Decomposition Theorems for Various Kinds of Languages Parallel in Nature*, S.I.A.M. J. Comput., Vol. 5, 1976, pp. 284-196.