Theoretical Computer Science 21 (1982) 341–349
North-Holland Publishing Company

NOTE

SOME REMARKS ON NON-ALGEBRAIC ADHERENCES

Sorin ISTRAIL

Department of Mathematics and Computer Center, University "Al.I.Cuza" Iaşi, 6600 Iaşi, Romania

Communicated by M. Nivat
Received October 1981
Revised April 1982

Abstract. We present some properties of non-algebraic adherences of languages, which can be consistently called ***context-sensitive adherences***, because it is proved that in the Chomsky hierarchy of adherences, context-sensitive and recursive-enumerable adherences coincide. Concerning the center-mapping the traversing from algebraic to non-algebraic takes simple recursive to not necessarily recursive-enumerable. Except for the last equality, the Chomsky hierarchy of adherences is proper: $R_{\text{Adh}} \subset CF_{\text{Adh}} \subset CS_{\text{Adh}} = RE_{\text{Adh}}$ and moreover, it can be refined in the non-algebraic case by considering the McNaughton–Nivat adherences.

Regarding the programming context, adherences occur when we consider the set of divergent computations 'shapes' of a program or a program scheme. In this respect, the 'interpretation' is powerful enough to translate R_{Adh} to CS_{Adh}.

1. Preliminaries

Given a finite alphabet A and $\mathbb{N}_+ = \{1, 2, \dots\}$ a word over A can be thought as a partial function $f : \mathbb{N}_+ \to A$, whose domain is

$$[n] = \{1, 2, \dots, n\}.$$

For any $m \leq n$ we shall write $f[m] = f(1)f(2)\cdots f(m)$.

An infinite word over A will be a total function $u : \mathbb{N}_+ \to A$. We shall denote by $u[n]$ its initial segment of length n, i.e.

Let be A^{ω} the set of infinite words over A and we put $A^{\infty} = A^* \cup A^{\omega}$. λ denotes the empty word.

The length of a finite word f is denoted $|f|$.

0304-3975/82/0000–0000/$02.75 © 1982 North-Holland

The operator 'left factors' is defined by

$$FG(f)=\{g\in A^*\mid g=f[n], n\leq|f|\}\quad \text{for } f\in A^*,$$

$$FG(u)=\{u[n]\mid n\in \mathbb{N}_+\}\quad \text{for } u\in A^\omega,$$

$$FG(L)=\bigcup_{x\in L} FG(x)\quad \text{for } L\subseteq A^\infty.$$

A basic notion in the theory of infinitary languages is that of *adherence* defined as follows: for $L\subseteq A^\infty$, $\text{Adh}(L)=\{u\in A^\omega\mid FG(u)\subseteq FG(L)\}$.

The *center* of a language $L\subseteq A^\infty$ is $L^c=FG(\text{Adh}(L))$.

For a family of languages $\mathscr{L}$, let $\mathscr{L}_{\text{Adh}}$ denote its family of adherences, i.e.

$$\mathscr{L}_{\text{Adh}}=\{L\mid \text{there in a finite alphabet } A \text{ and } L'\subseteq A^\infty, L'\in\mathscr{L} \text{ such that } L=\text{Adh}(L')\}.$$

The families of the Chomsky hierarchy will be denoted $\boldsymbol{R}$ (regular), $\boldsymbol{CF}$ (context-free), $\boldsymbol{CS}$ (context-sensitive), $\boldsymbol{RE}$ (recursive-enumerable).

Regarding adherence families we obtain $\boldsymbol{R}_{\text{Adh}}$, $\boldsymbol{CF}_{\text{Adh}}$, $\boldsymbol{CS}_{\text{Adh}}$, $\boldsymbol{RE}_{\text{Adh}}$.

The families $\boldsymbol{R}_{\text{Adh}}$ and $\boldsymbol{CF}_{\text{Adh}}$ are called of rational respectively algebraic adherences.

Let us denote $\boldsymbol{R}^c$, $\boldsymbol{CF}^c$, $\boldsymbol{CS}^c$, $\boldsymbol{RE}^c$ the corresponding families of centers of languages.

The family of rational adherences and that of algebraic adherences possess a representation theorem of the form:

$$L=\bigcup_{i=1}^{k} L_iL_i'^{\omega}, \tag{1}$$

where L_i, L_i' are languages in the corresponding family.

McNaughton Theorem ([6]). *Any rational adherence L can be represented as in* (1) *with* $L_i, L_i'\in\boldsymbol{R}$, $1\leq i\leq k$.

Nivat Theorem ([7]). *Any algebraic adherence L can be represented as in* (1) *with* $L_i, L_i'\in\boldsymbol{CF}$, $1\leq i\leq k$.

The theorem which follows will play an important role in the sequel. The accompanying name seems to to be meaningful for the obtained results.

It improves to 'unique completion' the theorem from [10, Theorem 9.9].

Let us consider a deterministic Turing machine M accepting L.

For any $w\in L$, let $\text{space}_M(w)$ be the space used in the computation of acceptance of w. Then the language

$$L''=\{wd^{\text{space}_M(w)-|w|}\mid w\in L\}$$

can be accepted by a linear bounded automaton which will simulate the Turing machine M using for the extension of its workspace only d-symbols.

Then we can state

Theorem (The Unique Completion Theorem). *For every r.e. set $L \subseteq V^*$, $d \notin V$ there exists a **CS** set L'' such that $L'' = \{wd^{i(w)} \mid w \in L, i(w) \geq 0\}$, where 'i' is a (partial) function.*

Note that i cannot be a computable function. For, let i be computable. Now $wd^{i(w)} \in L''$ iff $w \in L$ which implies that we can decide the membership for the (arbitrary) r.e. set L.

We suppose the reader familiar with the basic facts of the Nivat theory of infinitary languages [2, 7, 8, 9].

2. Properties of non-algebraic adherences

By 'non-algebraic' adherences we mean those adherences which are not algebraic but r.e.

As we shall see in this interpretation, that 'non-algebraic' can be replaced consistently, with 'context-sensitive'.

Let us present for the beginning two negative closure results for ***CS***. The first is well known, and is captured in

Proposition 1. *The family **CS** is not closed under FG.*

Remark. It is easy to see that ***RE*** is closed under FG, hence for $L \in \boldsymbol{CS}$, $FG(L) \in \boldsymbol{RE}$.

Proposition 2. *The centers of **CS** are not necessarily recursive-enumerable.*

Proof. Let us consider the well known not r.e. set $L \subseteq V^*$, $L = \{w \mid w \notin L(M_w)$, where M_w is the Turing machine encoded by $w\}$ [10].

We can consider only deterministic Turing machines without blocking, i.e. every non-final state has a successor.

Consider the set

$$L_1 = \{w \,\$\, w_1 \# w_2 \# \cdots \# w_n \mid n \geq 1,\ w_1 = w,\ w_i \vdash_{M_w} w_{i+1},\ w \in L\}.$$

Let us observe that L_1 is context-sensitive.

We have $\mathrm{Adh}(L_1) = \{w \,\$\, \mathrm{comp}_w \mid w \in L$ and $\mathrm{comp}_w \in (V \cup \{\#\})^\omega$ is the unique infinite computation of 'nonacceptance'$\}$.

Let us remark

$$L_1^c \cap V^*\$ = FG(\mathrm{Adh}(L_1)) \cap V^*\$ = L\$. \tag{2}$$

Suppose that L_1^c is in $\boldsymbol{RE}$. Then by (2) we have $L \in \boldsymbol{RE}$.

The only way out of this contradiction is to conclude that what were pretending is untenable. □

Corollary 1. *Every complement of a r.e. set can be represented as the intersection between a center of a* ***CS****-set and a regular set, i.e.*

$$\boldsymbol{RE}^{cc} \subset \boldsymbol{CS}^{c} \cap \boldsymbol{R}.$$

It is easy to observe that the above inclusion is proper.

A first result concerning non-algebraic adherences is that, in fact, we have only context-sensitive adherences.

Theorem 1.

$$\boldsymbol{CS}_{\mathrm{Adh}} = \boldsymbol{RE}_{\mathrm{Adh}}.$$

Proof. By the Unique Completion Theorem, for every $L \in \boldsymbol{RE}$, $L \subseteq V^*$ and $d \notin V$ there exist a $\boldsymbol{CS}$ set L'' given by $L'' = \{wd^{i(w)} \mid w \in L, i(w) \geqslant 0\}$.

Because of unicity of completion we have $\mathrm{Adh}(L) = \mathrm{Adh}(L'')$ which implies $\boldsymbol{RE}_{\mathrm{Adh}} \subseteq \boldsymbol{CS}_{\mathrm{Adh}}$. □

Corollary 2. $\boldsymbol{CS}^c = \boldsymbol{RE}^c$.

By the theorems of McNaughton and Nivat any rational [algebraic] adherence can be expressed in the form (1) where L_i, L_i' are regular [context-free] sets.

In contrast with the fact that rational and algebraic adherences are 'ω-explicit', we shall show that $\boldsymbol{CS}_{\mathrm{Adh}}$ $(= \boldsymbol{RE}_{\mathrm{Adh}})$ are 'ω-implicit', inplying *a fortiori* that a McNaughton–Nivat representation cannot exist for them.

We reformulate a result of Axel Thue [12] in our context.

A (finite or not) word is called *square-free* (or irreducible or nonrepetitive) if no subword of the form xx, with x a word $\neq \lambda$, occurs in it.

Theorem ((Thue 1912) [12]). *There are square-free (singleton) adherences.*

To give an example, we shall consider the homomorphism h, defined by $h(a) = abc$, $h(b) = ac$, $h(c) = b$ and the $\boldsymbol{CS}$-set

$$L = \{h^n(a) \mid n \geqslant 1\}. \tag{3}$$

It is easy to observe that $\mathrm{Adh}(L)$ is a context-sensitive singleton adherence which is square-free.

The ω-word $\mathrm{Adh}(L)$ is known as the Thue–Morse word [1] and the construction (3) is from [4] and gives a simple proof to the so-called 'square-freeness problem' [11], namely the construction of a square-free ω-word over a three letter alphabet.

As a consequence of Thue theorem, $\boldsymbol{CS}_{\mathrm{Adh}}$ cannot be represented by expressions with 'squares' which implies that explicit occurrence of 'ω' (i.e. infinite power) to represent them is impossible.

Corollary 3. *$\boldsymbol{CS}_{\mathrm{Adh}}$ (= $\boldsymbol{RE}_{\mathrm{Adh}}$) cannot be McNaughton–Nivat representable.*

A similar 'Rat = Rec?' problem can be formulated in the context of families of adherences.

Given $\mathscr{L}$ a family of languages, let us denote by $\mathscr{L}_{\mathrm{Adh}}$ the family of adherences of $\mathscr{L}$, and by MN-$\mathscr{L}_{\mathrm{Adh}}$ the family McNaughton–Nivat $\mathscr{L}$-adherences, i.e.

$$\text{MN-}\mathscr{L}_{\mathrm{Adh}} = \Big\{ L \mid \exists n,\ \exists L_0, L_1, \ldots, L_n, L'_1, \ldots, L'_n \in \mathscr{L}:$$
$$L = \mathrm{Adh}(L_0) = \bigcup_{i=1}^{n} L_i, L_i'^{\omega} \Big\}.$$

Note that sets of the form (1) may not be adherences. However the existence of L_0 assures consistency.

Problem. Given $\mathscr{L}$ a family of sets: $\mathscr{L}_{\mathrm{Adh}} = \text{MN-}\mathscr{L}_{\mathrm{Adh}}$?

For the Chomsky hierarchy the answers are given by the McNaughton theorem, the Nivat theorem and our Corollary 3.

Theorem 2. *Given a family $\mathscr{L}$ of languages closed under FG and intersection [intersection with regular sets] then $\mathscr{L}_{\mathrm{Adh}}$ is closed under intersection [intersection with $\boldsymbol{R}_{\mathrm{Adh}}$].*

Proof. Let be $L \in \mathscr{L}$, $E \in \boldsymbol{R}$. We consider the set $L' = FG(L) \cap FG(R) \in \mathscr{L}$.

We show that

$$\mathrm{Adh}(L') = \mathrm{Adh}(L) \cap \mathrm{Adh}(E). \tag{4}$$

Indeed,

$$u \in \mathrm{Adh}(L') \text{ iff } FG(u) \subseteq FG(L') = FG(FG(L) \cap FG(E))$$
$$\text{iff } \begin{cases} FG(u) \subseteq FG(L) \\ FG(u) \subseteq FG(E) \end{cases} \text{iff } \begin{cases} u \in \mathrm{Adh}(L) \\ u \in \mathrm{Adh}(E) \end{cases}$$
$$\text{iff } u \in \mathrm{Adh}(L) \cap \mathrm{Adh}(E).$$

So, given L and E we can construct $L' \in \mathscr{L}$ such that (4) holds.

The case $E \in \mathscr{L}$ and $\mathscr{L}$ closed under $\cap$ is proved in a similar way. □

Corollary 4. *The families $\boldsymbol{R}_{\mathrm{Adh}}$, $\boldsymbol{CS}_{\mathrm{Adh}}$ (= $\boldsymbol{RE}_{\mathrm{Adh}}$) are closed under intersection, while the family $\boldsymbol{CF}_{\mathrm{Adh}}$ is closed under intersection with rational adherences.*

Somewhat more general, we can prove the following result.

Corollary 5. *For every full trio $\mathcal{L}$ (i.e. a family of languages closed under homomorphism, inverse homomorphism and intersection with regular sets) the family $\mathcal{L}_{\mathrm{Adh}}$ is closed under intersection with rational adherences.*

Proof. Because every full trio is closed under *FG* the result follows from Theorem 2. □

3. The Chomsky hierarchy of adherences

If we consider the corresponding adherencence families of Chomsky hierarchy we observe that the hierarchy is preserved except for the last two families which collapse. In the non-algebraic case, the hierarchy can be refined by the consideration of McNaughton–Nivat adherences.

A simple technical result is presented in the next

Proposition 3. *If $\mathcal{L}$, $\mathcal{L}'$ are two families of languages satisfying:*

(i) *contain for any letter a, the singleton $\{a\}$;*

(ii) *are closed under $\cdot$, $\cup$, $*$;*

(iii) *$\exists L \in \mathcal{L}' \backslash \mathcal{L}$ such that* $\mathrm{Adh}(L) \in \mathrm{MN}\text{-}\mathcal{L}'_{\mathrm{Adh}}$ *then*

$$\mathrm{MN}\text{-}\mathcal{L}_{\mathrm{Adh}} \subset \mathrm{MN}\text{-}\mathcal{L}'_{\mathrm{Adh}}.$$

Proof. Take $L_0 = La^*$, $L \subseteq V^*$, $a \notin V$. Then $\mathrm{Adh}(La^*) = \mathrm{Adh}(L) \cup La^* = \bigcup_{i=1}^{n} L_i L_i'^{\omega} \cup La^* \in \mathrm{MN}\text{-}\mathcal{L}'_{\mathrm{Adh}}$ because of (iii) and the fact that $La^* \in \mathcal{L}'$.

Suppose towards contradiction that $\mathrm{Adh}(La^*) \in \mathrm{MN}\text{-}\mathcal{L}_{\mathrm{Adh}}$, that is, $\bigcup_{i=1}^{n} L_i L_i'^{\omega} \cup La^* = \bigcup_{j=1}^{p} E_j E_j'^{\omega} = \mathrm{Adh}(E)$, for some $E, E_i, E_i' \in \mathcal{L}$. Because '$a$' is 'fresh', it follows that there exists a subset J of $\{1, \ldots, p\}$ such that $j \in J$ implies $E_j' \backslash \{\lambda\} \subset a^+$ that is $E_j'^{\omega} = a^{\omega}$.

Now $La^* = \bigcup_{j \in J} E_j E_j'^{\omega} = (\bigcup_{j \in J} E_j) a^{\omega}$ which is equivalent to

$$L = \bigcup_{j \in J} E_j.$$

This contradicts $\mathcal{L} \subset \mathcal{L}'$. □

Theorem 3. *The following inclusions hold:*

$$\boldsymbol{R}_{\mathrm{Adh}} \subset \boldsymbol{CF}_{\mathrm{Adh}} \subset \mathrm{MN}\text{-}\boldsymbol{CS}_{\mathrm{Adh}} \subset \mathrm{MN}\text{-}\boldsymbol{RE}_{\mathrm{Adh}} \subset \boldsymbol{CS}_{\mathrm{Adh}} = \boldsymbol{RE}_{\mathrm{Adh}}.$$

Proof. The first three strict inclusions follow from Proposition 3 taking L of (iii) as follows:

$$\{b^n c^n \mid n \geqslant 0\}, \qquad \{b^{n^2} \mid n \geqslant 0\}$$

and respectively a subset of b^* which is r.e. but not context-sensitive. In all cases $\mathrm{Adh}(L) = b^\omega$. The last strict inclusion follows from Corollary 2. □

4. Adherences and programs

There are some connexions between divergences in (flowchart) program schemes and adherences, which justify again the interest in studying them.

Let us consider a flowchart program scheme S together with a labeling α of boxes such that each box have a different label and the labeling may be partial, i.e. there are some boxes without label (or more appropriate in our context, labeled by λ).

If Lab is the (finite) label alphabet, we can associate with finite computations and infinite computations over S, words and respectively ω-words, i.e. the sequence of labels of boxes visited by the computation.

In [5] such words are called *convergent computation shapes* and *divergent computation shapes*. Denote by $C(S, \alpha)$ and $D(S, \alpha)$ the corresponding sets.

Note that in case of partial labeling of S, it is possible to have a divergent computation with a finite 'shape'. However we shall include in $D(S, \alpha)$ only ω-words over Lab.

Let us consider total labelings of program schemes. A (flowchart) program scheme S is called *reduced* if for any box b, there is a path linking the START and STOP which passes through b.

It is easy to prove the following result.

Proposition 4. *For any reduced program scheme S and total assignment α, the set of divergent computation shapes $D(S,\alpha)$ equals the adherence of $C(S,\alpha)$, i.e. $D(S,\alpha) = \mathrm{Adh}(C(S,\alpha))$.*

Proof. Let $u \in D(S, \alpha)$. Then $u = u(1)u(2)\cdots$ is an infinite sequence of labels of boxes in S, describing a path in the graph of S.

For any n, let b_n be the box labeled by $u(n)$. Then S being reduced, it follows that for any n there exists a path linking b_n with STOP. Denote the corresponding word from Lab* by w_n.

Then, for any n, $u[n] \cdot w_n \in C(S, \alpha)$ which yields $u \in \mathrm{Adh}(C(S, \alpha))$. The converse inclusion follows similar. □

Corollary 6. *The family of divergent computation shapes of reduced total labeled program schemes equals $\boldsymbol{R}_{\mathrm{Adh}}$.*

When we consider *interpreted program schemes*, i.e. programs, the set of divergent computation shapes are not necessarily rational [5] and moreover not necessarily adherences.

However regarding partial labeled programs, every member of $\boldsymbol{CS}_{\mathrm{Adh}}$ $(=\boldsymbol{RE}_{\mathrm{Adh}})$ is the set of divergent computation shapes of some program.

Proposition 5. *The family of divergent computations shapes of partially labeled programs, which are adherences, equals* $\boldsymbol{CS}_{\mathrm{Adh}}$ $(=\boldsymbol{RE}_{\mathrm{Adh}})$.

Proof. Let $L=\mathrm{Adh}(L')$ be a member of $\boldsymbol{CS}_{\mathrm{Adh}}$. For simplicity we consider $L'\subset \{a,b\}^*$. Suppose that M is a Turing machine accepting L'.

Let us consider the following program P (having unrefined parts) which will provide us with the fact that considering a partial labeling α with $\mathrm{Lab}=\{a,b\}$ we have that $D(P,\alpha)=\mathrm{Adh}(L')=L$.

The input variable of P is x, which is initialized with an arbitrary member of $\{a,b\}^*$.

```
START
initialize $x$.
if $M$ accepts $x$
   then $y:=x; P'$;
      while true do
         $z:=\mathrm{Next}(x)$; $y:=\mathrm{tail}(z,x)$; $x:=z$;
         if $M$ accepts $x$ then $P'$
                              else
                     od
   else
STOP
```

where $\mathrm{Next}(x)$ is the next word after x in the lexicographic order; tail (z,x) is defined only if $z=xz'$ with $z'\neq\lambda$, in whcih case tail $(z,z)=z'$.

In P above the fragment P' is

```
      while $x\neq\lambda$ do
         if first-is-$a(y)$ then
a                    $y:=$erase-first$(y)$
                     else if first-is-$b(y)$ then
b                          $y:=$erase-first$(y)$
                                          else
                  od
```

We remark that α labels only two boxes in P'.

Our meaning for 'first-is-a' is 'if the first letter in y is a'; the function 'erase-first' erases the first letter of the nonempty word to which it is applied.

The program P diverges in two situations:

(i) for a word $w\in L'$ such that $w\in(L')^c=FG(\mathrm{Adh}(L'))$, i.e. is a prefix of a member of $L=\mathrm{Adh}(L')$;

(ii) for a word $w\in L'$ such that $w\notin L'^c$.

By the construction of α, $D(P, \alpha)$ will describe only the first type of divergence, namely

$$L = \text{Adh}(L') = D(P, \alpha). \quad \square$$

Acknowledgment

I want to thank very much Maurice Nivat for his encouragements to write this paper and one of the referees for detailed comments and pertinent suggestions which improved the paper.

References

[1] J. Berstel, Sur la construction de mots sans carré, *Sem. Théorie Nombres 1978–1979* (1979) 18.01–18.15.

[2] L. Boasson and M. Nivat, Adherences of languages, *J. Comput. System. Sci.* **20**(3) (1980) 285–309.

[3] S. Eilenberg, *Automata, Languages and Machines, Vol. A* (Academic Press, New York, 1974).

[4] S. Istrail, On irreducible languages and nonrational numbers, *Bull. Math. Soc. Math. Roumaine* **21**(3/4) (1977) 301–308.

[5] S. Istrail, On the complexity of program divergence, *Found. Control Eng.* 4(1) (1979) 19–26.

[6] R. McNaughton, Testing and generating infinite sequences by a finite automaton, *Information and Control* **9** (1966) 521–530.

[7] M. Nivat, Mots infinis engendrés par une grammaire algébrique, *R.A.I.R.O. Inform. Théor.* **11** (1977) 311–327.

[8] M. Nivat, Sur les ensembles de mots infinis engendrés par une grammaire algebrique, *R.A.I.R.O. Inform. Théor.* **12** (1978) 259–278.

[9] M. Nivat, Infinite words, infinite trees, infinite computations, in: J. de Bakker and J. van Leuwen, Eds., *Foundations of Computer Science*, Mathematical Centre Tracts **109** (Mathematical Centre, Amsterdam, 1979) 3–52.

[10] A. Salomaa, *Formal Languages* (Academic Press, New York, 1973).

[11] A. Salomaa, *Jewels of Formal Languages Theory* (Computer Science Press, Potomac, 1981).

[12] A. Thue, Über unendliche Zeichenreichen, *Norske Vid. Selsk. Skr. I Mat. Nat. Kl. Christiania* **7** (1906) 1–22.