REVUE FRANÇAISE D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. SÉRIE ROUGE

C. PAIR

Brèves communications. Application de la théorie des ramifications au problème de l'équivalence structurale de deux *C*-grammaires

Revue française d'informatique et de recherche opérationnelle. Série rouge, tome 5, n° R2 (1971), p. 130-136

<http://www.numdam.org/item?id=M2AN_1971__5_2_130_0>

R.I.R.O.
(5e année, R-2, 1971, p. 130-136)

APPLICATION DE LA THÉORIE DES RAMIFICATIONS AU PROBLEME DE L'EQUIVALENCE STRUCTURALE DE DEUX C-GRAMMAIRES

par C. PAIR (1)

Résumé. — *On emploie la théorie des ramifications et des bilangages :*
— pour donner une démonstration simple de la décidabilité de l'équivalence structurale de deux grammaires,
— pour indiquer et justifier un algorithme résolvant ce problème.

1. INTRODUCTION

[8] montre que le problème de l'équivalence structurale de deux C-grammaires est décidable et indique un algorithme de décision. Ce problème est identique à celui de l'équivalence de deux grammaires parenthésées, étudié par [5] et [4].

Le problème se pose fort naturellement en termes de ramifications et de bilangages [7] [6] (2), et les résultats connus sur les bilangages conduisent alors à une démonstration simple de la décidabilité comme nous le verrons au paragraphe 4. Ils permettent de justifier simplement un algorithme résolvant le problème; cet algorithme (paragraphe 5) n'est pas exactement celui auquel conduit l'étude [4], mais il lui est analogue. L'article n'a donc pas pour but d'indiquer des résultats nouveaux, mais de montrer la clarté et la simplicité apportées par l'étude algébrique des ramifications, sur un problème dont Mac-Naughton [5] dit qu'il n'est pas trivial et sur des démonstrations qui motivent cette phrase de Knuth [5] : « this fact is almost obvious; it is hoped that the reader will see (after the long explanation which follows) why it is obvious ».

(1) Université de Nancy 2.
(2) Une remarque de M. J. Bordier est à l'origine de cette étude.

2. RAPPELS

Les notions employées sont définies dans [7] et [6].

La notion de *ramification sur un ensemble* V formalise l'idée d'arborescence orientée de gauche à droite, avec une ou plusieurs racines, étiquetée par des éléments de V. Nous notons $\hat{V}$ l'ensemble des ramifications sur V : il possède deux lois de composition, l'une interne, la concaténation $+$, l'autre externe, l'enracinement $\times$ par un élément de V; c'est un *binoïde sur* V (ou *V-binoïde*) (1). Étant donné deux ensembles V et V', toute application t de V dans V' se prolonge en une application $\hat{t}$ de $\hat{V}$ dans $\hat{V}'$ qui, intuitivement, consiste à remplacer toute étiquette a d'une ramification sur V par l'étiquette $t(a) \in V'$: $\hat{t}$ s'appelle une *transcription*.

Une partie de $\hat{V}$ est un bilangage sur V et parmi les bilangages figurent les *bilangages grammaticaux* : nous notons $S(G)$ le bilangage grammatical engendré par une grammaire G (2), c'est-à-dire l'ensemble des « marqueurs » des phrases engendrées par G. Nous emploierons aussi les *bilangages réguliers*, généralisation des langages réguliers : ce sont les images inverses $\Psi^{-1}(\mathcal{B}')$ d'une partie $\mathcal{B}'$ d'un V-binoïde fini $\mathcal{B}$ par un homomorphisme Ψ de V-binoïdes, de $\hat{V}$ dans $\mathcal{B}$.

Nous utiliserons les résultats suivants :

— les bilangages réguliers sont les images par transcription des bilangages grammaticaux, et eux seuls;

— le complémentaire d'un bilangage régulier, la réunion et l'intersection de deux bilangages réguliers sont des bilangages réguliers.

Ces résultats sont démontrés dans [6] de manière constructive, c'est-à-dire qu'on sait effectivement passer d'un triplet $(\mathcal{B}, \mathcal{B}', \Psi)$, définissant un bilangage régulier $\Psi^{-1}(\mathcal{B}')$, à un couple formé d'une grammaire G et d'une transcription $\hat{t}$ tel que $\Psi^{-1}(\mathcal{B}') = \hat{t}[S(G)]$, et inversement; de même, on sait passer de manière effective de deux triplets $(\mathcal{B}_1, \mathcal{B}'_1, \Psi_1)$ et $(\mathcal{B}_2, \mathcal{B}'_2, \Psi_2)$ définissant deux bilangages réguliers L_1 et L_2 à ceux qui définissent $\complement L_1$, $L_1 \cup L_2$, $L_1 \cap L_2$.

3. LE PROBLEME

Soit une C-grammaire G, d'alphabet terminal T, d'alphabet non terminal N; on pose $V = T \cup N$. Les structures engendrées par G sont intuitivement obtenues en « oubliant » les étiquettes non terminales des ramifications engen-

(1) $\hat{V}$ peut être nommé *V-binoïde universel*. Dans [7], il est appelé binoïde libre sur V, mais cette dénomination n'est pas bonne et il est préférable de nommer *N-binoïde libre* sur T l'ensemble des ramifications sur $N \cup T$ où les éléments de T n'apparaissent qu'aux feuilles : elles sont en effet librement engendrées par T et la ramification vide, à l'aide des deux lois de composition concaténation et enracinement par un élément de N.

(2) C-grammaire généralisée, voir [7] [6].

drées par G. Pour les définir, introduisons un ensemble T' formé de T et d'un unique élément $\sigma \notin T$; soit $\hat{t}$ la transcription de $\hat{V}$ dans $\hat{T}'$ définie par :

$$t(a) = a \quad \text{pour} \quad a \in T, \qquad t(a) = \sigma \quad \text{pour} \quad a \in N.$$

Les *structures* engendrées par G sont les transformées par $\hat{t}$ des ramifications engendrées par G : leur ensemble $\hat{t}[\mathcal{S}(G)]$, qui est une partie de $\hat{T}'$, sera noté $\mathcal{S}'(G)$.

Le problème posé est alors : étant donné deux grammaires G_1 et G_2 de même alphabet terminal T, a-t-on $\mathcal{S}'(G_1) = \mathcal{S}'(G_2)$?

4. DECIDABILITE DU PROBLEME

$\mathcal{S}'(G_1) = \hat{t}[\mathcal{S}(G_1)]$ et $\mathcal{S}'(G_2) = \hat{t}[\mathcal{S}(G_2)]$ sont des bilangages réguliers. Il en est de même de leur différence symétrique

$$\mathcal{S}'(G_1) \Delta \mathcal{S}'(G_2) = [\mathcal{S}'(G_1) \cap \complement\, \mathcal{S}'(G_2)] \cup [\mathcal{S}'(G_2) \cap \complement\, \mathcal{S}'(G_1)]$$

Ce bilangage est donc le transcrit $\hat{t}'[\mathcal{S}(G)]$ d'un bilangage grammatical. De plus on sait passer de manière effective des grammaires G_1, G_2 aux triplets définissant $\mathcal{S}'(G_1)$, $\mathcal{S}'(G_2)$, puis leur différence symétrique, et enfin à la grammaire G.

Le problème se ramène à décider si $\hat{t}'[\mathcal{S}(G)]$ est vide, c'est-à-dire si $\mathcal{S}(G)$ est vide, c'est-à-dire si le langage engendré par G est vide. Or ce dernier problème est décidable [1].

5. UN ALGORITHME DE RESOLUTION

Pour obtenir un algorithme de décision pour le problème proposé, on peut suivre pas à pas la démonstration précédente, ce qui conduit, étant donné deux grammaires H et H', à construire deux grammaires $\bar{H}$ et H_i telles que $\mathcal{S}'(\bar{H}) = \complement\, \mathcal{S}'(H)$ et $\mathcal{S}'(H_i) = \mathcal{S}'(H) \cap \mathcal{S}'(H')$. On obtient l'algorithme qui va être exposé, que nous préférons cependant justifier directement.

Nous notons :

— V_1 et V_2 les alphabets de G_1 et G_2, réunion de leur alphabet terminal commun T et de leur alphabet non terminal;

— $\xrightarrow{1}$ et $\xrightarrow{2}$ leurs relations de production;

— x_1 et x_2 leurs axiomes.

Nous serons amenés à considérer, comme intermédiaires, des ramifications dont les racines ne sont pas les axiomes. Nous dirons qu'une ramification est

quasiment engendrée (en abrégé *qe*) par une grammaire G si elle est engendrée par G où on a remplacé l'axiome par un élément quelconque de l'alphabet de G.

Le principe général de l'algorithme consiste à chercher d'abord si la différence $\mathcal{S}'(G_1) - \mathcal{S}'(G_2)$ est vide, c'est-à-dire s'il n'existe aucune structure engendrée par G_1 et non engendrée par G_2. *Si* $\mathcal{S}'(G_1) - \mathcal{S}'(G_2) \neq \varnothing$, *les deux grammaires ne sont pas structuralement équivalentes. Si* $\mathcal{S}'(G_1) - \mathcal{S}'(G_2) = \varnothing$, *le problème est ramené à savoir si* $\mathcal{S}'(G_2) - \mathcal{S}'(G_1) = \varnothing$.

Étude de $\mathcal{S}'(G_1) - \mathcal{S}'(G_2)$: A toute ramification r quasiment engendrée par G_1, associons :

— sa racine $\rho(r) \in V_1$,

— le sous-ensemble $e(r)$ de V_2 formé des racines des ramifications s quasiment engendrées par G_2 de même structure que r, c'est-à-dire telles que $\hat{t}(s) = \hat{t}(r)$ (1).

Nous construirons l'ensemble E de tous les couples $(\rho(r), e(r))$.

Pour que $\mathcal{S}'(G_1) - \mathcal{S}'(G_2) \neq \varnothing$, *c'est-à-dire pour qu'il existe une structure engendrée par* G_1 *et pas par* G_2, *il faut et il suffit que l'ensemble* E *contienne un couple* (x_1, F) *avec* $x_2 \notin F$.

Détermination de $e(r)$: Toute ramification r à une racine a est de la forme

$$r = a \times (r_1 + r_2 + \ldots + r_p)$$

où $p \geqslant 0$ (pour $p = 0$, $r = a$) et où les r_j sont des ramifications à une racine b_j. De plus

$$r \; qe \text{ par } G_1 \Leftrightarrow r_1, \ldots r_p \; qe \text{ par } G_1 \text{ et } a \xrightarrow{1} b_1 \ldots b_p.$$

Cherchons les ramifications s *qe* par G_2 telles que $\hat{t}(s) = \hat{t}(r)$, sous la même forme

$$s = c \times (s_1 + \ldots + s_{p'}), \qquad \text{avec} \qquad \rho(s_j) = d_j :$$

$$s \; qe \text{ par } G_2 \Leftrightarrow s_1, \ldots, s_{p'} \; qe \text{ par } G_2 \text{ et } c \xrightarrow{2} d_1 \ldots d_{p'}$$

$$\hat{t}(r) = \sigma \times (\hat{t}(r_1) + \ldots + \hat{t}(r_p)) \qquad \hat{t}(s) = \sigma \times (\hat{t}(s_1) + \ldots + \hat{t}(s_{p'}))$$

$$\hat{t}(s) = \hat{t}(r) \Leftrightarrow p = p' \text{ et } \hat{t}(s_j) = \hat{t}(r_j) \text{ pour } j = 1, 2, \ldots, p.$$

D'autre part

$$s_j \; qe \text{ par } G_2 \text{ et } \hat{t}(s_j) = \hat{t}(r_j) \Leftrightarrow d_j \in e(r_j).$$

Comme $e(r)$ est l'ensemble des racines c de ces ramifications s :

$$(1) \qquad e(r) = \{ c \mid (\exists\, d_1 \in e(r_1), \ldots, \exists\, d_p \in e(r_p)) \; c \xrightarrow{2} d_1 \ldots d_p \}.$$

(1) L'ensemble $e(r)$ peut être vide : dans ce cas, et si la grammaire G_1 est réduite [1], on peut affirmer que $\mathcal{S}'(G_1) - \mathcal{S}'(G_2) \neq \varnothing$.

Construction de l'ensemble E : On peut définir par récurrence la *hauteur* $h(r)$. d'une ramification r sur V_1, de manière que :

$$h(a) = 0 \quad \text{pour} \quad a \in V_1, \qquad h[a \times (r_1 + \dots r_p)] = \max [h(r_1), \dots, h(r_p)] + 1.$$

La formule (1) permet alors de se ramener, pour la détermination de $e(r)$ à des ramifications de hauteurs inférieures.

Désignons par E_i l'ensemble des couples $(\rho(r), e(r))$ associés aux ramifications *qe* par G_1 de hauteur i ou plus. Par définition,

$$E_0 = \{ (a, a) \mid a \in T \}. \tag{2}$$

Montrons que :

$$E_{i+1} = E_0 \cup \{ (a, F) \mid (\exists\, b_1, \dots b_p, F_1, \dots, F_p) \tag{3}$$

$$[a \xrightarrow{1} b_1 \dots b_p \qquad \text{et} \qquad (b_j, F_j) \in E_i \text{ pour } j = 1, \dots p$$

$$\text{et} \qquad F = \{ c \mid (\exists\, d_1 \in F_1, \dots, \exists\, d_p \in F_p)\ c \xrightarrow{2} d_1 \dots d_p \}] \}.$$

a) Soit une ramification *qe* par G_1, de hauteur non nulle et au plus $i + 1$:

$$r = a \times (r_1 + \dots + r_p) \qquad \text{avec} \qquad \rho(r_j) = b_j.$$

Nous savons que $a \xrightarrow{1} b_1 \dots b_p$.

D'autre part les r_j sont *qe* par G_1 et de hauteur i au plus : posons $F_j = e(r_j)$ et $F = e(r)$; par définition $(b_j, F_j) \in E_i$; F est donné par (1). (a, F) appartient donc au second membre de (3). E_{i+1} est contenu dans ce second membre.

b) Réciproquement, envisageons a, F, $b_1 \dots, b_p$, $F_1, \dots F_p$ tels que

$$a \xrightarrow{1} b_1 \dots b_p\ , (b_j, F_j) \in E_i \text{ pour } j = 1, \dots p,$$

$$F = \{ c \mid (\exists\, d_1 \in F_1, \dots, d_p \in F_p)\ c \xrightarrow{2} d_1 \dots d_p \}$$

$(b_j, F_j) \in E_i$ signifie qu'il existe r_j *qe* par G, de hauteur i au plus, telle que

$$\rho(r_j) = b_j \qquad \text{et} \qquad e(r_j) = F_j.$$

Envisageons

$$r = a \times (r_1 + \dots + r_p).$$

r est *qe* par G_1, de hauteur $i + 1$ au plus et, d'après (1), $e(r) = F$: par suite (a, F) appartient à E_{i+1}.

D'après leur définition les ensembles E_i sont croissants et E est leur réunion. Ils sont contenus dans l'ensemble fini $V_1 \times \mathfrak{P}(V_2)$, donc ne peuvent être tous différents : il existe un i tel que $E_{i+1} = E_i$, et alors $E = E_i$.

Algorithme : L'algorithme consiste donc à appliquer les formules (2) et (3) jusqu'à ce que $E_{i+1} = E_i$, à regarder alors si E_i contient un couple (x_1, F) avec $x_2 \notin F$: dans ce cas G_1 et G_2 ne sont pas structuralement équivalentes; sinon, en échangeant G_1 et G_2, on applique de nouveau (2) et (3) jusqu'à ce que $E_{i+1} = E_i$: G_1 et G_2 sont structuralement équivalentes si, et seulement si, E_i ne contient aucun couple (x_2, F) avec $x_1 \notin F$.

Cet algorithme risque d'être long, notamment dans le cas où G_1 et G_2 sont bien structuralement équivalentes; dans le cas contraire, on s'arrête dès qu'on trouve dans E_i un couple (x_1, F) avec $x_2 \notin F$, ou, si G_1 est réduite, un couple $(a, \varnothing)$. On utilisera d'autre part des améliorations pratiques, usuelles dans ce genre d'algorithme, visant à éviter de retrouver à chaque étape les couples déjà rencontrés. Il sera de plus intéressant de coupler les règles de G_1 et G_2 qui ont même longueur et mêmes terminaux aux mêmes places, autrement dit qui ne diffèrent que par les non terminaux : elles seules peuvent être des règles $a \xrightarrow{1} b_1 \ldots b_p$ et $c \xrightarrow{2} d_1 \ldots d_p$ associées dans l'égalité (3).

6. RESOLUTION D'AUTRES PROBLEMES

Un argument analogue à celui du paragraphe 4 prouve qu'il est décidable de savoir si $\mathcal{S}'(G_1) \cap \mathcal{S}'(G_2)$ est vide.

Ici l'algorithme est beaucoup plus simple : on cherche l'ensemble I des couples $(a, c) \in V_1 \times V_2$ tels qu'il existe des ramifications r *qe* par G_1, s *qe* par G_2 vérifiant

$$\hat{t}(r) = \hat{t}(s) \qquad \rho(r) = a \qquad \rho(s) = c.$$

Pour que $\mathcal{S}'(G_1) \cap \mathcal{S}'(G_2) \neq \varnothing$, il faut et il suffit que $(x_1, x_2) \in I$.

En appelant I_i le sous-ensemble de I obtenu en se limitant aux ramifications r et s de hauteur i au plus, on peut écrire :

$$I_0 = T \times T$$

$$I_{i+1} = I_0 \cup \{ (a, c) \mid (\exists\, b_1, \ldots, b_p, d_1, \ldots d_p)[a \xrightarrow{1} b_1 \ldots b_p \text{ et } c \xrightarrow{2} d_1 \ldots d_p \text{ et } (b_j, d_j) \in E_i \text{ pour } j = 1, 2, \ldots p] \}.$$

On applique cette formule jusqu'à ce que $I_{i+1} = I_i$; alors $I = I_i$.

On peut aussi décider si $\mathcal{S}'(G_1) - \mathcal{S}'(G_2)$ ou $\mathcal{S}'(G_1) \cap \mathcal{S}'(G_2)$ est infini : on se ramène de la même façon à décider si un bilangage régulier $\hat{t}[\mathcal{S}(G)]$ est infini; on voit facilement que pour qu'il en soit ainsi, il faut et il suffit que $\mathcal{S}(G)$ soit infini puis, après réduction de la grammaire G, qu'il existe un symbole non terminal a et deux mots μ et μ' tels que $\mu\, a\, \mu'$ dérive strictement de a.

Les *bracketed languages* de [2] peuvent être aussi identifiés à des bilangages grammaticaux et on prouve donc de la même manière que les problèmes correspondants sont décidables.

D'ailleurs, on peut préciser l'un des résultats signalés au paragraphe 2 : tout bilangage régulier est image d'un bilangage grammatical par une transcription qui laisse invariants les symboles terminaux. Si on considère un bilangage régulier formé de structures sur un alphabet T (c'est-à-dire une partie de $\hat{T}'$,cf. paragraphe 3), c'est donc un $\hat{t}[\mathcal{S}(G)]$ où t laisse invariant les symboles terminaux de G et transforme nécessairement les symboles non terminaux en σ : donc, $\hat{t}[\mathcal{S}(G)] = \mathcal{S}'(G)$. On en déduit :

Etant donné deux C-grammaires G_1 et G_2 ,d'alphabet terminal T, la réunion, l'intersection, la différence de $\mathcal{S}'(G_1)$ et $\mathcal{S}'(G_2)$, le complémentaire de $\mathcal{S}'(G_1)$ dans $\hat{T}'$ sont des ensembles de structures engendrées par des C-grammaires.

BIBLIOGRAPHIE

1. Y. BAR-HILLEL, M. PERLES et E. SHAMIR, « On formal properties of simple phrase structure grammars », *Z. Phon., Sprachwiss., Komm.*, 1961, 14, 143-172.
2. S. GINSBURG et M. A. HARRISSON, « Bracketed context-free languages », *J. Computer and System Science*, 1967, 1, 1-23.
3. M. GROSS et A. LENTIN, *Notions sur les grammaires formelles*, Gauthier-Villars, Paris, 1967.
4. D. KNUTH, « A Characterization of parenthesis languages », *Inf. and Control*, 1967, 11, 269-289.
5. R. Mc NAUGHTON, « Parenthesis grammars », *J.A.C.M.*, 1967, 14, 490-500.
6. C. PAIR, « Sur des notions algébriques liées à l'analyse syntaxique », Centre d'Automatique de l'École des Mines, Fontainebleau (1969) et *Rev. F. Inf. et R.O.*, R/3, 1970.
7. C. PAIR et A. QUERE, « Définition et étude des bilangages réguliers », *Inf. and Control*, 1968, 13, 565-593.
8. M. PAULL et S. UNGER, « Structural equivalence of context-free grammars », *J. Computer and System Science*, 1968, 2, 427-463.
9. J. THATCHER, *Characterizing derivation trees of context-free grammars through generalized finite automata theory*, IBM Research note NC719, 1967.