J. M. AUTEBERT
J. BEAUQUIER
L. BOASSON
M. NIVAT

Quelques problèmes ouverts en théorie des langages algébriques

RAIRO. Informatique théorique, tome 13, nº 4 (1979), p. 363-378
<http://www.numdam.org/item?id=ITA_1979__13_4_363_0>

R.A.I.R.O. Informatique theorique/Theoretical Informatics
(vol. 13, n° 4, 1979, p. 363 à 379)

QUELQUES PROBLÈMES OUVERTS EN THÉORIE DES LANGAGES ALGÉBRIQUES (*)

par J. M. AUTEBERT ([1]), J. BEAUQUIER ([2])
L. BOASSON ([3]) et M. NIVAT ([3])

Communiqué par J. BERSTEL

Résumé. — *On présente un certain nombre de problèmes ouverts en théorie des langages algébriques.*

Abstract. — *We present here some open questions about the context-free languages.*

1. INTRODUCTION

L'article que voici est un recueil de conjectures et questions ouvertes concernant les langages algébriques.

Tout le monde sait que certaines questions très simples et formulées très tôt à leur sujet ne sont pas encore résolues malgré le grand nombre de travaux qui leur ont été consacrés par des chercheurs éminents. C'est le cas des deux suivantes :

— l'équivalence de deux automates à pile déterministes (*dpda*) est-elle décidable ?

— trouver une borne supérieure précise, c'est-à-dire atteignable, du temps de reconnaissance d'un langage algébrique par une machine de Turing.

Les résultats les plus fins concernant ces deux questions se trouvent dans les œuvres de Leslie Valiant [54, 55]. Il n'y a aucune raison semble-t-il de pronostiquer une prompte réponse à ces deux questions. Les conjectures que nous proposons ci-dessous à l'attention du lecteur sont moins connues, certaines n'ont même jamais été publiées.

(*) Reçu janvier 1978, révisé juin 1978.
([1]) Institut de Programmation, Université Paris-VI.
([2]) Institut d'Informatique, U.E.R. de Mathématiques, Université de Picardie.
([3]) U.E.R. de Mathématiques, Université Paris-VII.

Si nous cherchons à les caractériser nous dirons qu'elles sont toutes liées à la *complexité structurelle* des langages algébriques, ce que nous définirons comme suit :

Un langage algébrique, comme tout langage, peut être considéré de trois points de vue :

— on peut *l'engendrer* au moyen de quelque grammaire;

— on peut le *reconnaître* au moyen de tel ou tel type d'automate ou de machine;

— on peut essayer de le *construire* à partir de langages connus au moyen d'un certain nombre d'opérations sur les langages parmi lesquelles certaines apparaissent comme particulièrement naturelles et ont d'ailleurs beaucoup été utilisées : morphismes et morphismes inverses, intersection avec un rationnel, opérations booléennes, produit, étoile de Kleene, substitution, etc.

Ces trois points de vue ne sont naturellement pas étrangers les uns aux autres. Chacun d'entre eux induit une certaine notion de complexité : c'est la troisième que nous qualifierons de structurelle.

Un certain nombre des conjectures ci-dessous sont du même type que : « Le langage de Dyck est-il image dans un morphisme alphabétique de l'intersection de trois langages algébriques linéaires ? » [16]. Question qui d'ailleurs se reformule en termes de machines de Turing : « Le langage de Dyck est-il reconnaissable en temps linéaire par une machine de Turing ne faisant qu'un nombre k de demi-tours, borné à l'avance ? »

Problèmes de construction donc où la difficulté vient le plus souvent de morphismes qui effacent tout ou partie de l'information contenue dans un mot : la tactique est alors d'éclater le mot, de le truffer de marqueurs, de le factoriser en réécrivant les divers facteurs sur des alphabets disjoints jusqu'à faire apparaître un mot dont la structure reflète fidèlement la nature composite. Ceci est du moins la tactique pour établir le résultat positif : oui tel langage L se construit de telle façon à partir de quelques langages primitifs donnés.

Pour établir le contraire d'une telle assertion on n'a jusqu'à présent rien trouvé de mieux que les lemmes d'itération : il y a toute une famille de tels lemmes qui sont tous du type :

« Si L est un langage de la famille $\mathscr{L}$, tout mot assez long f de L se factorise en $f=f_1 f_2 \ldots f_n$ tel que l'intersection de L avec le langage borné $f_1^* f_2^* \ldots f_n^*$ a la propriété (P). »

Pour ce qui est des langages algébriques on prend $n=5$ et l'on tombe sur la notion de paire itérante, une paire itérante étant un quintuplet $(\alpha, u, \beta, v, \gamma)$ tel que $L \cap \alpha u^* \beta v^* \gamma$ ait une propriété (P) variable avec le qualificatif de la paire itérante.

Luc Boasson a systématiquement étudié les énoncés du type : « si L admet une paire de tel type et si Φ est une opération de telle nature sur L alors $\Phi(L)$ contient une paire du même type ».

On conçoit que de tels énoncés soient très précieux pour démontrer que L' ne peut être construit à partir de L au moyen de Φ. J. Beauquier a étendu ces techniques aux systèmes de paires et réussi à caractériser les générateurs du cône des algébriques comme étant les langages algébriques « structurellement les plus compliqués », c'est-à-dire admettant tous les systèmes de paires admissibles par un langage algébrique. Et l'on arrive ainsi aux questions essentielles de cet article.

« Peut-on caractériser certaines sous-classes de la famille des langages algébriques par une propriété de ses paires itérantes ? » ce qui permettrait de rendre nécessaires et suffisants certains lemmes d'itération. En particulier on aimerait pouvoir caractériser par un tel procédé les langages rationnels au sein des langages algébriques. L'outil très puissant déjà que constituent les paires et systèmes de paires deviendrait beaucoup plus puissant si l'on pouvait conclure de l'existence d'une paire à celle d'une paire plus courte, de longueur inférieure à une borne dépendant du langage.

Mais on tombe là sur des problèmes difficiles : passer d'une propriété des paires d'un langage, c'est-à-dire des intersections de L avec les langages bornés de la forme $\alpha u^* \beta v^* \gamma$, à une propriété de L, c'est passer d'une propriété locale à une propriété globale. On sait que c'est toujours délicat, en l'occurrence on ne sait pas le faire si L n'a pas une structure assez apparente (langage parenthétique, intersection d'un langage de Dyck et d'un rationnel, etc.).

Nous publions ces conjectures pour deux raisons :

— la première est que nous nous sommes persuadés qu'elles sont au cœur de tous les problèmes qu'on ne sait pas résoudre actuellement sur les langages algébriques y compris les deux sus-mentionnés : la résolution par l'affirmative d'une quelconque de ces conjectures ouvrirait une voie nouvelle pour l'étude des langages algébriques;

— la seconde est que le fil conducteur qui nous a conduit à rassembler ces conjectures, à savoir une certaine façon commune aux quatre auteurs, forgée en 6 ou 7 ans de collaboration continue, d'imaginer la résolution des problèmes ouverts en théorie des langages algébriques n'est peut-être pas le bon. Déjà beaucoup de conjectures ont cédé à la simple découverte d'un langage inconnu jusqu'alors; et l'expérience nous a appris que la famille des langages algébriques est très vaste et encore largement inexplorée : la clé des problèmes ouverts que l'on trouvera ci-dessous est peut-être dans l'observation de quantité d'exemples de langages algébriques. L'algorithmique, qui a elle aussi ses problèmes ouverts ($P=NP$?), ne se nourrit-elle pas essentiellement de l'analyse de quantité d'algorithmes, de même que la géométrie s'est longtemps nourrie de l'étude fine de mille et une courbes ou surfaces ?

Quant à l'intérêt de l'étude des langages algébriques nous n'en dirons rien : nous pensons que le lien intime qui unit les langages algébriques aux arbres et aux piles d'un maniement si universel dans tous les domaines de l'informatique justifie amplement le temps que l'on peut consacrer à la résolution d'un des problèmes évoqués dans cet article.

2. BREF HISTORIQUE ET GLOSSAIRE

Les concepts de langage et de grammaire algébriques (context-free) ont été introduits par Chomsky en 1959, dans l'idée de fournir un modèle mathématique aux langues naturelles. Très rapidement, les langages algébriques ont été caractérisés en termes d'automate à pile [18, 23]. Diverses sous-familles des langages algébriques, définies par des contraintes sur le processus de génération (grammaire) ou de reconnaissance (automate à pile) ont ensuite été étudiées. Citons la famille des langages linéaires [19], c'est-à-dire la famille des langages engendrés par des grammaires algébriques ayant au plus un non-terminal dans chaque membre droit de règle, la famille des langages déterministes [29] (c'est-à-dire la famille des langages reconnus par les automates à pile fonctionnant de manière déterministe) ou la famille des langages à un compteur [49, 24] (c'est-à-dire la famille des langages reconnus par les automates à pile utilisant un seul symbole de pile.)

C'est à partir des grammaires algébriques que fut étudié le phénomène de l'ambiguité [25, 46] (Une grammaire algébrique est ambiguë si elle engendre un mot par deux arbres de dérivation distincts; un langage algébrique est inhérement ambigu si toute grammaire qui l'engendre est ambiguë.) Parallèlement, furent définies des opérations entre langages. L'une des plus puissantes, la substitution, peut être facilement étendue aux familles de langages [34] et permet d'obtenir de nouvelles familles de langages à partir de familles connues. La première construction de ce type a été réalisée à partir des langages linéaires. En notant $QR(1)$ la famille des langages linéaires (appelés aussi langages quasi-rationnels d'ordre 1), on peut définir inductivement la famille des langages quasi-rationnels d'ordre n, $QR(n)$ par : $QR(n)=QR(n-1)\hat{\sigma}\, QR(1)$, où $\hat{\sigma}$ désigne la substitution entre familles de langages. La famille QR des langages quasi-rationnels (ou des langages non expansifs, ou encore des langages à dérivations bornées) est alors l'union pour $n \geqslant 1$ des familles $QR(n)$.

Il apparut, vers 1966, que de nombreuses familles de langages possédaient les mêmes propriétés de fermeture, vis-à-vis d'opérations comme les morphismes, les morphismes inverses, les intersections avec un langage rationnel. Une opération, la transduction rationnelle, apparue dans un article d'Elgot et Mezeï [22] sous une forme particulière, s'avéra grâce à la caractérisation due à Nivat [44] l'outil privilégié de l'étude de ces propriétés :

Caractérisation [44] :

Soient X et Y deux alphabets. La transduction T de 2^{X^*} dans 2^{Y^*} est rationnelle si et seulement si il existe un alphabet Z, un rationnel R de Z^* et deux morphismes Φ et Ψ de Z^* dans X^* et de Z^* dans Y^*, tels que, pour tout langage L de X^* : $T(L)=\Psi(\Phi^{-1}(L)\cap R)$.

On s'orienta ainsi vers la recherche de propriétés générales, valables pour toute famille de langages satisfaisant certaines conditions de fermeture, pour obtenir ensuite, comme corollaires, les propriétés relatives à des familles particulières. C'est ce point de vue qui a conduit à l'étude des *AFL* [30] et des cônes rationnels [20].

DÉFINITIONS : Un cône rationnel est une famille de langages fermée par transduction rationnelle.

Un full-*AFL* est un cône rationnel fermé par union, produit et étoile.

C'est également ce point de vue qui, très récemment, a conduit à l'étude des cylindres [2, 13, 17, 36].

DÉFINITION [2] : Un cylindre est une famille de langages fermée par intersection rationnelle et morphisme inverse.

Ces notions brièvement présentées ci-dessus nous permettront d'énoncer agréablement un certain nombre des conjectures que nous présentons dans la suite. Nous avons regroupé celles-ci sous quatre rubriques : la première concerne le problème d'une réciproque du lemme de l'étoile fournissant une caractérisation des langages algébriques qui sont rationnels; la deuxième traite des problèmes posés par le lemme de la double étoile et ses variantes; la troisième rubrique s'attache aux questions soulevées par les opérations entre langages et particulièrement par la substitution; la quatrième et dernière regroupe enfin quelques conjectures qui ne sont pas directement reliées aux grands axes précédents.

I. Réciproque du lemme de l'étoile

Cette première série de problèmes relève pour l'essentiel de l'étude des langages algébriques qui sont « presque rationnels ». Il convient de noter que le plus souvent, ces questions en rejoignent de fort anciennes : par exemple, parmi les plus classiques, celle de décider si un langage algébrique est rationnel. On sait que ceci est impossible en général [48], mais l'est dans le cas des langages déterministes [52]. Le caractère indécidable du problème général (et même du problème de l'égalité d'un langage algébrique et d'un langage rationnel donnés [48]), s'il explique que les problèmes ci-dessous soient difficiles, n'interdit pas d'espérer y apporter des réponses, éventuellement partielles.

L'idée essentielle est de tenter de caractériser les langages algébriques qui sont rationnels. Typiquement, les résultats cherchés ressemblent à ceux de Boasson [7] dont le plus simple peut s'énoncer ainsi :

Un langage L sur l'alphabet X vérifie la condition (P_*) si et seulement si pour tout α, u, β, v, γ dans X^* tels que $\{\alpha u^n \beta v^n \gamma \mid n \geqq 1\} \subseteq L$, on a $\alpha u^* \beta v^* \gamma$.

Proposition [7] : *Un langage algébrique vérifiant (P_*) est rationnel.*

On peut ainsi tenter d'établir quelques résultats voisins plus forts : le langage L sur X vérifie la condition (P_+) [resp. (P_{k*}), resp. (P_{k+})] si et seulement si pour tous α, u, β, v, γ de X^* tels que $\{\alpha u^n \beta v^n \gamma \mid n \geqq 1\} \subseteq L$, on a $\alpha u^+ \beta v^+ \gamma \subseteq L$ [resp. $\alpha (u^k)^* \beta (v^k)^* \gamma \subseteq L$; resp. $\alpha (u^k)^+ \beta (v^k)^+ \gamma \subseteq L$].

Conjecture 1 :

1 *a*. Un langage algébrique vérifiant (P_+) est rationnel.

1 *b*. Un langage algébrique vérifiant (P_{k*}) pour quelque entier k est rationnel.

1 *c*. Un langage algébrique vérifiant (P_{k+}) pour quelque entier k est rationnel.

Remarque 1 : Prouver 1 *c* donnerait en fait une caractérisation des langages algébriques qui sont rationnels puisque tout langage rationnel vérifie (P_{k+}) pour un entier k.

Remarque 2 : Les preuves de 1 *b* et 1 *c*, si elles sont possibles, sont sûrement délicates puisqu'il existe des langages algébriques non rationnels vérifiant ce type de propriétés avec *k variable* (suivant le mot $\alpha u \beta v \gamma$ considéré) *mais borné* (i.e. k prenant un nombre fini de valeurs) ! On vérifiera facilement que tel est le cas du langage [33] :

$$G = \{ a^{n_1} b a^{n_2} b \ldots a^{n_p} b \mid p \geqq 1,\ n_i \geqq 1,\ \exists j \text{ tq } n_j \neq j \}.$$

Dans le même ordre d'idées, on peut proposer une question tout à fait analogue mais dans un cadre très différent :

Conjecture 2 [47] : Le monoïde syntaxique d'un langage algébrique contient toujours un élément d'ordre infini (on trouvera une définition du monoïde syntaxique dans le traité d'Eilenberg [21]).

Ces questions d'aspects un peu combinatoire rejoignent les problèmes relatifs aux cônes rationnels minimaux [34], et plus particulièrement la :

CONJECTURE 3 : Il n'existe pas de cône rationnel minimal inclus dans celui des langages algébriques (i.e. il n'existe pas de cône rationnel dont tout élément algébrique non rationnel soit générateur).

Un problème plus simple de même nature apparaît naturellement dans ce cadre : un langage L est dit 1-localement linéaire (albrégé en *1-ll)* si et seulement si tout élément du cône rationnel qu'il engendre vérifie le lemme de l'étoile [56].

CONJECTURE 4 [3] : Un langage algébrique *1-ll* est rationnel (on sait que cette conjecture est fausse si le langage considéré n'est pas algébrique [3]).

II. Paires itérantes

Si l'on considère maintenant l'équivalent du lemme de l'étoile pour les langages algébriques, on peut considérer que l'on dispose de deux théorèmes d'itération : celui de Bar-Hillel, Perles, Shamir [4] et celui d'Ogden [45]. La première question, naturelle, est celle de leur caractère suffisant. On sait qu'il existe des langages non algébriques vérifiant ces théorèmes d'itération (*Voir* [39] et [14]). Néanmoins, au-delà de leur usage comme condition nécessaire, ils suggèrent l'importance de la notion de paire itérante dégagée récemment dans Boasson [9]; celle-ci apparaît comme un outil commode pour étudier les langages algébriques. Rappelons qu'une paire itérante d'un mot f dans le langage L sur X est un quintuplet $\eta=(\alpha, u, \beta, v, \gamma)$ de mots de X^* vérifiant :

- $f=\alpha u \beta v \gamma$;
- $\{\alpha u^n \beta v^n \gamma \mid n \geqq 1\} \subseteq L$;
- $uv \neq 1$.

Notons que les conditions (P) de la conjecture 1 pourraient s'énoncer en terme de « dégénérescence » des paires itérantes d'un langage. On peut aussi énoncer alors le lemme d'Ogden sous la forme :

LEMME d'OGDEN : *Étant donné un langage algébrique L sur l'alphabet X, il existe un entier N_0 tel que si dans un mot f de L on distingue N_0 (ou plus) occurrences de lettres, f admette une paire itérante $\eta=(\alpha, u, \beta, v, \gamma)$ dans L telle que soit α, u et β, soit β, v et γ contiennent des occurrences distinguées, le facteur $u\beta v$ en contenant au plus N_0.*

Les paires itérantes ont été classifiées suivant les propriétés de l'ensemble $\{(n, m)\in N^2 \mid \alpha u^n \beta v^m \gamma \in L\}$. On trouvera dans Boasson [9] les principaux types de paires étudiés. Retenons ici seulement que la paire η est dite très stricte si et seulement si cet ensemble se réduit à la diagonale de N^2, i.e. à $\{(n, n) \mid n \in N\}$. En outre, très vite, on se rend compte que l'étude des mots d'un langage L contenant plusieurs paires itérantes simultanées disjointes est nécessaire. On peut d'ailleurs considérer, quoique menés dans un formalisme différent, que les travaux de Ginsburg [26] sur les langages algébriques bornés relèvent de cette étude. Il est aussi bien clair que l'étude des langages à un compteur qui a commencé très tôt [49] a, en fait, débouché aussi très vite sur des problèmes de cette nature (*voir* [8]).

L'intérêt majeur des paires itérantes vient de ce que, informellement, une image par transduction rationnelle n'admet de système de paires très strictes d'un certain type que si l'objet admettait déjà un système du même type. (On trouvera dans Beauquier [5] des résultats précis de cette nature ainsi qu'une illustration de leur usage.)

A. Nous commencerons par un problème un peu technique (question 2) : la raison en est que la plupart des conjectures de ce même paragraphe pourraient être résolues (ou, bien avancées) si celui-ci recevait une réponse positive. Considérant une paire itérante très stricte $\eta = (\alpha, u, \beta, v, \gamma)$ d'un langage algébrique L, on définit sa longueur $|\eta| = |u| + |v|$.

QUESTION 1 : Étant donné un langage algébrique L, existe-t-il un entier l_0 (ne dépendant que de L) tel que si $\eta = (\alpha, u, \beta, v, \gamma)$ est une paire itérante très stricte de longueur supérieure à l_0, il existe une paire itérante très stricte de L de longueur inférieure à l_0 $\eta' = (\alpha, u', \beta, v', \gamma)$?

Intuitivement, étant donnée une paire itérante « très longue », peut-on en trouver une plus courte ayant le même contexte.

Le problème technique annoncé est de prouver un résultat de cette nature pour les systèmes de paires. Considérant un système de k paires itérantes très strictes (disjointes) du langage L, on peut en définir la longueur comme le maximum de la longueur des k paires qui le constituent. On peut alors poser la :

QUESTION 2 : Étant donné un langage algébrique L admettant un système de paires itérantes très strictes S, existe-t-il un entier l_0 (dépendant de L et peut-être du type du système, mais pas de S lui-même) tel que si S est de longueur supérieure à l_0, L admette un système S' de même type plus court que l_0 ?

Ces deux questions sont en fait des formulations modernes de problèmes très anciens concernant les langages algébriques : comment peut-on construire tous les langages algébriques à partir d'éléments simples (voire finis) et d'opérations de nature combinatoire ? Sans que l'on ne puisse attribuer vraiment à qui que ce soit la paternité à cette question, il est clair qu'elle court à travers bon nombre de publications concernant les langages algébriques. Par exemple, c'est bien de celle-ci que relève la caractérisation des langages quasi-rationnels de Ginsburg et

Spanier [31], ou encore, plus récemment, celle établie encore par Ginsburg [58] pour les langages associés aux « grammar forms ». On peut aussi considérer le théorème de Parikh [46], surtout dans sa preuve originelle, comme une amorce de réponse, bien que donnée en terme de décompositions plutôt que de recompositions. On retrouvera plus loin (*voir* prop. 1) un résultat expliquant la difficulté de ce problème. Cependant, dans certains cas particuliers, il semble plus facile de répondre aux questions 1 et 2; ainsi, par exemple, la forme particulière des grammaires parenthétiques de McNaughton [43] et Knuth [40] ou encore des grammaires simples de Hopcroft et Korenjack [38] peuvent laisser espérer des résultats. De même les travaux de Havel-Harrison [37] sur les grammaires déterministes ouvrent des perspectives positives.

Outre leur intérêt propre (souligné ci-dessus), les questions 1 et 2 offrent une voie possible à la solution de la :

CONJECTURE 5 [34] : Le cône rationnel des langages algébriques qui ne sont pas générateurs est non-principal.

En effet, si la question 2 admet une réponse positive, il sera possible de prouver cette conjecture en établissant le résultat plus fort :

CONJECTURE 6 [12] : Étant donné un langage algébrique L sur l'alphabet X et trois lettres a, b, c ne figurant pas dans X :

6 *a* [12] : Si L domine $\{a^n f_1 c f_2 b^n \mid n \geqq 1; f_1 f_2 \in L\}$, L est générateur.

6 *b* [10] : Si L domine $\{a^n f_1 b f_2 b \ldots f_n b \mid n \geqq 1;\ f_1, \ldots, f_n \in L\}$, L est générateur.

A défaut d'établir ces résultats, on peut tenter de traiter les problèmes analogues dans le cas plus simple du cône des langages linéaires, qui offre le même type de difficultés dans son principe et soulève des questions semblables à notre question 2.

CONJECTURE 7 :

7 *a* : Les langages linéaires qui ne sont pas générateurs du cône des linéaires forment un cône non principal.

7 *b* : Étant donné un langage linéaire L sur l'alphabet X et deux lettres a et b ne figurant pas dans X, si L domine $\{a^n l b^n \mid n \geqq 1;\ l \in L\}$ L est générateur du cône des linéaires.

Parmi les diverses formulations de la conjecture 5, il en est une qui, faisant toujours apparaître les problèmes évoqués par notre question 2, mérite d'être rappelée ici :

CONJECTURE 8 : Un langage algébrique dominant tous les langages quasi-rationnels est générateur.

Soit encore : le plus petit cône rationnel principal contenant les quasi-rationnels est le cône des langages algébriques tout entier.

Toujours indépendamment de leur intérêt propre, les questions 1 et 2 sont également importantes dans les études concernant les cylindres. Ceux-ci, introduits récemment avec diverses variantes [2, 17, 35] ont pour objet de rendre compte des familles qui ne sont pas des cônes rationnels (langages déterministes, non ambigus, ...) ainsi que des problèmes de complexité de reconnaissance. (On trouvera ainsi des relations simples mais pas fortuites entre les langages linéaires construits par Boasson-Nivat dans [13] et ceux de Sudborough [53].) Parmi les questions toujours ouvertes, la plus intéressante semble bien être la :

CONJECTURE 9 : Le cylindre des langages algébriques non ambigus est non principal.

Notons que les langages d'ambiguïté bornée, ou ceux d'ambiguïté au plus k peuvent donner lieu à des conjectures analogues. Rappelons aussi que le cylindre des langages déterministes est non principal (Greibach [36]) alors que celui des langages algébriques l'est (Greibach [35] reformulant un théorème de Shamir [51]).

III. Substitutions

Parmi les opérations entre familles de langages, l'une des plus puissantes est certainement la substitution. C'est elle qui permet de définir facilement les langages quasi-rationnels ou à compteur-itéré. En outre, elle ne fait pas sortir de la famille des langages algébriques qui, de ce point de vue, est impossible à décrire. Plus précisément, notant $C_1 \hat{\sigma} C_2$ le cône rationnel $\{ L_1 \sigma L_2 \mid L_1 \in C_1$ et $L_2 \in C_2 \}$, on a :

PROPOSITION 1 [34] : *Si C_1 et C_2 sont deux cônes rationnels tels que $C_1 \hat{\sigma} C_2 = \mathrm{Alg}$, alors soit C_1 soit C_2 est déjà égal à Alg.*

Ce fait découle immédiatement du principal résultat concernant les substitutions que nous rappelons ci-dessous. Étant donné deux langages A et B définis sur des alphabets disjoints X_A et X_B, on définit la substitution syntaxique de B dans A par :

$$A \circ B = \{ x_1 f_1 x_2 f_2 \dots x_n f_n \mid x_i \in X_A ; x_1 \dots x_n \in A; f_i \in B \} .$$

On peut alors énoncer le :

LEMME SYNTAXIQUE [34, 6] : *Étant donnés deux cônes rationnels C_1 et C_2, si $A \circ B \in C_1 \hat{\sigma} C_2$, alors soit $A \in C_1$, soit $B \in C_2$.*

Parmi les premières questions que l'on peut se poser, l'une des plus naturelle est alors :

QUESTION 3 : Ce lemme est-il encore (même partiellement) vrai dans le cas des cônes rationnels fidèles ?

Cette question apparaît immédiatement lorsque l'on cherche à répondre à :

QUESTION 4 : Existe-t-il un sous-cône rationnel des langages algébriques qui soit principal, mais ne soit pas fidèlement principal ?

Cette question peut être ainsi posée : Existe-t-il un langage algébrique F tel que pour tout langage L de $\mathscr{C}(F)$, on ait $\mathscr{C}_f(L) \subsetneq \mathscr{C}(F)$?

Il est en effet naturel de se poser une telle question puisque l'on sait non seulement que ceci est impossible dans les cônes classiques mais même que tout générateur de ceux-ci est fidèle; tel est le cas des familles des langages algébriques [11], linéaires [27], à un compteur [28] ... On connaît cependant des cônes rationnels qui, s'ils sont fidèlement principaux, admettent des générateurs qui ne sont pas des générateurs fidèles [27, 9]. Toutes ces questions relatives aux transductions et cônes fidèles rejoignent les problèmes de reconnaissance en temps réel ou « quasi-réel » (*voir* [15]). De même ce sont elles qui permettraient d'espérer traiter des problèmes semblables dans le cadre des langages « context-sensitifs » (*voir* par exemple [41] ou, plus récemment [42]).

La majorité des questions évoquées dans les paragraphes II et III font apparaître, ainsi qu'il a déjà été signalé, l'intérêt de la question 2. Mais elles conduisent aussi à un autre type de problème qui, à certains égards, est une vision différente de la même difficulté. Il s'agit de définir une notion de limite pour les suites de langages qui soit adéquate aux questions traitées. En général, le problème se pose en ces termes : on considère une suite croissante de langages $L_1 \subseteq L_2 \subseteq \ldots \subseteq L_n \subseteq \ldots$ et un langage A dominant rationnellement chaque langage L_i. Peut-on alors définir une limite de la suite, notée L_∞, de telle sorte que A domine L_∞ ? Ou encore, la suite ayant une limite « naturelle », peut-on prouver que A domine cette limite ? Par exemple, on peut considérer la suite de langages algébriques $L_0, L_1, \ldots, L_n, \ldots$ engendrés par les grammaires $G_0, G_1, \ldots, G_n, \ldots$ d'axiome $S_0, S_1, \ldots, S_n \ldots$ dont les règles sont

$$G_0 : \{ S_0 \to b \},$$
$$G_1 : \{ S_0 \to b,\ S_1 \to aS_1 S_0 + S_0 \},$$
$$G_n = \{ S_0 \to b,\ \ldots,\ S_i \to a S_i S_{i-1} + S_{i-1},\ \ldots,\ S_n \to a S_n S_{n-1} + S_{n-1} \}.$$

Cette suite de langages admet une limite « naturelle » (qui n'est autre que l'union infinie des langages L_n) engendrée par $G_\infty : \{ S \to aSS + b \}$. Est-il vrai que si le langage A domine tous les langages L_n, il domine L_∞? Notons que ce problème particulier est très voisin de la conjecture 8 ci-dessus.

Notons aussi que ce cas particulier cache une partie de la difficulté en ce sens que la limite « naturelle » est ici algébrique ce qui n'est pas nécessairement le cas. Nous l'avons cependant choisi car c'est lui qui a permis pour la première fois d'établir l'inclusion stricte des langages quasi-rationnels dans la famille des langages algébriques [44, 57]. De plus, c'est le plus souvent l'existence de telles limites qui assurent la décidabilité de bon nombre de problèmes. On pourrait par exemple traiter de ce point de vue les algorithmes de Moore [32, 1] ou même de Stearns [52].

IV. Divers

Cette dernière section est consacrée à quelques problèmes ouverts dont les énoncés sont, soit plus particuliers, soit moins clairement reliés aux grands axes présentés ci-dessus. Cependant, ils ont été choisis parce que d'un point de vue technique, ils nous semblent révéler des difficultés de même nature que celles que l'on peut pressentir en travaillant sur les questions proposées jusqu'ici. Enfin, il ne faudrait pas croire que ces problèmes soient plus mineurs ou plus faciles : il n'en est rien.

Nous commencerons par quelques questions concernant les langages ambigus. Celles-ci, quoique sans rapports directs avec la conjecture 9, offrent le même type d'objet à étudier.

CONJECTURE 10 : Un langage algébrique d'ambiguïté bornée k peut toujours être décrit comme union de k langages non ambigus.

Notons que l'on sait d'une part qu'un tel langage est image dans une transduction rationnelle d'image finie (bornée par k) sur un langage déterministe (et non sur le monoïde libre tout entier) et d'autre part qu'une transduction rationnelle d'image finie bornée par k sur le monoïde libre tout entier est une union de k transductions rationnelles univoques. Le premier résultat n'est qu'une formulation particulière du théorème de Chomsky-Schützenberger [19] ou même de résultats de McNaugton [43]. Le second est un résultat récent de Schützenberger [50].

Toujours concernant les problèmes d'ambiguïté, il en est un plus particulier éclairant un aspect de la conjecture 1; reprenant le langage G de la remarque 2.

CONJECTURE 11 : Le langage algébrique G est ambigu.

Notons que les preuves classiques d'ambiguïté utilisant le lemme d'Ogden semblent ici particulièrement inadaptées et ce, parce que ce langage est presque rationnel ! On devra même noter que les preuves ainsi données ne sont le plus souvent adaptées qu'aux relations de type égalité entre exposant et qu'on ne sait plus les reproduire pour des inégalités : ainsi, par exemple, on sait que $\{a^n b^p c^q \mid n=p \text{ ou } p=q\}$ est ambigu. On ne le sait pas pour $\{a^n b^p c^q \mid n \neq p \text{ ou } p \neq q\}$!

Ce même lemme d'Odgen, utilisé cette fois pour prouver qu'un langage donné n'est pas algébrique, échoue de la même façon et pour les mêmes raisons quand on veut montrer :

CONJECTURE 12 : Le langage C_2 des mots contenant un carré n'est pas algébrique. (Ce langage vérifie la condition (P_+) de la conjecture 1 et n'est pas rationnel.)

On peut penser que devant ce genre de langages, il serait utile de savoir quelque chose du complémentaire d'un langage algébrique; ainsi, par exemple, on peut se poser la :

QUESTION 5 : Existe-t-il un langage algébrique L non rationnel dont le complémentaire ne contienne aucun carré ($L \supseteq C_2$) ?

Divers problèmes ouverts traitent, non pas du complémentaire, mais de l'intersection. Ainsi, parmi les plus anciens (bien que jamais publié) :

CONJECTURE 13 : Si deux langages quasi-rationnels ont une intersection algébrique, celle-ci est encore quasi-rationnelle.

L'intersection des langages est mal connue; celle des familles ne l'est guère mieux. Certaines questions de cette nature apparaissent de prime abord comme assez simples et pourtant elles n'ont pas reçu de réponse. Ainsi, si l'on désigne par Lin et Oct les familles des langages linéaires et à un compteur, on remarque facilement que tous les langages « connus » dans Lin$\cap$Oct sont dans le cône engendré par $S_1 = \{a^n b^n \mid n \geqq 0\}$, et pourtant on ne sait pas résoudre la :

CONJECTURE 14 : Lin$\cap$Oct est le cône rationnel engendré par S_1.

Enfin, toujours parmi les problèmes faisant intervenir l'intersection, rappelons que la famille BNP est définie comme la famille des langages qui sont images dans un morphisme continu (= « n'effaçant pas » ou « ε-free ») de l'intersection de langages linéaires. Book-Nivat-Paterson [16] ont montré que l'on peut toujours se borner alors à l'intersection de trois langages linéaires. On sait aussi que si l'homomorphisme est quelconque, on trouve tous les langages récursivement énumérables. Cependant, on ne sait pas si :

CONJECTURE 15 :

a) La famille BNP est strictement contenue dans celle des langages récursivement énumérables.

b) Les langages de Dyck restreint sur une lettre $D_1'^*$ et S_1^* ne sont pas dans la famille BNP.

BIBLIOGRAPHIE

1. A. V. Aho et J. D. Ullman, *The Theory of Parsing*, Translation and Compiling, vol. 1, Prentice Hall, 1972.
2. J. M. Autebert, *Opérations de cylindre et applications séquentielles gauches inverses*, Acta Informatica, vol. 11, 1979, p. 241-258.
3. J. M. Autebert, L. Boasson et G. Cousineau, *A Note on 1-Locally-Linear Languages*, Information and Control, vol. 37, 1978, p. 1-4.
4. Y. Bar-Hillel, M. Perles et E. Shamir, *On Formal Properties of Simple Phrase Structure Grammars*, Z. Phonetik., Sprach. Kommunikation Forsch., vol. 14, 1961, p. 143-172.
5. J. Beauquier, *Générateurs algébriques et systèmes de paires itérantes*, Theoretical Computer Sc., vol. 8, 1979, p. 293-323.
6. J. Beauquier, *A Remark About the Syntactic Lemma*, soumis à Mathematical Systems Theory.
7. L. Boasson, *Un critère de rationalité des langages algébriques*. In Automata, Programming and Languages, M. Nivat, éd., North Holland, 1972, p. 359-365.
8. L. Boasson, *Two Iteration Theorems for Some Families of Languages*, J. Comput. System Sc., vol. 7, 1973, p. 583-596.
9. L. Boasson, *Langages algébriques, paires itérantes et transductions rationnelles*, Theoretical Computer Sc., vol. 2, 1976, p. 209-223.
10. L. Boasson, B. Courcelle et M. Nivat, *A New Complexity Measure for Languages*, *in* A Conference on Theoretical Computer Science, Waterloo, 1977, p. 130-138.
11. L. Boasson, J. P. Crestin et M. Nivat, *Familles de langages translatables et fermées par crochet*, Acta Informatica, vol. 2, 1973, p. 383-393.
12. L. Boasson et M. Nivat, *Sur diverses familles de langages fermées par transductions rationnelles*, Acta Informatica, vol. 2, 1973, p. 180-188.
13. L. Boasson et M. Nivat, *Le cylindre des langages linéaires*, Mathematical Systems Theory, vol. 11, 1977, p. 147-155.
14. L. Boasson et S. Horvath, *On Languages Satisfying Ogden's Lemma*, R.A.I.R.O., Informatique Théorique, vol. 12, 1978, p. 201-202.
15. R. V. Book et S. A. Greibach, *Quasi Realtime Languages*, Mathematical Systems Theory, vol. 4, 1970, p. 97-111.
16. R. V. Book, M. Nivat et M. Paterson, *Reversal Bounded Acceptors and Intersection of Linear Languages*, S.I.A.M. J. Comput., vol. 3, 1974, p. 283-297.
17. W. J. Chandler, *Abstract Families of Deterministic Languages*, Proceedings du 1er A.C.M. Symposium of Theory on Computing, Marina del Rey, 1969, p. 21-30.
18. N. Chomsky, *Context-Free Grammars and Push-Down Storage*, M.I.T. Res. Lab. Electron. Quart. Prog. Rep., vol. 65, 1962.
19. N. Chomsky et M. P. Schützenberger, *The Algebraic Theory of Context-Free Languages*, *in* Computer Programming and Formal Systems, North Holland, 1963, p. 118-161.
20. S. Eilenberg, Communication au congrès international des mathématiciens, Nice, 1970.

21. S. EILENBERG, *Automata, Languages and Machines*, vol. A, Academic Press, New York, 1974.

22. C. C. ELGOT et J. F. MEZEI, *On Relations Defined by Generalized Finite Automata*, I.B.M. J. Res. Dev., vol. 9, 1962, p. 47-68.

23. R. J. EVEY, *The Theory and Application of Push-Down Store Machines*, Mathematical Linguistics and Automatic Translation, Harvard University, Computation Lab. Rep., N.S.F. 10, mai 1963.

24. P. C. FISCHER, A. R. MEYER et A. L. ROSENBERG, *Counter Machines and Counter Languages*, Mathematical Systems Theory, vol. 2, 1968, p. 265-283.

25. R. W. FLOYD, *On Ambiguity in Phrase-Structure Languages*, Comm. Assoc. Comput. Mach., vol. 5, 1962, p. 526-534.

26. S. GINSBURG, *Algebraic and Automata-Theoretic Properties of Formal Languages*, North Holland, 1975.

27. S. GINSBURG, J. GOLDSTINE et S. A. GREIBACH, *Uniformly Erasable AFL*, J. Comput. System Sc., vol. 10, 1975, p. 165-182.

28. S. GINSBURG, J. GOLDSTINE et S. A. GREIBACH, *Some Uniformly Erasable Families of Languages*, Theoretical Computer Science, vol. 2, 1976, p. 29-44.

29. S. GINSBURG et S. A. GREIBACH, *Deterministic Context Free Languages*, Information and Control, vol. 9, 1966, p. 620-648.

30. S. GINSBURG et S. A. GREIBACH, *Abstract Families of Languages*, *in* Memoirs of the Amer. Math. Soc., vol. 87, 1969, p. 1-32.

31. S. GINSBURG et E. H. SPANIER, *Derivation-Bounded Languages*, J. Comp. Syst. Sc., vol. 2, 1968, p. 228-250.

32. A. GINZBURG, *Algebraic Theory of Automata*, Academic Press, New York, 1968.

33. J. GOLDSTINE, *Substitution and Bounded Languages*, J. Comput. System Sc., vol. 6, 1972, p. 9-29.

34. S. A. GREIBACH, *Chains of Full AFL's*, Mathematical Systems Theory, vol. 4, 1970, p. 231-242.

35. S. A. GREIBACH, *The Hardest Context Free Language*, S.I.A.M. J. Comput., vol. 2, 1973, p. 304-310.

36. S. A. GREIBACH, *Jump PDA's and Hierarchies of Deterministic Context-Free Languages*, S.I.A.M. J. Comput., vol. 3, 1974, p. 111-127.

37. I. HAVEL et M. HARRISON, *Strict Deterministic Grammars*, J. Comput. System Sc., vol. 7, 1973, p. 237-277.

38. J. E. HOPCROFT et A. J. KORENJAK, *Simple Deterministic Languages*, I.E.E.E. Conf. Rec. 7th Ann. Symp. Switching and Automata Theory, 1966, p. 36-46.

39. S. HORVÁTH, *The Family of Languages Satisfying Bar Hillel's Lemma*, R.A.I.R.O.-Informatique théorique, vol. 12, 1978, p. 192-200.

40. D. E. KNUTH, *A Characterisation of Parenthesis Languages*, Information and Control, vol. 11, 1967, p. 269-289.

41. P. LANDWEBER, *Three Theorems on Phrase-Structure Grammars of Type* 1, Information and Control, vol. 6, 1963.

42. M. LATTEUX, *Langages commutatifs*, Thèse Sc. Math, Université Lille-I, 1978.

43. R. MACNAUGTHON, *Parenthesis Grammars*, J. Assoc. Comput. Mach., vol. 14, 1967, p. 490-500.

44. M. NIVAT, *Transductions des langages de Chomsky*, Thèse Sc. Math., Paris, 1967.

45. W. OGDEN, *A Helpful Result for Proving Inherent Ambiguity*, Mathematical Systems Theory, vol. 2, 1967, p. 191-194.

46. R. J. PARIKH, *On Context-Free Languages*, J. Assoc. Comput. Mach., vol. 13, 1968, p. 570-580.

47. J. F. PERROT, *Introduction aux monoïdes syntactiques des langages algébriques*, *in* Langages Algébriques, J.P. CRESTIN et M. NIVAT, éds., 1973, p. 167-222.

48. A. SALOMAA, *Formal Languages*, Academic Press, New York, 1973.

49. M. P. SCHÜTZENBERGER, *Finite Counting Automata*, Information and Control, vol. 5, 1962, p. 91-107.

50. M. P. SCHÜTZENBERGER, *Sur les relations rationnelles entre monoïdes libres*, Theoretical Computer Sc., vol. 3, 1976, p. 243-259.

51. E. SHAMIR, *A Representation Theorem for Algebraic and Context Free Power Series in Non-Commuting Variables*, Information and Control, vol. 11, 1967, p. 239-254.

52. R. E. STEARNS, *A Regularity Test for Push-Down Machines*, Information and Control, vol. 11, 1967, p. 323-340.

53. I. H. SUDBOROUGH, *A Note on Tape-Bounded Complexity Classes and Linear Context-Free Languages*, J. Assoc. Comput. Mach., vol. 22, 1975, p. 499-500.

54. L. VALIANT, *Regularity and Related Problems for Deterministic Push-Down Automata*, J. Assoc. Comput. Mach., vol. 22, 1975, p. 1-10.

55. L. VALIANT, *General Context-Free Recognition in Less than Cubic Time*, J. Comput. System Sc., vol. 10, 1975, p. 308-315.

56. A. P. J. VAN DER WALT, *Locally-Linear Families of Languages*, Information and Control, vol. 32, 1976, p. 27-32.

57. N. K. YNTEMA, *Inclusion Relations Among Families of Context-Free Languages*, Information and Control, vol. 10, 1967, p. 572-597.

58. A. B. CREMERS et S. GINSBURG, *Context Free Grammars Forms*, *in* Automata, Languages and programming, 2nd I.C.A.L.P., Saarbrücken, 1974, Lecture Notes in Comput. Sc., n° 14, p. 364-382.