Le lemme de Parikh via les grammaires de graphes

DINH Trong Hieu
Institut de la Francophonie pour l'Informatique - IFI
Hanoï, Vietnam.
Email: dthieu@ifi.edu.vn

***Résumé* - Le lemme de Parikh nous donne les conditions nécessaires pour la distribution des lettres dans les mots du langage. Dans cet article, on présente une démonstration géométrique du lemme de Parikh via les grammaires de graphes.**

***Mots cles* - Langages formels, grammaire de graphes, lemme de Parikh.**

I. INTRODUCTION

Depuis les travaux initiaux de Chomsky, bien des résultats ont été établis sur les langages algébriques et de nombreux ouvrages de référence existent sur le sujet (voir entre autres [3], [5]).

Depuis 1985, Muller et Schupp [6] ont apporté une vision géométrique des langages algébriques. Plus exactement, ils ont montré l'identité effective entre un automate à pile et une grammaire déterministe de graphes : tout automate à pile peut être transformé en une grammaire déterministe de graphes engendrant le graphe des transitions de l'automate, et inversement toute grammaire déterministe de graphes peut être transformée en un automate à pile dont le graphe des transitions est engendré par la grammaire. Le fait de pouvoir décrire, à l'aide d'une grammaire de graphes, la structure géométrique d'un ensemble algébrique de mots, n'est pas anodin. Bien des résultats classiques sur les langages algébriques deviennent naturels si on les aborde via les grammaires de graphes. Par exemple, la détermination des configurations accessibles à partir d'une configuration donnée se traduit par le calcul d'un plus petit point fixe sur la grammaire de graphes. Le fameux lemme des paires itérantes s'établit simplement et géométriquement par le biais de grammaires de graphes.

Le lemme de Parikh a été introduit depuis 1966 [7]. Il a attiré beaucoup d'attentions, et depuis presque quarante ans, différentes approches pour le montrer ont été présentées (voir [4] et [1]). Dans ce papier, on présente une démonstration géométrique du lemme de Parikh.

II. GRAMMAIRE ALGÉBRIQUE ET LEMME DE PARIKH

Une *grammaire algébrique* (ou une *grammaire hors contexte*)est un triplet $G = (T, N, P)$ où

- T est un alphabet de *lettres terminales*,
- N est un alphabet disjoint de T de *lettres non-terminales* ou *variables*,
- P est un sous-ensemble fini de $N \times (N \cup T)^*$ de *règles* ou de *productions*.

Pour tout $(S, m) \in P$, S est le *membre gauche* et $m \in (T \cup N)^*$ est le *membre droit* de cette règle, et la règle est écrite sous la forme $S \to m$.

Soit $G = (T, N, P)$ une grammaire algébrique. Un mot $u \in (T \cup N)^*$ *se récrit* en un mot $v \in (T \cup N)^*$ si $u = u_1.S.u_2$, $v = u_1.m.u_2$ et $(S, m) \in P$, et on note $u \to v$. La fermeture réflexive et transitive, $\to^*$ de la récriture $\to$ est l'opération de *dérivation*:

$u \to^* v$ si $\exists n \geq 0, \exists u_0, u_1, ..., u_n, u_0 \to u_1 ... \to u_n$ avec $u_0 = u$ et $v_0 = v$.

Le langage engendré par $G = (T, N, P)$ en partant de *l'axiome* $S \in N$ est le langage $L_G(S)$ des mots terminaux dérivant de S:

$$L_G(S) = \{u \in T^* | S \to^* u\}.$$

Les langages engendrés par les grammaires algébriques sont appelés les *langages algébriques*. La famille des langages algébriques sur l'alphabet T est notée $Alg(T^*)$. Ce sont aussi les langages reconnus par les *automates à pile* (voir [3] ou [5]).

Un ensemble de la forme

$$\{\alpha_0 + k_1\alpha_1 + ... + k_n\alpha_n | k_1, ..., k_n \in \mathbb{N}\}$$

où $\alpha_0, ..., \alpha_n$ sont les élément de $\mathbb{N}^m$, est dit un sous-ensemble *linéaire* de $\mathbb{N}^m$. Un *ensemble semi-linénaire* est une union finie d'ensembles linéaires.

Comme N^m est pour $+$ un monoïde commutatif, ses ensembles linéaire sont ses parties rationnelles. De plus, ils forment une algebre de Boole [2] : la famille des ensembles semi-linéaire de N^m et fermée par union, intersection et complémentation.

Etant donné un alphabet $T = \{a_1, ..., a_m\}$, nous définissons l'application ψ de T^* dans $\mathbb{N}^m$ comme suit:

$$\psi(w) = (|w|_{a_1}, ..., |w|_{a_m})$$

où $|w|_{a_i}$ désigne nombre d'occurrences du caractère a_i dans le mot w.

On dit que $\psi(w)$ est l'image de Parikh de w.

Par exemple, avec $T = \{a, b, c\}$, on a: $\psi(abacac) = (3, 1, 2)$.

Soient $x, y \in T^*$ deux mots quelconques, nous avons

$$\begin{aligned} \psi(x.y) &= \psi(x) + \psi(y) \\ &= \psi(y) + \psi(x) = \psi(y.x) \end{aligned}$$

Ainsi ψ est un morphisme de T^* dans $\mathbb{N}^m$.

L'image de Parikh de tout langage $L \subseteq T^*$ est définie par union:

$$\psi(L) = \{\psi(x) | x \in L\}$$

Comme la rationalité est préservée par morphisme, on a $\psi(L)$ rationnel pour tout langage rationnel L.

Par exemple, le langage $L = \{a^n b^{2n+1} | n \geq 0\}$ sur $\{a, b\}$ a l'image de Parikh $\psi(L) = \{(0,1) + k.(1,2) | k \geq 0\}$ qui est aussi ensemble semi-linéaire (et même linéaire) de $\mathbb{N}^2$. Cette propriété se généralise à tout langage algébrique [7].

Lemme de Parikh : *L'image de Parikh $\psi(L)$ de tout langage algébrique L est de façon effective*[1] *un ensemble semi-linéaire.*

Le lemme de Parikh donne les conditions nécessaires de la distribution des lettres dans les mots du langage.

III. Grammaire de graphes

On nomme *graphe* un couple (V, E) où V est un ensemble quelconque et E est une partie de $V \times T \times V$ (produit cartésien d'ensembles).

Les éléments de V sont appelés les *sommets*, et T est l'ensemble des étiquettes.

Tout élément (s, a, t) de E est appelé un *arc* du graphe de *source* s, de *but* t et d'*étiquette* a, et est aussi noté $s \xrightarrow{a} t$.

Un *chemin* d'un sommet s à un sommet t dans un graphe est une suite $s_0 \xrightarrow{a_1} s_1 ... \xrightarrow{a_n} s_n$ consécutive d'arcs avec $s_0 = s$ et $s = t$.

Maintenant, on généralise la notion d'arc, dont le nombre de sommets peut être différent de deux: un *hyperarc* est un élément $as_1...s_n$ de TV^* d'étiquette a et de sommets ordonnés $s_1, ..., s_n$. Le nombre n de sommets d'un hyperarc est appelé son *arité*.

On nomme *hypergraphe* un couple (V, E) où V est un ensemble quelconque de sommets et E est un ensemble d'hyperarcs: $E \subseteq TV^*$.

Comme pour une grammaire algébrique (de mots), une *grammaire de graphes* est un quadruplet $R = (T, N, V, P)$ où

- T est un alphabet d'étiquettes *terminales*,
- N est un alphabet d'étiquettes *non-terminales* ou *variables*,
- V est un ensemble quelconque de *sommets*,
- P est un sous-ensemble fini de *règles* (x, H) où le membre gauche $x \in NV^*$ est un hyperarc non-terminal et $H \subseteq NV^* \cup TVV$ est un hypergraphe fini d'hyperarcs non-terminaux et d'arcs terminaux.

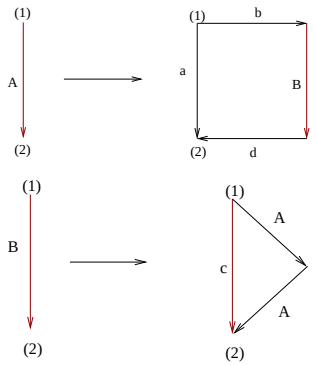

Fig. 1. Une grammaire déterministe de graphes.

[1]L'image $\psi(L)$ peut être trouvée par un algorithme déterministe

On note V_H l'ensemble des sommets d'un hypergraphe H.

Une grammaire de graphes est *déterministe* si deux membres gauches ont des étiquettes différentes.

La *récriture* $\xrightarrow[R]{}$ selon une grammaire R de graphes est un remplacement du membre gauche par le membre droit d'une règle de la grammaire:

$$G \xrightarrow[R]{} (G - \{Bs_1...s_n\}) \cup \{Cf(t_1)...f(t_m) | Ct_1...t_m \in H\}$$

où $(Bx_1...x_n, H) \in R$ et f est une injection de V_H dans $(V - V_B) \cup \{s_1, ..., s_n\}$ avec $f(x_i) = s_i$.

Par exemple, à partir une hypergraphe G de la figure 2 :

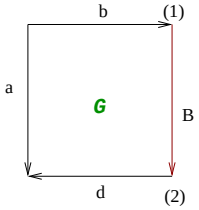

Fig. 2. Un hypergraphe G.

on a une récriture de G selon la grammaire de graphe R de la figure 1 comme indiqué dans la figure 3 :

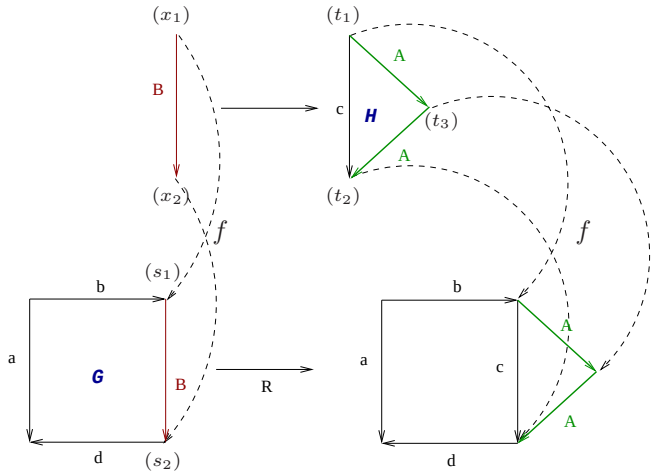

Fig. 3. Une récriture selon la grammaire de graphes de la figure 1.

A partir d'un axiome $A \in N.V^*$ d'une grammaire déterministe de graphes $R = (T, N, V, P)$, on remplace parallèlement tous les hyperarcs non-terminaux par leurs membres droits dans P. Cette récriture parallèle est répétée itérativement. Cela s'arrête quand tous les hyperarcs non-terminaux sont remplacés et il n'y a plus que les arcs terminaux dans le graphe obtenu Tr, ou bien la répétition est infinie, dans ce deuxième cas, le graphe engendré Tr est infini. On dit que Tr est *engendré* par la grammaire de graphes R à partir de A.

Un graphe Tr qui est engendré par une grammaire de graphes est dit *graphe régulier*.

On suppose que l'hyperarc axiome A possède un sommet de départ $A_{départ}$ et un sommet d'arrivée A_{fin}. Si β est

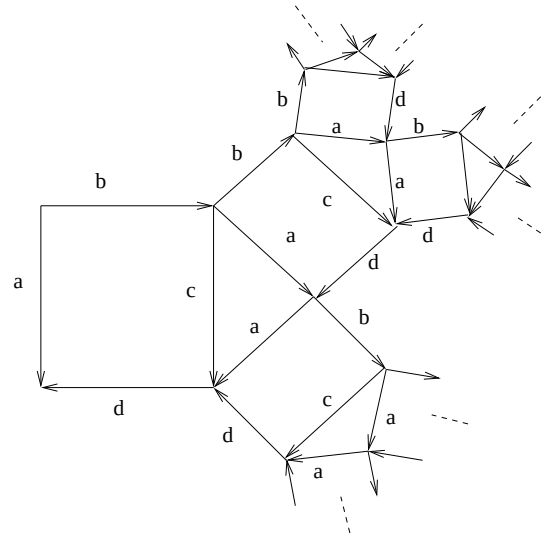

Fig. 4. Le graphe des transitions engendré par la grammaire de graphes de la figure 1.

un chemin de $A_{départ}$ à A_{fin} sur Tr, β est dit un *chemin engendré* par R à partir de A. Toute suite u des étiquettes des arcs de β est un mot reconnu par R. L'ensemble des mots reconnus par R à partir de l'axiome A est dit le *langage reconnu* par R à partir de A et est noté $L(A)$.

Il est aisé de vérifier que tout langage $L(A)$ reconnu par un graphe régulier Tr est algébrique.

Pour simplifier, tout chemin β d'un graphe des transitions Tr est un chemin qui commence par le sommet de départ et termine par le sommet d'arrivée de l'axiome.

On appelle Tr_i la nouvelle partie de Tr qui est ajoutée pour la $i^{ième}$ récriture parallèle. Donc, Tr_0 est réduit aux sommets de l'axiome A.

Remarquons que les sommets de sortie des Tr_i sont les sommets d'entrée de Tr_{i+1}: on les appelle les *sommets frontières* dont l'ensemble est noté F_i. En plus, tout chemin u d'un sommet x de Tr_i à un sommet y de Tr_{i+1} doit passer par la frontière F_i, c'est-à-dire qu'il comprend au moins un sommet de l'ensemble F_i.

Par conséquent, tout chemin d'un sommet x de Tr_i à un sommet y de Tr_j avec $i < j$ doit passer par toutes les frontières intermédiaires F_t pour $t \in [i, j-1]$.

A toute grammaire algébrique G, on peut associer de façon effective une grammaire déterministe de graphes R engendrant le graphe des transitions $Tr(G)$ de G.

Proposition : *On peut transformer toute grammaire algébrique $G = (T, N, P)$ en une grammaire de graphes $R = (T, N, V, P')$ où $P' = \{(A_{1,2}, H_A)|A \in N\}$ de sorte que pour tout nonterminal $A \in N$, $L_G(A) = L_R(A)$.*

On va montrer que la grammaire de graphes $R = (T, N, V, P')$ peut être construire effectivement a partir d'une grammaire algébrique $G = (T, N, P)$ quelconque.

Pour tout $A \in N$, on définit:

$$\begin{aligned} H_A \quad := \quad & \{1 \xrightarrow{a} a.v | (A, av) \in P \wedge a \in N \cup T \wedge v \neq \epsilon\} \\ \cup \quad & \{u.a \xrightarrow{a} 2 | (A, ua) \in P \wedge a \in N \cup T \wedge u \neq \epsilon\} \\ \cup \quad & \{1 \xrightarrow{a} 2 | (A, a) \in P \wedge a \in N \cup T\} \\ \cup \quad & \{u.av \xrightarrow{a} ua.v | (A, uav) \in P \wedge a \in N \cup T \\ & \wedge u, v \neq \epsilon\} \end{aligned}$$

et l'ensemble de règles P' est défini :

$$P' := \{A \rightarrow H_A | A \in N\}.$$

L'ensemble des sommets V est défini :

$$V := \{u.v | uv \in Im(P) \wedge u, v \neq \epsilon\} \cup \{1, 2\}.$$

Par exemple, on considère la grammaire algébrique $G = (T, N, P)$ où $T = \{a, b, c, d\}$, $N = \{A, B\}$ et les règles sont:

$$\begin{aligned} A &\rightarrow Bb \\ A &\rightarrow aAB \\ B &\rightarrow c \\ B &\rightarrow BBb \end{aligned}$$

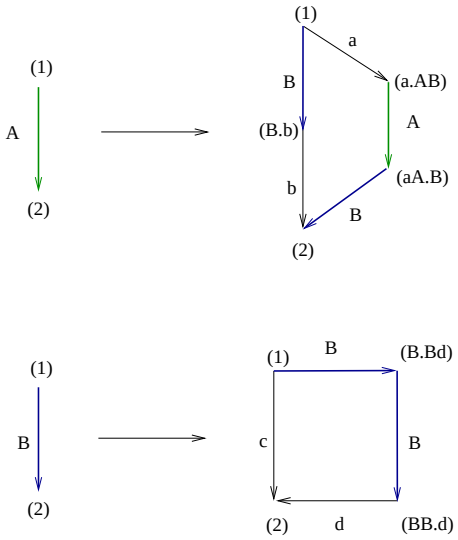

Fig. 5. L'ensemble P' des règles pour un grammaire de graphe R.

IV. La démonstration du lemme de Parikh via les grammaires de graphes

L'ensemble des étiquettes non-terminaux récrits pour engendrer un chemin β est noté $\Lambda(\beta)$.

Soit F un ensemble de chemins engendrés par une grammaire de graphes R . *Une réduction* d'un chemin α de F est l'opération dont on supprime une partie pour obtenir un autre chemin α' de F, comme indiqué à la figure 7.

On dit que α *se réduit* en (α', β_B), et on note:

$$\alpha \xrightarrow{réduire} (\alpha', \beta_B);$$

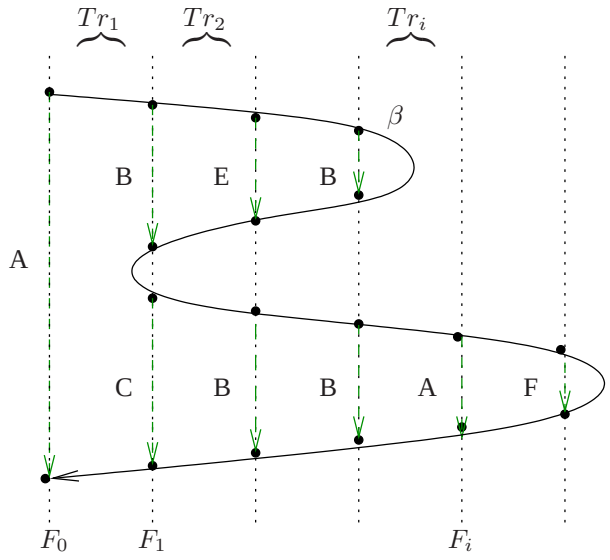

Fig. 6. $\Lambda(\beta) = \{A, B, C, E, F\}$.

dans le cas où β_B n'est pas pertinent, on peut l'ignorer et écrire simplement $\alpha \xrightarrow{réduire} \alpha'$. On dit que α est *réductible* selon l'ensemble F. Si α ne peut pas se réduire, on dit que α est *irréductible* (selon F).

Un ensemble F de chemins est dit *réductible* s'il contient au moins un chemin réductible. Sinon, F est dit *irréductible*.

Remarque Tout chemin irréductible engendré par une grammaire de graphes R est de longueur bornée.. En conséquence, tout ensemble irréductible de chemins engendrés par une grammaire de graphes est fini.

Dorénavant, on suppose que pour toute grammaire de graphes $R = (T, N, V, P)$, on a $N = \{B_1, ..., B_{|N|}\}$.

Pour tout sous-ensemble N' de N, on considère le graphe $Tr_{N'}$ qui ne contient que les chemins α qui satisfont la condition:

$$\Lambda(\alpha) = N'$$

Le langage reconnu par $Tr_{N'}$ est noté $L_{N'}$. On a :

$$L = \bigcup_{N' \subseteq N} L_{N'}$$

Évidemment, si l'axiome $A \notin N'$, $L_{N'}$ est vide. Alors

$$L = \bigcup_{N' \subseteq N, A \in N'} L_{N'}$$

En conséquence :

$$\psi(L) = \bigcup_{N' \subseteq N, A \in N'} \psi(L_{N'})$$

Comme N est fini, c'est une union finie.

Sans perte de généralité, nous pouvons restreindre N' à N.

Pour chaque non-terminal B_i, nous définissions un ensemble noté R_{B_i} des chemins β_{B_i} engendrés par R à partir de B_i et qui satisfont les conditions suivantes :

- β ne contient qu'un seul B_i comme un non-terminal qui n'est pas encore remplacé;

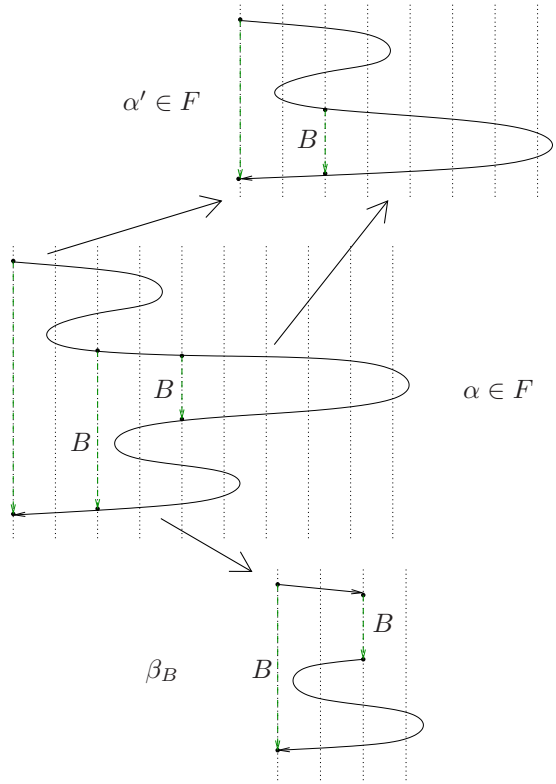

Fig. 7. Une réduction d'un chemin en un autre en retirant une répétition.

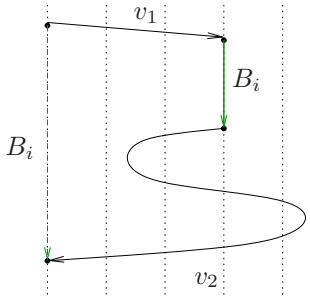

Fig. 8. Un membre β_{B_i} de l'ensemble R_{B_i}.

- β est irréductible.

C'est-à-dire, tout $\beta_{B_i} \in R_{B_i}$ a la suite correspondante u des étiquettes dans la forme $u = v_1 B_i v_2$ où $v_1 \in T^*, v_2 \in T^*$.

On note L_{B_i} le langage:

$$L_{B_i} = \{v_1 v_2 | v_1 B_i v_2 \in R_{B_i}\}.$$

Nous définissons un ensemble Ts de chemins α engendrés par R à partir de l'axiome A, de sorte que:

- tous les non-terminaux sont présents dans $\Lambda(\alpha)$;
- α est irréductible selon Ts.

Propriété d'un chemin α dans Ts:

Toute répétition contient au moins un non-terminal qui n'apparait pas ailleurs. On note L_A le langage reconnu par le graphe Ts.

Avec les conditions ci-dessus, les L_{B_i} et L_A sont des langages finis.

Nous allons montrer que:

$$\psi(L_N) = \psi(L_A.L^*_{B_1}.L^*_{B_2}...L^*_{B_{|N|}}) \quad (1)$$

Nous montrons d'abord que :

$$\psi(L_A.L^*_{B_1}.L^*_{B_2}...L^*_{B_{|N|}}) \subseteq \psi(L_N) \quad (2)$$

et puis :

$$\psi(L_N) \subseteq \psi(L_A.L^*_{B_1}.L^*_{B_2}...L^*_{B_{|N|}}). \quad (3)$$

i) Vérifions la propriété (2):
Tous d'abord, nous avons $L_A \subseteq L_N$, donc:

$$\psi(L_A) \subseteq \psi(L_N) \quad (4)$$

En plus, soient β un chemin dans R_{B_i} ($B_i \in N$), α un chemin dans Tr_N et $v \in L_{B_i}$ et $u \in L_N$ sont les mots correspendants. Par la définition de Tr_N, $\Lambda(\alpha) = N$, c'est-à-dire $B_i \in \Lambda(\alpha)$.

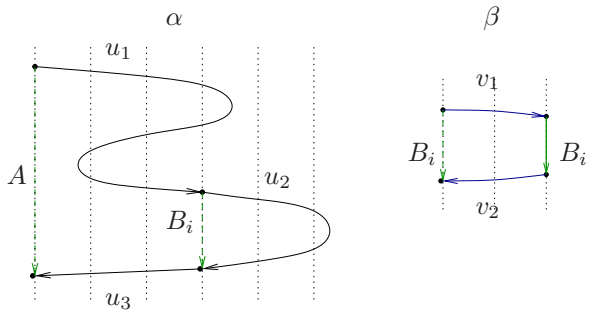

Fig. 9. $\alpha \in Ts$ et $\beta \in R_{B_i}$.

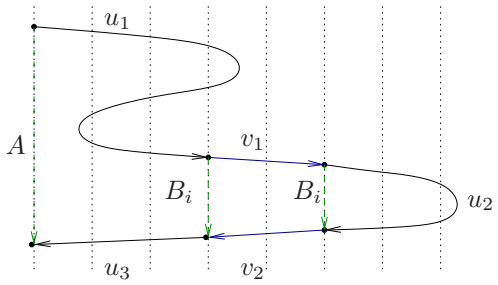

Fig. 10. α' obtenu à partir de α par insertion de β.

Nous avons $v = v_1.B_i.v_2$ et $u = u_1.u_2.u_3$.
Donc, nous pouvons construire un chemin α' obtenu à partir de α par insertion de β comme indique dans la figure 10.
On appelle w le mot reconnu par le chemin α'.
Alors $w = u_1.v_1.u_2.v_2.u_3$.
Ainsi

$$\begin{aligned}\Lambda(\alpha') &= \Lambda(\alpha) \cup \Lambda(\beta) \\ &= N \cup \Lambda(\beta) = N.\end{aligned}$$

Alors $\alpha' \in Tr_N$, c'est-à-dire $w \in L_N$

Donc,

$$\begin{aligned}\psi(uv) &= \psi(u_1.u_2.u_3.v_1.v_2) \\ &= \psi(u_1.v_1.u_2.v_2.u_3) \\ &= \psi(w) \in \psi(L_N)\end{aligned}$$

Alors $\forall u \in L_N, v \in L_{B_i}, \forall B_i \in N$, on a :

$$\psi(uv) \in \psi(L_N).$$

Donc

$$\psi(L_N.L_{B_i}) \subseteq \psi(L_N), \forall B_i \in N. \quad (5)$$

En conséquence,

$$\psi(L_N.L^*_{B_1}.L^*_{B_2}...L_{B_{|N|}}) \subseteq \psi(L_N).$$

Avec (4), nous obtenons (2):

$$\psi(L_A.L^*_{B_1}.L^*_{B_2}...L^*_{B_{|N|}}) \subseteq \psi(L_N).$$

ii) Vérifions l'inclusion (3): Soient u un mot quelconque dans L_N, et $\alpha \in Tr_N$ le chemin correspondant à u.
Il n'y a que deux possibilités.

Soit α est irréductible, alors $\alpha \in Ts$ et

$$u \in L_A \quad (6)$$

Soit α est réductible ($\beta \notin Ts$).

Dans le deuxième cas, il existe au moins une répétition de remplacements dans α qu'on peut supprimer. Evidemment, on peut choisir toujours une répétition qui ne contient pas d'autre répétition dedans. On suppose que c'est la répétition du non-terminal B_i. Donc, α peut se réduire en (α_1, β_1):

$$\exists \alpha_1 \in Tr_N, \beta_1 \in R_{B_i} : \alpha \overset{réduire}{\longrightarrow} (\alpha_1, \beta_1)$$

Pour α_1, il y a aussi deux cas comme α. Soit $\alpha_1 \notin T_S$, ou soit α_1 est réductible et peut être réduit en une paire (α_2, β_2).
Parce que le mot u est fini, il existe une suite finie des paires $(\alpha_1, \beta_1), ..., (\alpha_k, \beta_k)$. Dont $\beta_p \in R_{B_{i_p}}$, avec$B_{i_p} \in N, 1 \leq p \leq k-1$ tel que:

$$\alpha_p \overset{réduire}{\longrightarrow} (\alpha_{p+1}, \beta_{p+1}), 1 \leq p \leq k-1$$

et

$$\alpha_k \in T_S.$$

Par définition de $R_{B_{i_p}}$, le mot u_p correspondant au chemin $\beta_{B_{i_p}}$ dans la forme $u_p = u_{p,1}.B_{i_p}.u_{p,2}$, avec $u_{p,1}, u_{p,2} \in T^*$, pour $1 \leq p \leq k-1$. Le mot correspondant à β_k est noté v.
Nous avons:

$$u = u_{1,1}.u_{2,1}...u_{k-1,1}.v.u_{k-1,2}...u_{1,2}$$

et

$$\begin{aligned}\psi(u) &= \psi(u_{1,1}.u_{2,1}...u_{k-1,1}.v.u_{k-1,2}...u_{1,2}) \\ &= \psi(u_{1,1}.u_{1,2}...u_{k-1,1}.u_{k-1,2}.v) \\ &= \psi(u_{1,1}.u_{1,2}) + ... + \psi(u_{k-1,1}.u_{k-1,2}) + \psi(u_k).\end{aligned}$$

Mais, nous avons aussi:

$$\begin{aligned} \psi(u_{p,1}.u_{p,2}) &\in \psi(L_{B_{i_p}}) \\ \psi(u_k) &\in \psi(L_A). \end{aligned}$$

Donc, pour tout $u \in L_N$ nous avons

$$\psi(u) \in \psi(L_A.L^*_{B_1}.L^*_{B_2}...L^*_{B_{|N|}}).$$

En conséquence $\psi(L_N) \subseteq \psi(L_A.L^*_{B_1}.L^*_{B_2}...L^*_{B_{|N|}})$.
En définitive,

$$\psi(L_N) = \psi(L_A.L^*_{B_1}.L^*_{B_2}...L^*_{B_{|N|}})$$

Donc, $\psi(L_N)$ est semi-linéaire. Nous pouvons généraliser à tous les $L_{N'}$ qui sont semi-linéaires.
Nous avons le résultat:

$$\psi(L) = \bigcup_{N' \subseteq N, S \in N'} \psi(L_{N'})$$

est semi-linéaire. □

V. Conclusion

Cette démonstration géométrique du lemme de Parikh n'est qu'une première étape pour établir de façon géométrique via les grammaires de graphes bien des résultats connus sur les langages algébriques. On pourra ensuite utiliser les grammaires de graphes pour résoudre par point fixe de nouveaux problèmes sur les langages algébriques.

References

[1] L. Aceto, Z. Esik, A. Ingolfsdottir, *A fully equational proof of Parikh's theorem*, BRICS Report Series, 2001.
[2] S. Ginsburg, E. Spanier *Bounded Algol like languages*, Trans. Amer. Math. Soc. 113, pp. 333-365, 1964.
[3] M. Harrison, *Introduction to formal language theory*, Addison - Wesley, 1978.
[4] M. Hopkins, D. Kozen, *Parikh's theorem in commutative Kleene algebra*, Proc. IEEE Conf. Logic in Computer Science (LICS'99), IEEE Press, pp.394-401, 1999.
[5] J. Hopcroft, J. Ullman, *Introduction to automata theory, languages, and compilation*, Addison - Wesley, 1979.
[6] D. Muller, P. Schupp, *The theory of ends, pushdown automata, and second-order logic*, TCS 37, pp.51-75, 1985.
[7] R. Parikh, *On context-free languages*, J. Assoc. Comput. Mach., Vol. 13(4):570-581, 1966.