

C. PAIR

**Sur des notions algébriques liées à
l'analyse syntaxique**

Revue française d'informatique et de recherche opérationnelle. Série rouge, tome 4, n° R3 (1970), p. 3-29

http://www.numdam.org/item?id=M2AN_1970__4_3_3_0

© AFCET, 1970, tous droits réservés.

L'accès aux archives de la revue « Revue française d'informatique et de recherche opérationnelle. Série rouge » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

SUR DES NOTIONS ALGÈBRIQUES LIÉES À L'ANALYSE SYNTAXIQUE (1)

par C. PAIR (2)

Résumé. — On définit les principaux êtres mathématiques rencontrés dans l'analyse syntaxique des langages de Chomsky (« context-free »), en insistant particulièrement sur les notions de ramification et de binoïde universel. On montre que les méthodes habituelles d'analyse ramènent toutes à un même problème qui est de déterminer, pour le mot α analysé, l'ensemble $h^{-1}(\alpha) \cap K \cap P$ où h est un homomorphisme alphabétique, K un langage régulier et P un langage de parenthèses.

1. INTRODUCTION

Il existe dans la littérature de nombreuses méthodes d'analyse syntaxique pour les grammaires de Chomsky, dont certaines se ressemblent étrangement ; grossièrement, on peut distinguer des méthodes d'analyse descendante et des méthodes d'analyse ascendante. Mais on tente rarement de faire une théorie unifiée de l'analyse syntaxique, ni même de préciser les êtres mathématiques utilisés.

Lorsqu'on effectue l'analyse syntaxique d'un mot α pour une grammaire de Chomsky, on associe à α un certain nombre de *marqueurs de phrase*, qui sont les *structures arborescentes* de α , ou encore on parenthèse α , obtenant ainsi un certain nombre de mots d'un langage de parenthèses ou langage restreint de Dyck. Au cours de l'analyse, on construit une *pile* ou, plus précisément, on emploie un *transducteur à pile* dont la donnée est le mot α à analyser et dont les résultats caractérisent d'une certaine manière les marqueurs de phrase cherchés.

Une théorie de l'analyse syntaxique doit donc préciser ce que sont les êtres dont nous venons de dresser la liste et quelles relations ils ont entre eux. C'est

(1) Exposé fait le 10 mars 1969 au Centre d'Automatique de l'École des Mines, Fontainebleau.

(2) Institut Universitaire de Calcul Automatique, Nancy.

Pour pouvoir étudier le langage P , indiquons d'abord quelques propriétés simples de l'application τ .

Proposition 1 :

Soient α, α', β des mots sur W .

a) $\tau(\alpha) = \tau(\alpha')$ entraîne $\tau(\alpha\beta) = \tau(\alpha'\beta)$.

b) $\tau(\alpha) = 0$ entraîne $\tau(\alpha\beta) = 0$.

c) $\tau(\alpha) \neq 0$ et $\tau(\beta) \neq 0$ entraînent $\tau(\alpha\beta) = \tau(\alpha)\tau(\beta)$.

Les démonstrations sont immédiates par récurrence sur la longueur de β .

La partie (a) a pour cas particulier, avec $\alpha' = \wedge$:

$$\tau(\alpha) = \wedge \Rightarrow \tau(\alpha\beta) = \tau(\beta).$$

De la partie (c), on déduit que si $\alpha \in V^*$, alors $\tau(\alpha) = \alpha$. De la partie (b) résulte que tout mot non vide de P commence par une lettre appartenant à V .

Théorème 1 :

Soit α un mot non vide appartenant à P . Il existe $a \in V, \beta \in P, \gamma \in P$, uniques, tels que $\alpha = a\beta\bar{a}\gamma$.

Soit $a \in V$ la première lettre de α : $\alpha = a\alpha'$.

Si $\alpha = a\beta\bar{a}\gamma$, où $\beta \in P$, β est le plus grand facteur gauche de α' tel que $\tau(\beta) \neq 0$: en effet $\tau(\beta\bar{a}) = 0$ et il suffit d'appliquer la proposition 1, b. La décomposition envisagée est donc unique.

Réciproquement, soit β le plus grand facteur gauche de α' tel que $\tau(\beta) \neq 0$. Si β était égal à α' , d'après la proposition 1, c, on aurait $\tau(\alpha) = a\tau(\beta) \neq \wedge$. Donc $\beta \neq \alpha'$ et on peut écrire :

$$\alpha = a\beta x\gamma \quad \text{où} \quad x \in W \text{ et } \gamma \in W^*.$$

Par définition de β , $\tau(\beta x) = 0$, donc il existe $b \in V$ tel que $x = \bar{b}$ et que $\tau(\beta)$ ne soit pas terminé par b . Mais $\tau(a\beta x) \neq 0$, donc $\tau(a\beta) = a\tau(\beta)$ est terminé par b . Par suite $\tau(\beta) = \wedge, b = a$ et $\tau(a\beta x) = \wedge$. Il en résulte que $\tau(\gamma) = \tau(\alpha) = \wedge$, ce qui achève la démonstration.

EXEMPLES : Pour $\alpha = ab\bar{b}a\bar{b}\bar{b}$, $\beta = \gamma = b\bar{b}$; pour $\alpha = abb\bar{a}$, $\beta = b\bar{b}$, $\gamma = \wedge$.

La définition donnée de P est dissymétrique car elle procède de « gauche à droite ». Mais désignons par $\tilde{\alpha}$ le mot réfléchi, ou image-miroir, du mot α , et par $\bar{\alpha}$ le mot obtenu en remplaçant dans $\alpha \in W^*$ toute parenthèse ouvrante a par la parenthèse fermante \bar{a} et inversement ; plus précisément $\bar{\alpha}$ est l'image de α dans l'homomorphisme h défini par :

$$(\forall a \in V)(h(a) = \bar{a} \quad \text{et} \quad h(\bar{a}) = a);$$

Enfin, soit
$$c(\alpha) = \tilde{\bar{\alpha}} = \tilde{\tilde{\alpha}}$$

Par exemple, si $\alpha = ab\bar{b}\bar{a}\bar{b}\bar{b}$, $c(\alpha) = b\bar{b}ab\bar{b}\bar{a}$. $c(\alpha)$ s'appelle le *conjugué* du mot α et :

$$c(\alpha\beta) = c(\beta) c(\alpha).$$

Pour tout $\alpha \in W^*$, $c[c(\alpha)] = \alpha$; autrement dit, l'application c est une involution de W^* .

Proposition 2 :

L'application c est une involution de W^ qui laisse P invariant.*

Il suffit de montrer que si $\alpha \in P$, alors $c(\alpha) \in P$, ce que nous ferons par récurrence sur la longueur de α , à l'aide du théorème 1.

Si $\alpha = \wedge$, $c(\alpha) = \wedge \in P$. Si $\alpha \neq \wedge$, il s'écrit

$$\alpha = a\beta\bar{a}\gamma \quad \text{avec} \quad \beta \in P, \gamma \in P,$$

d'où

$$c(\alpha) = c(\gamma) a c(\beta) \bar{a}$$

et $c(\beta) \in P$, $c(\gamma) \in P$ par l'hypothèse de récurrence. D'après la proposition 1,

$$\tau[c(\alpha)] = \tau[ac(\beta)\bar{a}] = \wedge.$$

Du théorème 1 et de la proposition 2 résulte que, pour tout mot α non vide appartenant à P , existent $a \in V$, $\beta \in P$, $\gamma \in P$, uniques, tels que $\alpha = \beta a \gamma \bar{a}$.

3. NOTION DE PILE

Envisageons un mot de P ,

$$\alpha = x_1 x_2 \dots x_n$$

Pour vérifier que α appartient à P , on calcule successivement les transformés par τ des facteurs gauches de α . Si α_i désigne le facteur gauche de α dont la longueur est i ($0 \leq i \leq n$), on obtient $\tau(\alpha_i)$ soit en ajoutant à la fin de $\tau(\alpha_{i-1})$ une lettre de V , soit en enlevant la dernière lettre de $\tau(\alpha_{i-1})$. Nous dirons que la suite des mots $\tau(\alpha_i)$ est une pile.

EXEMPLE : $\alpha = ab\bar{b}\bar{a}\bar{b}\bar{b}$

$$\tau(\alpha_0) = \wedge ; \tau(\alpha_1) = a ; \tau(\alpha_2) = ab ; \tau(\alpha_3) = a ;$$

$$\tau(\alpha_4) = \wedge ; \tau(\alpha_5) = b ; \tau(\alpha_6) = \wedge.$$

Définition 2 [11] :

On appelle pile sur un ensemble V toute suite finie $\beta_0, \beta_1, \dots, \beta_n$ de mots sur V , telle que :

$$- \beta_0 = \beta_n = \wedge ;$$

$$- \text{pour } i = 1, \dots, n, \text{ il existe } a_i \in V \text{ tel que } \beta_i = \beta_{i-1} a_i \text{ ou } \beta_{i-1} = \beta_i a_i ;$$

dans le premier cas, on dit que i est une entrée de a_i , dans le second i est une sortie de a_i .

Ainsi, à tout mot $\alpha = x_1 \dots x_n$ appartenant à P est associée la pile $\tau(\alpha_0), \tau(\alpha_1), \dots, \tau(\alpha_n)$; i est une entrée de a dans la pile si, et seulement si, $x_i = a$; i est une sortie de a si, et seulement si, $x_i = \bar{a}$.

Inversement, étant donné une pile $\beta_0, \beta_1, \dots, \beta_n$, envisageons le mot $\alpha = x_1 \dots x_n$ où, si i est une entrée de a dans la pile, $x_i = a$, et si i est une sortie de a , $x_i = \bar{a}$: par récurrence sur i , il est immédiat que $\tau(\alpha_i) = \beta_i$; en particulier $\tau(\alpha) = \tau(\alpha_n) = \wedge$; le mot α appartient donc à P et la pile lui est associée. On a ainsi démontré :

Théorème 2 :

L'application du langage de parenthèses déduit de V dans l'ensemble des piles sur V , qui transforme tout mot en sa pile associée, est une bijection.

Il n'y a donc pas de différence essentielle entre ces deux ensembles.

4. RAMIFICATIONS

4.1. Introduction

Les structures arborescentes qui constituent les marqueurs de phrase définis par une grammaire de Chomsky peuvent être décrites comme des arborescences, munies d'une orientation dans le sens de la lecture, dont les nœuds sont étiquetés par des symboles de l'alphabet V réunion de l'alphabet terminal et de l'alphabet non terminal de la grammaire. Il est commode d'accepter plusieurs racines, pour permettre des axiomes non nécessairement réduits à une lettre, et surtout pour faciliter la théorie algébrique qui va être faite.

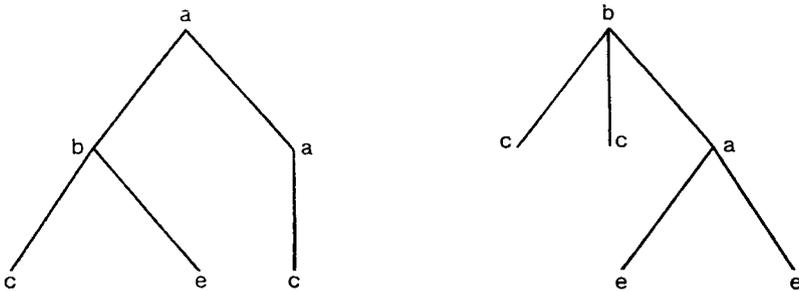


Figure 1

Nous nommerons *ramifications sur V* les êtres que nous venons de décrire intuitivement et \hat{V} leur ensemble. Nous en donnerons une définition axiomatique après avoir dégagé les propriétés qui permettent de les caractériser.

Deux lois de composition opèrent naturellement sur les ramifications :

a) Une loi de composition interne (notée $+$), la *concaténation*, qui juxtapose deux ramifications à m et n « racines » en une ramification à $m + n$ racines ; ainsi la ramification de la figure 1 est obtenue par concaténation de deux ramifications à une seule racine ; cette loi est associative et on introduit la *ramification vide*, notée \wedge , qui en est élément neutre.

b) Une loi de composition externe (notée \times), ou *enracinement*, qui à un élément a de V et une ramification r , associe la ramification obtenue en ajoutant à r une racine étiquetée a ; la figure 2, par exemple, schématise $a \times r$ lorsque r est la ramification de la figure 1.

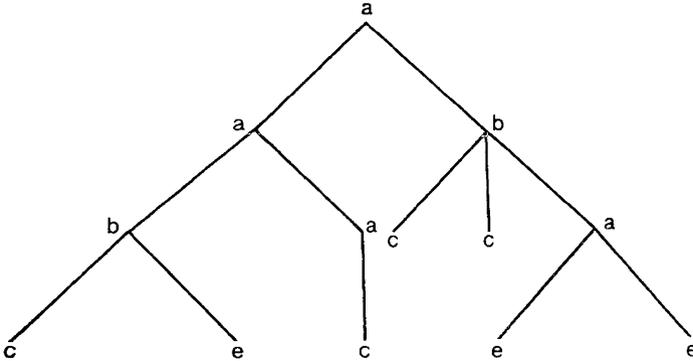


Figure 2

En particulier $a \times \wedge$ est la ramification réduite à un seul point étiqueté a ; on peut identifier $a \times \wedge$ et a . Avec cette convention et l'habituelle convention de priorité de \times sur $+$, la ramification de la figure 1 s'écrit :

$$a \times (b \times (c + e) + a \times c) + b \times (c + c + a \times (e + e))$$

Le mode de construction des ramifications au moyen de ces deux lois est précisé par le fait que toute ramification non vide sur V , r , s'écrit de manière unique

$$r = a \times s + t \quad \text{avec} \quad a \in V, s \in \hat{V}, t \in \hat{V}.$$

De plus, les ramifications ont une « taille » finie, que les opérations font croître.

4.2. Définition axiomatique de \hat{V}

Définition 3 :

On appelle binoïde sur V tout ensemble muni

- a) d'une loi de composition interne associative avec un élément neutre ;
- b) d'une loi de composition externe à opérateurs dans V .

EXEMPLES : 1. L'ensemble \mathbf{N} des entiers naturels, muni de l'addition et de la loi externe notée $.$ définie par :

$$(\forall a \in V) (\forall n \in \mathbf{N}) (a . n = n + 1)$$

est un binoïde sur V .

2. Le monoïde libre V^* engendré par V , muni de la concaténation et de la loi externe notée $.$ définie par

$$(\forall a \in V) (\forall \alpha \in V^*) (a . \alpha = a)$$

est un binoïde sur V .

Dans la suite, les deux lois d'un binoïde seront toujours notées respectivement $+$ et \times , l'élément neutre de $+$ sera noté \wedge .

Axiomes :

- 1) \hat{V} est un binoïde sur V :
- 2) Soit $r \neq \wedge$ une ramification de \hat{V} ; il existe $a \in V, s \in V, t \in V$, uniques, tels que $r = a \times s + t$;
- 3) Il existe une application ν de \hat{V} dans l'ensemble des entiers naturels, telle que :
 - $\nu(\wedge) = 0$,
 - $\nu(r + s) > \nu(r)$ si $s \neq \wedge$, $\nu(r + s) > \nu(s)$ si $r \neq \wedge$,
 - $\nu(a \times r) > \nu(r)$.

Il résulte de (2) et (3) que $\nu(r)$ est strictement positif lorsque $r \neq \wedge$.

D'autre part, d'après (2), $a \times \wedge = b \times \wedge$ entraîne $a = b$: on pourra donc identifier $a \times \wedge$ et a ; ainsi V sera une partie de \hat{V} .

4.3. Récurrences dans \hat{V}

L'outil essentiel pour l'étude de \hat{V} est le raisonnement par récurrence (proposition 3) ainsi que la définition d'une fonction par récurrence (proposition 4).

Proposition 3 :

Soit \mathcal{F} un prédicat dans \hat{V} . Si

- a) $\mathcal{F}(\wedge)$ est vrai,
 - b) $\mathcal{F}(s)$ et $\mathcal{F}(t) \Rightarrow (\forall a \in V) \mathcal{F}(a \times s + t)$,
- alors $\mathcal{F}(r)$ est vrai pour tout $r \in \hat{V}$.

Il suffit de raisonner par récurrence sur $\nu(r)$, en utilisant les axiomes 2 et 3.

On peut remplacer la partie (b) de cette proposition 3 par

$$b') \quad \mathcal{F}(s) \text{ et } \mathcal{F}(t) \Rightarrow \mathcal{F}(s + t)$$

et

$$b'') \quad \mathcal{F}(s) \Rightarrow (\forall a \in V) \mathcal{F}(a \times s).$$

A titre d'application, démontrons que dans \hat{V} n'est strictement contenu aucun sous-binoïde, c'est-à-dire aucune partie où se trouve \wedge et qui soit stable pour $+$ et \times : en effet soit B un sous-binoïde de \hat{V} ; par récurrence, toute ramification r de \hat{V} appartient à B car :

$$\begin{aligned} \wedge &\in B \\ s \in B \quad \text{et} \quad t \in B &\quad \text{entraînent} \quad s + t \in B \quad \text{et} \quad a \times s \in B ; \\ \text{donc } B &= \hat{V}. \end{aligned}$$

En particulier, l'ensemble formé de \wedge et des combinaisons finies des éléments de V par $+$ et \times est un sous-binoïde de \hat{V} , donc \hat{V} lui-même : toute ramification sur V est vide ou combinaison finie d'éléments de V par $+$ et \times (cf. l'exemple de 4.1); \hat{V} est engendré par V .

Proposition 4 :

Soient un ensemble E , un élément e_0 de E et, pour tout $a \in V$, une application ψ_a de $\hat{V}^2 \times E^2$ dans E . Il existe une application f de \hat{V} dans E , et une seule, telle que :

- a) $f(\wedge) = e_0$
- b) $(\forall a \in V, s \in \hat{V}, t \in \hat{V}) f(a \times s + t) = \psi_a(s, t, f(s), f(t))$.

En effet, d'après les axiomes 2 et 3, ces conditions définissent f , de manière unique, successivement sur les ensembles de ramifications r telles que $v(r)$ soit égal à $0(r = \wedge)$, 1, 2, etc...

Donnons des exemples importants de fonctions définies grâce à la proposition 4.

Mot des racines d'une ramification : soit ρ , application de \hat{V} dans V^* , définie par :

$$\rho(\wedge) = \wedge ; \rho(a \times s + t) = a\rho(t).$$

$\rho(r)$ s'appelle *mot des racines* de r . Pour la ramification de la figure 1, le mot des racines est ab .

Mot des feuilles : soit φ , application de \hat{V} dans V^* , définie par :

$$\varphi(\wedge) = \wedge ; \varphi(a \times s + t) = \text{si } s \neq \wedge \text{ alors } \varphi(s)\varphi(t) \text{ sinon } a\varphi(t).$$

$\varphi(r)$ s'appelle *mot des feuilles* de r . Pour la ramification de la figure 1, le mot des feuilles est *cecccee*.

Familles : Soit $b \in V$. On appelle *familles de prédécesseur* b dans une ramification r les mots de l'ensemble $F_b(r)$, où F_b est l'application de \hat{V} dans l'ensemble des parties de V^* définie par :

$$F_b(\wedge) = \emptyset$$

$$F_b(a \times s + t) = \text{si } a = b \text{ alors } F_a(s) \cup F_a(t) \cup \{ \rho(s) \} \text{ sinon } F_b(s) \cup F_b(t).$$

La ramification de la figure 1 possède deux familles de prédécesseur b qui sont ce et cca et une seule famille de prédécesseur c , le mot vide \wedge .

Transcription : soit un ensemble fini V' et une application g de V dans V' . L'application \hat{g} de \hat{V} dans \hat{V}' , définie ci-dessous, remplace toute « étiquette » $a \in V$ d'une ramification $r \in \hat{V}$ par $g(a) \in V'$:

$$\hat{g}(\wedge) = \wedge ; \hat{g}(a \times s + t) = g(a) \times \hat{g}(s) + \hat{g}(t).$$

\hat{g} s'appelle une *transcription* de \hat{V} dans \hat{V}' .

L'application g de V dans V' se prolonge également en un (unique) homomorphisme g^* du monoïde libre V^* dans le monoïde libre V'^* :

$$g^*(a_1 \dots a_n) = g(a_1) \dots g(a_n) ;$$

nous dirons aussi qu'un tel homomorphisme est une *transcription* entre monoïdes libres. Montrons que g^* transforme le mot des racines de r en celui $\hat{g}(r)$:

$$\rho[\hat{g}(r)] = g^*[\rho(r)].$$

L'application $\rho' = \rho \circ \hat{g}$ vérifie :

$$\rho'(\wedge) = \wedge$$

$$\rho'(a \times s + t) = \rho[g(a) \times \hat{g}(s) + \hat{g}(t)] = g(a)\rho'(t).$$

D'autre part, l'application $\rho'' = g^* \circ \rho$ vérifie :

$$\rho''(\wedge) = \wedge$$

$$\rho''(a \times s + t) = g^*(a\rho(t)) = g(a)\rho''(t).$$

D'après la proposition 4, ces applications sont donc égales.

On montrera de même que

$$\varphi[\hat{g}(r)] = g^*[\varphi(r)]$$

et que les familles de prédécesseur a dans $\hat{g}(r)$ sont déduites par g^* des familles de r dont le prédécesseur b vérifie $g(b) = a$.

Plus généralement, la proposition 4 est la base de la théorie des fonctions récursives primitives dans \hat{V} [13]. Les fonctions qu'elle permet de construire peuvent être employées pour définir des grammaires transformationnelles [3] [5], ou des traductions et en particulier des compilateurs.

4.4. Binoïde universel

La proposition 4 va aussi nous permettre de démontrer :

Proposition 5 :

Soit B un binoïde sur V . Il existe un, et un seul, homomorphisme de binoïdes sur V , de \hat{V} dans B .

Un homomorphisme de binoïdes sur V , de \hat{V} dans B , est une application h de \hat{V} dans B qui vérifie, pour tous $a \in V, r \in \hat{V}, s \in \hat{V}$:

$$(1) h(\wedge) = \wedge ; (2) h(r + s) = h(r) + h(s) ; (3) h(a \times r) = a \times h(r).$$

De ces égalités résulte

$$(4) h(a \times s + t) = a \times h(s) + h(t).$$

D'après la proposition 4, il existe une application h unique vérifiant (1) et (4). Cette application h vérifie aussi (3) : il suffit, dans (4), de faire $s = r$ et $t = \wedge$. Montrons que h vérifie également (2); nous procéderons par récurrence sur r :

— si $r = \wedge$, (2) est évident ;

— si $r = a \times r' + r''$,

$$\begin{aligned} h(r + s) &= h(a \times r' + r'' + s) \\ &= a \times h(r') + h(r'' + s) && \text{d'après l'associativité et (4)} \\ &= a \times h(r') + h(r'') + h(s) && \text{(hyp. de récurrence)} \\ &= h(a \times r' + r'') + h(s) && \text{d'après l'associativité et (4)} \\ &= h(r) + h(s). \end{aligned}$$

EXEMPLES : 1. B est l'ensemble des entiers naturels, avec les lois indiquées au paragraphe 4.2, exemple 1. L'homomorphisme h de V dans B vérifie donc

$$h(\wedge) = 0 \quad h(r + s) = h(r) + h(s) \quad h(a \times r) = h(r) + 1.$$

$h(r)$ s'appelle la *taille* de r .

2. $B = V^*$, avec les lois indiquées en 4.2, exemple 2. L'homomorphisme de \hat{V} dans V^* n'est autre que l'application ρ « mot des racines ».

3. $B = \hat{V}$ avec sa loi externe \times et, comme loi interne \oplus :

$$r \oplus s = s + r.$$

L'homomorphisme h vérifie :

$$h(\wedge) = \wedge \quad h(r + s) = h(s) + h(r) \quad h(a \times r) = a \times h(r).$$

$h(r)$ s'appelle *ramification réfléchie* de r : intuitivement, il s'agit de l'image de r dans un miroir vertical ; on la notera \tilde{r} . La figure 3 schématise \tilde{r} si r est la ramification de la figure 1.

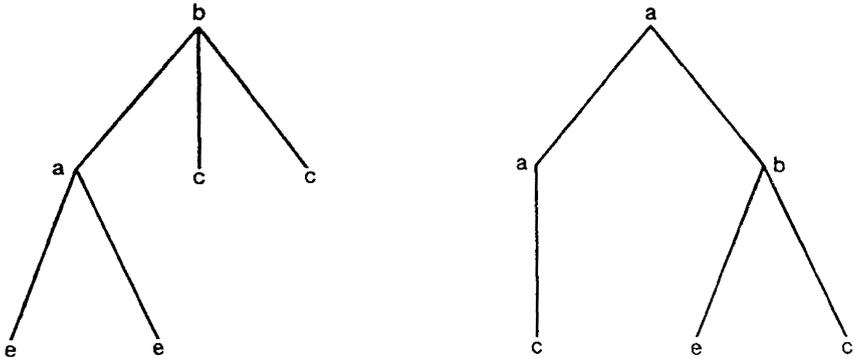


Figure 3

En suivant la même méthode que plus haut dans le cas des transcriptions, on montre que le mot des racines de \tilde{r} est le réfléchi du mot des racines de r , que le mot des feuilles de \tilde{r} est le réfléchi du mot des feuilles de r et que les familles de prédécesseur a dans \tilde{r} sont les mots réfléchis des familles de prédécesseur a dans r .

4. Une partie de \hat{V} est un « langage à deux dimensions », ce qu'on peut appeler un *bilangage* ; la notion d'homomorphisme de binoïdes permet de généraliser la notion de langage régulier [6] en celle de *bilangage régulier* [12] : un bilangage régulier est une partie K de \hat{V} telle qu'il existe un binoïde fini B pour lequel K est l'image réciproque $h^{-1}(B')$ d'une partie B' de B par l'homomorphisme h de \hat{V} dans B .

La proposition 5 conduit à la définition :

Définition 4 :

On appelle binoïde universel sur V (ou engendré par V) un ensemble \hat{V} qui vérifie les axiomes 1, 2, 3.

4.5. Unicité et existence du binoïde universel sur V

Théorème 3 :

Étant donné un ensemble V , il existe un binoïde universel sur V , unique à un isomorphisme près.

L'unicité résulte de la proposition 5 : soient \hat{V} et B deux binoïdes universels sur V ; il existe un homomorphisme h de \hat{V} dans B et un homomorphisme h' de B dans \hat{V} ; $h'oh$ est un homomorphisme de \hat{V} dans \hat{V} , mais l'identité de \hat{V} , $I_{\hat{V}}$, en est un autre et d'après l'unicité de l'homomorphisme de \hat{V} dans \hat{V}

$$h'oh = I_{\hat{V}};$$

de même

$$hoh' = I_B.$$

h est par suite une bijection, donc un isomorphisme, de \hat{V} sur B : B est isomorphe à \hat{V} .

D'autre part, d'après le théorème 1, le langage de parenthèses P (paragraphe 2) vérifie les axiomes si on y définit deux lois de composition par :

$$\beta + \gamma = \beta\gamma \quad a \times \beta = a\beta\bar{a}$$

pour $\beta, \gamma \in \hat{V}$ et $a \in V$; l'application ν de l'axiome 3 pourra être la longueur.

Les binoïdes universels sur V sont donc tous isomorphes. Dans la suite, \hat{V} désigne l'un d'eux dont les éléments sont nommés ramifications et intuitivement considérés comme étant les structures arborescentes telles que celles des figures 1 et 2. Comme P , muni des deux lois ci-dessus, est aussi un binoïde universel sur V , \hat{V} et P sont isomorphes ; autrement dit :

Théorème 4

Il existe une bijection p de \hat{V} sur P , telle que, pour $a \in V, r \in \hat{V}, s \in \hat{V}$:

$$p(\wedge) = \wedge \quad p(r + s) = p(r)p(s) \quad p(a \times r) = ap(r)\bar{a}$$

On peut appeler P *binoïde universel parenthétique* et $p(r)$ *représentation parenthétique* de la ramification r .

Il s'agit de la représentation la plus naturelle d'une ramification à l'aide de parenthèses. Pour la ramification r de la figure 1,

$$p(r) = abc\bar{c}\bar{e}\bar{e}\bar{b}acc\bar{c}\bar{a}\bar{b}cc\bar{c}\bar{c}\bar{a}\bar{e}\bar{e}\bar{e}\bar{e}\bar{a}\bar{b}.$$

Par récurrence, on obtient immédiatement les résultats qui suivent.

Proposition 6 :

Soit r une ramification non vide sur V , a et b deux éléments de V .

- a) $p(r)$ et $\varphi(r)$ commencent par la même lettre ;
- b) si $\varphi(r)$ se termine par a , $p(r)$ se termine par \bar{a} ;
- c) si $p(r)$ contient le facteur ab , une famille de prédécesseur a commence par b ;
- d) si $p(r)$ contient le facteur $\bar{a}\bar{b}$, nécessairement $b = a$ et $\varphi(r)$ contient a ;
- e) si $p(r)$ contient le facteur $\bar{a}\bar{b}$, une famille de prédécesseur b se termine par a ;
- f) si $p(r)$ contient le facteur $\bar{a}\bar{b}$, une famille admet ab comme facteur.

5. AUTRES REPRÉSENTATIONS BIJECTIVES DE \hat{V} DANS P

5.1. Représentation descendante droite

Ci-dessus, le choix des deux lois de composition de P , $+$ et \times , pour vérifier l'axiome 2, a été guidé par le théorème 1 : il s'agissait de mettre tout mot α de P , non vide, de manière unique, sous la forme

$$\alpha = a \times \alpha' + \alpha'' ;$$

on est parti pour cela de la décomposition

$$\alpha = a\beta\bar{a}\gamma$$

et on a pris $\alpha' = \beta$ et $\alpha'' = \gamma$.

On peut aussi prendre $\alpha' = \gamma$ et $\alpha'' = \beta$; on est alors conduit aux deux lois de composition :

$$a \times \gamma = a\bar{a}\gamma$$

$$\delta + \beta = \Gamma(\delta)\beta\Delta(\delta)$$

où, pour tout $\delta \in \mathcal{W}^*$, $\Gamma(\delta)$ est le plus grand facteur gauche de δ qui est un mot sur \mathcal{V} et $\Delta(\delta)$ le quotient de δ par $\Gamma(\delta)$

$$\delta = \Gamma(\delta)\Delta(\delta).$$

$+$ est bien une loi de composition interne dans P car, d'après la proposition 1, si $\delta, \beta \in P$:

$$\tau[\Gamma(\delta)\beta] = \tau[\Gamma(\delta)] \quad \tau(\delta + \beta) = \tau[\Gamma(\delta)\Delta(\delta)] = \tau(\delta) = \wedge ;$$

le mot vide \wedge est élément neutre et $+$ est associative car :

$$(\delta + \beta) + \gamma = \Gamma(\delta)\Gamma(\beta)\gamma\Delta(\beta)\Delta(\delta) = \delta + (\beta + \gamma).$$

Les axiomes 1 et 2 sont donc satisfaits et, ici encore, on peut choisir la longueur comme application v de l'axiome 3. Ainsi P , muni de ces deux lois, est encore un binoïde universel sur \mathcal{V} . Il est donc isomorphe à \hat{V} ; autrement dit il existe une bijection dd de \hat{V} sur P telle que

$$dd(\wedge) = \wedge \quad , \quad dd(r + s) = \Gamma[dd(r)]dd(s)\Delta[dd(r)] \quad , \quad dd(a \times r) = a\bar{a}dd(r).$$

EXEMPLES :

$$dd(c) = c\bar{c}$$

$$dd(c + e) = ce\bar{c}$$

$$dd[b \times (c + e)] = b\bar{b}ce\bar{c}$$

$$dd(a \times c) = a\bar{a}c\bar{c}$$

envisageons la ramification r de la figure 1 ; $r = r_1 + r_2$ où r_1 et r_2 ont chacune une racine :

$$dd(r_1) = a\bar{a}ba\bar{a}c\bar{c}b\bar{c}e\bar{e}c\bar{c}$$

$$dd(r_2) = b\bar{b}cca\bar{a}e\bar{e}e\bar{e}c\bar{c}$$

$$dd(r) = ab\bar{b}cca\bar{a}e\bar{e}e\bar{e}c\bar{c}a\bar{b}a\bar{a}c\bar{c}b\bar{c}e\bar{e}c\bar{c}$$

Théorème 5 :

a) Il existe une bijection dd de \hat{V} sur P , telle que, pour $a \in V$, $r \in \hat{V}$, $s \in \hat{V}$:

$$dd(\wedge) = \wedge, \quad dd(r + s) = \Gamma[dd(r)] dd(s) \Delta[dd(r)], \quad dd(a \times r) = a\bar{a} dd(r).$$

b) Pour toute ramification r de taille n :

$$dd(r) = \rho(r)\bar{a}_1\alpha_1\bar{a}_2\alpha_2 \dots \bar{a}_n\alpha_n \quad \text{où}$$

— pour $i = 1, 2, \dots, n$, $\alpha_i \in V^*$ est une famille de prédécesseur $a_i \in V$ et, réciproquement, si α est une famille de prédécesseur a dans r , il existe un i tel que $a_i = a$ et $\alpha_i = \alpha$,

— la suite extraite de $a_1a_2 \dots a_n$, formée des a_i tels que $\alpha_i = \wedge$, est le mot réfléchi du mot des feuilles de r .

Notons qu'en particulier le mot α_n est nécessairement vide puisque $dd(r)$ appartient à P .

Il reste à démontrer la partie (b) du théorème, ce que l'on peut faire par récurrence sur r :

— elle est évidente pour $r = \wedge$;

— si $r = s + t$: $\rho(r) = \rho(s)\rho(t)$, $\varphi(r) = \varphi(s)\varphi(t)$, les familles de prédécesseur a dans r sont celles de s et celles de t , la taille de r est la somme de celles de s et t ; enfin, si

$$dd(s) = \rho(s)\bar{a}_1\alpha_1 \dots \bar{a}_m\alpha_m \quad \text{et} \quad dd(t) = \rho(t)\bar{b}_1\beta_1 \dots \bar{b}_n\beta_n,$$

alors $dd(s + t) = \rho(s)\rho(t)\bar{b}_1\beta_1 \dots \bar{b}_n\beta_n\bar{a}_1\alpha_1 \dots \bar{a}_m\alpha_m$.

— si $r = a \times s$: $\rho(r) = a$, les familles de prédécesseur b dans r sont celles de s et, pour $b = a$, $\rho(s)$; la taille de r est celle de s augmentée de 1,

$$dd(s) = \rho(s)\bar{a}_1\alpha_1 \dots \bar{a}_m\alpha_m \quad \text{entraîne} \quad dd(r) = a\bar{a}\rho(s)\bar{a}_1\alpha_1 \dots \bar{a}_m\alpha_m ;$$

si $s \neq \wedge$, $\rho(s) \neq \wedge$ et $\varphi(r) = \varphi(s)$; si $s = \wedge$, $dd(r) = a\bar{a}$ et $\varphi(r) = a$.

Avec les notations du théorème, a_1 est la dernière racine de r ; si $\alpha_1 \neq \wedge$, a_2 est le dernier élément de la famille α_1 ; si $\alpha_1 = \wedge$, a_2 est l'avant-dernière racine, ... : $dd(r)$ présente les familles de r après leur prédécesseur, en allant

des racines vers les feuilles et en commençant par la « droite » de r (cf. les exemples donnés). Nous dirons que $dd(r)$ est la *représentation descendante droite* de r et nous nommerons P , muni des lois $+$ et \times indiquées dans ce paragraphe, *binoïde universel descendant droit*.

5.2. Représentation descendante gauche

Il est souvent plus agréable pour les applications d'obtenir les familles et le mot des feuilles d'une ramification, « de gauche à droite ».

Posons, pour toute ramification r ,

$$dg(r) = dd(\tilde{r}) \quad (\text{cf. 4.4, exemple 3})$$

Du théorème 5 résulte aisément :

Théorème 6 :

a) Il existe une bijection dg de \hat{V} sur P , telle que pour $a \in V$, $r \in \hat{V}$, $s \in \hat{V}$:

$$dg(\wedge) = \wedge, \quad dg(r + s) = \Gamma[dg(s)]dg(r)\Delta[dg(s)], \quad dg(a \times r) = a\bar{a}dg(r).$$

b) Pour toute ramification r de taille n :

$$dg(r) = \tilde{\rho}(r)\bar{a}_1\tilde{\alpha}_1\bar{a}_2\tilde{\alpha}_2 \dots \bar{a}_n\tilde{\alpha}_n \quad \text{où}$$

— pour $i = 1, 2, \dots, n$, $\alpha_i \in V^*$ est une famille de prédécesseur $a_i \in V$ et, réciproquement, si α est une famille de prédécesseur a dans r , il existe un i tel que $a_i = a$ et $\alpha_i = \alpha$,

— le mot des feuilles de r est la suite extraite de $a_1 a_2 \dots a_n$ formée des a_i tels que $\alpha_i = \wedge$.

$dg(r)$ s'appelle *représentation descendante gauche* de r . dg est l'isomorphisme de \hat{V} dans P , binoïde universel pour les deux lois $+$ et \times définies par :

$$\beta + \gamma = \Gamma(\gamma)\beta\Delta(\gamma) \quad \text{et} \quad a \times \beta = a\bar{a}\beta.$$

P muni de ces lois sera nommé *binoïde universel descendant gauche*.

On vérifiera que pour la ramification r de la figure 1 :

$$dg(r) = ba\bar{a}abb\bar{b}ec\bar{c}e\bar{a}c\bar{c}b\bar{a}cc\bar{c}c\bar{a}ee\bar{e}e.$$

5.3. Représentation ascendante gauche

Pour obtenir les familles d'une ramification en allant des feuilles vers les racines, on est amené à « retourner » $dd(r)$. Le réfléchi d'un mot de P n'est pas un mot de P , mais le conjugué d'un mot de P appartient à P (proposition 2).

Posons donc

$$ag(r) = c[dd(r)].$$

Les propriétés de ag se déduisent immédiatement de celles de dd (théorème 5). Pour tout mot α de W^* , $c[\Gamma(\alpha)]$ est le plus grand facteur droit de $c(\alpha)$ qui est un mot sur \bar{V} ; désignons par $\Delta'(\beta)$ le plus grand facteur droit d'un mot β qui appartient à \bar{V}^* et par $\Gamma'(\beta)$ le quotient de β par $\Delta'(\beta)$ à droite :

$$\beta = \Gamma'(\beta)\Delta'(\beta) ;$$

$$c[\Gamma(\alpha)] = \Delta'[c(\alpha)] ; \quad c[\Delta(\alpha)] = \Gamma'[c(\alpha)].$$

Le théorème 5 entraîne alors immédiatement :

Théorème 7 :

a) Il existe une bijection ag de \hat{V} sur P , telle que pour $a \in V$, $r \in \hat{V}$, $s \in \hat{V}$:

$$ag(\wedge) = \wedge, \quad ag(r + s) = \Gamma'[ag(r)] ag(s) \Delta'[ag(r)], \quad ag(a \times r) = ag(r) a \bar{a}.$$

b) Pour toute ramification r de taille n :

$$ag(r) = \bar{\beta}_1 b_1 \bar{\beta}_2 b_2 \dots \bar{\beta}_n b_n \tilde{r} \quad \text{où}$$

- pour $i = 1, 2, \dots, n$, $\beta_i \in V^*$ et $\tilde{\beta}_i$ est une famille de prédécesseur $b_i \in V$; réciproquement, si α est une famille de prédécesseur a dans r , il existe un i tel que $b_i = a$ et $\tilde{\beta}_i = \alpha$,
- le mot des feuilles de r est la suite extraite de $b_1 b_2 \dots b_n$ formée des b_i tels que $\beta_i = \wedge$.

En particulier, le mot β_1 est nécessairement vide puisque $ag(r)$ appartient à P : b_1 est la première feuille. $ag(r)$ présente les familles de r des feuilles vers les racines et « de gauche à droite » : on appelle $ag(r)$ *représentation ascendante gauche* de la ramification r . ag est l'isomorphisme de \hat{V} dans P , binoïde universel pour les deux lois $+$ et \times définies par :

$$\beta + \gamma = \Gamma'(\beta)\gamma\Delta'(\beta) \quad \text{et} \quad a \times \beta = \beta a \bar{a}.$$

P muni de ces deux lois sera appelé *binoïde universel ascendant gauche*.

A l'aide de la définition de ag ou de la première partie du théorème 7, on vérifiera que, pour la ramification r de la figure 1 :

$$ag(r) = ce\bar{e}cb\bar{c}c\bar{a}\bar{a}\bar{b}accee\bar{e}\bar{e}\bar{e}\bar{a}\bar{a}\bar{c}\bar{c}\bar{b}\bar{b}\bar{a}.$$

On pourrait définir de la même façon une représentation ascendante droite.

En conclusion, on pourrait identifier \hat{V} et P puisqu'ils sont isomorphes, mais cette identification peut se faire de plusieurs manières dont aucune n'est a priori préférable aux autres.

6. GRAMMAIRES DE CHOMSKY

Définition 5 :

Nous appellerons *grammaire de Chomsky (ou simplement grammaire) un quadruplet* (T, N, \rightarrow, X) où :

- T est un ensemble fini (alphabet terminal);
- N est un ensemble fini disjoint de T (alphabet non terminal); on pose $V = T \cup N$ (alphabet);
- \rightarrow est une relation binaire entre N et V^* (relation de production) telle que, pour tout $a \in N$, les mots α vérifiant $a \rightarrow \alpha$ forment un langage régulier ne contenant pas le mot vide ;
- X est un langage régulier ne contenant pas le mot vide (ensemble des axiomes).

Une ramification r sur V est engendrée par cette grammaire si, dans r :

- a) pour toute famille α de prédécesseur $a \in N$, $a \rightarrow \alpha$;
- b) toute famille de prédécesseur $a \in T$ est vide ;
- c) le mot des racines appartient à X .

Le langage engendré par une grammaire est l'ensemble des mots des feuilles des ramifications qu'elle engendre.

EXEMPLE :

$$T = \{ +, \times, (,), x, y \}; \quad N = \{ e, t, f \}; \quad X = \{ e \};$$

les couples de la relation de production sont en nombre fini :

$$e \rightarrow e + t \quad e \rightarrow t \quad t \rightarrow t \times f \quad t \rightarrow f \quad f \rightarrow x \quad f \rightarrow y \quad f \rightarrow (e).$$

La figure 4 représente une ramification engendrée par la grammaire. $(x + y) \times y + x$ est un mot du langage engendré, qui est formé des « expressions arithmétiques » employant les opérateurs $+$ et \times et les variables x et y .

Dans la suite, nous noterons $\mathcal{R}(G)$ l'ensemble (bilangage) des ramifications engendrées par la grammaire G et $L(G)$ le langage engendré par G :

$$L(G) = \varphi[\mathcal{R}(G)].$$

La définition 5 diffère de la définition habituelle [2], [6] de deux manières essentielles. D'abord, elle est plus générale en ce que l'ensemble des couples de la relation de production et celui des axiomes peuvent être infinis. En réalité, on montre que, pour toute grammaire G , on peut trouver une grammaire G'

où les couples a, α tels que $a \rightarrow \alpha$ sont en nombre fini (avec même $|\alpha| \leq 2$), qui possède un seul axiome réduit à une lettre, et qui est fortement équivalente à G , c'est-à-dire qu'il existe une bijection de $\mathcal{R}(G)$ sur $\mathcal{R}(G')$ qui conserve le

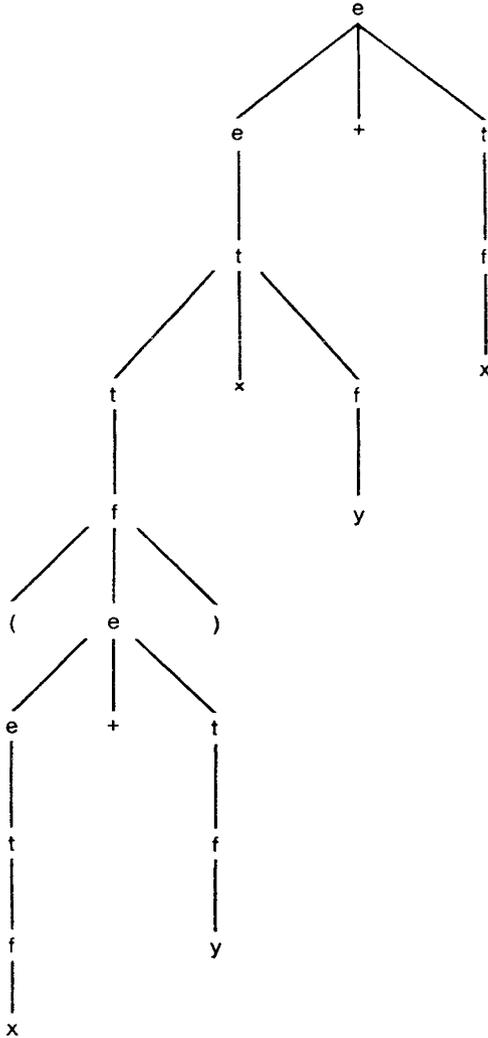


Figure 4

mot des feuilles. D'autre part, cette généralisation, souvent utile, ne complique pas la plupart des résultats sur les grammaires.

L'autre différence concerne la définition du langage engendré, pour laquelle on utilise habituellement les *dérivations* : il s'agit des suites $\delta_0, \delta_1, \dots, \delta_n$ de

mots sur V telles que l'on passe de δ_{i-1} à δ_i en remplaçant $a \in N$ par $\alpha \in V^*$ vérifiant $a \rightarrow \alpha$; le langage engendré est l'ensemble des mots β sur T tels qu'existe une dérivation où δ_0 est un axiome et $\delta_n = \beta$. Nous ne démontrerons pas ici l'équivalence des deux points de vue, nous bornant à signaler la liaison entre une dérivation et, par exemple, $dd(r)$ pour une ramification r engendrée par la grammaire :

$$dd(r) = \rho(r)\bar{a}_1\alpha_1 \dots \bar{a}_n\alpha_n ;$$

pour $i = 0, 1, \dots, n$, posons

$$\beta_i = \tau[\rho(r)\bar{a}_1\alpha_1 \dots \bar{a}_i\alpha_i]$$

où l'application τ est celle qui a été définie au paragraphe 2 (en particulier $\beta_0 = \rho(r)$ et $\beta_n = \wedge$ car $dd(r) \in P$), et

$$\gamma_i = h[\rho(r)\bar{a}_1\alpha_1 \dots \bar{a}_i\alpha_i]$$

où h est l'homomorphisme de W^* dans T^* tel que

$$h(\bar{a}) = a \text{ si } a \in T, \quad h(\bar{a}) = \wedge \text{ si } a \in N, \quad h(a) = \wedge \text{ pour } a \in V ;$$

en particulier $\gamma_0 = \wedge$ et $\gamma_n = \tilde{\varphi}(r)$ d'après le théorème 5. Lorsque $a_i \in T$, on passe de β_{i-1} à β_i en enlevant a_i et de γ_{i-1} à γ_i en ajoutant a_i :

$$\beta_i\tilde{\gamma}_i = \beta_{i-1}\tilde{\gamma}_{i-1} ;$$

lorsque $a_i \in N$, $\gamma_i = \gamma_{i-1}$ et on passe de β_{i-1} à β_i en remplaçant a_i par α_i tel que $a_i \rightarrow \alpha_i$; en résumé, la suite des mots $\beta_i\tilde{\gamma}_i$ tels que $a_i \in N$ est une dérivation qui conduit de $\beta_0\tilde{\gamma}_0 = \rho(r) \in X$ à $\beta_n\tilde{\gamma}_n = \varphi(r)$, mot des feuilles de r .

7. L'ANALYSE SYNTAXIQUE POUR LES GRAMMAIRES DE CHOMSKY

Soit une grammaire $G = (T, N, \rightarrow, X)$. Effectuer l'analyse syntaxique d'un mot α sur l'alphabet T , pour la grammaire G , c'est déterminer les ramifications engendrées par G dont α est le mot des feuilles, autrement dit l'ensemble

$$\varphi^{-1}(\alpha) \cap \mathcal{R}(G).$$

Jusqu'à présent, nous n'avons pas précisé le binoïde universel \hat{V} où se trouvent les ramifications engendrées par G , puisque les binoïdes universels sur V sont tous isomorphes. Mais la résolution pratique du problème de l'analyse syntaxique dépend du choix de ce binoïde libre : il est commode de prendre P , muni de lois de composition convenables ; ainsi, on aura à déterminer des mots, ce qui est habituel pour un algorithme.

Nous étudierons φ et $\mathcal{R}(G)$ selon le binoïde universel choisi.

7.1. Analyse descendante prédictive

On détermine les ramifications par leur représentation descendante gauche. Plus précisément, le binoïde universel choisi est le binoïde universel descendant gauche; dg , unique isomorphisme de ce binoïde universel dans lui-même, est l'identité : $dg(r) = r$.

D'après le théorème 6, si r est engendrée par G , $\varphi(r)$ est déduit de $dg(r) = r$ par l'homomorphisme h de W^* dans T^* déjà envisagé au paragraphe 6 :

$$h(\bar{a}) = a \text{ si } a \in T, h(\bar{a}) = \wedge \text{ si } a \in N, h(a) = \wedge \text{ pour } a \in V;$$

$$\varphi(r) = h(r).$$

h est un *homomorphisme alphabétique* entre monoïdes libres, c'est-à-dire qu'il transforme une lettre en une lettre ou en le mot vide.

Étudions maintenant $\mathcal{R}(G)$. Ici et dans toute la suite, pour $a \in N$, nous désignerons par K_a le langage régulier des mots α tels que $a \rightarrow \alpha$; pour $a \in T$, nous poserons $K_a = \{ \wedge \}$; enfin, soit

$$K_0 = \bigcup_{a \in V} \bar{a} \tilde{K}_a;$$

K_0 est un langage régulier [6]. D'après le théorème 6, si $r(= dg(r))$ est engendrée par G ,

$$r \in \tilde{X} K_0^* \cap P.$$

Réciproquement, tout mot de $\tilde{X} K_0^* \cap P$ s'écrit

$$\tilde{\alpha}_0 \bar{a}_1 \tilde{\alpha}_1 \dots \bar{a}_n \tilde{\alpha}_n \text{ avec } \alpha_0 \in X \text{ et } \alpha_i \in K_{a_i} \text{ pour } 1 \leq i \leq n;$$

comme il appartient à P , il s'agit d'une ramification ; d'après le théorème 6, son mot des racines est $\alpha_0 \in X$ et, pour toute famille α de prédécesseur a existe un i tel que $a_i = a$ et $\alpha_i = \alpha$: si $a \in N$, $a \rightarrow \alpha$; si $a \in T$, $\alpha = \wedge$. Cette ramification est engendrée par G . Donc

$$\mathcal{R}(G) = \tilde{X} K_0^* \cap P.$$

$\tilde{X} K_0^*$ est un langage régulier [6].

Théorème 8 :

L'ensemble des ramifications engendrées par une grammaire dans le binoïde universel descendant gauche est l'intersection d'un langage régulier avec le langage de parenthèses qui constitue ce binoïde.

Le langage engendré par la grammaire est l'image de cette intersection par un homomorphisme alphabétique (cf. [4], [15]).

L'analyse syntaxique d'un mot α détermine l'ensemble

$$h^{-1}(\alpha) \cap K \cap P$$

où K est le langage régulier $\tilde{X}K_0^*$.

Cette analyse syntaxique, qui détermine en bloc chaque famille, des racines vers les feuilles et de gauche à droite, peut être appelée analyse descendante prédictive de gauche à droite [7]. L'emploi du binoïde universel descendant droit aurait conduit à une analyse analogue, de droite à gauche.

7.2. Analyse ascendante

Le binoïde universel choisi est le binoïde universel ascendant gauche : $ag(r)$; autrement dit, on détermine les ramifications par leur représentation ascendante gauche. On utilise le théorème 7 pour étudier φ et $\mathcal{R}(G)$.

Si r est engendrée par G , $\varphi(r) = h'(r)$, où h' est l'homomorphisme alphabétique de W^* dans T^* défini par :

$$h'(a) = a \text{ si } a \in T, \quad h'(a) = \wedge \text{ si } a \in N, \quad h'(\bar{a}) = \wedge \text{ pour } a \in V.$$

Posons

$$K_1 = \bigcup_{a \in V} \tilde{K}_a a :$$

K_a est défini en 7.1 et \tilde{K}_a est l'ensemble des mots $\tilde{\alpha}$, pour $a \in K_a$. Par une démonstration analogue à celle de 7.1, on voit que

$$\mathcal{R}(G) = K_1^* \tilde{X} \cap P.$$

$K_1^* \tilde{X}$ est un langage régulier.

Théorème 9 :

L'ensemble des ramifications engendrées par une grammaire dans le binoïde universel ascendant gauche est l'intersection d'un langage régulier avec le langage de parenthèses qui constitue ce binoïde. Le langage engendré par la grammaire est l'image de cette intersection par un homomorphisme alphabétique.

L'analyse syntaxique d'un mot α détermine ici l'ensemble

$$h'^{-1}(\alpha) \cap K' \cap P$$

où K' est le langage régulier $K_1^* \tilde{X}$. Elle peut être appelée analyse ascendante de gauche à droite.

7.3. Analyse descendante parenthétique

Choisissons maintenant le binoïde universel parenthétique ($p(r) = r$), autrement dit déterminons les ramifications par leur représentation parenthétique.

Soit r une ramification de \hat{V} où toute famille de prédécesseur $a \in T$ est vide, alors qu'aucune famille de prédécesseur $a \in N$ n'est vide. Dans ce cas (donc en particulier si r est engendrée par G),

$$\varphi(r) = h(r) = h'(r)$$

où h et h' sont les homomorphismes de W^* dans T^* définis aux paragraphes 7.1 et 7.2.

Pour démontrer ce résultat, on raisonne par récurrence sur r : il est évident pour $r = \wedge$.

Si $r = a \times s + t = as\bar{a}t$, deux cas sont possibles :

$a) s \neq \wedge$, d'où $a \in N$:

$$\varphi(r) = \varphi(s)\varphi(t) \quad , \quad h(r) = h(s)h(t) \quad , \quad h'(r) = h'(s)h'(t).$$

$b) s = \wedge$, d'où $a \in T$: $\varphi(r) = a\varphi(t)$, $h(r) = ah(t)$, $h'(r) = ah'(t)$.

Il suffit alors d'utiliser l'hypothèse de récurrence.

Mais, ici, il n'est pas vrai en général que $\mathcal{R}(G)$ soit l'intersection avec P d'un langage régulier : pour que ce résultat soit exact, on devra faire sur la grammaire G une hypothèse supplémentaire. Donnons d'abord quelques définitions.

7.4. Grammaires locales

Soient deux sous-ensembles D et F de V et un sous ensemble de V^2 (c'est-à-dire un ensemble de mots de deux lettres sur V); le langage L des mots $a_0 \dots a_n$ dont la première lettre a_0 appartient à D , la dernière lettre a_n à F et tout facteur $a_{i-1}a_i$ ($1 \leq i \leq n$) à I , c'est-à-dire le langage

$$L = (DV^* \cap V^*F) \cap (V^*(V^2 - I)V^*)$$

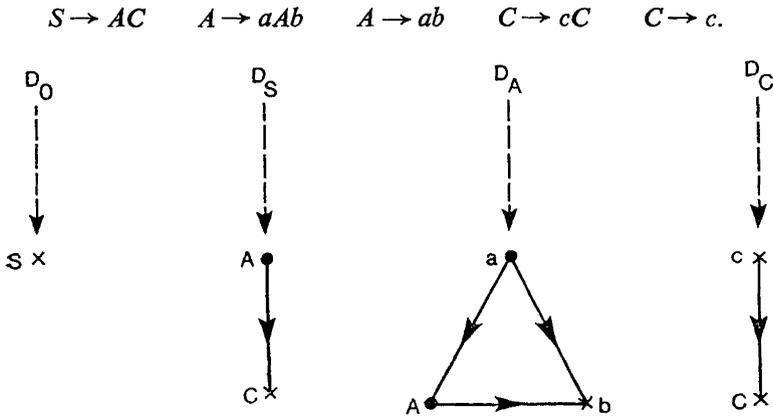
est régulier : on dit que L est le langage local d'ensemble initial D , d'ensemble final F , d'ensemble de transitions I ; L n'est autre que l'ensemble des chemins du graphe (V, I) joignant un point de D à un point de F . On montre [9] que tout langage régulier est biunivoquement déduit d'un langage local par une transcription entre monoïdes libres (4.3).

Définition 6 :

Une grammaire $G = (T, N, \rightarrow, X)$ est dite locale si les langages K_a (7.1), pour tout $a \in N$, et le langage X sont des langages locaux avec le même ensemble I de transitions.

Pour représenter cette grammaire, on peut donner le graphe (V, I) , accompagné des ensembles initial et final des divers langages K_a pour $a \in N$ et du langage X . L'intérêt des grammaires locales vient de la proposition qui suit ; elle entraîne que toute grammaire peut être représentée grâce un graphe dont on

étiquette les points avec des lettres de son alphabet (voir par exemple [16]) ; la figure 5 représente ainsi la grammaire d'axiome S , dont la relation de production est définie par



Les points de l'ensemble $F_0 = F_S = F_A = F_C$ sont représentés par des croix.

Figure 5

Proposition 7 :

Pour toute grammaire G d'alphabet V , il existe une grammaire locale G' d'alphabet V' et une transcription \hat{g} du binoïde universel \hat{V}' dans le binoïde universel \hat{V} , qui transforme biunivoquement $\mathcal{R}(G')$ en $\mathcal{R}(G)$.

Soit $G = (T, N, \rightarrow, X)$: les langages réguliers X et K_a , pour $a \in N$, sont déduits, biunivoquement, de langages locaux X' et K'_a , par des transcriptions de monoïdes libres h_0 et h_a respectivement. On peut toujours supposer deux à deux disjoints les alphabets de X' et des divers langages K'_a . Soient alors V' la réunion de ces alphabets, D_0 et F_0 l'ensemble initial et l'ensemble final de X' , D_a et F_a ceux de K'_a , D la réunion de D_0 et des $D_a (a \in N)$, F celle de F_0 et des F_a , I la réunion des ensembles de transitions de X' et des K'_a . Comme l'alphabet de X' est disjoint de celui des langages K'_a , X' est aussi le langage local d'ensemble initial D_0 , d'ensemble final F_0 et d'ensemble de transitions I . De même K'_a est le langage local d'ensemble initial D_a , d'ensemble final F_a et de même ensemble de transitions I .

Définissons une application g de V' dans l'alphabet V de G :

$$\begin{aligned} g(b) &= h_0(b) & \text{si } b \text{ appartient à l'alphabet de } X' \\ g(b) &= h_a(b) & \text{si } b \text{ appartient à l'alphabet de } K'_a. \end{aligned}$$

g se prolonge en une transcription g^* de V'^* dans V^* qui transforme biunivoquement X' en X et K'_a en K_a , et en une transcription \hat{g} de \hat{V}' dans \hat{V} .

Envisageons la grammaire locale G' d'alphabet terminal $g^{-1}(T)$, d'alphabet non terminal $g^{-1}(N)$, d'ensemble d'axiomes X' , dont la relation de production \rightarrow' est définie par

$$b \rightarrow' \beta \Leftrightarrow \beta \in K'_{g(b)}$$

Nous allons montrer que \hat{g} transforme biunivoquement $\mathcal{R}(G')$ en $\mathcal{R}(G)$.

a) Soit $r \in \mathcal{R}(G')$: une famille α de prédécesseur a dans $\hat{g}(r)$ est déduite par g^* d'une famille β de r dont le prédécesseur b vérifie $g(b) = a$ (4.3) :

$$\text{si } a \in N, b \in g^{-1}(N), \beta \in K'_{g(b)} = K'_a, \text{ d'où } \alpha \in K_a;$$

$$\text{si } a \in T, b \in g^{-1}(T), \beta = \wedge, \text{ d'où } \alpha = \wedge.$$

D'autre part, toujours d'après 4.3 :

$$\rho[\hat{g}(r)] = g^*[\rho(r)] \in g^*(X') = X.$$

Donc, si $r \in \mathcal{R}(G')$, $\hat{g}(r) \in \mathcal{R}(G)$.

b) Avant de démontrer la réciproque, nous démontrerons d'abord un lemme. Soit $\mathcal{L}(G)$ l'ensemble des ramifications sur V dont toute famille α de prédécesseur $a \in N$ vérifie $a \rightarrow \alpha$ et dont toute famille de prédécesseur $a \in T$ est vide ; on définit de la même façon $\mathcal{L}(G')$. Les ramifications engendrées par G sont les ramifications de $\mathcal{L}(G)$ dont le mot des racines appartient à X .

Lemme :

Étant donné une ramification r appartenant à $\mathcal{L}(G)$ et un mot α' tel que $g^(\alpha') = \rho(r)$, il existe dans $\mathcal{L}(G')$ une ramification r' et une seule telle que $\hat{g}(r') = r$ et $\rho(r') = \alpha'$.*

Faisons une récurrence sur r . Le résultat est évident pour $r = \wedge$ (d'où $\alpha' = \wedge$).

Soit $r = b \times s + t$. La ramification r' cherchée s'écrit de manière unique

$$r' = b' \times s' + t'.$$

$$r = \hat{g}(r') \Leftrightarrow g(b') = b \text{ et } \hat{g}(s') = s \text{ et } \hat{g}(t') = t$$

$r' \in \mathcal{L}(G') \Leftrightarrow s' \in \mathcal{L}(G') \text{ et } t' \in \mathcal{L}(G') \text{ et } [(b \in N \text{ et } \rho(s') \in K'_b) \text{ ou } (b \in T \text{ et } s' = \wedge)]$

$$\rho(r') = \alpha' \Leftrightarrow b' \rho(t') = \alpha'.$$

Cherchons d'abord à déterminer $s' : \hat{g}(s') = s$ entraîne $g^*(\rho(s')) = \rho(s)$. Si $s \neq \wedge$, $\rho(s) \in K_b$, donc il existe dans K'_b un mot β' unique tel que $g^*(\beta') = \rho(s)$; d'après l'hypothèse de récurrence, il existe une ramification $s' \in \mathcal{L}(G')$, et une seule, telle que $\hat{g}(s') = s$ et $\rho(s') = \beta'$. Si $s = \wedge$, $b \in T$ et d'autre part $\hat{g}(s') = s$ équivaut à $s' = \wedge$. Dans les deux cas, s' est ainsi déterminée de manière unique.

Le mot α' donné s'écrit $c\gamma'$, avec $g(c) = b$ et $g^*(\gamma') = \rho(t)$: l'égalité $b'\rho(t') = \alpha'$ équivaut à $b' = c$ (ce qui détermine b') et $\rho(t') = \gamma'$. D'après l'hypothèse de récurrence, il existe t' , unique, dans $\mathfrak{L}(G)$, telle que $\hat{g}(t') = t$ et $\rho(t') = \gamma'$.

c) Étant donné une ramification r de $\mathfrak{R}(G)$, il existe un mot $\alpha' \in X'$ et un seul tel que $g^*(\alpha') = \rho(r)$, puisque $\rho(r) \in X$. Le lemme entraîne alors qu'il existe une ramification r' et une seule appartenant à $\mathfrak{R}(G')$ telle que $\hat{g}(r') = r$, ce qui achève de démontrer la proposition 7.

7.5. Retour à l'analyse descendante parenthétique

Théorème 10 :

L'ensemble des ramifications engendrées par une grammaire locale dans le binoïde universel parenthétique (sur son alphabet) est l'intersection d'un langage régulier local avec le langage de parenthèses qui constitue ce binoïde.

Soit la grammaire locale $G = (T, N, \rightarrow, X)$, $V = T \cup N$, D_0 et F_0 l'ensemble initial et l'ensemble final du langage X , D_a et F_a ceux du langage K_a ($a \in N$), I l'ensemble de transitions commun à ces langages.

a) D'après la proposition 6, si $r \in \mathfrak{R}(G)$, le mot $p(r) = r$ appartient au langage local K sur l'alphabet $\mathcal{W} = V \cup \bar{V}$ d'ensemble initial D_0 , d'ensemble final \bar{F}_0 et d'ensemble de transitions

$$I' = \left(\bigcup_{a \in N} aD_a \right) \cup \{ a\bar{a} \mid a \in T \} \cup \left(\bigcup_{b \in N} \bar{F}_b\bar{b} \right) \cup \{ \bar{a}b \mid ab \in I \}.$$

b) Considérons maintenant un mot de P dont tous les facteurs de longueur 2 appartiennent à I' . Il s'agit d'une ramification r .

Montrons d'abord, par récurrence sur r , que tout facteur de longueur 2 dans $\rho(r)$ appartient à I : c'est évident si $r = \wedge$. Sinon, $r = a \times s + t = a\bar{s}at$, $\rho(r) = a\rho(t)$: le seul facteur de longueur 2 qui se trouve dans $\rho(r)$ sans être dans $\rho(t)$ est, si $t \neq \wedge$, ab où b est la première racine de t ; d'après la proposition 6, t commence par b , $\bar{a}b$ est donc un facteur de r , $\bar{a}b$ appartient à I' et par suite ab à I .

Toujours par récurrence, montrons que r appartient à $\mathfrak{L}(G)$, c'est-à-dire que toute famille de prédécesseur $a \in N$ appartient à K_a et que toute famille de prédécesseur $a \in T$ est vide. Pour $r = a \times s + t = a\bar{s}at$, les familles de r sont celles de s et t , qui par l'hypothèse de récurrence possèdent ces propriétés, ainsi que $\rho(s)$, famille de prédécesseur a . Si $a \in N$, $\rho(s)$ commence par la première lettre de s (proposition 6), qui appartient à D_a , et se termine par c si s se termine par \bar{c} : comme $\bar{c}a \in I'$, $c \in F_a$; enfin, nous avons vu que les facteurs de longueur 2 dans $\rho(s)$ appartiennent à I ; donc $\rho(s) \in K_a$. Si $a \in T$, nécessairement $s = \wedge$, car dans r la lettre a est suivie de \bar{a} .

c) Si r appartient à l'intersection de P et du langage K défini plus haut en (a), r appartient donc à $\mathcal{L}(G)$, $\rho(r)$ commence dans D_0 , se termine dans F_0 , alors que ses facteurs de longueur 2 sont dans I : $\rho(r) \in X$ et par suite $r \in \mathcal{R}(G)$, ce qui achève la démonstration.

Dans le cas d'une grammaire locale, l'analyse syntaxique d'un mot α doit donc déterminer l'ensemble $h^{-1}(\alpha) \cap K \cap P = h'^{-1}(\alpha) \cap K \cap P$. La proposition 7 permet de ramener à ce cas le cas général : à une grammaire G on associe une grammaire locale G' , $\mathcal{R}(G)$ est biunivoquement déduit de $\mathcal{R}(G')$ par une transcription \hat{g} entre binoïdes libres ; le mot des feuilles de $\hat{g}(r)$ étant déduit de celui de r par la transcription g^* entre monoïdes libres, l'analyse syntaxique d'un mot α pour la grammaire G revient à déterminer

$$\hat{g}[h^{-1} \circ g^{*-1}(\alpha) \cap K \cap P].$$

$h^{-1} \circ g^{*-1} = (g^* \circ h)^{-1}$, et $g^* \circ h$ est encore un homomorphisme alphabétique. De plus, $L(G) = g^*[\varphi(K \cap P)] = g^* \circ h(K \cap P)$.

Le mode d'analyse qui vient d'être étudié part de la première racine pour aller vers les feuilles qui en descendent, puis recommence avec la seconde racine, ... On peut l'appeler *analyse descendante parenthétique*.

8. CONCLUSION

Dans tous les modes d'analyse syntaxique que nous avons présentés, le problème à résoudre est le même : associer à tout mot α sur l'alphabet terminal, l'ensemble

$$h^{-1}(\alpha) \cap K \cap P$$

où P est un langage de parenthèses, K un langage régulier et h un homomorphisme alphabétique. C'est ce *problème fondamental de l'analyse syntaxique* qui est à étudier. Nous nous bornerons à décrire grossièrement son étude.

Les mots de K sont engendrés par un automate fini; un mot de P est engendré grâce à une pile; l'usage simultané d'un automate fini et d'une mémoire-pile permet d'engendrer les mots de l'intersection de K et P ; on a ainsi un modèle simple d'un *automate à pile* [6], qui peut être perfectionné en mêlant le fonctionnement de l'automate fini à celui de la pile pour rendre la génération plus efficace en évitant des fausses pistes. De plus, cet automate est pourvu d'une donnée, le mot α ; au fur et à mesure que sortent les lettres d'un mot engendré, on doit vérifier que son image par h est α , ce qui est facile puisque l'homomorphisme h est alphabétique : lorsque h transforme la lettre qui sort en une lettre, on compare cette dernière à la première lettre non encore lue dans la donnée α . La machine ainsi obtenue comme modèle d'un algorithme d'analyse syntaxique est un *transducteur à pile* [6]; ici encore, son fonctionnement peut être rendu plus efficace en lisant à l'avance un certain nombre de caractères de la donnée α pour prévoir et éviter des fausses pistes.

L'autre face d'une théorie de l'analyse syntaxique est l'étude des modes d'analyse. On peut appeler *mode d'analyse* (de gauche à droite), pour une grammaire G d'alphabet V , un couple d'une loi de composition interne et d'une loi de composition externe faisant du langage de parenthèses P un binoïde universel sur V , tel que, dans ce binoïde universel, $\mathcal{R}(G)$ soit l'intersection de P et d'un langage régulier et que la restriction à $\mathcal{R}(G)$ de l'application φ (mot des feuilles) soit aussi celle d'un homomorphisme alphabétique. Nous avons présenté plusieurs modes d'analyse; on peut se poser le problème, pour une grammaire G , de trouver tous ses modes d'analyse, de les comparer et de les classer.

REFERENCES

- [1] C. CHEVALLEY, *Fundamental concepts of algebra*, Academic Press, New York (1956).
- [2] N. CHOMSKY, *On certain formal properties of grammars*, Information and Control 2 (1959), p. 137-167.
- [3] N. CHOMSKY, *Aspects of the theory of syntax*, The M.I.T. Press, Cambridge (1965).
- [4] N. CHOMSKY et M. P. SCHUTZENBERGER, *The algebraic theory of context-free languages. Computer programming and formal systems* (Braffort, Hirschberg, editors), North Holland, Amsterdam (1963), p. 118-161.
- [5] K. CULIK, *On some transformations in context-free grammars and languages*, Czechoslovak Mathematical Journal 17 (1967), p. 278-311.
- [6] S. GINSBURG, *The mathematical theory of context-free languages*, McGraw-Hill, New York (1966).
- [7] S. KUNO et O. OETTINGER, *Multiple-path syntactic analyzer*, Information Processing 62 (Popplewell, editor), North Holland, Amsterdam (1963), p. 306-312.
- [8] R. MCNAUGHTON, *Parenthesis grammars*, Journal of the ACM 14 (1967), p. 490-500.
- [9] J. MYHILL, *Finite automata and the representation of events*, Wright Air Development Command Technical Report, 57-624 (1957), p. 112-137.
- [10] M. NIVAT, *Transductions des langages de Chomsky*. Thèse, Annales de l'Institut Fourier, Université de Grenoble (1968).
- [11] C. PAIR, *Étude de la notion de pile, application à l'analyse syntaxique*. Thèse, Faculté des Sciences de l'Université de Nancy (1965).
- [12] C. PAIR et A. QUERE, *Définition et étude des bilangages réguliers*, Information and Control 13 (1968), p. 565-593.
- [13] R. PETER, *Die Pair-Quere-schen freien Binoïden als Spezialfälle der angeordneten freien holomorphen Mengen*, Eötvös Lorand Universität Budapest (1968).
- [14] M. O. RABIN et D. SCOTT, *Finite automata and their decision problems*, IBM Journal of Research and Development 3 (1959), p. 114-125.
- [15] M. P. SCHUTZENBERGER, *On context-free languages and pushdown automata*, Information and Control 6 (1963), p. 246-264.
- [16] S. H. UNGER, *On syntax directed translators*, R. C.A. Laboratories (1963).