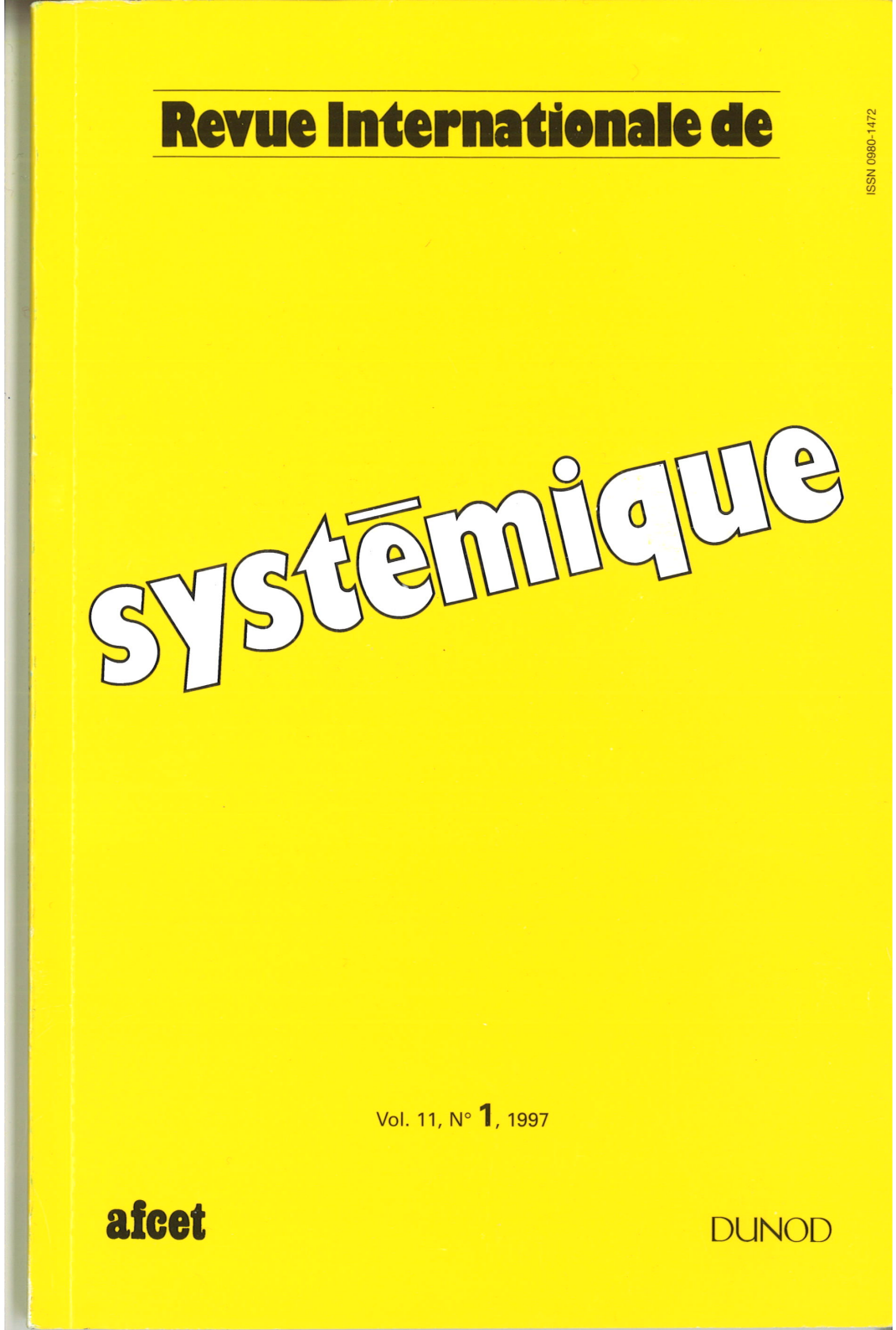

AFSCET

Revue Internationale de systēmique

Revue
Internationale
de Sytémique

volume 11, numéro 1, pages 49 - 67, 1997

Un calcul de préférence en syntaxe

Marcel Cori et Jean-Marie Marandin

Numérisation Afscet, mars 2016.

REVUE INTERNATIONALE DE SYSTÉMIQUE
Vol. 11, N° 1, 1997, pp. 49 à 67

UN CALCUL DE PRÉFÉRENCE EN SYNTAXE

Marcel CORI [1] et Jean-Marie MARANDIN [2]

Résumé
L'objet de cet article est la résolution des ambiguïtés locales en langue naturelle. On admet que cette résolution s'effectue à l'aide de préférences, et on traite un cas de préférence au niveau syntaxique : le fait qu'une structure canonique est préférée à une structure non canonique. La dichotomie est illustrée par des exemples en français; elle est représentée dans le formalisme des Grammaires d'arbres polychromes (GAP) qui est brièvement introduit. Après avoir défini ce qu'est la représentation d'un énoncé, on définit une relation de préférence entre les différentes représentations d'un énoncé donné. On montre comment calculer la représentation préférée : le calcul se fonde sur la composition d'arbres élémentaires et utilise l'opposition entre la nature canonique ou non canonique des sous-arbres.

Abstract
The paper addresses the problem of uncertainty bound to local structural ambiguity in Natural Language. Assuming that local ambiguity resolution involves preferences, we deal with only one case of preference on the syntactic level: the preference between canonical *vs* non canonical phrases. The dichotomy is illustrated with examples in French; it is represented in the framework of Polychrome tree grammars (PTG) which is briefly outlined. We define a notion of representation of an utterance and a preference relation between representations of a given utterance. We show how to compute the preferred representation: the computation operates on combinations of elementary trees and exploits the canonical *vs* non canonical nature of the subtrees.

1. Université Paris-VII, Case 7003, 2, place Jussieu, 75251 Paris Cedex 05, mco@ccr.jussieu.fr.
2. CNRS URA 1028, Case 7003, 2, place Jussieu, 75251 Paris Cedex 05, marandin@linguist.jussieu.fr

Rev. intern. systémique. 0980-1472 Vol. 11/97/01/ $ 7.00/

I. INTRODUCTION

Il existe un cas particulier d'incertitude dans le traitement automatique des langues : l'incertitude provoquée par des ambiguïtés syntaxiques. Une portion d'énoncé est dite ambiguë quand différentes analyses syntaxiques peuvent être produites à partir d'une suite de formes lexicales. Pour traiter le problème, il a été proposé d'ordonner les analyses selon leur probabilité. Les analyseurs se sont fondés sur trois types de principes : (a) principes relevant de la perception, (b) calculs statistiques, (c) principes d'ordonnancement intrinsèquement syntaxiques ("structural ranking") [1].

Dans cet article, nous considérons un cas d'ambiguïté que l'on peut clairement traiter par une préférence intrinsèquement syntaxique : cette préférence met en jeu la distinction entre syntagmes canoniques et syntagmes non canoniques.

Nous étudions d'abord la corrélation entre syntagmes canoniques et syntagmes préférés (§ 2). Nous analysons ensuite (§ 3) la différence entre les syntagmes canoniques et les syntagmes non canoniques. Au § 4 nous introduisons les grammaires d'arbres polychromes qui constituent un cadre adéquat pour exprimer cette analyse. Enfin (§ 5), nous montrons comment calculer la préférence sur des structures complexes obtenues par la composition d'arbres élémentaires canoniques et non canoniques.

II. LES DONNÉES

II.1. Syntagmes canoniques et syntagmes non canoniques

Considérons les exemples suivants :

(1) a. Les rayonnements magnétiques perturbent *les électriques.*

b. Il a mangé *les pourries.*

c. *Le parler vrai* du ministre lui a causé des ennuis.

(2) a. Il a un veston *très sport.*

b. Paul est *très sieste.*

(3) a. *Que tu viennes* m'ennuie.

b. *Le frapper* pourrait nous valoir des ennuis.

Les syntagmes en italique partagent une caractéristique : « un constituant de catégorie X apparaît alors que l'on attend un constituant de catégorie Y ». Il n'y a pas de nom dans les syntagmes nominaux de (1), mais un adjectif (*électrique, pourries*) ou un verbe à l'infinitif (*parler*). Dans (2), un nom (*sport, sieste*) apparaît, quand on attend un adjectif. De la même manière, une proposition apparaît en (3) au lieu d'un syntagme nominal. Nous appellerons *syntagmes non canoniques* les syntagmes avec un constituant non attendu.

II.2. Syntagmes préférés

La distinction entre syntagmes canoniques et syntagmes non canoniques est corrélée à une préférence. Considérons les exemples suivants dans lesquels les formes lexicales sont ambiguës quant à leur catégorie :

(4) Il a mangé *les mûres* (mûres : N ou A).

(5) Il est *très calme* (calme : A ou N).

(6) *Le manger cru* pourrait avoir des vertus thérapeutiques (le : pronom ou déterminant).

Quand les énoncés (4)-(6) sont interprétés isolément, l'interprétation fondée sur la structure canonique est préférée à l'interprétation fondée sur la structure non canonique.

L'énoncé (4) est interprété en analysant *mûres* comme un nom et non comme un adjectif. De même, l'interprétation qui sera préférée pour (5) prendra *calme* pour un adjectif et non pour un nom. En (6), *le manger cru* est un syntagme nominal plutôt qu'un syntagme verbal, bien que *le* puisse être analysé comme un pronom à l'accusatif, *manger* étant un verbe transitif.

On notera que l'interprétation fondée sur une structure non canonique s'obtient sans difficulté quand la tête lexicale du syntagme ne présente pas d'ambiguïté catégorielle (comparer (4) et (1.b), (5) et (2.b)), ou quand le verbe est intransitif (comparer (6) et (1.c)).

III. ANALYSE DES SYNTAGMES NON CANONIQUES

III.1. Travaux antérieurs

De très nombreuses études ont été consacrées à l'analyse des syntagmes non canoniques. Ces syntagmes ont été considérés comme le produit d'une ellipse (tradition grammaticale), d'une translation (Corblin, 1991), (Tesnière, 1959) ou d'une distorsion entre deux modules (Milner, 1989). Dans les grammaires X-bar, certains syntagmes non canoniques sont analysés comme des structures régulières comportant une catégorie vide. Par exemple, le GN

de (1.a) est analysé comme :

$$[_{GN}\ [_{DET}\ \text{les}]\ [_{N}\ \text{e}]\ [_{GA}\ \text{électriques}]].$$

Nous n'entendons pas présenter ici une critique complète de chacune des analyses; nous nous contentons de présenter notre propre analyse.

III.2. L'analyse

III.2.1. *Analyse du syntagme nominal*

Considérons les syntagmes nominaux de (1). Nous faisons l'hypothèse que l'adjectif ou le verbe occupe la position noyau du GN. L'argument majeur sur lequel se fonde cette analyse est le suivant : la constitution interne du GN et ses relations externes dépendent de propriétés de la tête lexicale (adjectif ou verbe). Nous renvoyons le lecteur à Kerleroux (1990, 1991) et Marandin (à paraître) pour une analyse détaillée. Tirons en les conséquences pour l'analyse de la non-canonicité :

– (i) il n'y a pas de différence structurelle entre un GN qui admet une tête nominale et un GN qui admet une tête verbale ou adjectivale : N, A ou V apparaissent exactement dans la même position. Ceci peut être illustré par les arbres suivants (Figure 1).

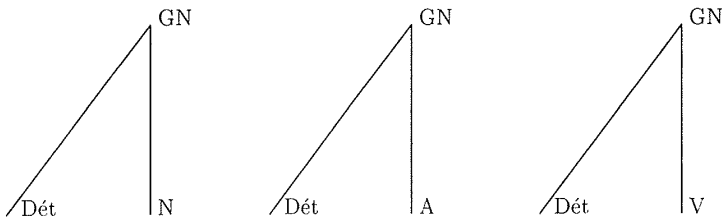

Figure 1.

– (ii) le contraste entre les syntagmes canoniques et les syntagmes non canoniques n'est pas corrélé à une différence de structure, il est lié au fait qu'un N est plus naturel qu'un A ou un V dans la position noyau d'un GN[2]. Le contraste se reflète uniquement dans la syntaxe.

III.2.2. *Analyse du syntagme adjectival et de la proposition*

Nous proposons une analyse similaire pour les syntagmes adjectivaux.

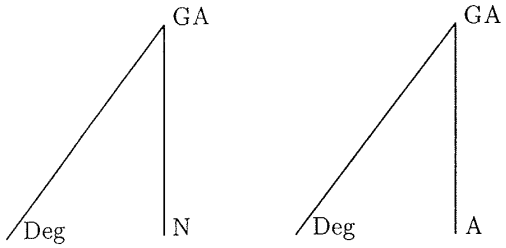

Figure 2.

Un N peut occuper la position noyau d'un syntagme adjectival. L'argument majeur sur lequel se fonde cette analyse est qu'une règle de conversion "N → A" (appartenant à la composante morphologique du français) serait *ad hoc*.

De façon analogue, pour S nous admettons l'analyse suivante [3] :

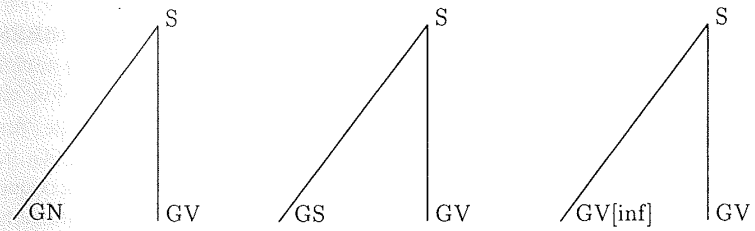

Figure 3.

Dans tous ces cas, le contraste entre syntagmes canoniques et syntagmes non canoniques n'est pas corrélé avec une différence de structure. On peut juste dire qu'un A est plus naturel qu'un N dans la position noyau de GA et qu'un GN est plus naturel qu'un GS ou un GV dans la position sujet.

III.3. Représentation

Les conséquences pour la représentation sont les suivantes :

– (i) les positions doivent être définies sans référence à la catégorie du constituant qui les occupent ; la position noyau ne fait pas exception (Cori et Marandin, 1993). Ceci est l'une des principales motivations pour l'introduction de couleurs dans la définition des arbres syntaxiques. Par exemple, la position noyau du GN est définie comme étant $\langle \mathrm{GN}, 3 \rangle$, où 3 désigne une couleur.

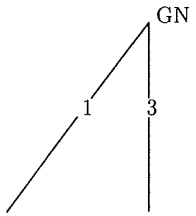

Figure 4.

– (ii) le fait que certaines tournures soient plus naturelles que d'autres doit être pris en compte par la grammaire. C'est pourquoi les ensembles d'arbres élémentaires qui constitueront nos grammaires seront partitionnés en deux sous-ensembles [4].

IV. ARBRES POLYCHROMES ET GRAMMAIRES D'ARBRES POLYCHROMES

Les grammaires d'arbres polychromes appartiennent à la famille des grammaires syntagmatiques généralisées (telles que GPSG, HPSG, LFG...). Leur principale originalité réside dans la définition de la dimension syntaxique où les positions sont définies indépendamment de la catégorie des constituants qui les occupent.

IV.1. Arbres polychromes

Notations : X^* est le monoïde libre sur X, c'est-à-dire l'ensemble des suites finies composées d'éléments de X ; ε est la suite vide (mot vide du monoïde libre).

IV.1.1. Définition : Soit p un entier strictement positif. Soit *Cat* un ensemble, l'ensemble des catégories. Un arbre polychrome est défini comme étant un triplet $A = \langle X, \delta, L \rangle$, où :

(i) X est un ensemble fini non vide, l'ensemble des sommets,

(ii) $L : X \rightarrow Cat$ est une application qui attribue une étiquette à chaque sommet [5],

(iii) $\delta = \langle \delta_1, \delta_2, ..., \delta_p \rangle$ est une suite d'applications : $\delta_i : X \rightarrow X^*$,

(iv) $\langle X, \delta_{1p} \rangle$ est un arbre [6].

p est le nombre des *couleurs* : si, pour un x donné, y a une occurrence dans $\delta_i(x)$, cela signifie qu'il y a un arc $\langle x, y \rangle$ de l'arbre auquel la couleur i est attribuée. L'arbre $\langle X, \delta_{1p} \rangle$ est, par conséquent, un arbre « monochrome », obtenu en « oubliant » les couleurs de l'arbre polychrome. C'est un arbre ordonné, en ce sens que l'ordre entre les fils d'un sommet donné est significatif. Nous revenons plus en détail sur cette question d'ordre en IV.1.3.

IV.1.2. Étant donné un arbre polychrome $A = \langle X, \delta, L \rangle$, on définit l'application $\delta^* : X \rightarrow \mathcal{P}(X)$ par :

(i) $\forall x \in X \; x \in \delta^*(x)$

(ii) $\forall x, y, z \in X \; \forall i \in \{1, ..., p\} \; x \in \delta_i(y)$ et $y \in \delta^*(z) \Rightarrow x \in \delta^*(z)$

$\delta^*(x)$ est l'ensemble des *descendants* d'un sommet x donné, qui inclut le sommet x lui-même.

IV.1.3. Les couleurs sont ordonnées, selon l'ordre entre les nombres entiers qui les désignent. Par ailleurs, les fils reliés à un sommet père donné par des arcs de même couleur sont ordonnés. En effet, $\delta_i(x)$ est une *suite* de sommets. On peut par conséquent définir un ordre strict sur tous les fils d'un même sommet, et induire un ordre sur tous les sommets d'un arbre. Cet ordre de *précédence* sera un ordre partiel strict.

Étant donné un arbre polychrome $A = \langle X, \delta, L \rangle$, l'ordre de précédence P est défini comme suit :

(i) si $\delta_{1p}(x) = x_1 ... x_q$ et $i < j$ alors $\langle x_i, x_j \rangle \in P$

(ii) si $x \in \delta^*(u)$, $y \in \delta^*(v)$ et $\langle u, v \rangle \in P$ alors $\langle x, y \rangle \in P$

IV.1.4. Les *feuilles* d'un arbre polychrome $A = \langle X, \delta, L \rangle$ sont tous les sommets qui n'ont pas de fils, autrement dit tous les sommets z tels que $\delta_{1p}(z) = \varepsilon$.

L'ordre de précédence induit un ordre total sur les feuilles d'un arbre.

feuilles $(A) = z_1 z_2 ... z_s$ est la suite formée de toutes les feuilles de A selon l'ordre de précédence.

La *racine* de A est l'unique sommet de l'arbre qui n'est fils d'aucun autre sommet, autrement dit le sommet x tel que $\delta^*(x) = X$.

IV.1.5. On prend $p = 5$. En Figure 5, on présente un exemple d'arbre polychrome qui représente l'énoncé (1.a).

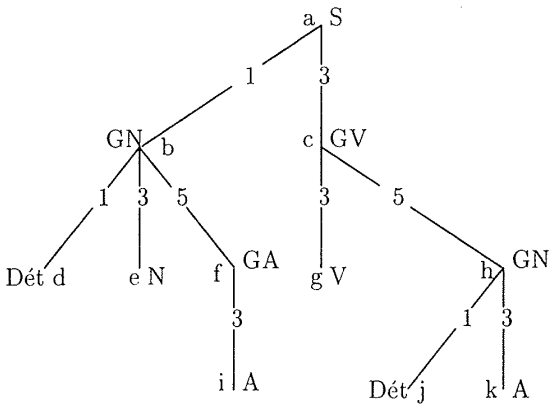

Figure 5.

IV.1.6. Étant donnés deux arbres polychromes $A = \langle X, \delta, L\rangle$ et $A' = \langle X', \delta', L'\rangle$, et une bijection $f : X \to X'$, A est *isomorphe* à A' si :

(i) $\forall x \in X \;\; L'(f(x)) = L(x)$,

(ii) $\forall x \in X \;\forall i \in \{1, ..., p\} \;\; \delta_i(x) = x_1 x_2 ... x_q$

$\Rightarrow \delta'_i(f(x)) = f(x_1) f(x_2) ... f(x_q)$.

Dans ce qui suit, on considèrera les classes d'équivalence d'arbres modulo la relation d'isomorphisme. Si A est isomorphe à A', on pourra écrire $A = A'$. Ceci revient à dire qu'on ne distingue pas deux arbres qui ne diffèrent que par les noms des sommets.

IV.2. La composition des arbres

IV.2.1. Composabilité : Soient $A_1 = \langle X_1, \delta_1, L_1\rangle$ et $A_2 = \langle X_2, \delta_2, L_2\rangle$ deux arbres polychromes tels que $X_1 \cap X_2 = \varnothing$; soit a la racine de A_2 et b une feuille de A_1 [7].

A_1 et A_2 sont composables selon b si et seulement si $L_1(a) = L_2(b)$.

C'est-à-dire que deux arbres sont composables si la racine de l'un reçoit la même étiquette catégorielle qu'une feuille de l'autre.

IV.2.2. Composition : La composition s'effectue en « fusionnant » ces deux sommets.

Formellement, l'arbre composé selon b de A_1 et A_2 est l'arbre polychrome $A = \langle X, \delta, L\rangle$ (écrit $[A_1, b, A_2]$) vérifiant :

(i) $X = X_1 \cup (X_2 - \{a\})$,

(ii) $\forall x \in X_1 - \{b\} \quad \delta(x) = \delta_1(x)$

$\forall x \in X_2 - \{a\} \quad \delta(x) = \delta_2(x)$

$\delta(b) = \delta_2(a)$,

(iii) $\forall x \in X_1 \quad L(x) = L_1(x)$

$\forall x \in X_2 - \{a\} \quad L(x) = L_2(x)$.

Il a été montré dans Cori et Marandin (1993, 1994) que l'ordre de composition des arbres n'est pas significatif [8].

IV.3. Grammaires d'arbres

IV.3.1. Définition : Un arbre polychrome est *pauvre* si et seulement si pour tout x et pour toute couleur i, $\delta_i(x)$ a une longueur égale à 0 ou à 1.

De chaque sommet, il ne part qu'un arc d'une couleur donnée.

IV.3.2. Définition : Un arbre polychrome est *élementaire* si et seulement si c'est un arbre pauvre dont un seul sommet (la racine) a au moins un fils.

C'est un arbre de profondeur 1.

IV.3.3. Une *grammaire d'arbres polychrome (GAP)* est donnée par un ensemble fini d'arbres élémentaires :

$$G = \{A_1, A_2, ..., A_m\}$$

IV.3.4. Une grammaire G *engendre* un ensemble d'arbres polychromes, par composition. Cet ensemble, noté $T(G)$, est défini par :

(i) si A appartient à G, alors A appartient à $T(G)$;

(ii) si A appartient à $T(G)$ et A' à G, alors chaque $[A, b, A']$ appartient à $T(G)$.

Propriété : Les arbres de $T(G)$ sont des arbres polychromes pauvres.

IV.3.5. Exemple de grammaire : Les arbres de la grammaire sont donnés en Figure 6.

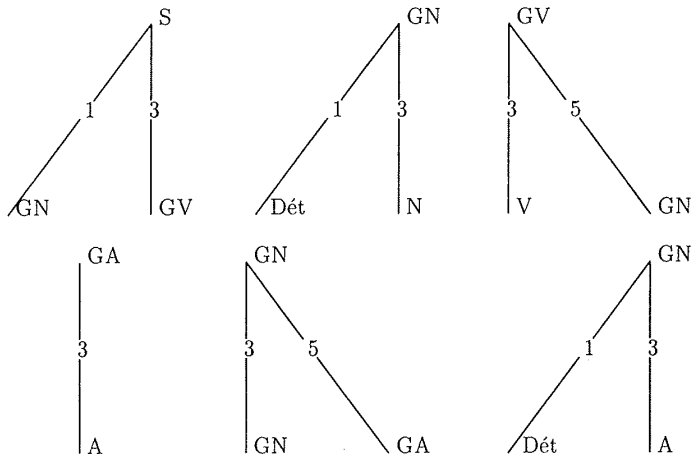

Figure 6.

Cette grammaire engendre [9] notamment l'arbre de la Figure 5.

V. PRÉFÉRENCE EN SYNTAXE

V.1. Préliminaires

V.1.1. Soit $A = \langle X, \delta, L\rangle$ un arbre polychrome :

– pour chaque sommet $x \in X$, $A(x)$ est le sous-arbre dont la racine est x, c'est-à-dire le sous-arbre dont l'ensemble des sommets est $\delta^*(x)$;

– pour chaque sommet $x \in X - \textit{feuilles}\ (A)$, $A[x]$ est le sous-arbre élémentaire dont la racine est x, autrement dit le sous-arbre dont l'ensemble des sommets est $\{x\} \cup \delta_{1p}(x)$.

Propriété : Étant donnée une GAP G, pour chaque arbre polychrome $A = \langle X, \delta, L\rangle$ de $T(G)$, on a :

$$\forall x \in X - \textit{feuilles}\,(A) \quad A[x] \in G$$

Cette propriété garantit que les arbres complexes se décomposent en arbres élémentaires qui appartiennent à la grammaire.

V.1.2. Une *forêt polychrome* est définie comme étant un quadruplet $A = \langle X, \delta, L, R\rangle$, où X, δ et L reçoivent la même définition qu'en 4.1.1, $R = a_1 ... a_m$ est une suite formée de tous les sommets de X tels que

$$\forall i \forall y \in X \quad a_i \notin \delta_{1p}(y)$$

et tous les $A(a_i)$ sont des arbres polychromes disjoints.

Autrement dit, une forêt est une suite d'arbres polychromes. La forêt n'a pas une seule racine, mais plusieurs qui sont ordonnées dans la suite R.

On obtient encore un ordre total sur les feuilles en écrivant :

$$\textit{feuilles}\,(A) = \textit{feuilles}\,(A(a_1)) ... \textit{feuilles}\,(A(a_m))$$

V.1.3. Étant donné un arbre $A = \langle X, \delta, L\rangle$ ou une forêt $A = \langle X, \delta, L, R\rangle$, pour chaque sommet $x \in X$ on définit la *portée* de x comme étant un ensemble de nombres entiers consécutifs. Si $\textit{feuilles}\,(A) = z_1 ... z_n$ et $\textit{feuilles}\,(A(x)) = z_i ... z_k$, cet ensemble se définit par : $\textit{portée}\,(x) = \{i, \ldots, k\}$.

La portée est l'ensemble des feuilles qui sont « vues » d'un sommet, autrement dit l'ensemble des feuilles qui sont des descendants de ce sommet. On décrit cet ensemble par les indices des feuilles afin de pouvoir comparer les portées de deux sommets dans deux arbres différents (mais qui comportent un même nombre de feuilles).

V.1.4. Si $A = \langle X, \delta, L, R\rangle$ est une forêt avec $R = a_1 ... a_m$, la définition précédente peut être étendue à une suite $a_r ... a_s$ extraite de R :

$$\textit{portée}\,(a_r ... a_s) = \textit{portée}\,(a_r) \cup ... \cup \textit{portée}\,(a_s)$$

On obtient encore un ensemble de nombres entiers consécutifs.

V.1.5. Exemple : Une forêt polychrome [10] est donnée en Figure 7 avec $R = abcde$.

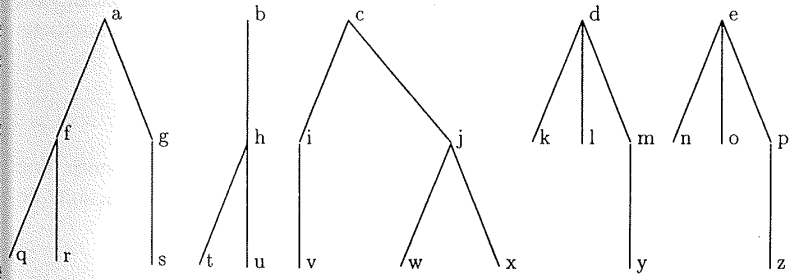

Figure 7.

On a :

$feuilles\,(A) = q\,r\,s\,t\,u\,v\,w\,x\,k\,l\,y\,n\,o\,z$

$feuilles\,(A\,(b)) = t\,u$

$feuilles\,(A\,(c)) = v\,w\,x$

$feuilles\,(A\,(d)) = k\,l\,y$

et, par conséquent :

$portée\,(b\ c\ d) = \{4, 5, 6, 7, 8, 9, 10, 11\}$

V.2. Représentation d'énoncés

V.2.1. Soit U un ensemble de formes lexicales. Une fonction de catégorisation associe un ensemble de catégories à chaque forme lexicale :

$$cat : \quad U \rightarrow \mathcal{P}\,(Cat)$$

Par exemple, l'ensemble {N, A} est associé à la forme lexicale *calme* (exemple (5)).

V.2.2. Soit G une GAP et *cat* une fonction de catégorisation. Soit $u_1\,u_2 ... u_n \in U^*$ une suite de formes lexicales; une représentation (avant compactage) de cette suite est donnée par un arbre polychrome $A = \langle X, \delta, L\rangle$ appartenant à $T\,(G)$ tel que :

$$\forall\, i \leq n \quad L\,(z_i) \in \mathrm{cat}\,(u_i)$$

avec $feuilles\,(A) = z_1\,z_2 ... z_n$.

Chaque feuille de l'arbre correspond à une forme lexicale de la suite. L'étiquette de la feuille est une catégorie légitime de la forme lexicale.

V.2.3. Une *représentation généralisée* d'une suite de formes lexicales est donnée par une forêt polychrome $A = \langle X, \delta, L, R\rangle$ qui vérifie les conditions ci-dessus et telle que, si $R = a_1 ... a_m$,

$$\forall\, i \quad a_i \notin feuilles\,(A) \Rightarrow A\,(a_i) \in T\,(G)$$

On définit, à des fins techniques, pour une forêt polychrome donnée, une application $\gamma : X \rightarrow X^*$ qui associe à chaque sommet de la forêt la suite de ses fils si elle n'est pas vide, et lui-même sinon. Formellement :

(i) $a_i \in feuilles\,(A) \Rightarrow \gamma\,(a_i) = a_i$,

(ii) $a_i \notin feuilles\,(A) \Rightarrow \gamma\,(a_i) = \delta_{1p}\,(a_i)$.

Si $A = \langle X, \delta, L, R\rangle$ est la représentation généralisée d'une suite u, la *sous-représentation immédiate* de la suite lexicale est la forêt polychrome $\sigma\,(A) = \langle X', \delta', L', R'\rangle$, où

(i) $X' = X - (R - feuilles\,(A))$,

(ii) δ' et L' sont les restrictions de δ et L à X',

(iii) $R' = \gamma\,(a_1) ... \gamma\,(a_m)$.

La sous-représentation immédiate d'une suite est encore une représentation généralisée de cette suite. Elle est obtenue en supprimant un niveau de la forêt, par le haut.

V.2.4. Un exemple de représentation généralisée

La Figure 8*a* fournit une représentation généralisée de l'énoncé (1a). La sous-représentation immédiate est donnée en 8*b* et toutes les sous-représentations suivantes reviennent au graphe dégénéré de la Figure 8*c*.

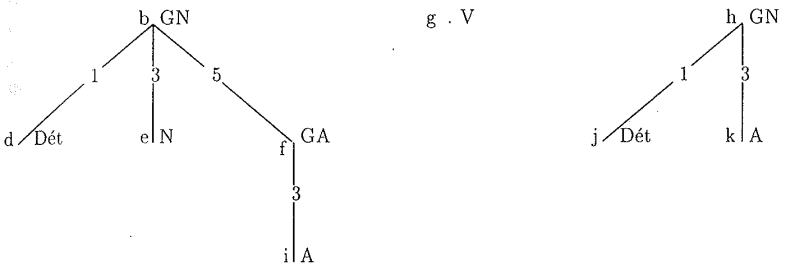

Figure 8a.

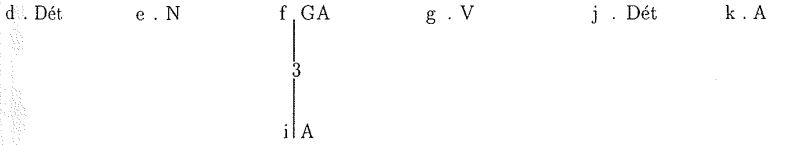

Figure 8b.

Figure 8c.

V.3. Nouvelle définition des grammaires

V.3.1. Une grammaire est constituée de deux sous-ensembles disjoints, C l'ensemble des structures canoniques et N l'ensemble des structures non canoniques : $G = C \cup N$.

V.3.2. Exemple de grammaire

Les arbres de C apparaissent en Figure 9*a*; les arbres de N en 9*b*.

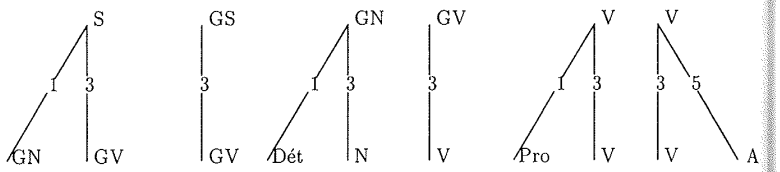

Figure 9a.

Figure 9b.

V.4. Calcul de préférence

V.4.1. Étant donné un ensemble U, une grammaire G, une fonction de catégorisation *cat* et une suite u de U^*, on définit une relation de préférence entre des représentations (généralisées) de u.

Soient $A_1 = \langle X_1, \delta_1, L_2, R_1 \rangle$ et $A_2 = \langle X_2, \delta_2, L_2, R_2 \rangle$ deux représentations de u; soient $R_1 = a_1 ... a_m$ et $R_2 = b_1 ... b_q$.

On dit que A_1 est *immédiatement préférée* à A_2, et on écrit $A_1 \geq A_2$, si et seulement si pour tout a_i appartenant à R_1 tel que $A_1[a_i]$ appartienne à N, il existe une suite $b_s ... b_t$ extraite de R_2 telle que :

(i) $portée(a_i) \subset portée(b_s ... b_t)$,

(ii) $\forall j \in \{s, ..., t\}$ $A_2[b_j] \in N$.

On dit que A_1 est *préférée* à A_2, et on écrit $A_1 \gtrdot A_2$, si et seulement si l'une des trois conditions suivantes est vérifiée :

(i) $\forall x \in X_1 - feuilles(A_1) \quad A_1[x] \notin N$,

(ii) $A_1 \geq A_2$ et non $A_2 \geq A_1$,

(iii) $A_1 \geq A_2$ et $A_2 \geq A_1$ et $\sigma(A_1) \gtrdot \sigma(A_2)$.

V.4.2. Propriété : La relation de préférence est une relation de préordre.

Preuve : Une relation de préordre est une relation qui est réflexive et transitive.

La préférence immédiate est, de manière évidente, réflexive. Étant donné un arbre A, il existe nécessairement un n tel que $\sigma^{(n)}(A)$ vérifie :

$$\forall x \in X - feuilles(\sigma^{(n)}(A)) \quad A[x] \notin N$$

(en considérant que $\sigma^{(0)}(A) = A$). Ce qui démontre, en vertu des conditions (i) et (iii), la réflexivité de la relation de préférence.

Démontrons à présent la transitivité de la préférence immédiate. Supposons que nous ayons $A_1 \geq A_2$ et $A_2 \geq A_3$, et supposons $A_3 = \langle X_3, \delta_3, L_3, R_3 \rangle$ avec $R_3 = c_1 ... c_q$. Pour chaque i tel que $A_1[a_i] \in N$ et pour chaque $j \in \{s, ..., t\}$ il existe une suite $c_l ... c_r$ telle que $portée(a_i) \subset portée(c_l ... c_r)$ et

$$\forall k \in \{l, ..., r\} \quad A_3[c_k] \in N$$

La transitivité de la préférence est démontrée par induction sur la profondeur des représentations.

Supposons que nous ayons $A_1 \gtrdot A_2$ et $A_2 \gtrdot A_3$.

(1) Quand $A_1 \gtrdot A_2$ d'après (i), on a aussi $A_1 \gtrdot A_3$. (A_1 est préférée, au sens large, à toute autre représentation généralisée.)

(2) Quand $A_1 \gtrdot A_2$ d'après (ii), on ne peut avoir $A_2 \gtrdot A_3$ d'après (i), car sinon on aurait aussi $A_2 \geq A_1$.

Si $A_2 \geq A_3$ (cas (ii) ou (iii) de $A_2 \gtrdot A_3$), alors $A_1 \geq A_3$ d'après la transitivité de la préférence immédiate, et on n'a pas $A_3 \geq A_1$ qui aurait impliqué $A_2 \geq A_1$. On a donc $A_1 \gtrdot A_3$ d'après (ii).

(3) Quand $A_1 \gtrdot A_2$ d'après (iii) :

(3.1) Si $A_2 \gtrdot A_3$ d'après (i), on a :

$$\forall x \in X_1 \quad A_1[x] \notin N$$

et donc $A_1 \gtrdot A_3$ d'après (i).

(3.2) Si $A_2 \gg A_3$ d'après (ii), on a $A_1 \geq A_3$ d'après la transitivité de la préférence immédiate, et pas $A_3 \geq A_1$ qui impliquerait $A_3 \geq A_2$. On a par conséquent $A_1 \gg A_3$ d'après (ii).

(3.3) Si $A_2 \gg A_3$ d'après (iii), on déduit, sur la base de la transitivité de la préférence immédiate et l'hypothèse de récurrence, que :

$$A_1 \geq A_3 \quad \text{et} \quad A_3 \geq A_1 \quad \text{et} \quad \sigma(A_1) \gg \sigma(A_3)$$

On a par conséquent $A_1 \gg A_3$ d'après (iii).

V.4.3. Exemples de calculs de préférence : Commençons par un exemple abstrait. La Figure 10 représente une grammaire, avec ses arbres canoniques et ses arbres non canoniques. Nous admettons que les arbres polychrome de la Figure 11 sont les représentations d'une même suite [11].

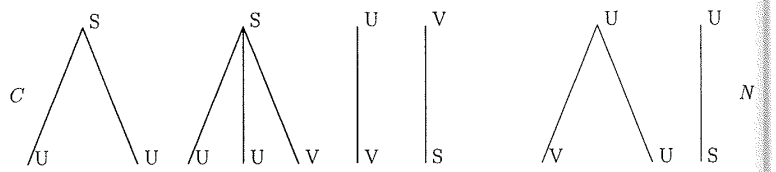

Figure 10.

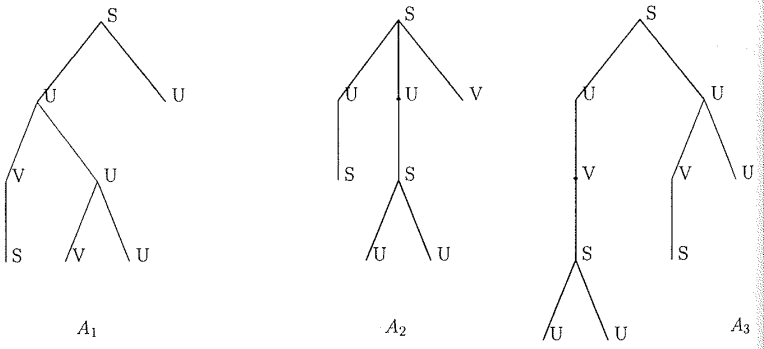

Figure 11.

On a :

$$A_1 \geq A_2 \quad \text{et} \quad A_2 \geq A_1 \quad \text{et} \quad A_2 \geq A_3 \quad \text{et} \quad A_3 \geq A_2.$$

$\sigma(A_1) \geq \sigma(A_2)$ et $\sigma(A_2) \geq \sigma(A_1)$, mais ni $\sigma(A_1) \geq \sigma(A_3)$, ni $\sigma(A_3) \geq \sigma(A_1)$. Il n'y a par conséquent pas de relation entre A_1 et A_3.

Cet exemple montre que la relation de préférence est un préordre partiel. $\sigma(\sigma((A_2))$ vérifie la condition (i) de V.4.1. D'où $\sigma(\sigma((A_2)) \gg \sigma(\sigma((A_1))$ et $A_2 \gg A_1$.

Considérons à présent les deux arbres (Figure 12) qui donnent les deux représentations possibles de l'énoncé *le manger cru aurait des vertus thérapeutiques* à l'aide de la grammaire donnée en V.3.2 (Figure 9).

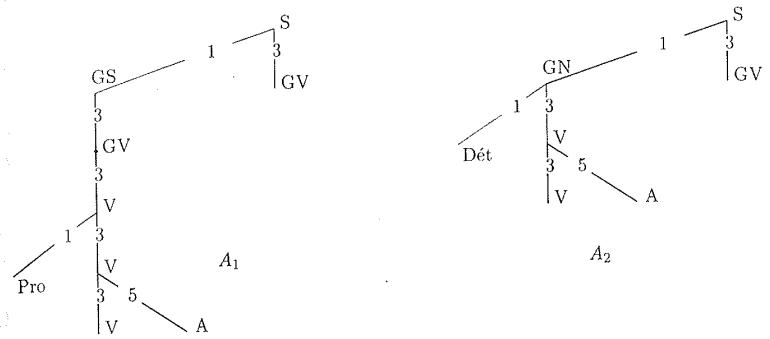

Figure 12.

L'arbre A_2 est préféré à l'arbre A_1 : cela provient de la condition (ii) de V.4.1. Intuitivement, cela signifie que la préférence liée à la canonicité des structures enchâssantes l'emporte sur la préférence liée à la canonicité des structures enchâssées.

VI. CONCLUSION

Reprenons les exemples (4) et (6) en les insérant dans des contextes explicites :

(7) a. Pierre a apporté les pommes. Marie a mangé les mûres, Paul a jeté les vertes.

b. Il n'y a plus de feu pour cuire le bifteck. Le manger cru pourrait avoir des vertus thérapeutiques.

On observe que l'interprétation préférée est celle qui est corrélée avec une interprétation anaphorique. En (7a), le GN *les mûres* dénote une partie

de l'ensemble introduit par *les pommes* : (7a) est interprété comme *elle a mangé les pommes mûres*. En (7b), *le* est analysé comme un pronom plutôt que comme un déterminant et le segment *le manger cru* est analysé comme un GV : *manger cru [un objet] pourrait avoir des vertus thérapeutiques*. Ceci peut s'expliquer par le « Principle of Referential Success » (Crain et Steedman, 1985) : « if there is a reading that succeeds in referring to an entity already established in the hearer's mental model of the domain of discourse, then it is favored over one that does not ». En (7), la lecture anaphorique (anaphore partitive en (7a) et anaphore pronominale en (7b)) correspond à l'analyse non canonique : la préférence au niveau syntaxique s'efface derrière la préférence au niveau discursif.

La résolution des ambiguïtés locales peut être vue comme la résolution d'un conflit entre préférences liées à différentes propriétés des énoncés. Ces préférences n'ont pas toutes le même poids. La préférence syntaxique s'applique quand la lecture anaphorique ne convient pas. Il semble que l'on ait là une caractérisation générale de toutes les préférences structurelles.

Notes et références

1. Entre autres Kimball (1973), Frazier et Ford (1978), Marcus (1980), Crain et Steedman (1985).

2. Nous laissons de côté l'explication de cette naturalité qui invoquerait une incompatibilité entre le type sémantique de l'unité lexicale attendue dans la position noyau et le type effectif du terme occupant.

3. GS est analogue au S′ de la notation X-bar.

4. Nous avons introduit le cas du syntagme nominal en détail. Une argumentation semblable peut être avancée pour la position noyau du syntagme adjectival ainsi que pour les configurations de positions de la phrase afin de rendre compte des exemples (2) et (3).

5. L'étiquetage est réduit à une catégorie dans cet article. Un étiquetage plus complexe, incluant des traits, est en réalité nécessaire. Voir Cori et Marandin (1993).

6. δ_{1p} est l'application de X dans X^* telle que $\forall x \in X\ \delta_{1p}(x) = \delta_1(x)\,\delta_2(x)...\delta_p(x)$.

7. Si la condition $X_1 \cap X_2 = \varnothing$ n'est pas vérifiée, les sommets de A_1 ou de A_2 peuvent être renommés afin qu'elle le soit.

8. Notre présentation de la composition des arbres est simplifiée ici; une présentation plus explicite doit inclure l'opération de compactage qui a pour effet de produire des structures plus « plates » en fonction de certaines conditions vérifiées par la position noyau. Nous renvoyons le lecteur à Cori et Marandin (1993).

9. L'opération de compactage est mise en jeu pour produire la structure plate du GN. Voir note 8.

10. Nous n'indiquons pas les couleurs sur cet exemple abstrait, dans la mesure où elles n'ont aucun effet sur la définition de la portée.

11. Nous omettons encore les couleurs, car elles n'ont aucun effet sur la relation de préférence.

F. CORBLIN, Les Éléments de syntaxe structurale, de L. TESNIÈRE, in *La grammaire française entre comparatisme et structuralisme,* H. HUOT ed., Paris : Colin, 1991.

M. CORI et J.-M. MARANDIN, Grammaires d'arbres polychromes, Paris : *TAL*, p. 34-1, 1993.

M. CORI et J.-M. MARANDIN, Polychrome Tree Grammars (PTG): a formal approach in *Current Issues in Mathematical Linguistics*, C. MARTIN-VIDE ed., Elsevier, 1994.

S. CRAIN et M. STEEDMAN, On not being led up the garden path: the use of context by the psychological syntax processor, D. DOWTY *et al.* eds., *Natural Language Parsing,* Cambridge: Cambridge UP, 1985.

L. FRAZIER et J. FORD, The sausage-machine: A new two-stage parsing model, *Cognition,* 6 : p. 291-325, 1978.

F. KERLEROUX, Du mode d'existence de l'infinitif substantivé en français contemporain, Cahiers de grammaire 15 : p. 57-99, Toulouse : U. de Toulouse-Le Mirail, 1990.

F. KERLEROUX, Il est d'un calme! et d'un élégant! Un phénomène de distorsion (...), *BULAG* 17: p. 79-116, Besançon : Université de Franche-Comté, 1991.

J. KIMBALL, Seven principles of surface structure parsing in natural language, *Cognition*, 2 : p. 15-47, 1973.

J.-M. MARANDIN, à paraître, Pas d'entité sans identité. L'analyse du groupe nominal Dét+A, *Mots et grammaires*, Paris : Klincksieck.

M. MARCUS, *A Theory of Syntactic Recognition for Natural Language*, Cambridge: MIT Press, 1980.

J.-C. MILNER, *Introduction à une science du langage,* Paris : Le Seuil, 1989.

L. TESNIÈRE, *Éléments de syntaxe structurale,* Paris : Klincksieck, 1959.