LES GRAMMAIRES DE CONSTITUANTS GENERAUX.

par R. D. Guedj.

Il est deux faits linguistiques d'une grande importance qui nous ont conduits à définir un nouveau type de grammaire formelle, les grammaires de constituants généraux (GCG):

- la présence dans les langues naturelles de mots non connexes
- la distinction entre le niveau de la composition et le niveau de l'expression dans la génération ou la reconnaissance d'une phrase.

0.1 . Mots non connexes.

Si nous voulons analyser une phrase aussi simple que

" the boy switches the light off "

En tout premier lieu nous pouvons dire que cette phrase est formée, dans l'ordre, d'un article, d'un substantif, d'un verbe, d'un article, d'un substantif, d'une postposition; Cependant il est bien clair que le verbe switches et la postposition off sont intimement liés pour former le verbe switches off, un tel mot sera appelé mot non connexe à deux insertions.

Les GCG vont nous permettre d'accepter des mots non connexes à plusieurs insertions comme des entités; c'est ainsi que l'on représente un mot à 2 insertions par le lien qui unit les deux parties qui le constituent:

(que l'on appellera aussi peigne à deux dents).

0.2. Expression et Composition.

Il est un fait essentiel sur lequel nous nous permettrons d'insister dès à présent. Dans la génération ou la reconnaissance d'une phrase nous distinguons deux niveaux: le niveau de l'expression et le niveau de la composition. Tout d'abord un exemple choisi hors du domaine linguistique nous permettra de rendre plus clair cette distinction. Lorsqu'une entreprise désire construire un immeuble de n appartements elle a, au préalable, constitué un stock de briques, de sable, de tuiles, de verre, de bois,... c'est ce stock de matériaux que nous appelons la composition de l'immeuble. Puis l'architecte utilise les éléments qui constituent ce stock pour bâtir un immeuble. Cependant il pourra disposer les appartements de manières différentes, et c'est la disposition des appartements que nous appelons l'expression de l'immeuble. Ainsi deux immeubles peuvent différer au niveau de la composition (par la nature ou l'importance du stock) et de plus ayant même compostions il peuvent différer par l'expression (par la disposition des appartements). Revenons à notre objet, en linguistique ces deux niveaux apparaissent clairement. Au niveau de la compostion nous nous bornons à énoncer les types syntaxiques nécessaires à la formation d'une phrase (remarquons que l'ordre dans lequel nous les disposons a peu d'importance ici). Au niveau de l'expression, nous commençons par donner des valeurs aux différents types syntaxiques de la liste fournie par la composition (ces valeurs pouvant être des mots à plusieurs insertions), ensuite nous indiquons la disposition relative de ces mots. Dans le cas de mots non connexes il s'agit d'imbriquer les peignes qui représentent ces mots, avec l'aide des "morphismes de peignes".

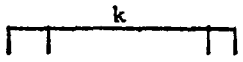
Dans cet article nous donnons la définition de la catégorie SEG, des arbres biordonnés et des GCG, cette dernière définition suppose connues les notions fondamentales de la théorie des catégories. Ensuite nous donnons une propriété algébrique des langages engendrés par les GCG, qui nous permettra de les comparer aux types de grammaires déjà existants.

§ 1. CATEGORIE SEG.

1.1. Simplexe. Soit $k \in \mathbb{N}$. On appellera simplexe k (noté $\rceil k \lceil$):

$$\rceil k \lceil = \{ 1, 2, \ldots, k \},$$

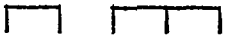
on le représente aussi par le peigne à k dents

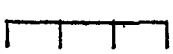


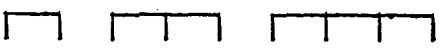
1.2. Simplexe segmenté. On appelle simplexe segmenté une suite de simplexes.

On appelle produit des simplexes segmentés $S_1, \ldots, S_n$ le simplexe segmenté représenté par le peigne obtenu en juxtaposant de gauche à droite les peignes qui représentent successivement $S_1 \ldots, S_n$ (que l'on note $S = \underset{i=1 \ldots n}{*} S_i$)

Ex:

S_1 : 

S_2 : 

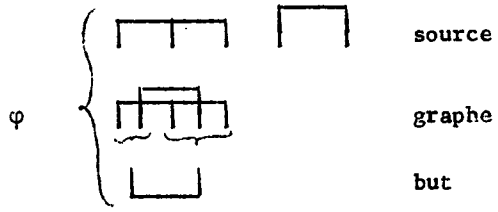
$S_1 * S_2$: 

1.3. Morphismes de peignes.

Soit un simplexe segmenté (source) un morphisme de peignes

assemble les peignes qui constituent ce simplexe segmenté (graphe pour construire un nouveau simplexe segmenté (but). Il nous faut donc préciser comment les peignes de la source sont assemblés pour former les peignes du but.

Ex: le morphisme φ :



la source . φ est le simplexe segmenté composé dans l'ordre d'un peigne n_1 à trois dents et d'un peigne n_2 à deux dents.

le graphe de φ indique que l'on insère la première dent du peigne n_2 entre la $1^{ère}$ et la $2^{ème}$ dent du peigne n_1 et que l'on insère la deuxième dent du peigne n_2 entre la $2^{ème}$ et la $3^{ème}$ dent du peigne n_1 .

le but de φ est un peigne à 2 dents, la première est formée, dans l'ordre, par la $1^{ère}$ de n_1 et la $1^{ère}$ de n_2 , la deuxième par la $2^{ème}$ de n_1 , la $2^{ème}$ de n_2 et la $3^{ème}$ de n_1 (on retourne le peigne du but pour marquer plus clairement la manière de l'obtenir).

1.4. Composition de morphismes.

Soit deux morphismes φ et π tels que source [illegible] but φ

on définit le morphisme ($\pi \circ \varphi$) de la manière suivante:

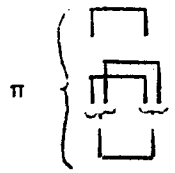
source ($\pi \circ \varphi$) = source φ

but ($\pi \circ \varphi$) = but φ

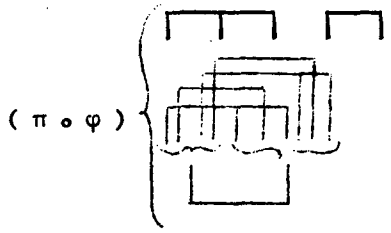
et graphe de ($\pi \circ \varphi$) est déduit de graphe φ et graphe π de la manière que rendra clair l'exemple suivant:

φ est le morphisme défini plus haut

π est le morphisme



alors ($\pi \circ \varphi$) est le morphisme:



1.5. Catégorie SEG.

La catégorie SEG est une catégorie à produit direct dont les objets sont les simplexes segmentés, les flèches sont les morphismes de peignes et la composition des flèches est la composition des morphismes de peignes.

§ 2. ARBRES BIORDONNÉS.

2.1. Définition.

Un arbre biordonné est un ensemble $\mathcal{A}$ fini muni de deux relations d'ordre: $\mathcal{H}$ (ordre hiérarchique) et $\mathcal{S}$ (ordre séquentiel) assujetties aux conditions suivantes:

A_1 : le prédécesseur immédiat pour $\mathcal{H}$ de tout élément de $\mathcal{A}$ lorsqu'il existe, est unique.

A_2 : $\mathcal{S}$ est une relation d'ordre total.

A_3 : $\forall$ A, B $\in \mathcal{A}$; (A,B) $\Longrightarrow$ $\mathcal{S}$(A,B)

A_4 : $\forall$ A, A', B, B' $\in \mathcal{A}$,

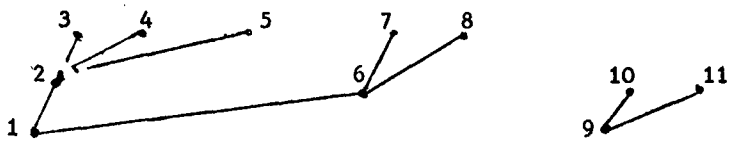
$$\left(\mathcal{S}(A,B),\ \mathcal{H}(A,A'),\ \mathcal{H}(B,B')\right) \longrightarrow \mathcal{S}(A',B')$$

Un élément sans prédécesseur pour $\mathcal{H}$ est appelé sommet de $\mathcal{A}$, un arbre biordonné qui admet plusieurs sommets est dit non connexe.

2.2. Listes.

Etant donné un ensemble fini (l'alphabet), on appelle liste sur Λ le couple $(\mathcal{A}, \mathcal{E})$ d'un arbre biordonné et d'une application étiquetage $\mathcal{E}$ de $\mathcal{A}$ dans le monoïde libre construit sur Λ.

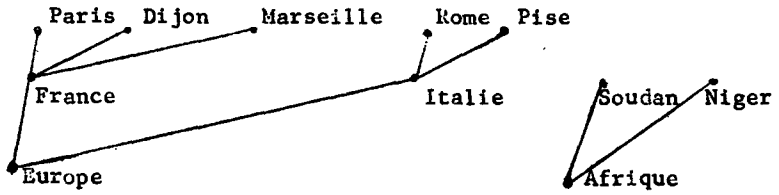
Exemple: Soit l'arbre biordonné suivant, les nombres attachés à chaque noeud sont leurs numéros dans l'ordre $\mathcal{S}$:



et soit l'alphabet Λ = {Europe, France, Italie, Afrique, Soudan, Niger, Rome, Pise, Dijon, Paris, Marseille}

et $\mathcal{E}(1)$ = Europe, $\mathcal{E}(2)$ = France, $\mathcal{E}(3)$ = Paris, $\mathcal{E}(4)$ = Dijon, $\mathcal{E}(5)$ = Marseille, $\mathcal{E}(6)$ = Italie, $\mathcal{E}(7)$ = Rome, $\mathcal{E}(8)$ = Pise, $\mathcal{E}(9)$ = Afrique, $\mathcal{E}(10)$ = Soudan, $\mathcal{E}(11)$ = Niger

On forme la liste $(\mathcal{A}, \mathcal{E})$:



§ 3. GRAMMAIRES DE CONSTITUANTS GENERAUX.

Une GCG est la donnée du 5-upple suivant:

$$\mathcal{G} = \{ \Lambda_N , \mathcal{R}, \mathcal{T}; [\Lambda_T] , \xi , \alpha' \}$$

les 3 premières composantes constituent la composition et les trois dernières contituent l'expression.

3.1. Composition: $C = (\Lambda_N , \mathcal{R}, \mathcal{T})$.

Λ_N : ensemble fini appelé alphabet nonterminal ou alphabet des catégories syntaxiques (dont les éléments sont notés par une capitale latine), Λ_N comprend un élément distingué, l'élément initial S.

$\mathcal{R}$: ensemble fini de règles syntaxiques r, muni de deux applications α source et β but :

$$\mathcal{R} \xrightarrow{\alpha} \Lambda_N^* \quad \text{et} \quad \mathcal{R} \xrightarrow{\beta} \Lambda_N$$

une règle r sera notée aussi:

$$r : \beta(r) \longrightarrow \alpha(r) \text{ ou } r^{\alpha(r)}_{\beta(r)}$$

$\mathcal{T}$: ensemble fini de règles lexicales t, muni d'une application β' but,

$$\beta' : \mathcal{T} \xrightarrow{\beta'} \Lambda_N .$$

$\forall A \in \Lambda_N$ on définit l'ensemble $\mathcal{T}_A$:

$$\mathcal{T}_A = \{ t \in \mathcal{T} \mid \beta'(t) = A \}$$

éventuellement un $\mathcal{T}_A$ peut être vide.

Soit Λ'_N l'image par β' de $\mathcal{T}$, c'est l'ensemble des catégories syntaxiques auxquelles on fera correspondre dans l'expression, à l'aide de l'application source α' de $\mathcal{T}$, une expression terminale.

Disons maintenant quelles sont les constructions élaborées au niveau de la composition.

On définit une catégorie avec produit tensoriel SYN qui a pour objet les mots du monoïde Λ_N^* , le produit des objets n'étant autre que le produit de juxtaposition du monoïde. Les morphismes de SYN sont

construits à partir du système $\mathcal{R}$ de morphismes générateurs.

Pour $B \in \Lambda_N$ et $\hat{A} \in \Lambda_N^*$, Soit $Hom_{SYN}(\hat{A},B)$ l'ensemble des morphismes de SYN ayant $\hat{A}$ pour source et B pour but, cet ensemble n'est autre que l'ensemble des arbres biordonnés de sommet B et de base $\hat{A}$, construits en prenant pour liens les règles de $\mathcal{R}$ (à chaque noeud de l'arbre correspond une lettre de Λ_N, à chaque noeud non situé à la base correspond une règle de $\mathcal{R}$ dont le but est la lettre du noeud et dont la source est le mot de Λ_N^* formé, dans l'ordre séquentiel, par la suite des lettres des noeuds immédiatement inférieurs).

Un élément $s_B^{\hat{A}} \in Hom_{SYN}(\hat{A},B)$ est appelé <u>un schéma de syntagme de type B</u> (construit à partir de $\hat{A}$). Dans la terminologie classique $s_B^{\hat{A}}$ serait la donnée d'une certaine B-dérivation de $\hat{A}$.

On définit ensuite un syntagme composé de type B, $(s_B^{\hat{A}}, \hat{t})$ comme un couple d'un schéma de syntagme de type B, et d'un mot $\hat{t} \in \mathcal{C}^*$ tel que $\beta'(\hat{t}) = \hat{A}$, c'est-à-dire qu'un schéma de syntagme de type B ne peut permettre la construction d'un syntagme composé de type B que si $\hat{A} \in (\Lambda_N')^*$.

Un syntagme composé de type B pourra aussi être notée $s_B^{\hat{A}}(\hat{t})$.

3.2. <u>Expression</u>. $E = \left(\left[\Lambda_T \right], \xi, \alpha' \right)$.

Nous supposerons ici que la catégorie d'expression est toujours SEG.

$\left[\Lambda_T \right] = \left[\Lambda_T^1, \ldots, \Lambda_T^n \right]$ alphabet terminal gradué,

Λ_T^i : ensemble des mots terminaux à i-insertions. On considère le SEG-oïde (ou échelle libre) construit sur $\left[\Lambda_T \right]$.

ξ : foncteur de SYN dans SEG commutant ave le produit tensoriel. On désigne par $\xi(A)$ l'entier strictment positif, nombre d'insertions de

l'expression des syntagmes de catégorie syntaxique A. On désigne par $\xi(r)$ le morphisme de SEG correspondant à la règle r : $\xi(r)$ aura pour but le peigne à $\xi(\psi(r))$ dents et pour source la liste de peignes à $\xi(A_i)$ dents (si $\alpha(r) = A_1 \ldots A_n$).

α' : est une famille d'applications $\left\{ \alpha'_A \right\}_{A \in \Lambda_N}$ qui envoient $\mathcal{C}_A$ dans $(\Lambda^*_T)^{\xi(A)}$, c'est-à-dire que à chaque terme on fait correspondre un mot ayant le nombre d'insertions convenable...

On définit maintenant un syntagme exprimé. Soit un syntagme composé $s_B^{\hat{A}}(\hat{t})$, α' associe à $\hat{t}$, considéré comme liste de termes t_i, une liste mots à plusieurs insertions $\alpha'(t_i)$

$$\hat{t} \rightsquigarrow^{\alpha'} \left\{ \alpha'(t_i) \right\} = \alpha'(\hat{t})$$

avec $\forall\ i\ \ \alpha'(t_i) \in (\Lambda_T^*)^{\xi(\beta'(t_i))}$

sur la liste $\left\{ \alpha'(t_i) \right\}$ opère le morphisme de SEG, $\xi(s_B^{\hat{A}})$ le résultat de cette opération sera un syntagme exprimé de type B, expression du syntagme composé $s_B^{\hat{A}}(\hat{t})$. On pourra écrire ce syntagme exprimé $\xi(s_B^{\hat{A}})\left(\alpha'(\hat{t})\right)$. On note ξ_B l'ensemble des syntagmes exprimés de type B, en particulier ξ_S sera le langage engendré par la grammaire de constituants généraux

$L(\mathcal{G}) = \xi_S$

et si $\xi(S) = 1$ on a $t(\mathcal{G}) \subset \Lambda_T^*$

On remarque qu'un même syntagme exprimé peut être l'expression de plus syntagmes composés, c'est pourquoi on définit l'ambiguité de structure d'un syntagme exprimé $\xi(s_B^{\hat{A}})\left(\alpha'(t)\right)$ comme le cardinal de l'ensemble:

$$\left\{ \left(s_B^{\hat{A}^i}, \hat{t}^i \right) \mid \xi\left(s_B^{\hat{A}}\right)' \alpha'(\hat{t}^i) = \xi\left(s_B^{\hat{A}}\right)\left(\alpha'(\hat{t})\right) \right\}$$

si l'on caractérise les règles de types CF comme ayant un membre de gauche appartenant à Λ_N et non Λ_N^*, on peut considérer les GCG comme des grammaires

de type CF plus générales puisque $\forall r \in \mathcal{R}$ et $t \in \mathcal{T}$, $\beta(r) \in \Lambda_N$ et $\beta'(t) \in \Lambda_N$.

- Une GCG soumise aux restrictions:

α - seul $\Lambda_T^1 \neq \emptyset$

β - $\xi(S) = 1$

est appelée Grammaire Context Free à Peignes (GCFP)

- Une GCG (resp. GCFP) dans laquelle $\forall r \in \mathcal{R}$, $\xi(r)$ est un morphisme de SEG tel que tout peigne de la source $\xi(\alpha(r))$ est utilisé une fois et une seule dans le graphe, est appelée GCG linéaire (GCGL) (resp. GCFPL).

Nous allons montrer que les GCFP sont une généralisation des grammaires CF. Soit une GCF $\mathcal{G}' = \{ V', V_T, \mathcal{R}, S \}$. On sait qu'à tout GCF on peut faire correspondre une GCF syntaxique équivalente. Une grammaire syntaxique est une grammaire dans laquelle l'ensemble $\mathcal{R}$ des règles est la réunion disjointes des ensembles $\mathcal{R}_T$ et $\mathcal{R}_N$, définis ainsi:

$\mathcal{R}_T = \{ r: A \longrightarrow \hat{B} \mid \hat{B} \in (V_T)^* \}$ et $\mathcal{R}_N = \{ r: A \longrightarrow \hat{B} \mid \hat{B} \in (V - \{V_T\})^* \}$

Ainsi soit une GCGL $\mathcal{G}$ dans laquelle nous faisons, au niveau de l'expression, la restriction γ

γ - $\forall A \in \Lambda_N \quad \xi(A) = 1$

(On remarque que γ est plus puissante que β)

la restriction γ et le fait que la grammaire $\mathcal{G}$ soit linéaire impliquent que $\forall r \in \mathcal{R}$, $\xi(r)$ est un produit de juxtaposition (au sens monoïde). Et d'après α, puisque l'alphabet $[\Lambda_T]$ n'est plus un alphabet gradué, l'échelle libre sur $[\Lambda_T]$ se réduit au monoïde libre sur Λ_T^1.

Et ainsi à tout règle $r_N \in \mathcal{R}_N$ (de $\mathcal{G}'$) $r_N: B \longrightarrow \hat{A}$ on fait correspondre d'une manière unique une règle $r \in \mathcal{R}$ (de $\mathcal{G}$)

$r: B \longrightarrow \hat{A}$ (et réciproquement).

A toute règle $r_T \in \mathcal{R}_T$ (de $\mathcal{G}'$) $r_T : B \longrightarrow \varphi \in V_T^*$ on fait correspondre d'une manière unique une règle $t \in \mathcal{T}$ (de $\mathcal{G}$) $t : B \longrightarrow$ telle que au niveau de l'expression $\alpha'(t) = \hat{\varphi}$ (et réciproquement).

Lemme

Ainsi à toute GCF $\mathcal{G}'$ on peut faire correspondre une GCFPL $\mathcal{G}$, soumises aux deux restrictions α et γ, telle que $\mathcal{G}'$ et $\mathcal{G}$ soient fortement équivalentes.

Nous allons maintenant donner l'exemple d'une GCFP (de peu d'intérêt linguistique puisqu'elle n'engendre qu'une phrase).

Exemple:

Soit la GCFP $\mathcal{G}_A = \{ \Lambda_N , \mathcal{R} , \mathcal{T} ; \Lambda_{T_A} , \xi_A , \alpha'_A \}$.

Données de Composition.

- $\Lambda_N = \{ S , Art , N , V , S_N \}$
- $\mathcal{R} = \{ r_o , r_1 \}$

  $r_o : S \longrightarrow S_N.V.S_N$, $S_N \longrightarrow Art.N$
- $\mathcal{T} = \{ t_o, t_1 , t_2, t_3 \}$

  $t_o : V \longrightarrow$, $t_1 : N \longrightarrow$, $t_2 : N \longrightarrow$, $t_3 : Art \longrightarrow$

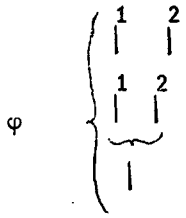
Données d'expression.

- $\Lambda_{T_A} = \{$ the, off, switches, boy, light $\}$
- ξ_A est défini par :

  $\xi_A(S) = 1$, $\xi_A(V) = 2$, $\xi_A(N) = 1$, $\xi_A(Art) = 1$.

  $\xi_A(r_o) = \psi$

$$\psi = \left\{ \begin{array}{l} 1 \quad 2 \quad 3 \\ 1 \quad 3 \; 2 \end{array} \right.$$

$\xi_A(r_1) = \varphi$ 

- α'_A est défini par:

$\alpha'_A(t_o)$ = switches-off

$\alpha'_A(t_1)$ = boy, $\alpha'_A(t_2)$ = lights, $\alpha'_A(t_3)$ = the

- le seul syntagme composé construit avec $\mathcal{G}_A$ est:

($s = s_S^{Art.N.V.Art.N}$; $\hat{t} = t_3 t_1 t_o t_3 t_2$) .

et le seul syntagme exprimé est: $\xi_A(s)\{\alpha'_A(\hat{t})\}$

c'est-à-dire la phrase: the boy switches the light off.

§ 4. UNE PROPRIÉTÉ ALGEBRIQUE DES GCF.

Soit une GCG $\mathcal{G}$ telle que l'alphabet terminal est réduit à un seul terme x qui a une insertion: $\Lambda_T = \{ x \}$

On démontre le théorème suivant:

Théorème. Tout langage L sur un alphabet $\Lambda_T = \{ x \}$ contient soit une progression arithmétique sur x, soit une progression géométrique de la forme:

$$L = \{ x^W + uV^k \mid u, V, W, \text{ fixés} \in \mathbb{N}, \text{ avec } u \text{ et } V \neq 0 \text{ et } k \in \mathbb{N} \}$$

c'est ainsi que l'on démontre que le langage L(!),

$$L(!) = \{ x^{n!} \mid n \in \mathbb{N} \}$$

n'est pas du type GCG, alors que Friant montre que L(!) est engendré par une grammaire CS.

C'est pourquoi on peut donner la classification suivante:

$$CF \subset CFP \subset CS$$

les inclusions sont toutes strictes.

B I B L I O G R A P H I E

- o - O - o -

J. P. Benzécri, Modo Imago Nati (Inst. Stat. Univ. Paris, 1966)

J. P. Benzécri, Linguistique Mathématique, algèbre des constituants non connexes (Inst. Stat. Univ. Paris, 1966).

J. Friant, Thèse de 3° cycle: Les langages CS (Inst. Stat. Univ. Paris, 1966).

R. D. Guedj, Thèse de 3° cycle: Les grammaires de constituants généraux (Inst. Stat. Univ. Paris, 1966)