

ANALYSE DE DISCOURS ET INFORMATIQUE

Utilisation du Langage DEREDEC dans des études sur l'argumentation
et de raisonnement en langue naturelle.

Alain Lecomte

Table des matières

Avant-propos	37
1. Introduction. Problèmes de la catégorisation d'un texte en termes d'une logique naturelle	37
2. Topologie des séquences	41
2.1 Interprétation et topologie	41
2.2 L'espace de la séquence	43
2.3 Construction de l'espace d'une séquence: anaphore et coréférence	44
3. Calcul sur les segments et les séquences	49
4. Les automates DEREDEC et le problème de la constitution des séquences	56
4.1 Ouvertures des segments	57
4.2 Fermetures des segments	58
4.3 Automates	59
Conclusion linguistique et informatique	60
Notes	62
Annexe	65

ANALYSE DE DISCOURS ET INFORMATIQUE.

Utilisation du Langage DEREDEC dans des études sur l'argumentation et de raisonnement en langue naturelle.

Alain LECOMTE

AVANT-PROPOS

Cet article doit beaucoup aux avis pertinents de J.M. Marandin et de C. Péquegnat pour la partie linguistique et de P. Plante pour la partie informatique. Nous sommes conscients du fait que deux ensembles de concepts s'y côtoient et souvent entrent en conflit: des concepts linguistiques d'une part (anaphore, coréférence...) et des concepts logico-informatiques d'autre part ("logique naturelle" pour la partie logique et concepts du langage DEREDEC pour la partie informatique). Cet entrelacs de références hétérogènes ne rend pas très facile la lecture de ce texte. Nous pensons néanmoins qu'un effort devrait être continué dans la voie d'une confrontation toujours plus poussée de ces différents langages. C'est à ce prix en effet qu'un sens nouveau peut être donné à la notion de formalisation en linguistique et en Analyse du Discours. Indiquons ainsi que c'est par parti-pris que nous avons adopté en de nombreux endroits la terminologie DEREDEC plutôt que la terminologie classique lorsque celle-ci était peut-être plus adéquate: nous éprouvions le besoin de parler en termes de ce que nous savions pouvoir réaliser en DEREDEC. (Par exemple la fonction "THEME", la relation de "DETERMINATION" ou la notion de "COMPLEMENT" (P1 ou P2)).

1. INTRODUCTION - PROBLEMES DE LA CATEGORISATION D'UN TEXTE EN TERMES D'UNE "LOGIQUE NATURELLE".¹⁾

Le travail qui est présenté ici s'inscrit dans un projet concernant l'utilisation de DEREDEC dans le domaine de l'argumentation et du raisonnement en langue naturelle. Comme on sait²⁾, le logiciel de Pierre Plante a le très gros avantage d'admettre directement en entrée

des procédures qu'il permet, du texte écrit en langue naturelle. Son fonctionnement, basé sur des reconnaissances de structures (modèles d'exploration) et sur des automates à états finis (possédant quelque parenté avec les ATN de Woods mais autorisant davantage d'opérations) permet de traiter le texte d'entrée de différentes façons: application d'une grammaire (telle que la GDS de P. Plante), introduction (sous forme d'"EXFAL" - expressions admissibles au lexique) d'informations de type sémantique ou textuel³⁾, projection de certaines de ces informations en des endroits appropriés (possibilité de représenter des mécanismes d'inférence), indexation et segmentation. Un avantage important, par ailleurs, est de ne pas faire dépendre l'analyse d'un texte d'une théorie fixée définitivement (grammaire, découpage morpho-syntaxique...) mais de permettre de réaliser des algorithmes qui soient la concrétisation de théories variables supposées adaptées à divers problèmes qu'on se pose.

Ainsi dans l'optique de J.-B. Grize et de ses collaborateurs (trices) du Centre de Recherches sémiologiques de Neuchâtel⁴⁾, peut-on viser à concevoir une batterie d'automates qui déposeraient à la surface d'un texte les traces des opérations dites "de la logique naturelle" qui se réalisent dans l'occurrence des items dont le texte est composé.

Soit par exemple le court texte suivant, extrait d'un article paru dans la Recherche no 106, 1979.

Exemple 1

Si de nombreuses larves de papillons peuvent être utilisées pour leur production de soie, seul le ver à soie est élevé systématiquement. Domesticqué depuis des origines fort anciennes, il n'existe plus à l'état sauvage. Pour se transformer en papillon la larve s'abrite en tissant un cocon formé d'un seul fil ininterrompu dont la longueur peut dépasser 1000 mètres. Déroulé, le cocon donnera la soie grège utilisée dans les filatures. Aujourd'hui, le ver à soie est devenu un instrument d'étude privilégié pour les chercheurs en raison de sa domestication et de ses caractéristiques biologiques.

On souhaiterait parvenir à introduire une catégorisation des unités lexicales usant des symboles α_i , γ , θ et δ , ces symboles pouvant être interprétés de la façon suivante:

α_i : opération d'ancrage (mise en place d'"objets"⁵⁾ autour desquels le texte s'organise); i désigne le rang de l'objet ancré et sert à

coindexer d'autres unités (dans le sens de "tokens") du texte (par exemple les répétitions lexicales, mais aussi les liens de coréférence).

- γ : opération d'ingrédient (enrichissement des objets ancrés). Par γ se trouve introduit un ensemble de traits appartenant à l'objet ancré par α .
- θ : opération de spécification. Le texte précise dans quel contexte est pris l'objet ancré par α (exemples: "pour se transformer en papillon", "dont la longueur dépasse 1000 mètres" - signifie que "le fil" est pris sous l'aspect de sa longueur - "Aujourd'hui", etc.).
- δ : opération de détermination. δ sert à fabriquer, à l'aide d'un objet préalablement ancré et d'un prédicat, ce qu'on appelle une détermination: il s'agit en somme d'un noyau d'énoncés à partir duquel les autres énoncés du texte peuvent être mis en relation (de paraphrase, de présupposition, d'opposition, etc.).

Cette catégorisation étant effectuée, on obtiendrait ensuite en extrayant, à l'aide de "modèles d'exploration", les syntagmes affectés de chacun de ces symboles ou d'une concaténation de certains d'entre eux, des représentations de ce que sont les objets de discours pris dans ce texte, sous forme de classes-objets structurées:

1. α_1 : *<le ver à soie >* δ : *<est élevé systématiquement>*
 γ : *<larves de papillon,>*
<production de soie >
<domestiqué depuis des origines fort anciennes>
<ne...plus état sauvage>
2. α_2 : *<l'état sauvage>*
 γ : *<ne...plus>*
<domestiqué depuis des origines fort anciennes>
<il (le ver à soie)>
3. $\theta\alpha_1$: *<la larve >* δ : *<. s'abrite>*
 γ : *<un cocon>*
 $\gamma\gamma$: *<un seul fil ininterrompu>*
 $\theta\gamma\gamma$: *<longueur>*
 $\gamma\theta\gamma\gamma$: *<1000 mètres>*

4. α_3 : <le cocon> δ : <déroulé, donnera>
 γ : <la soie grège>
 $\gamma\gamma$: <les filatures>
5. $\theta\alpha_1$: <le ver à soie> δ : <est devenu>
 γ : <un instrument d'étude>
 $\gamma\gamma$: <les chercheurs>
 <sa domestication>
 <ses caractéristiques biologiques>

Cette liste (ou liste de listes) pourrait alors se lire ainsi: le premier objet ancré est <le ver à soie>, il est composé d'un certain nombre d'ingrédients ou "aspects de l'objet" (glosables en "être une larve de papillon", "être producteur de soie", "être domestiqué...", "ne plus être à l'état sauvage"... autant de traits pouvant entrer dans une sorte de définition contextuelle de l'objet envisagé) et il supporte une opération de détermination: "est élevé systématiquement" (il s'agit de l'information considérée comme pertinente à cet endroit du texte, le reste, à savoir les ingrédients, participant seulement de la définition de l'objet sur lequel une information est apportée).

Un des aspects peut être considéré comme un objet de discours qui est ancré à son tour: <l'état sauvage>. Il possède alors lui-même comme ingrédients des aspects appartenant déjà au premier objet, ainsi qu'une reprise anaphorique de cet objet. Ceci marque une bipolarité du texte à cet endroit, lequel peut être vu comme parlant de <ver à soie> ou de l'objet <état sauvage/domestiqué>. (Interprétation: il est concevable qu'à ce moment du texte, celui-ci enchaîne soit sur l'un, soit sur l'autre de ces deux objets... soit sur les deux). $\theta\alpha_1$ indique qu'une spécification est introduite par-dessus un ancrage. La spécification sera ici: "pour se transformer en papillon", dans ce contexte, l'objet <ver à soie> devient l'objet <larve> ou encore "la larve" doit être conçu comme reprise anaphorique de "le ver à soie". Toujours dans ce contexte, l'objet ainsi spécifié reçoit une détermination: "s'abrite". La liste qui suit donne les ingrédients, les ingrédients des ingrédients et des spécifications de ces ingrédients. Ainsi " $\theta\sigma\gamma\sigma$: <longueur>" signifie que l'ingrédient "un seul fil ininterrompu" est vu, à cet endroit du texte sous l'aspect de sa "longueur", laquelle recevra alors un ingrédient particulier ("1000 mètres"). Et ainsi de sui-

te. Si on tentait de parcourir la classe-objet structurée associée à *<ver à soie>* on obtiendrait ainsi une suite d'enrichissements, de déterminations et de spécifications traduisant autant d'étages dans un fil argumentatif, autant de transformations "cognitives" d'un thème de discours pour parvenir à un état conclusif.

Bien sûr, un tel repérage est encore très intuitif et on nous objectera la difficulté d'isoler quelques critères formels permettant de distinguer un γ d'un α voire un γ d'un θ . En fait, il s'agira, dans ce qui suit, d'éclaircir quelques mécanismes de reconnaissance auxiliaires dans la perspective de production de telles analyses.

2. TOPOLOGIE DES SEQUENCES

2.1 Interprétation et topologie

Nous partirons de l'hypothèse que la forme "raisonnement" d'un texte (comme d'ailleurs la forme "récit", ou la forme "description", ou la forme "biographie", etc.) n'est pas directement inscrite dans le texte à la façon dont sont inscrits les mots et leurs rapports grammaticaux. Une telle forme n'est pas donnée à voir comme l'est par exemple le phénomène d'accord en genre et en nombre: ceci signifie que son repérage ne ressort pas d'une quelconque "grammaire de texte" (au sens donné à ce terme par exemple par Van Dijk). Il y a raisonnement parce qu'il y a lecture et interprétation. Pour reprendre les expressions du sémioticien Par Aage Brandt (1973), nous dirons avec lui que "tout discours inclut à titre d'actants constitutifs un DIRE qui est du côté du destinataire et un VOIR qui est du côté du destinataire". Une "forme" (raisonnement, récit, description...) se détache donc à partir d'un rapport entre ces actants qu'on peut dire "rapport d'interprétation". Pour Brandt, l'énoncé est un transfert d'un DIRE dans un VOIR, ce qui implique pour nous la possibilité d'un modèle topologique de l'interprétation: ce qui est dit occurrent alors comme un ensemble de signaux déclenchant pour le récepteur des opérations d'ouverture, de fermeture, de compactification et de mise en relation dans des espaces. Dans ces opérations d'interprétation, il se produit un effet de séquentialité qui est alors identifié à un

raisonnement ou à un récit ou à autre chose encore.

Il va de soi cependant que les rapports topologiques qui se tissent à partir des signaux fournis par les mots, d'une part, ne sont pas déterminés avec unicité (plusieurs ensembles sont possibles) et d'autre part ne se tissent pas à partir de rien: il y a toujours déjà un espace informé. Comme le dit Pêcheux (1982: 35, 45), "lire c'est toujours réinscrire un énoncé dans l'espace d'une problématique". Cet espace-réceptacle où se projette un texte pour qu'il fasse sens, cela pourra être aussi bien celui d'un inconscient individuel ("poinçonné" d'un fantasme particulier (cf. Lacan)) que celui d'une idéologie. Dans un texte d'historien sur les placements d'enfants au XVIIIe siècle (LEROY-LADURIE, 1979: 15-21) la commutation quasi automatique de <mise en nourrice> et de <allaitement mercenaire> en dira long par exemple sur les catégories d'interprétation requises par un texte qui se voudrait tout d'"objectivité scientifique".

Dans un cas pareil, c'est une similitude de place dans des formes syntaxiques récurrentes qui fait apparaître comme "naturelle" la commutation des deux expressions. Ou bien encore, ce sera un enchaînement particulier de places dans des formes syntaxiques qui provoquera un effet semblable, effet qu'on peut qualifier de coréférentiel. Ainsi dans notre exemple traitant du ver à soie, nous avons une "forme" qui consiste en: un SN ("le ver à soie") dont on a supposé qu'il portait la marque d'un ancrage, suivi d'une détermination ("est élevé systématiquement"). Cette forme est suivie d'une autre forme comportant d'abord une spécification ("pour se transformer en papillon") puis un SN du même type que le précédent ("la larve") Puis enfin une détermination ("s'abrite"). On pourra alors supposer, par exemple, que l'enchaînement de ces deux formes crée pour les SN en position similaire une relation de coréférence⁶⁾.

Ce genre de schéma caractérisera alors ce qu'on appellera: une règle d'interprétation textuelle.

La reconnaissance de tels liens de coréférence est fondamentale pour l'analyse des discours, en tant que celle-ci se propose justement de repérer ces points singuliers où un signifiant glisse vers un autre (soit sous l'effet d'une métonymie: <mise en nourrice>→<allaitement> ou <ver à soie>→<larve>, soit sous celui d'une condensation (métaphore)

<nourrice>><mercenaire>) et qui sont autant de points autour desquels s'organisent les textes (et les polémiques). Pour y parvenir, il nous faudra d'abord définir les unités de segmentation pertinentes du texte, à l'intérieur desquelles s'organiseront les expressions lexicales, dans un réseau de places.

2.2 L'espace de la séquence

Nous définissons d'abord un segment comme une chaîne de mots, de longueur arbitraire contenant une expression nominale en position de thème. La notion de thème est reprise ici aux travaux de Halliday notamment. Est thématiqué ce qui est "mis en relief" du point de vue informationnel, avant le groupe verbal, voire même comme constituant une unité autonome, s'il n'y a pas de groupe verbal. (Exemples: "Pierre est venu hier soir"; "Hier soir, Pierre est venu"; "Oh! catastrophe!"...) sous sa forme actuelle, DEREDEC permet de définir comme thèmes tous les GN ("Groupes nominaux") occurrents avant un GV ("Groupe verbal") dans un GP ("Groupe propositionnel"). Cette définition pourrait être raffinée, de façon à obtenir par exemple comme thèmes, des GN figurant dans un complément circonstanciel antéposé, ou bien des GN-sujets ayant été déplacés. Une expression nominale en position thématique possède une expansion sous forme d'une suite d'"aspects" (les γ^j d'une série α dans l'exemple que nous avons donné plus haut); ceux-ci sont formés de syntagmes nominaux munis de déterminants (indéfinis, possessifs, démonstratifs...) instaurant avec l'expression nominale thème (E.N.T.) des relations anaphoriques ou coréférentielles.

Repérer des segments est ainsi un préalable: à l'étape actuelle de notre recherche, nous nous centrons sur la fonction nominale. On peut raisonnablement penser en effet que celle-ci s'exerce en des lieux de stabilité référentielle. (On retrouve ici la notion travaillée par Thom de stabilité structurelle d'une forme, une des conditions de cette stabilité étant justement la nomination). Le discours produit un réseau de relations entre des unités (des "noeuds") stabilisées. D'où notre démarche: définir les segments (et au-delà: les séquences) à partir des liens anaphoriques et coréférentiels entre expressions nominales et entre expressions nominales et pronominales. Le réseau formé par ces liens constitue l'espace de la séquence, son système de coordonnées.

Nous définissons ensuite une séquence comme un arrangement de segments soit sous forme emboîtée, soit sous forme juxtaposée. Une séquence fait donc l'unité d'un faisceau d'expressions nominales et va l'orienter dans le cas d'un raisonnement, vers un point de convergence, qui sera une sorte de thème d'un niveau supérieur aux thèmes des segments.

2.3 Construction de l'espace d'une séquence: anaphore et coréférence

Avant de parvenir à un calcul des segments et des séquences, il nous faut préciser la nature de ces liens qui les constituent. Ces liens s'expriment en termes de coréférence, d'anaphore et d'anaphore associative, toutes relations présentes dans les phénomènes dits de reprise immédiate qui ont donné lieu déjà à une littérature abondante (cf. en particulier F. Corblin (1983), J.C. Milner (1976), C. Blanché-Benveniste et A. Chervel (1966)).

Nous tenterons dans ce qui suit de caractériser ces relations à partir d'une approche utilisant notre notion de découpage en segments et séquences et de voir en quoi elles sont connectées avec les notions introduites précédemment (ancrage, détermination, spécification, ingrédient...). Afin de respecter la terminologie usuelle au moins approximativement (car les définitions de ces relations sont très variables selon les auteurs), nous caractériserons la coréférence comme une relation symétrique entre des unités du texte.

(Cf. Milner (1976):

Il y a relation de coréférence entre deux unités référentielles A et B quand elles se trouvent avoir la même référence - ce qui peut arriver sans que l'interprétation de l'une soit affectée par l'interprétation de l'autre. (p. 32).

Milner donne l'exemple: "Napoléon/l'Empereur" où la paire est "constituée par des termes dont la référence actuelle est identifiée indépendamment" (p. 34)).

Nous caractériserons l'anaphore, au contraire, comme une relation asymétrique. Pour elle, en effet, l'interprétation d'un terme est dépendante de celle d'un autre terme, ainsi qu'il arrive dans le cas de l'anaphore pronominale, bien sûr, (un syntagme nominal est repris par un

pronom) mais aussi dans le cas où le deuxième SN d'une paire, étant par lui-même relativement indéfini, reçoit son interprétation d'un SN de tête qui, lui, est davantage défini.

Exemple (2)

Un ver à soie se transforme en papillon. La larve s'abrite alors dans un cocon.

Mais comme il a été déjà souvent remarqué (cf. Hawkins (1978), Corblin (1983)) des cas de reprise ont lieu sans qu'il y ait anaphore (au sens "strict") ni coréférence, par exemple dans:

Exemple (3)

L'avion décolle... Le palonnier vibrait...

où, là aussi, pourtant, l'interprétation du deuxième SN nécessite la connaissance du premier. On parle, en ce cas, suivant Hawkins, d'anaphore associative.

Parmi ces trois relations, un auteur comme Corblin ne considère comme autorisant la reprise immédiate que les deux espèces d'anaphore, à l'exclusion de la coréférence. Il caractérise en effet les phénomènes de reprise par deux propriétés:

- A - Il s'agit de rapports qui s'instaurent en vertu d'une succession: l'interprétation du SN défini exige la prise en compte d'un déjà-là du texte.
- B - La notion de phrase n'est pas pertinente pour repérer le domaine à l'intérieur duquel l'autre terme du rapport peut être trouvé. (p. 119).

Or ces deux propriétés sont absentes de la coréférence telle qu'elle a été définie plus haut.

Nous retenons également du travail de F. Corblin l'idée selon laquelle

l'anaphore consiste non à répéter quelque donnée du texte antérieur, mais à situer nécessairement l'interprétation de tout nouveau SN d'un texte relativement à la mémoire des interprétations antérieures, à doter nécessairement chaque SN d'un rapport explicite aux interprétations déjà produites. (p. 120).

La notion de mémoire paraît en effet centrale dans l'analyse de ces phénomènes. Il s'agit d'une mémoire à court terme et propre à un texte qui se constitue lorsque se structurent ce que nous avons appelé les objets de discours par l'effet d'une sériation d'expressions nominales.

Il apparaît toutefois quelques difficultés lorsque nous voulons distinguer anaphore et coréférence. A ce propos, les notions de "plus ou moins défini" ou de "plus ou moins identifié" qui interviennent entre autres dans la caractérisation faite par Milner de ces relations, semblent garder un certain flou. Si, usuellement, dans

Exemple (4)

Le ver à soie se transforme en papillon. La larve s'abrite dans un cocon,

le deuxième SN paraît anaphoriser le premier, on peut toutefois tout aussi bien accepter une lecture différente d'une telle séquence, si on considère par exemple que suivant le critère proposé par Milner, "*la larve*" est un terme dont la référence actuelle est identifiée indépendamment de celle du syntagme "*le ver à soie*". Le cas serait encore plus frappant dans une suite du genre :

Exemple (5)

Bombyx muri est élevé pour produire de la soie. Le ver à soie...

Supposons en effet qu'il y ait plusieurs espèces de ver à soie et que *Bombyx muri* en soit une, l'interprétation anaphorique ou seulement coréférentielle de cette séquence dépendrait alors du fait qu'on possède ou qu'on ne possède pas cette connaissance.

Dans ce cas en effet, "*le ver à soie*" apparaît comme un indéfini relativement à *Bombyx muri* (cas d'anaphore) dans l'autre au contraire, ils sont aussi définis l'un que l'autre et leur relation est symétrique (cas de coréférence sans anaphore).

De fait, il apparaît assez nettement lorsqu'on considère ces phénomènes de reprise caractérisés par la non-pertinence de la notion de phrase, qu'ici encore plus qu'en tout autre domaine linguistique d'investigation, on ne saurait se contenter de raisonner sur des exemples tirés hors de tout contexte, et qu'il est nécessaire de construire ces notions transphrastiques que nous baptiserons ici "segment" et "séquence".

Si, à ce propos, les concepts précédemment introduit (p. et provenant des essais de logique naturelle tentés par J.-B. Grize et al. ont quelque pertinence, alors on peut les référer à la dimension de la séquence d'une part et on peut, d'autre part, tenter de les relier à ces phénomènes d'interprétation que sont l'anaphore et la coréférence. Dans le contexte d'une ou plusieurs séquences, en effet, un exemple tel

que (4) perd de son ambiguïté: on pourra en effet "calculer" que "le ver à soie" est le principal thème du discours et que le syntagme "la larve" occurrent en position thématique dans une phrase qui succède à l'introduction de "le ver à soie", ne pourra qu'en être une reprise au sens où on devra l'interpréter comme "la larve consistant en le ver à soie". Dans le court texte que nous avons choisi comme exemple, cette interprétation sera rendue encore plus évidente par la "mémoire" d'une séquence antérieure où "le ver à soie" est situé parmi "de nombreuses larves de papillons". Notons alors combien cet exemple illustre la thèse de Corblin sur les "domaines lexicaux structurés": nous avons bien un défini ("le ver à soie") dont le rôle est d'"opérer sur l'ensemble des domaines structurés par le texte depuis le début de celui-ci" (à savoir dans notre cas le domaine lexical constitué autour de "larves de papillons"), ce "travail (se faisant) à partir du contenu lexical du SN défini" (p. 121). Ceci étant, le défini exerce en effet son rôle sur le domaine précédemment introduit, mais ici, c'est-à-dire lorsque c'est la première fois où un défini "occure" dans le texte -ce rôle est de spécifier ce domaine sous l'aspect d'une introduction de ce dont le texte va parler. C'est là ce que nous avons appelé une opération d'ancrage. Il doit être possible de soutenir -bien que cela sorte des limites de cet article- que tout ancrage suppose du déjà-là textuel.

Autrement dit, on n'ancre un objet qu'à partir d'un domaine déjà introduit même s'il est relativement indéfini (ce qui se repère au système des déterminants, par exemple: "de nombreux...", "quelques...", "certains..."). Si cette introduction préalable n'a pas été explicite, l'effet de l'ancrage sera de créer rétrospectivement cette idée de déjà-là, en donnant une impression de raccord à d'autres textes antérieurs.

Lorsqu'un ancrage a eu lieu, l'unité lexicale supportant cet ancrage admet nécessairement une expansion: celle-ci consiste soit en des reprises anaphoriques ("il", "la larve"...) soit en des termes reliés par une relation associative, celle-ci étant alors entendue en un sens plus général que l'anaphore associative, (on y admettra des SN déterminés par un possessif ou des SN indéfinis figurant éventuellement en position de complément d'objet, par exemple: "le ver à soie" → "la larve" → "un cocon" constitue une expansion).

L'expansion se termine lorsque apparaît un nouvel ancrage.

3. CALCUL SUR LES SEGMENTS ET LES SEQUENCES

Nous proposons ici de jeter quelques bases de ce calcul. Il doit évidemment partir de quelques principes (même si ceux-ci sont destinés à être modifiés en cours de route) et des marques matérielles que le texte comporte. De ce point de vue, il est inévitable d'accorder l'attention aux déterminants nominaux. Pour simplifier, nous classifions les expressions nominales en quatre groupes:

- l-expressions: celles dont le déterminant est un défini ("le", "la", "les") leur occurrence indique qu'un objet est identifié "actuellement" (c'est-à-dire par la l-expression elle-même: cas d'ancrage) ou bien a été identifié antérieurement (cas de reprise).
- u-expressions: celles dont le déterminant est un indéfini ("un", "une", "des") leurs occurrences se répartissent en plusieurs cas qui ont trait à la distinction générique/situation. Exemples (non exhaustifs):
 1. "un chat, ça griffe" (générique)
 2. "un chat est entré hier dans ma cave" (situation)
 - 2'. "un chat est entré hier dans ma cave. L'animal y a renversé toutes mes bouteilles" (situation).1 et 2' comportent une identification, mais elle est conférée à posteriori par les reprises anaphoriques ("ça", "l'animal") 2 n'en comporte pas. En anticipant quelque peu sur la suite, nous pouvons dire que "un chat", dans ces exemples, reçoit le statut d'"expression nominale thématique" en 1 et en 2', mais non en 2 (où le "thème" sera davantage la situation en elle-même, plutôt que l'objet désigné⁷).
- c-expressions: celles dont le déterminant est un démonstratif ("ce", "cet", "cette", "ces"). Ce sont des expressions qui n'ont pas l'autonomie référentielle des l- et des u-expressions, puisque leur référence leur est toujours conférée par un segment antérieur. Le problème est alors de déterminer le champ de cette référence: s'applique-t-elle à une l- ou une u-expression seulement (cf.: "un chat...cet animal...") ou bien à un segment entier, voire à une séquence (cf.: "la lumière est blanche. Quand elle traverse un prisme, elle montre toutes les couleurs de l'arc-en-ciel. Ce phénomène a été étudié par Newton").
- p-expressions: celles dont le déterminant est un possessif. Nous parlerons de lien de coréférence associative pour désigner le lien exis-

tant entre une telle expression nominale qu'elle reprend sous l'aspect d'une unité dépendante.

Comme pour les u-expressions, les c- et les p-expressions reçoivent une interprétation thématique lorsqu'elles sont reprises dans la suite du texte soit par une l-expression (ex.: "...un cocon. Le cocon...") soit par un anaphorique (ex.: "...J'ai souvent rencontré cette femme. Elle m'a toujours paru bizarre" ou bien: "Paul a mangé son chapeau. Il était dur à digérer".)

Détermination des points-origines

Nous appellerons désormais Expression Nominale Générique (E.N.G.) une expression nominale supportant un ancrage et figurant donc comme tête de liste d'une expansion. Les liens de coréférence ou d'anaphore entre les expressions nominales quelconques et une E.N.G. se détermineront, en même temps que celle-ci, par un principe de hiérarchie.

Règle 1 - Un nom propre en position thème (ou complément si le thème n'est pas une expression nominale) est une E.N.G.

Règle 2 - Une l-expression en position thème (ou complément si le thème n'est pas une E.N.G. déterminée par une règle antérieure) munie d'au moins une autre détermination nominale que l'article est une E.N.G. si elle ne suit pas immédiatement une E.N.G. déterminée par la règle 1.

(Commentaire: il s'agit de ce que la littérature logique appelle "les descriptions définies". La clause restrictive a pour but d'identifier la description définie avec le nom propre qu'elle suit immédiatement, comme dans l'exemple: "A New York, j'ai rencontré Chomsky. Le célèbre linguiste américain faisait une conférence").

Règle 3 - Une u- ou une p-expression en position thème (ou complément si le thème n'est pas une expression nominale) est une E.N.G. si elle est immédiatement suivie par une c- ou une p-expression elle-même en position thème ou par un pronom-sujet anaphorique "il", "elle"... ou par un pronom relatif sujet.

(Exemple: "...des veuves prennent en charge un nouveau-né, elles ne peuvent l'allaiter...")

ou bien: "...des veuves prennent en charge un nouveau-né. Ces femmes..."

ou bien: "... nous avons posé la question à des ouvriers. Leurs témoignages..."

Règle 4 - Une l-expression en position thème ou contenue dans une expression nominale en position-thème ou en position complément s'il n'y a pas d'expression nominale en position thème est une E.N.G. si on peut construire à partir d'elle une expansion propre⁸⁾ non vide ne contenant pas d'E.N.G. du type des règles 1, 2 ou 3, soit vers l'avant du texte, soit vers l'après, soit dans les deux directions, soit dans l'espace restreint de la relative où elle figure comme sujet.

(exemples: "La lumière du soleil est blanche. Après avoir traversé un prisme, elle montre toutes les couleurs qui existent dans le monde visible. La nature reproduit le même résultat dans la belle gamme des couleurs de l'arc-en-ciel".

Soit l'expression: "la nature". Son expansion, s'il s'agissait d'une E.N.G. contiendrait:

soit (vers l'avant): "toutes les couleurs qui existent dans le monde visible", il s'agit-là d'une E.N.G. d'après la règle 2;

soit (vers l'après): "le même résultat" et "la belle gamme des couleurs de l'arc-en-ciel" qui sont des E.N.G. d'après la règle 2⁹⁾ cette expansion serait donc vide. D'où il résulte que "la nature" n'est pas une E.N.G. (donc le sera l'expression nominale "la belle gamme des couleurs de l'arc-en-ciel") (le cas de "le même résultat" devrait faire l'objet d'un traitement particulier en tant que constituant une unité fondant ce qu'ultérieurement nous caractériserons comme "thème transéquentiel", c'est-à-dire thème qui traverse deux séquences au moins. Si nous avons eu "la nature reproduit ce résultat...", il en eût été de même puisque alors "ce résultat" aurait appartenu à une autre expansion, déjà constituée).

"...un bord de l'arc-en-ciel est toujours rouge et l'autre toujours violet; entre eux se trouvent rangées toutes les autres couleurs. L'explication que Newton donne de ce phénomène est celle-ci..."

Ici "l'arc-en-ciel" est contenu dans une expression nominale thème ("un bord de l'arc"), on peut construire à partir d'elle une expansion propre qui contient "rouge", "l'autre", "violet", mais aussi: "un bord". C'est donc une E.N.G.).

2) effectueraient les opérations suivantes:

a) une u-expression qui est thème mais n'est pas suivie d'une c-expression, ni d'une p-expression, n'est pas retenue (3);

b) une l-expression qui est thème et peut recevoir une expansion non vide ne contenant pas d'E.N.G. de niveau supérieur (1), (2) ou (3) est une E.N.G. par (4). Et son expansion contient la u-expression et la p-expression précédentes, ainsi que la u-expression qui suit, jusqu'à l'occurrence d'une E.N.G. de type (2). ("L'état sauvage" apparaît comme l-expression en position de complément dans un GP ("Groupe propositionnel") dont le thème est un pronom sujet);

c) une l-expression qui est de nouveau thème et peut recevoir une expansion non vide ("papillon" et "cocon"), est une E.N.G. par 4 ("la larve"). La l-expression "la longueur" étant prise dans une relative est une E.N.G. (par (4)) dont l'expansion se limite à: "1000 mètres";

d) autres l-expressions qui sont des E.N.G. par (4): "le ver-à-soie".

Fermeture d'un segment

En principe, il y a fermeture d'un segment donc d'une expansion d'une expression générique, lorsqu'on passe à une nouvelle expression générique, dont la référence est disjointe par rapport à celle de la première (deux noms propres distincts par exemple). En pratique, la reconnaissance de cette disjonction pose problème, on notera même, souvent, son ambiguïté. Exemple: "Monsieur Machin a rencontré son collègue. Le célèbre philosophe a entonné le troisième couplet de l'Hymne à la Joie". Qui est "le célèbre philosophe"? "Si Monsieur Machin" détermine le premier segment, que dire du "célèbre philosophe"? continue-t-il la séquence ou inaugure-t-il un nouveau segment? Ce genre d'ambiguïté nous paraît inéliminable et nous semble même faire partie de façon constitutive du procès de lecture d'un texte. C'est pourquoi nous admettons qu'un segment puisse ne pas être refermé "actuellement", ce qui laisse supposer que l'expansion de son générique se poursuive dans le texte jusqu'à englober certains expansions de génériques ultérieurs. Nous tenterons cependant d'ébaucher une liste de marques de fermeture de seg-

ment. Ce seront, en général, des GP (Groupes propositionnels) dont l'effet est de conclure une séquence, ou tout au moins, d'indiquer une "résolution". Ces GP pourront se signaler par:

- certains adverbes ou circonstanciels ou groupes propositionnels antéposés (cf. "clairement...", "finalement...", "aujourd'hui...", "le 15 mai 1846",...);
- certains déterminants verbaux "ne...que...", ou "ne...plus...";
- un morphème ou une locution "condenseur" ("en bref", "en résumé", "il ressort que", "il résulte que...").

Nous signalons d'autre part le rôle des spécifications dans l'opération de bornage à droite du segment antérieur gauche. Ce rôle sera souvent rempli par des groupes infinitifs antéposés ("Pour se transformer en papillon...") ou par des circonstanciels antéposés ("Aujourd'hui...", "le 15 mai...", etc.).

Exemple 2 (début de "La Sarrasine" de Balzac, dont Barthes a produit un découpage séquentiel dans S/Z)¹⁰⁾

(J'étais plongé dans [une de ces rêveries profondes] qui saisissent
ENG ③
tout le monde, même un homme frivole au sein des [fêtes les plus tumultueuses]). ([Minuit] venait de sonner à l'horloge de l'Elysée
ENG ② ENG ①
Bourbon). (Assis dans [l'embrasure d'une fenêtre] et caché sous [les
ENG ②
plis onduleux d'un rideau de moire], je pouvais contempler à mon aise le jardin de l'hôtel où je passais la soirée). ([Les arbres], imparfaitement couverts de neige se détachaient faiblement du fond grisâtre que formait un ciel nuageux, à peine blanchi par ([la lune])
ENG ② ENG ④
ENG ④
(Vus au sein de [cette atmosphère fantastique], ils ressemblaient vaguement à des spectres mal enveloppés de leurs linceuls, image gi-

gantesque de ([la fameuse dans des morts]). Puis (en me retournant
de [l'autre côté] (je pouvais admirer [la danse des vivants]! (Là
fourmillaient, s'agitaient et papillonnaient [les plus jolies fem-
mes de Paris]...))

Remarques:

- "fêtes les plus tumultueuses" est pris comme une l-expression. Ceci est très en accord avec les modèles d'exploration DEREDEC: il s'agit de rechercher les GN dans lesquels figurent "le", "la" ou "les":
- "Minuit" est posé comme nom propre.

Segments et séquences :

1. (J'étais plongé dans une de ces rêveries profondes (qui saisissent tout le monde même un homme frivole au sein des fêtes les plus tumultueuses).

Commentaire: le premier objet ancré (E.N.G.) est "rêveries profondes". Ceci parce que c'est une u-expression, qu'elle est en position complément dans un GP où le thème est "J'" et qu'elle est reprise par un pronom anaphorique: le relatif "qui".

Le second est "fêtes les plus tumultueuses" parce que c'est une l-expression de type (2).

Un segment (relatif à "fêtes les plus tumultueuses") est emboîté dans l'autre (relatif à "rêveries profondes") .

2. (Minuit venait de sonner à l'horloge de l'Elysée Bourbon)
3. (Assis dans l'embrasure d'une fenêtre et (caché sous les plis onduleux d'un rideau de moire (je pouvais contempler à mon aise le jardin de l'hôtel où je passais la soirée
4. (Les arbres, imparfaitement couverts de neige, se détachaient faiblement du fond grisâtre que formait un ciel nuageux (à peine blanchi par la lune)
5. (Vu au sein de cette atmosphère fantastique, ils ressemblaient vaguement à des spectres mal enveloppés de leurs linceuls (image gigantesque de la fameuse danse des morts)

6. Puis (en me retournant de l'autre côté (je pouvais admirer la danse des vivants))

7. (Là, fourmillaient, s'agitaient et papillonnaient les plus jolies femmes de Paris)

Commentaires: ② ferme une parenthèse du segment gauche ① à cause de l'introduction d'un nom propre ("Minuit")

③ ferme le segment ② à cause du groupe antéposé ("spécification") "Assis dans..." et "caché sous..."

⑤ ferme ④ pour la même cause ("Vus au sein de...")

⑥ ferme ⑤ par le GP antéposé au participe présent ("en me retournant...") et en même temps se ferme elle-même à cause du morphème "puis" interprétable comme conclusif

⑦ ferme ⑥ à cause de l'adverbe antéposé "Là" et se ferme elle-même .

Comme on le voit, une indétermination demeure quant à la fermeture de nombreux segments. Ainsi se demandera-t-on si "rêveries" traverse, comme objet, tout le texte, auquel cas les autres expressions nominales en formeraient une expansion, ou bien s'il s'arrête avant "Minuit"? Par contre, l'anaphorique "ils" recolle les segments 4 et 5 et fait des expressions nominales occurrentes en eux, une expansion de l'objet ancré "arbres". Ce découpage est très proche, comme on pourra le constater, de celui opéré par Barthes.

4. LES AUTOMATES DEREDEC ET LE PROBLEME DE LA CONSTITUTION DES SEQUENCES

Dans cette section, nous tentons d'aborder le problème de l'automatisation du découpage séquentiel à partir d'une simplification des hypothèses introduites en 3. Cette simplification a pour but de faciliter un premier essai de mise à contribution du logiciel DEREDEC.

Il est clair que, par ses propriétés supposées ici connues, ce logiciel est capable:

- a) de repérer les expressions des différents types répertoriés en (3), et ce grâce à des modèles d'exploration;
- b) de les introduire comme conditions d'application de règles au sein d'automates qui pourront déposer des marques nouvelles (ouverture et fermeture) à la surface du texte;

- c) d'indexer à chaque lexème entrant dans la composition d'une expression nominale portant l'indice d'un ancrage, une liste représentant son expansion (sous la forme d'une "EXFAL");
- d) d'engendrer au moyen de modèles d'exploration fouillant la représentation du texte comme ensemble d'EXFAD et d'EXFAL les objets propres au texte (c'est-à-dire ses points de stabilité ou d'ancrage dans l'espace des discours).

4.1 Ouvertures de segments

Nous avons travaillé avec les modèles suivants:

GNTH : (GP((X)(TP-)(=GN((X)(DET+)(GN((DL2(("LE"))(("LA"))(("LES"))))))))))))

Commentaire: recherche des GN -préfixation du signe "■"- en position de THEME à l'intérieur d'un GP -relation (TP-)- qui contiennent une relation de détermination -relation (DET+)- entre quelque chose (X) et un GN contenant un article (DL2) consistant en "le", "la" ou "les".

GNQ : (=X((GN((QUA))))((N22((QUA))))))

Commentaire: ce modèle disjonctif recherche les expressions contenant un GN contenant un quantificateur -surcatégorisé "QUA", ou bien un quantificateur pronominal en position sujet -"tout", "quelques-uns"...

GNEQUI: (GP((GV((ET)))(TP-)(GN((DL2))((=NL))))))

Commentaire: ce modèle recherche les noms -NL- qui figurent dans les GN contenant également un article (DL2) et qui sont thèmes des groupes verbaux contenant une occurrence du verbe "être", (surcatégorisé "ET").

Remarques :

- Ces modèles ne sont pas conformes aux règles énoncées dans le "principe hiérarchique" du 3. Nous voulions ici ouvrir des parenthèses: devant les l-expressions munies de déterminations nominales autres que l'article (les "descriptions définies") en position thème, devant les quantificateurs et devant les formulations équatives (l'hypothèse étant qu'elles ancrent un nouvel objet, celui qui apparaît comme premier terme);
- l'application de ces modèles implique une catégorisation supplémentaire des expressions atomiques, notamment par "QUA" (quantificateur) et par "ET" (verbe "être").

4.2 Fermetures de segments

Nous avons travaillé avec les modèles suivants:

GNSPE : (GP((X)(TP+)(=GN)(DET+)(GN((X(("UN"))(("UNE"))(("DES"))))))))

Commentaire: Ce modèle recherche les u-expressions, déterminées, dans un propos.

EQUIGN : (GP((N22)(TP+)(GV((GV((ET)))(P1-)(=GN))))))

Commentaire: Ce modèle recherche les GN qui apparaissent dans le développement du propos d'un groupe verbal contenant "être" à l'intérieur d'un groupe verbal relié par une relation de propos à thème à un pronom comme le démonstratif "ce" ou "c'".

CONDA : (=X((CONDA)))

Commentaire: Ce modèle recherche les expressions surcatégorisées comme "condenseurs".

Remarques:

- L'hypothèse retenue était de fermer une parenthèse après un développement de propos, après un GN introduit par une expression déictique du genre "c'est...". et après une expression catégorisée comme condenseur (adverbe antéposé, locution du genre "en bref", "en résumé", "en fin de compte"...).
- La catégorisation d'une expression comme condenseur peut elle-même se faire au moyen d'un automate.

(Celui-ci utilise un modèle MBREF défini par:

(=X((C22(("EN"))(("PAR"))(("POUR"))))(("BREF"))(("DONC"))

recherche des expressions contenant ou bien une préposition du genre "en", "par", "pour", ou bien le mot "bref" ou bien le mot "donc" ¹¹⁾

et procède de la façon suivante:

si un GP contient une expression conforme à ce modèle, il est demandé de l'imprimer au terminal, l'utilisateur répond alors par OUI s'il y voit une expression "qui condense", auquel cas chaque expression atomique est envoyée au terminal et l'utilisateur catégorise "CONDA" l'une d'elles).

4.3 Automates

Le premier automate AUTOUV part d'un état où il se met au début du texte. Dans un second état, s'il reconnaît un GP où l'un des modèles "ouvreurs" est vérifié, il dépose à gauche du GP une parenthèse ouvrante catégorisée "PO". A la ponctuation finale, il appelle AUTOV (automate fermant les parenthèses). Des états intermédiaires sont rajoutés afin de régler les différentes variables qui interviennent lors de l'application des modèles. (Par exemple: le modèle est-il conjonctif? disjonctif?) Dans le détail AUTOUV se présente ainsi:

```
(S1 (PHI S2 D)(X S1 G))
(S2 (GNTH*S4 G)
  (GNQ*S4 G)
  (ET *(SETQ CONJUNCTION T) S3)
  (C321 (E AUTOV S1 G) (RETURN))
  (X S2 D))
(S3 (GNEQUI*(SETQ CONJUNCTION NIL) S4 G)
  (X (SETQ CONJUNCTION NIL) S2 D))
(S4 (C32 (OD)(F D PO) (ALLER D 1) REMPLACER PO NIL "("))
  S2 D)
  (PO (ALLER D 2) S2 D)
  (PHI (OD)(F D PO) (ALLER D1) (REPLACER (PO NIL "("))
  S2 D)
  (X S4 G)))
```

Un autre automate, AUTOBJ, a été construit: son but était de construire les expansions des expressions génériques. Dans un premier temps, il n'a pas été tenu compte des marques d'ouverture et de fermeture des segments. Le principe en a été le suivant:

1. Extraire un certain nombre d'expressions déterminantes déjà catégorisées (D12, C21, C22, plusieurs catégories temporaires...) en les surcatégorisant "ANA".

2. Provoquer une nouvelle catégorisation en contexte pour tous les mots surcatégorisés "ANA", au terme de laquelle ils reçoivent des "étiquettes":

"QUA" (pour quantificateurs et numéraux) "DEM" (pour démonstratifs)
"DEF" (pour articles définis) "ANAF" (pour indéfinis et possessifs).

3. Introduire une ponctuation finale de texte (C321).

4. Appliquer AUTOBJ, qui fonctionne ainsi:

Soit les modèles:

MOD* dépistant les noms déterminés par un ANAF (correspondant à nos u- et p-expressions)

MOD2* dépistant les noms déterminés par un DEF dans un GN figurant dans un propos (correspondant à nos l-expressions)

MOD3* dépistant les noms déterminés par un DEM (nos c-expressions)

et NTH* dépistant les noms déterminés par un DEF en position de thème on fait:

dans les cas respectifs où un GP contient une réalisation de MOD*, de MOD2* ou de MOD3* raccorder à l'expression atomique conforme à l'un de ces modèles une EXFAL (liste) comportant les expressions atomiques vérifiant le modèle NTH* dans l'avant du texte.

Ce raccordement se fait au moyen de la fonction LEXIDEC¹²⁾.

Nous ne donnons pas ici le détail de cet automate. On trouvera en annexe les résultats auxquels ont conduit les applications de AUTOUV et de AUTOBJ à un court texte sur "la robotisation" (cf. C. PEQUEGNAT, 1984). On excusera par ailleurs le côté très rudimentaire de ces réalisations effectives d'automates qui n'ont pour autre but, encore une fois, que d'indiquer une voie de travail en DEREDEC.

CONCLUSION LINGUISTIQUE ET INFORMATIQUE

Même si nos automates sont encore rudimentaires et nos modèles d'exploration simplistes, nous souhaitons pouvoir montrer par ce travail l'apport d'un langage informatique aux études sur le langage. Cet intérêt réside en plusieurs points.

D'abord le langage informatique (tel que DEREDEC) permet de formuler une hypothèse linguistique sous la forme d'une opération effectuable (celle qui consiste par exemple à appliquer une batterie d'automates à un texte) et donc, par là même, de poser les bases d'une démarche authentiquement expérimentaliste en linguistique.

Ensuite la réalisation concrète d'algorithmes permet de traiter à titre d'exemples non plus seulement des phrases isolées mais des séquences en contexte.

Par ailleurs, lorsque le langage informatique est assez riche, comme l'est DEREDEC, il fournit un ensemble de possibilités très large pour représenter ces notions que le linguiste rencontre sur son passage sans toujours pouvoir les définir, telles que par exemple: mémoire intratextuelle ou savoir intertextuel. A ce propos, il nous semble opportun d'utiliser la possibilité qu'on a en DEREDEC de simuler l'apprentissage du "sens" des termes d'un texte (lorsqu'on peut transférer par "EXFALISATION" les propriétés repérées au cours de la lecture sur une occurrence particulière d'une unité lexicale, en lui adjoignant cette liste de propriétés au moyen des fonctions LEXIDEC ou EXPOZ). Cette simulation est susceptible de représenter assez bien la constitution d'une mémoire intratextuelle, alors que la possibilité offerte de travailler simultanément sur plusieurs canaux peut être proposée comme représentation du savoir intertextuel. De même, la notion de MODELE D'EXPLORATION nous paraît bien représenter celle de "forme syntaxique".

Enfin on doit souligner ici la différence d'approche entre des méthodes classiques en Intelligence Artificielle qui partent d'univers sémantiques déterminés à priori sans trop se soucier des effets propres aux textes ni de leurs fonctionnements et les méthodes qu'il est possible de développer à l'aide d'un langage tel que DEREDEC. De nombreuses objections ont été faites à l'Intelligence Artificielle (cf. DREYFUS, 1984) faisant résider son inefficacité à simuler une intelligence dans son impossibilité à appréhender une dimension paradigmatique du savoir.

Or cette dimension paradigmatique est fondamentale dans l'appréhension et la compréhension des faits de discours. La possibilité qu'offre DEREDEC de décrire et d'explorer un ensemble quasi illimité de textes et donc, de les mettre en parallèle, semble répondre dans une petite partie, à ces objections.

NOTES

- (1) "Logique naturelle" est pris ici dans un sens distinct de celui qu'il revêt par exemple chez Lakoff. Chez ce dernier, en effet, la logique est dite naturelle en tant qu'elle est un assouplissement de la logique formelle (essentiellement celle des prédicats du premier ordre). De nombreux auteurs (cf. M.-J. Borel, 1978) se sont interrogés sur cette conception, leur argument pouvant se résumer par l'alternative suivante: ou bien un assouplissement de la logique formelle est apte à capter des phénomènes langagiers propres à la langue naturelle mais alors cet "assouplissement" est tel qu'il fait perdre tout l'intérêt de qualifier de "logique" un ensemble de schémas et de règles sur ces schémas, ou bien on veut préserver les avantages d'une logique (cohérence et complétude par exemple) mais alors la langue décrite n'est plus la langue naturelle... Le Centre de Recherches Sémiologiques de l'Université de Neuchâtel a développé une autre conception de la logique dite naturelle, qui s'inscrit dans la lignée des travaux de Jean Piaget concernant l'étude des structures opératoires à des niveaux pré-formels. Il s'agit de dégager des opérateurs dont l'action ne possède pas la réversibilité ni l'universalité des opérations formelles, mais qui sont étroitement dépendants des contenus qu'ils manipulent. Un début de formalisation a conduit les chercheurs de ce centre à s'inspirer des travaux de S. Lesniewski. C'est de cette conception qu'il est ici question. Elle est exposée en particulier dans Borel, Grize, Miéville (1983).

- (2) Le logiciel DEREDEC est un langage de programmation évolué écrit à partir de LISP. Il est constitué de fonctions descriptives et de fonctions exploratrices à appliquer sur des données textuelles ne nécessitant aucun codage particulier. Ces fonctions peuvent être utilisées comme opérations à l'intérieur d'automates. Le traitement d'un texte suppose la catégorisation de ses lexèmes ("expressions atomiques") au moyen de symboles qui renvoient à des catégories morphosyntaxiques (N1 pour "nom", V1 pour "verbe conjugué", V21 pour "verbe à l'infinitif", V23 pour "participe passé", D12 pour "article", D13 pour "adjectif", etc. (voir la liste exhaustive dans le manuel de DEREDEC établi par P. Plante)). Celle-ci se fait d'abord hors contexte (automatiquement, par un dictionnaire de formes non ambiguës d'une part, et interactivement pour des formes non encore enregistrées dans le dictionnaire) puis en contexte (pour les ambiguïtés). (L'utilisateur peut bien sûr toujours changer ses catégories, comme son dictionnaire, en fonction de ses besoins). Il est possible alors, par exemple, de programmer des automates oeuvrant dans une perspective "bottom-up", c'est-à-dire de regroupement d'unités syntaxiques vers le niveau le plus élevé. La Grammaire de Surface de P. Plante est ainsi une batterie de 60 automates, incluant environ 1600 règles, qui analyse un texte, vu comme concaténation de GP ("Groupes propositionnels") dans une optique proche de celle de Chomsky (1965). Ces automates déposent par ailleurs des relations du genre: relation "thème-propos", relation de détermination dans un groupe nominal, relation de développement du propos (P1, P2...), etc. Ce sont des relations dites "de dépendance contextuelle" (ayant donné leur nom au logiciel).

- (2) Des fonctions permettent d'autre part de brancher sur chaque noeud des arbres syntagmatiques issus de l'application d'une grammaire (et formant une "DDT" -description de texte-), des arborescences supplémentaires appelées EXFAL susceptibles de transporter des informations paradigmatiques. Les fonctions exploratrices reposent sur un principe de "pattern-matching" permettant d'isoler dans un texte toutes les expressions vérifiant tel ou tel "modèle" défini à priori (à partir d'une syntaxe proche de celle d'une DDT). Les principes de ce logiciel sont exposés dans P. Plante (1981) et dans A. Lecomte, J. Leon, J.M. Marandin (1983).
- (3) Nous entendons par information de type textuel les informations qui sont dépendantes du texte, c'est-à-dire qui sont relatives par exemple à la place d'un item lexical dans un texte, à sa fréquence d'apparition reportée au total ou à une zone quelconque du texte, aux liens grammaticaux qu'il possède avec d'autres items, etc.
- (4) Cf. en particulier: BOREL, GRIZE, MIEVILLE, 1983.
- (5) En logique classique des prédicats, on admet des "variables d'objet". Ce sont simplement des symboles sémantiquement interprétables comme des éléments d'ensembles. Pour l'approche "logique naturelle", cette notion liée à une idée du quelconque, est abandonnée pour être remplacée par une notion d'objet en un sens particulier. Un objet est propre à un discours. Deux discours différents n'auront ainsi jamais les mêmes objets. Une des tâches de l'Analyse des discours est de construire les objets d'un discours, en tant que points de stabilités à partir desquels s'organisent ses formes pertinentes. La représentation d'un objet sera donnée par le concept de classe méréologique, concept des travaux de Lesniewski (cf. note 1), la matérialisation d'un objet par une liste d'expressions, en général nominales, liées entre elles par des relations que nous tenterons de caractériser plus loin en termes de coréférence et d'anaphore. Les opérateurs de la logique naturelle notée α , γ , θ , ω ont pour champ des objets. Ainsi α "ancree" un objet, γ en produit une expansion, θ le spécifie et ω le "dénivelle". (Par dénomination ou par nominalisation par exemple). Par le jeu de ces opérations, un objet peut donc être transformé en un autre. En particulier, θ est une "transformation d'objet".
- (6) On verra en fait plus loin que dans le cas présent, cette interprétation sera renforcée par le lien déjà existant entre "larves" et "ver-à-soie", "le ver-à-soie" étant une expression définie qui structure un domaine lexical déjà introduit (lien "...de nombreuses larves de papillons... seul le ver-à-soie").
- (7) Ceci introduit une dichotomie entre objets et événements qui nous paraît fondamentale dans les études sur l'argumentation.
- (8) Nous appellerons expansion-propre d'une E.N.G. la liste des termes de son expansion qui n'appartiennent pas déjà à une autre expansion.

- (9) Ici, la clause de la règle 2 concernant une l-expression en position complément si le thème n'est pas une E.N.G. déterminée par une règle antérieure doit s'interpréter comme se référant à une règle antérieure à celle qu'on applique actuellement en l'occurrence: la règle 4.
- (10) Il peut paraître étonnant que, s'agissant de "raisonnement" en langue naturelle nous donnions en exemple un texte qui n'est pas un raisonnement mais un fragment de récit. En fait, l'interprétation "raisonnement", comme nous l'avons dit plus haut (supra) n'est pas conférable directement à un discours. Au seul niveau de la segmentation où nous nous plaçons il n'y a pas de critère permettant de distinguer "raisonnement" et "récit". Cette différence se dégagera ensuite, elle viendra essentiellement, à notre avis, du système des pronoms sujets et du système des temps (et d'autres marques portées par les verbes, en particulier des marques aspectuelles). Avant cette bifurcation, les deux formes se confondent en un tronc commun qui est la possibilité de découper en séquences. Dans un récit, comme dans un raisonnement, il y a des "condensations" et des "déplacements". La raison majeure d'avoir choisi ici cet extrait de "La Sarrasine" est la possibilité de confronter notre segmentation, obtenue à partir de critères "formels" de la page avec une segmentation produite intuitivement par l'un des plus fins parmi les analyseurs de textes: Roland Berthes.
- (11) Ce modèle peut paraître bien simpliste... nous ne l'avons fait que pour montrer qu'à condition de procéder à une analyse plus poussée du lexique il est toujours possible, quitte à allonger considérablement les listes, de repérer des groupes propositionnels candidats à l'opération de "condensation".
- (12) En DEREDEC, la fonction LEXIDEC est une fonction admettant comme premier argument: un premier modèle d'exploration (ou une suite d'expressions atomiques) comme deuxième argument: une suite de modèles d'exploration puis comme troisième et quatrième arguments: une indication que la recherche selon les deuxièmes modèles doit s'effectuer avant l'occurrence de l'expression détectée par le premier argument, ou après ou dans la totalité du texte, et une indication portant sur le caractère récursif de l'application de la fonction. Appelée de l'intérieur d'un automate, LEXIDEC s'applique au texte décrit par cet automate en munissant chaque expression détectée par le premier argument d'une arborescence dont les branches portent le nom d'un modèle figurant dans le deuxième argument et dont les extrémités sont les expressions détectées par ces modèles du deuxième argument.
Les modèles d'exploration contenant la variable PATRON voient celle-ci prendre automatiquement pour valeur soit une expression atomique détectée par le premier argument soit une expression détectée par un modèle du deuxième argument lorsque l'application est récursive.

ANNEXE

1. Texte en input

Peut-on déjà se rendre compte à quel point la robotisation change le travail en atelier? Nous avons posé la question à deux conducteurs de presses, à deux conducteurs de machines d'usinage et de soudure, à deux agents d'entretien qui vivent cette expérience. Leurs témoignages suffisent pour se faire une première idée. Il semble bien que s'amorce là une révolution aussi profonde que celle qu'ont connue, voilà dix ans, les repousseurs à la main lorsqu'ils ont dû lâcher la plane et la cuillère, car les presses façonnaient plus vite qu'eux des disques plus épais ou revêtus d'anti-adhésif. De même, à l'enduction, au polissage, à la soudure, à l'électroménager, des machines de plus en plus perfectionnées ont réussi à reproduire, à grande cadence, ce que le coup de main de l'ouvrier parvenait à accomplir, à force d'expérience et de fatigue physique. Les conducteurs approvisionneurs de ces machines ont acquis, à leur tour, une qualification reconnue non seulement en tenant compte des cadences, mais aussi en assurant certains réglages et en contrôlant la qualité de leur production. Voilà que maintenant leur travail est effectué par des bras articulés munis de mains en forme de ventouses ou de pinces et dirigés par des cerveaux cachés dans des coffrets énigmatiques. L'homme n'a plus à manipuler des milliers de pièces, à longueur de journée. Tant mieux! Mais cependant il reste là, auprès des machines, pour veiller à ce que tout marche régulièrement, pour parer aux incidents imprévisibles, pour contrôler en permanence les produits, pour prévoir l'approvisionnement en pièces détachées, pour changer de séries sans perdre de temps. Il peut se comparer alors au pilote d'un avion de ligne: en vol automatique, s'il lâche le manche et le palonnier, il garde les yeux rivés sur son tableau de bord pour renseigner par radio la tour de contrôle. Et pourquoi n'appellerait-on pas pilotes de lignes ceux qui font marcher ces chaînes de machines automatisées? Cela indiquerait clairement qu'ils gardent la maîtrise de ces automates, dont ils se servent comme jadis l'ouvrier se servait de ses outils rudimentaires.

2. Automates I

(DEFPROP AUTOUV

(AUTOUV AUTOUV

(S1 (PHI S2 D) (X S1 G))

(S2 (GNTH* S4 G)

(GNQ* S4 G)

(ET* (SETQ CONJUNCTION T) S3)

(C321 (E AUTOV S1 G) (RETURN))

(X S2 D))

(S3 (GNEQUI* (SETQ CONJUNCTION NIL) S4 G)

(X (SETQ CONJUNCTION NIL) S2 D))

(S4 (C32 (O D) (F D PO) (ALLER D 1) (REPLACER (PO NIL "()) S2 D)

(PO (ALLER D 2) S2 D)

(PHI (O D) (F D PO) (ALLER D 1) (REPLACER (PO NIL "()) S2 D)

(X S4 G))

VALUE)

```
(DEFPROP AUTOOF
(AUTOOF AUTOOF
  (S1 (GNSPE* S2 D)
      (EQUIGN* S2 D)
      (CONDA* S3 D)
      (PHI (RETURN))
      (X S1 G))
  (S2 ((C32 C321) (O G) (F G PF) (ALLER G 1 ) (REPLACER (PF NIL
    "}") S1 G)
      (PF (ALLER G 2 ) S1 G)
      (X S2 G))
  (S3 ((C32 C321) (O G) (F G PF) (ALLER G 1 ) (REPLACER (PF NIL
    "}") S1 G)
      (PF (ALLER G 2.) S1 G)
      (X S2 D)))
VALUE)
```

3. Modèles I

```
(DEFV GNIH* (GP ((X) (TP -) (=GN ((X) (DET +) (GN ((D12 ((LE))
  ((LA)) ((LES))))))))))
(DEFV GNQ* (=X ((GN ((QUA)))) ((N22 ((QUA))))))
(DEFV GNEQUI* (GP ((GV ((ET))) (TP -) (GN ((D12)) ((=NL))))))
(DEFV GNSPE* (GP ((X) (TP +) (=GN ((GN) (DET +) (GN ((X) ((UN))
  ((UNE)) ((DES))))))))))
(DEFV EQUIGN* (GP ((N22) (TP +) (GV ((GV ((ET))) (P1 -) (=GN))))))
(DEFV CONDA* (=X ((CONDA))))
(DEFV MODQUE* (=X ((QUE))))
(DEFV ET* (=X ((ET))))
```

4. Texte en output

Peut-on déjà se rendre compte à quel point la robotisation change le travail en atelier

?

(

Nous avons posé la question à deux conducteurs de presses, à deux onducteurs de machines d'usinage et de soudure, à deux agents d'entretien qui vivent cette expérience

Leurs témoignages suffisent pour se faire une première idée

.

(

Il semble bien que s'amorce là une révolution aussi profonde que celle qu'ont connue, voilà dix ans, les repousseurs à la main lorsqu'ils ont dû lâcher la plan et la cuillère, car les presses façonnaient plus vite qu'eux des disques plus épais ou revêtus d'antiadhésif

.

(

De même, à l'enduction, au polissage, à la soudure, à l'électroménager, des machines de plus en plus perfectionnées ont réussi à reproduire, à

grande cadence, ce que le coup de main de l'ouvrier parvenait à accomplir, à force d'expérience et de fatigue physique

Les conducteurs, approvisionneurs de ces machines, ont acquis à leur tour, une qualification reconnue non seulement en tenant compte des cadences, mais aussi en assurant certains réglages et en contrôlant la qualité de leur production

(
Voilà que maintenant leur travail est effectué par des bras articulés munis de mains en forme de ventouses ou de pinces et dirigées par des cerveaux cachés dans des coffrets énigmatiques

L'homme n'a plus à manipuler des milliers de pièces, à longueur de journée

Tant mieux

! Mais cependant il reste là, auprès des machines, pour veiller à ce que tout marche régulièrement, pour parer aux incidents imprévisibles, pour contrôler en permanence les produits, pour prévoir l'approvisionnement en pièces détachées, pour changer de séries sans perdre de temps

Il peut se comparer alors au pilote d'un avion de ligne

: en vol automatique, s'il lâche le manche et le palonnier, il garde les yeux rivés sur son tableau de bord pour renseigner par radio la tour de contrôle

Et pourquoi n'appellerait-on pas pilotes de lignes ceux qui font marcher ces chaînes de machines automatisées

?

Cela indiquerait clairement qu'ils gardent la maîtrise de ces automates, dont ils se servent comme jadis l'ouvrier de ses outils rudimentaires

)

NIL

[Commentaire: le texte ressort comme une suite de GP (groupes propositionnel) et de points de ponctuation. Intercallée entre eux, parfois, une parenthèse ouvrante ou une parenthèse fermante. La première indique une ouverture de segment, la seconde une fermeture].

5. Automates II

```
(AUTOBJ (S1 (PHI S2 D) (X S1 G))
  (S2 (MOD* (LEXIDEC MOD* (NIH*) AVANT NON) S3)
    (MOD2* (LEXIDEC MOD2* (NIH*) AVANT NON) S3)
    (MOD3* (LEXIDEC MOD3* (NIH*) AVANT NON) S3)
    (C32 (RETURN))
    (X S2 D))
  (S3 (MOD2* (LEXIDEC MOD2* (NIH*) AVANT NON) S4)
    (MOD3* (LEXIDEC MOD3* (NIH*) AVANT NON) S4)
    (C32 (RETURN))
    (X S2 D))
  (S4 (MOD3* (LEXIDEC MOD3* (NIH*) AVANT NON) S2 D)
    (C32 (RETURN))
    (X S2 D)))
```

6. Modèles II

```
(DEFV MOD* ((ANAF)) ((NL ((=EA))))))
(DEFV MOD2* (GP ((X) (TP +) (GV ((X) (F1 -) (GN ((DEF)) ((NL (
    (=EA))))))))))
(DEFV MOD3* (GGN ((DEM)) ((NL ((=EA))))))
(DEFV NIH* (GP ((X) (TP -) (GN ((DEF)) ((NL ((=EA))))))))
```

7. Description de texte (début)

```
(GPO NIL
  (N211 ((TP 1 +)) (T18 NIL "NOUS"))
  (GV1 ((TP -1 -))
    (GV1 ((P2 7 -) (P2 4 -) (P1 1 -)) (V1 NIL "AVONS"))
    (GV23 ((P1 -1 +))
      (GV23 ((p1P2? 5 -) (P2? 3 -) (P1 1 -)) (GV23 NIL (V23 NIL
"POSE"))))
      (GN ((P1 -1 +) (DET? 2 -))
        (D12 NIL (T11 NIL (DEF NIL "LA")))
        (N1 NIL
          ("QUESTION" (NIH* ("ROBOTISATION")))))
        (C22 NIL "A")
        (GN ((P2? -3 +) (DET? -2 +) (DET? 2 -))
          (D12 NIL (QUA NIL "DEUX"))
          (N1 NIL "CONDUCTEURS"))
        (C211 NIL "DE")
        (GN ((P1P2? -5 +) (DET? -2 +)) (N1 NIL "PRESSES"))
      (C31 NIL ",")
      (C22 NIL "A")
      (GN ((P2 -4 +))
        (GN ((DET 2-))
          (GN NIL (D12 NIL "DEUX")) (N1 NIL "CONDUCTEURS")))
        (C211 NIL "DE")
        (GN ((DET -2 +))
          (GN ((DET? 5 -) (DET 2 -)) (GN NIL (N1 NIL "MACHINES")))
          (C211 NIL "D")
          (GN ((CO? 3 -) (DET -2 +)) (N1 NIL "USINAGE"))
          (C1 NIL "ET")
          (C211 NIL "DE")
          (GN ((DET? -5 +) (CO? -3 +)) (N1 NIL "SOUDURE"))
        (C31 NIL ",")
        (C22 NIL "A")
        (GN ((P2 -7 +))
          (GN ((DET? 3 -) (DET 2-))
            (GN NIL
              (D12 NIL (QUA NIL "DEUX"))
              (N1 NIL ("AGENTS" (NIH* ("ROBOTISATION")))))
            (C211 NIL "D")
            (GN ((DET? 1 -) (DET -2 +))
              (N1 NIL ("ENTRETIEN" (NIH* ("ROBOTISATION"))))))
```

```

(GP11 ((DET? -1 +) (DET? -3 +))
(CP11 NIL (N23 NIL "QUI"))
(GV1 NIL
(GV1 ((P1 1 -)) (V1 NIL "VIVENT"))
(GN ((P1 -1 +))
(D12 NIL (DEM NIL "CETTE"))
(NL NIL ("EXPERIENCE" (NTH* ("ROBO=
TISATION")))))))))))
(C32 NIL ".")

```

8. Liste de noms-têtes de listes (dénominations d'objets)

"ROBOTISATION" "PRESSES" "COUP" "MAIN" "OUVRIER" "CONDUCTEURS" "MACHINES"
"HOMMES"

9. Ingrédients ("spécifiques") des objets (ou: expansion propre de chaque objet)

(Procédure) Exemple 1:

```

(PROJEC (ROB.OBJ) (CL.OB1) (NI(((=X(NTH* ("HOMME"))))))))
(TYPE(CL.OB1))

```

"PRODUITS" "APPROVISIONNEMENT" "MANCHE" "PALONNIER" "YEUX" "TOUR" "CHAINES"
"MACHINES" "MAITRISE" "AUTOMATES" "OUVRIER" "OUTILS"

Exemple 2:

expansion propre de "ROBOTISATION" :

"QUESTION" "AGENTS" "ENTRETIEN" "EXPERIENCE" "IDEE" "REVOLUTION"

10. Interprétation des résultats

Un nom-tête de liste correspond à la dénomination d'un objet ancré dans le discours. On le qualifie également d'"expression générique" en tant qu'il engendre une expansion. Celle-ci est faite d'"aspects de l'objet" matérialisés par des unités lexicales entretenant avec le nom-tête de liste des rapports d'anaphore ou de coréférence. En ce qui concerne l'exemple 2 ("ROBOTISATION") il sera intéressant de savoir, du point de vue de l'analyse des discours qui est le nôtre, que ROBOTISATION est anaphorisé par: QUESTION, EXPERIENCE, IDEE et REVOLUTION.

Autant de substituts contextuels au mot "ROBOTISATION" qui marquent, dans ce texte particulier les différents aspects sous lesquels est repris ce nom. On saura également qu'il existe un lien d'anaphore associative entre ROBOTISATION et AGENTS ou ENTRETIEN.

En ce qui concerne l'exemple 1, l'expansion de HOMME est beaucoup plus vaste, ce qui est assez attendu. Dans le contexte actualisé par ce discours, les aspects sous lesquels HOMME apparaît sont bien sûr liés à des activités de travail. Ainsi l'expansion de HOMME dans le texte recoupe l'expansion de HOMME dans la réalité du travail, puisqu'on y trouve:

PRODUITS, APPROVISIONNEMENT, MANCHE, PALONNIER, TOUR, CHAINES, MACHINES, AUTOMATES, OUTILS

autrement dit: des types d'activité ("approvisionnement") des résultats

de ces activités ("produits") et des moyens de ces activités ("outils", "machines"...).

Les relations ici implicites sont, bien sûr, des relations d'anaphore associative. L'expansion considérée contient également: YEUX, c'est le seul ingrédient de la classe-objet HOMME qui ne soit pas directement lié au monde du travail et fasse en quelque sorte partie des aspects "naturels" de l'objet. Cette expansion très particulière "explique" alors le seul lien de coréférence apparent dans cette série, qui est le lien entre HOMME et OUVRIER.

"Notre "analyse" automatique rate malheureusement par manque de finesse dans la détermination des modèles, une autre unité coréférentielle à HOMME et qui serait: PILOTES DE LIGNES (présents dans notre expansion seulement par l'intermédiaire de deux de ses ingrédients: MANCHE et PALONNIER).

Au prix d'une légère complexification de notre algorithme, nous pourrions obtenir cependant cette relation de coréférence qui nous permettrait d'en établir une autre; celle vers laquelle le texte converge et qui lui sert de conclusion: l'équivalence entre OUVRIER et PILOTES DE LIGNES, interprétable comme substitution orientée ("aujourd'hui grâce à la ROBOTISATION, l'OUVRIER devient PILOTE DE LIGNE").

Comme on le voit, les résultats obtenus par ces algorithmes programmables en DEREDEC peuvent donner lieu à des interprétations qui rappellent les essais d'analyse automatique du discours tels que AAD 69 de M. Pêcheux effectués dans la lignée des travaux de Z. Harris.

Ils s'en distinguent toutefois par une volonté de mise en relief de la structure argumentative des textes: les items mis en équivalence ne le sont plus à partir d'une identité de distribution dans des moules préétablis, mais à partir de critères de formes qui sélectionnent des syntagmes en fonction de leur rôle (ancrage, expansion, spécification,...).