

Marc CAVAZZA*

Sémiotique textuelle et contenu linguistique

Les rapports entre le contenu sémantique et les formalismes de description sont un problème majeur pour le développement d'une technologie informatique des objets culturels, dont le traitement automatique des textes n'est qu'une première étape. Dans cet article, en prenant comme cadre théorique la sémiotique du texte, nous passons en revue quelques problèmes classiques liés à la description du contenu sémantique. Ceci nous amène à avancer un certain nombre de directions de recherche pour aborder les problèmes de formalisation sous un angle différent, en inversant les rapports traditionnels entre forme et contenu, selon lesquels ce sont les formalismes qui déterminent leur instanciation, indépendamment de considérations globales sur les objets sémiotiques à traiter.

Mots-clés : sémiotique, sémantique textuelle, sémantique lexicale.

From Text Semiotics to Linguistic Content. A major issue for the development of humanities computing technologies is to determine the proper relationships between form and content in semiotic objects. In this paper, we review some current conceptions of semantic content description in the framework of textual semiotics. We then suggest some further research directions, which aim at subordinating the instanciation of specific formalisms to the actual description of underlying semantic content. These proposals challenge the traditional conceptions of Cognitive Science, where formalisms are instanciated without considering the global properties of semiotic objects.

Key Words : Semiotics, Text Semantics, Lexical Semantics.

1. SEMIOTIQUES, LINGUISTIQUE ET SCIENCES COGNITIVES

La sémiotique est le plus souvent présentée comme une généralisation de l'étude des systèmes de signes, effectuée à partir de

* Equipe «sémantique des textes», Centre de Linguistique Française, Université Paris IV, 1, Rue Victor Cousin, 75230 Paris CEDEX 05

l'étude des propriétés des signes linguistiques. Cette généralisation peut cependant adopter des approches très différentes, parmi lesquelles on peut distinguer trois tendances principales¹ :

— Une première généralisation se fait du signe vers le symbole, par une abstraction logique de la sémantique ; elle adopte le paradigme logique comme base de la sémiotique (Granger, 1979). Nous la désignerons ici par le terme de *sémiotique du symbole*. Cette sémiotique se pose souvent en théorie générale du signe, et se rapproche souvent de la philosophie du langage (en particulier dans ses aspects formels) ; la sémiotique peircienne et ses diverses interprétations en font également partie.

— La seconde procède transversalement entre différents systèmes de signes intervenant dans la communication ; toutefois, ce sont encore les contenus et les systèmes de signes qui seront privilégiés à ce niveau. Les travaux les plus connus dans ce domaine sont sans doute ceux de Barthes (1967) et Metz (1977). Le problème qui domine cette approche est celui de la fondation du sémiotique par le sémantique. Barthes (1985), en introduisant la notion de sémiotique *isologue*, installe, peut-être sans le vouloir, la sémantique au centre de cette sémiotique « multimédia ».

— La troisième, très différente des deux premières, étend le domaine du signe linguistique du mot au texte ; de Hjelmslev (1968) à Rastier (1989) cette *sémiotique du texte* va se fixer pour objet le texte. Il y a là une coupure épistémologique essentielle avec la sémiotique du signe-symbole, mais également une évolution par rapport au structuralisme traditionnel et à son approche des systèmes de signes en langue². En effet, en se fondant sur des objets attestés que sont les textes, l'étude des systèmes de signifiés peut être étendue au processus interprétatif sans pour autant perdre le contact avec les objets linguistiques ; il est ainsi possible d'étudier la dynamique des signes sans quitter la sphère sémiotique.

L'étude de la sémiotique du texte ouvre des perspectives importantes, tant sur le plan épistémologique, dans la discussion du problème du sens en linguistique et en sciences cognitives, que sur un

¹ Cette classification est bien entendu orientée par le propos de l'article et ne prétend pas à l'exhaustivité. On pourra en particulier lui reprocher de ne pas accorder de place à la sémiotique du récit (au sens de l'école française).

² La sémiotique du texte est en fait une *sémantique* du texte qui relève également de la sémiotique des cultures.

plan opératoire, pour la description des données linguistiques dans des applications de traitement automatique des textes.

Tirant leur importance d'enjeux théoriques ou des nécessités de l'implémentation informatique, un certain nombre de problèmes classiques en sciences cognitives ont été décrits, essentiellement pour des raisons historiques, dans le cadre de la sémiotique du symbole. En d'autres termes, on pourrait dire que le contenu a souvent été pensé par rapport aux exigences de la forme, qui bénéficiait d'une meilleure description. Dans cet article, nous proposons donc de passer en revue les implications de la sémiotique du texte pour la description du contenu sémantique, de façon à pouvoir proposer des directions de recherche permettant d'aborder les rapports entre forme et contenu à partir du contenu lui-même.

2. PRINCIPES DE LA SEMIOTIQUE DU TEXTE

Une description détaillée de la Sémantique Différentielle Unifiée (SDU), qui fonde la sémiotique du texte (Rastier, 1989), dépasse le cadre de cet article. Il est néanmoins utile d'en rappeler ici les principes et en particulier les relations dynamiques entre contenu sémiotique et interprétation. La Sémiotique du texte repose sur la Sémantique Différentielle Unifiée, théorie sémantique lexicale qui unifie la description des signifiés du mot au texte, en postulant, d'une part, une dynamique du contenu lexical et, d'autre part, la formation des structures sémantiques *textuelles* (isotopies et molécules sémiques), construites à partir des traits sémantiques lexicaux. En tant que théorie componentielle, elle est interprétative et le texte y est l'objet central de la description (d'où ses relations avec le programme de *l'herméneutique matérielle*). La possibilité de fixer des interprétations dans le formalisme sémantique, c'est-à-dire d'identifier des règles d'actualisation du contenu sémantique, confère à la théorie une capacité descriptive utilisable aussi bien pour l'analyse littéraire que pour des applications de traitement automatique des textes :

— Le texte est la réalité principale du processus d'interprétation dont il est à la fois le point de départ et le point d'arrivée ; la structure pivot du processus d'interprétation est le sémème, qui fournit les traits sémantiques de départ et dont le contenu est actualisé par les opérations interprétatives.

— La description unifiée, du lexique au texte, est ancrée sur des régularités elles-mêmes fondées sur des normes culturelles (genre et typologie des textes), et dont les textes sont un objet empirique.

— L'interprétation est basée sur une description figée de l'actualisation du contenu sémantique et de la construction, à différents niveaux, micro-, méso- et macrosémantique, de représentations du contenu sémantique ; elle n'est reliée à aucun sujet psychologique.

— La description du contenu sémantique se fonde sur le paradigme différentiel, qui comprend tout aussi bien l'opposition entre sémèmes que la démarche de constitution de classes élémentaires de signification.

Par rapport à la sémiotique du symbole, la sémiotique du texte est donc avant tout une sémiotique du *contenu* ; toutefois, ses aspects dynamiques et interprétatifs lui permettent de disputer à l'approche logico-symbolique ses aspects inférentiels, tout en conservant une continuité avec les objets sémiotiques empiriques que sont les textes.

3. DU SYMBOLE EN CONTEXTE A LA SEMIOTIQUE DU TEXTE

L'opposition entre sémiotique du texte et sémiotique du symbole se manifeste en premier lieu dans les modes de description et de gestion du contenu des signes. Pour une sémiotique du symbole, c'est en effet l'approche référentielle et hiérarchique qui domine, alors que la sémiotique du texte repose sur la description différentielle — donc *contextuelle* — des signifiés. La portée exacte de l'approche différentielle du lexique est encore mal comprise, faute de disposer de vastes lexiques différentiels (plusieurs milliers à plusieurs dizaines de milliers d'entrées). Toute la lexicographie repose sur des processus classificatoires ou définitoires et l'utilisation principale est faite d'éléments génériques hors contexte³ plutôt que d'éléments spécifiques.

Alors que la sémiotique du symbole est perpétuellement confrontée au problème de la référence du signe, qui est en fait son problème fondateur, la sémiotique du texte, notamment dans ses aspects différentiels, se trouve confrontée à un problème dual : la projection du non-linguistique dans le signe linguistique. En d'autres termes, comment une opposition entre signifiés donne-t-elle naissance à un sème précis dans un contexte interprétatif donné⁴? Cette projection, qui résulte du processus interprétatif, ou d'une version fixée de

³ Cette précision — hors contexte — a son importance ; nous verrons plus loin que la recherche d'une généralité minimale signe le paradigme différentiel au même titre que la relation d'opposition, qui est trop souvent privilégiée.

⁴ Rappelons que la notion d'interprétation admet des variantes faibles, qui se prêtent à des approches opératoires, contrairement à la *sémiosis illimitée*.

l'interprétation définie comme lecture, n'est pas accessible à la systématisation. En effet, si l'on peut projeter des éléments de connaissance dans le lexique, c'est bien une conséquence directe du rôle central du texte dans le processus interprétatif, en particulier du fait que les connaissances « encyclopédiques » activées lors de l'interprétation, peuvent se déposer dans le lexique et effectuer un retour vers le texte ; la seule systématisation possible est *a posteriori*, et se fonde sur des principes de régularité et des normes de genre.

La sémiotique du texte repose sur une conjonction de hiérarchies locales (des classes sémantiques minimales, imbriquées, et en petit nombre), de catégories spécifiques et de différences, le tout s'opposant à une approche purement hiérarchique basée sur des catégories générales. Il ne s'agit de rien de moins que d'une alternative à la conception transmise par les catégories d'Aristote et l'arbre de Porphyre, qui domine pratiquement sans partage la sémantique cognitive, comme on pourra s'en convaincre, par exemple, à la lecture de Sowa (1984) ; d'autres arguments critiques sur ces aspects de la philosophie du symbole sont développés dans Rastier (1990), à propos de la tripartition syntaxe-sémantique-pragmatique.

De son côté, Eco, dans un article intitulé *L'antiporphyre* (Eco, 1984), va fonder l'opposition entre approche différentielle et approche hiérarchique. Ce thème sera développé plus tard dans Eco (1988), mais dans une direction différente ; aussi nous en tiendrons nous ici à l'article original. Eco prend comme point de départ l'arbre de Porphyre, entendu comme un arbre unique censé recouvrir la totalité des espèces, organisé par genres et différences. Mais en développant la notion de différence, il met en évidence, à partir d'Aristote et de ses commentateurs, deux difficultés majeures :

— La première, que l'arbre ne définit les différences qu'entre certains types et pas certains autres (ce point sera largement repris par Eco par la suite).

— La seconde, que certaines différences ont une portée qui dépasse les types pour lesquels elles ont été identifiées, et sont applicables à de nombreux types, à différents nœuds de l'arbre et à différents niveaux de profondeur de cet arbre.

Ces constatations, déjà faites par Abélard, portent un coup fatal à l'existence d'un arbre unique et invariable, la seule solution aux problèmes décrits ci-dessus étant la duplication des différences à plusieurs niveaux de l'arbre. Mais ceci entraîne une nouvelle

conséquence, qui est que la hiérarchie peut et doit être modifiée selon le contexte à l'œuvre :

À partir du moment où «mortel» n'implique pas «rationnel», qu'est-ce qui interdit de placer «rationnel» sous «mortel» contrairement à ce qui arrive dans l'arbre classique⁵ ?

Et Eco de proposer une série d'arbres alternatifs, qui ne sont pas sans rappeler la constitution de taxèmes : “lait” et “perle” se distinguent ainsi, dans l'arbre des //objets blancs//, par les traits /solide/ et /liquide/, alors que “perle” et “ébène” se distingueraient dans l'arbre des //objets solides// par les traits /blanc/ et /noir/.

Sa conclusion tend à invalider la possibilité pour une approche hiérarchique de décrire la sémantique des langues, mais va surtout poser le problème du *contexte* :

Dans un arbre composé exclusivement de différences, celles-ci peuvent être réorganisées à loisir suivant le point de vue selon lequel un sujet donné est considéré. L'arbre est une structure sensible au contexte, et non un dictionnaire absolu.

Cependant, plus loin, Eco défend l'hypothèse selon laquelle les traits sémantiques possèdent un caractère indiciaire, qui servirait de base à une *sémiosis illimitée*, à la Peirce. Probablement rejette-t-il ainsi leur nature synthétique, brusquement perçue comme encombrante ? Cette *sémiosis illimitée*⁶ est cependant incompatible avec une approche opératoire des éléments différentiels, car elle débouche sur des descriptions arbitraires ou, pire encore, sur le retour d'un *sujet* interprétant. Ces considérations peuvent cependant expliquer qu'Eco puisse caractériser la Sémiotique du Texte comme étant une entreprise « encyclopédique ».

Nous préférons faire l'hypothèse qu'il est possible de conserver le caractère synthétique des traits sémantiques à condition de justifier empiriquement d'un contexte qui détermine ce caractère synthétique ; c'est là sans doute la supériorité d'une approche textuelle sur une approche encyclopédique où les éléments de l'encyclopédie restent à convoquer *ex nihilo*. En fait, il reste certainement à décrire ce que

⁵ Dans un des exemples classiques tirés de l'arbre de Porphyre, le genre « mortel » se subdivise en « rationnel » et « irrationnel ».

⁶ À toutes fins utiles, il peut être utile de préciser que la *sémiosis illimitée* ne doit pas être confondue avec la circularité de la description métalinguistique des traits sémantiques, qui découle de ce que les traits sémantiques sont eux-mêmes décrits en langage naturel.

serait une entreprise d'étude sémiotique des textes ; ni dictionnaire ni encyclopédie, il s'agirait sans doute de la reconstitution de classes sémantiques dans des corpus étendus et typés par genres. Des études de vocabulaires spécialisés en donnent un aperçu prometteur (Ballòn Aguirre *et al.*, 1992).

Il importe maintenant de préciser que la démarche différentielle consiste tout autant à constituer les classes sémantiques en contexte qu'à formuler les oppositions à l'intérieur des classes. Des oppositions ne sauraient à elles seules définir le paradigme différentiel, en particulier lorsque ces oppositions sont présentées en dehors de tout contexte. En ce qui concerne les classes de signification, on peut se reposer sur des analyses distributionnelles, des analyses de co-occurrences, voire sur la traditionnelle épreuve de commutation ; *in fine*, la capacité d'induction du lecteur semble toujours dépasser les capacités des traitements statistiques, particulièrement dans le cas de données hiérarchisées. Enfin, le paradigme différentiel n'est pas neutre par rapport à l'ontologie et en particulier la querelle des universaux (de Libéra, 1995). En définissant les différences spécifiques comme des « accidents essentiels », il porte aussi un coup sévère à la conception classique des catégories en linguistique⁷. On peut remarquer sur ce point la filiation directe qui existe entre les concepts analytiques et les catégories, filiation qui passe par les notions de Condition Nécessaire et Suffisante (CNS) ou de postulat de signification (*meaning postulate*). La nature synthétique des traits, fondée sur les contextes, entraîne leur prolifération mais délègue aux normes de genre le soin d'en limiter le nombre, ou plutôt de le rendre simplement fini. Le type est alors, dans les termes de F. Rastier, une « collection d'accidents ». De cette nature synthétique découle l'hypothèse *nominaliste* des traits sémantiques abordée plus loin, et qui constitue une autre divergence majeure entre sémiotique du symbole et sémiotique du texte.

4. LA DESCRIPTION : CONTENU ET CONTINUITÉ

4.1 Principes et méthode de la description

Nous avons dans des travaux antérieurs (Cavazza, 1994) étudié les problèmes de description du contenu lexical. Après un bref rappel des problèmes fondamentaux, nous poursuivons ici quelques unes des directions abordées alors, en les inscrivant dans le cadre de la sémiotique des textes.

⁷ Notre réserve à l'égard des catégories ne vise que l'usage qui en a été fait en linguistique, à l'exclusion des recherches psychologiques.

En premier lieu, si la sémiotique du symbole n'est certes pas incompatible avec une description du contenu, sa position reste le plus souvent lexicographique, là où la sémiotique du texte sera lexicologique, et il n'est pas inutile de revenir sur cette distinction. Si l'entreprise lexicographique est largement privilégiée dans le domaine de la sémantique informatique, comme en témoignent les travaux sur l'extraction automatique de lexiques à partir de dictionnaires (Boguraev & Briscoe, 1989), les approches componentielles et contextuelles du signifié ont été largement décriées. Il est pourtant paradoxal de rejeter la lexicologie comme *ad hoc*, là où l'on accepte que des principes similaires soient figés dans la lexicographie, sous la forme de définitions parfois contestables. Comme nous le verrons plus loin, ceux qui critiquent l'approche componentielle ne peuvent éviter de la réinventer, ne serait-ce que dans la dénomination des propriétés, attributs, fonctions ou valeurs de différentes structures et formalismes lexicaux.

La difficulté principale de la lexicologie demeure la description de traits à partir du texte, notamment la reconnaissance de classes de signification en contexte. On a pu décrire un certain nombre de phénomènes syntaxiques jouant un rôle majeur dans l'actualisation des traits et leur mise en relief, souvent en reproduisant dynamiquement au cours de l'interprétation des mécanismes d'opposition ou de contraste. On trouvera chez Rastier (1987 ; 1989) des exemples du rôle de structures syntagmatiques (parataxe, énumérations ...) permettant de former des interprétants syntaxiques. Rappelons simplement que les phénomènes susceptibles de manifester de façon statique l'existence d'un taxème sont les mêmes qui jouent un rôle dans la propagation dynamique du contenu. C'est par exemple le cas de l'énumération, qui d'une part linéarise certains taxèmes (Rastier *et al.*, 1994, p.62) et, d'autre part favorise la propagation des traits,

Dans un texte récent sur le cryptage des fichiers informatiques, traduit d'une revue américaine, on a pu ainsi lire que le cryptage est interdit *en Iran, en Iraq et en France*. La lecture évaluative du sémème "France" s'enrichit de la propagation du trait /péjoratif/, afférent à la description des deux premiers sémèmes dans la topique propre à cet article. La propagation de traits sémantiques le long des énumérations concerne tout aussi bien des sèmes inhérents que des sèmes afférents ; ses mécanismes sont cependant plus complexes que l'exemple ci-dessus ne le laisse supposer, puisque la possibilité de propagation peut être sous le contrôle d'interprétants de nature topique (cf. *infra*).

Le travail récent de Mettinger (1994) apporte un éclairage complémentaire sur les relations entre des structures syntaxiques particulières et l'opposition en contexte de sémèmes, qui joue un rôle aussi bien dans la description que dans la dynamique des traits sémantiques au cours du processus interprétatif. Mettinger recense en effet, à partir d'une étude de corpus portant sur une vingtaine de romans en langue anglaise, une douzaine de structures syntactico-argumentatives qui mettent des sémèmes en opposition⁸. Parmi ces structures on peut citer les constructions *X et Y*, *ni X ni Y*, *X ou Y*, *X plutôt que Y*, *de X à Y*, *X et en même temps Y*, etc. Ces structures peuvent être rattachées à des « fonctions textuelles », par exemple la *coexistence*, comme dans l'exemple suivant⁹ :

Cornelia was the prey of conflicting emotions. She was deeply *embarrassed* but at the same time pleasurably *thrilled* (Mettinger, 1994, p. 48).

Nous ne nous prononcerons pas ici sur la nature exacte du lien entre la sémantique différentielle et la théorie de l'argumentation. On peut toutefois risquer l'hypothèse d'un lien instrumental entre certaines formes argumentatives et, d'une part, la manifestation de taxèmes, et, d'autre part, la dynamique des contenus sémantiques (dans ce second cas, on serait en présence d'une nouvelle classe d'interprétants). Ces processus dynamiques étant stabilisés, ce qui constitue la méthode de description d'une *lecture*, on peut inclure la recherche de ces structures argumentatives dans l'arsenal méthodologique de la description sémantique.

Mettinger distingue encore les oppositions non contextuelles et les oppositions contextuelles. Les premières sont souvent difficiles à distinguer de l'antonymie pure et simple ; les secondes correspondent aux oppositions décrites par la sémantique interprétative, y compris dans leur variabilité, comme l'illustrent les deux exemples suivants¹⁰ :

She drew out from her handbag the letter she had received that morning from Lady Tamplin. Katherine was no fool. She understood the nuances of that letter as well as anybody and the reason of Lady Tamplin's show of affection towards a long-forgotten cousin was not lost upon her. It was for *profit* and not for *pleasure* that Lady Tamplin was so anxious for the company of her dear cousin.

⁸ Ce travail peut être interprété dans le cadre de la sémantique différentielle sans craindre de trahir la pensée de son auteur, dans la mesure où la référence à Coseriu y est fréquente et explicite.

⁹ Extrait de *Three Acts Tragedy*, d'A. Christie (1935).

¹⁰ Extraits de *Mystery of the Blue Train*, d'A. Christie (1928).

'It is as well you are in France', said Van Aldin. 'We are made of sterner stuff in the States. *Business* comes before *pleasure* there.' (Mettinger, 1994, p.72).

Le caractère contextuel¹¹ de l'opposition nécessite parfois l'introduction de connaissances encyclopédiques pour décrire les sèmes résultant de l'opposition ; ce qui ne pose pas de problème en soi, pourvu qu'une relation claire puisse être établie entre connaissances linguistiques et encyclopédiques. Un des avantages majeurs de l'approche différentielle tient précisément dans le fait que des connaissances encyclopédiques sont projetées dans un formalisme linguistique et utilisées pour construire des représentations sémantiques de niveau supérieur (isotopies et molécules sémiques). Mais pour maintenir la validité de cette description, il est souhaitable que ces données encyclopédiques puissent correspondre à des données linguistiques rencontrées dans d'autres contextes. Nous avons déjà décrit certains phénomènes qui peuvent concourir à la minimalité des sèmes (Cavazza, 1994), comme la permutation de classes de sèmes (c'est-à-dire le fait qu'un trait spécifique puisse apparaître ailleurs comme un trait générique), l'ensemble de ces phénomènes restant encore à recenser, ce qui constitue un des axes de recherche de la sémiotique du texte. L'autre aspect important, caractéristique de l'approche textuelle, est celui de la portée descriptive à l'intérieur du corpus. Les oppositions sont à interpréter en contexte, notamment quant au choix où à la dénomination des traits. Par exemple, une opposition relativement classique comme rêve-réalité pourra être décrite par le sème /illusion/, et ce avec d'autant plus de certitude que ce trait est capable de décrire d'autres parties du texte en tant que trait générique. Le phénomène de permutation de classe est, ici encore, susceptible de jouer un rôle dans la cohérence des descriptions.

Mais comment alors concilier la description des sèmes sur un corpus, qui pourrait apparaître *ad hoc*, avec la généralité de l'approche et une réutilisabilité des lexiques sémantiques ainsi construits ? À l'intérieur d'un genre textuel et d'une thématique (par exemple les commentaires sur la vie politique française), il existe une régularité sur

¹¹ Il est nécessaire de poser une limite supérieure aux oppositions contextuelles, quand celles-ci ne sont pas normées, comme le suggère par ailleurs Martin (1990), qui considère que l'opposition entre "homme" et "orgue", dans l'énoncé *ce n'est pas un homme, c'est un orgue*, ne saurait pour autant servir de base à la constitution d'un taxème.

les traits (par exemple ; //désaccords//¹², //alliances//, //succès électoraux// ...), qui relève de la topique propre au genre.

En l'absence de tout contexte, l'opposition, si elle demeure possible, n'en pose pas moins un certain nombre de problèmes. Ainsi n'est-il pas sacrilège de reprendre le taxème des sièges chez Pottier, pour s'interroger non pas sur la minimalité mais sur la pertinence des traits proposés. Greimas (1966, p. 37) avait suscité ce doute, en mettant en cause l'utilité pratique des sèmes /avec dossier/ et /avec bras/ dans ce taxème. La sémiotique du texte permet de mettre fin à ce type d'interrogations, d'une part grâce au contexte naturellement fourni, d'autre part, par la recherche de normes de genres pour guider la description. En d'autres termes, dans la sémiotique textuelle, *la description n'est que de l'interprétation figée*.

4.2. L'exemple de l'isotopie

La notion d'isotopie est caractéristique de la sémiotique textuelle dans la mesure où, d'une part, elle unifie différents paliers de description et, d'autre part, elle est exclusivement basée sur le signifié. Elle est représentative d'une approche non-compositionnelle et sans primat syntaxique¹³. L'unification de différents paliers est impossible à une sémiotique du symbole dont les partis pris vont déléguer à un niveau pragmatique les opérations contextuelles. Par exemple, la notion de cohésion textuelle a été le plus souvent placée dans le domaine de la pragmatique et déléguée notamment aux anaphoriques¹⁴, à qui on confère un statut particulier ; en restant dans la sémantique, la cohésion est traitée par la notion d'isotopie.

Le pouvoir de description de l'isotopie est généralement sous-évalué ; au-delà de la constitution d'isotopies linéaires à partir des contenus lexicaux, certains entrelacements d'isotopies peuvent directement manifester des structures de niveau supérieur, notamment un niveau élémentaire de la composante dialectique. Ainsi l'exemple ci-dessous (extrait du journal *Le Monde*) contient-il deux isotopies ; l'une, sur le trait /intellectuel/ (marquée en italique) et l'autre (soulignée) sur le trait /répression/. Nous avons volontairement choisi d'exprimer ces isotopies à un niveau générique.

¹² Suivant la convention, nous noterons //taxème// les désignations des classes sémantiques.

¹³ À l'exception bien sûr du niveau mésosémantique, qui recouvre une partie des fonctions de la syntaxe.

¹⁴ Certains anaphoriques réintroduisent sous une forme différente la problématique de la référence, extérieure au texte.

L'interdiction par l'Autorité Palestinienne des *livres* d'Edward Said, *auteur* américain d'origine palestinienne hostile au processus de paix israélo-palestinien défini par les accords d'Oslo et au gouvernement de Yasser Arafat, a été condamnée dans un communiqué, vendredi 6 septembre, par une trentaine d'*intellectuels* et *écrivains* de pays arabes, qui dénoncent une violation des libertés fondamentales.

Un certain nombre de travaux récents basés sur des corpus proposent des notions similaires à celles d'isotopie ; ainsi la notion d'*information mutuelle* (Church & Hanks, 1989) est-elle proche de la notion d'isotopie, mais dans la mesure où elle se base sur la simple co-occurrence de termes pris deux à deux, son pouvoir descriptif est faible. Plus récemment, Morris et Hirst (1991) ont utilisé les relations du *thesaurus* (qui sont des relations structurales figées, mais entre lexèmes et non entre sémèmes), pour décrire la cohésion textuelle. La relation de l'isotopie au contenu lexical et à d'autres structures textuelles lui confère cependant un pouvoir descriptif bien supérieur et possède l'avantage d'être utilisable à des niveaux de spécificité très variables, conformément aux principes énoncés plus haut.

4.3. Une sémantique de l'évaluation

Les textes narratifs simples ont été pris comme modèles, aussi bien pour décrire l'analyse par les sujets que pour définir des procédures de traitements automatiques. Curieusement, cette approche a été désignée sous le nom de sémiotique du récit, et la plupart des procédures d'analyse de récit qui ont constitué l'essentiel du TALN dans les années 70 (avec en particulier l'école de Yale) mettent en œuvre des concepts transposés de ceux de l'analyse structurale du récit. Mais le traitement explicite des textes, outre les difficultés auxquelles il se heurte, est assez éloigné de l'utilisation réelle qui est faite des documents. Si l'on considère par exemple la recherche d'informations dans une base textuelle, celle-ci se fait non pas par une lecture narrative du document cherchant à reconstituer les étapes du récit, mais bien plus par des lectures en diagonale, qui dégagent des éléments thématiques, topiques, éventuellement isolés mais pourtant interprétables en tant que tels et tout à fait significatifs. Le point de vue que nous défendons ici est le suivant : de très nombreux textes sont d'abord lus pour leur *évaluations*¹⁵. Comme l'écrivait Barthes (1985, p. 331) :

¹⁵ Nous préférons ce terme à celui de connotation, trop marqué par l'opposition connotation/dénotation.

La description physique d'un personnage, étendue à plusieurs phrases, peut n'avoir qu'un signifié de connotation, qui est la «nervosité» de ce personnage, bien que le mot ne figure pas au plan de la dénotation.

L'évaluation n'est généralement pas décrite par des données structurelles ; elle est au contraire fortement dépendante du lexique. Sur le plan de la technique linguistique, l'évaluation peut être représentée aussi bien par des dimensions évaluatives (par ex. /favorable/, /péjoratif/ ...), par des domaines que par des traits spécifiques en contexte ou des traits afférents¹⁶. Les niveaux de style, dans leur projection lexicale, relèvent tout autant de phénomènes d'évaluation. Mais plus que le sens de telle ou telle lexie, c'est l'usage du texte tout entier qui est régi par les phénomènes d'évaluation ; dès les premiers instants de la lecture d'un texte, on est en mesure de dire si son ton est positif, critique, dramatique ou emphatique. Nombre d'applications pratiques relèvent d'ailleurs de ce type de problématique, comme la détermination du caractère élogieux ou critique d'un article politique, du caractère de gravité d'un bulletin de santé, du caractère favorable d'une critique théâtrale ou littéraire ... La combinaison de certaines évaluations avec des structures thématiques données reste souvent la meilleure interprétation du texte.

Pour une sémiotique textuelle appliquée, la lecture du texte s'attache à des phénomènes de mise en relief des traits qui fondent la thématique et la dialectique. L'opposition tend en principe à produire des traits hautement spécifiques, pourtant, ils peuvent être appelés à jouer un rôle essentiel dans la description du texte, en particulier quand l'opposition se fait sur des dimensions évaluatives (le fait que l'opposition se fasse sur des dimensions évaluatives étant un cas particulier de permutation de classe). Sur le plan méthodologique, on pourra ainsi orienter les différences vers des dimensions évaluatives ou des traits pertinents compte tenu du type du texte (ou de la classe de texte) étudié(s). Ainsi, dans un texte sur la santé d'un chef d'état¹⁷, on cherchera en priorité les traits indiquant la gravité ou au contraire l'amélioration, comme dans la dépêche suivante, extraite du journal *Le Monde* :

L'absence de Boris Eltsine de la scène publique depuis plusieurs jours alimente de nouveau les *rumeurs* sur l'état de santé du président russe. Le porte-parole du Kremlin a démenti, lundi 19 août, que le chef de l'Etat s'apprêterait à

¹⁶ Il faut aussi souligner ici le rôle primordial de l'afférence, en particulier pour les dimensions évaluatives normées (ex. /intelligence/ Ø /positif/).

¹⁷ Ce type de texte tend à constituer une fraction significative des articles de la plupart des quotidiens.

subir une *opération chirurgicale* en Suisse, comme l'a affirmé lundi l'hebdomadaire américain Time. De telles publications, a déclaré Serguei Iastrjembski, s'inscrivent dans les campagnes de *bruits alarmistes* qui surviennent périodiquement concernant la *santé* du président. (...) Samedi, le quotidien allemand à grand tirage Bild Zeitung, citant des sources proches des *services secrets*, avait affirmé que le président russe souffrait de *troubles cardiaques graves* et qu'il était sur le point de subir une *transplantation du cœur* en Crimée.

4.4. Formalisation et représentation

Un inconvénient de la sémiotique du symbole appliquée à la linguistique est de maintenir la confusion entre représentation et formalisation. Cette distinction, que nous reprenons de Jayez (1988), joue un rôle essentiel dans la différenciation de la sémantique linguistique (donc des rapports synchroniques et diachroniques entre signifiés) par rapport à la sémantique logique. La représentation confère au formalisme un simple rôle de vecteur ; les manipulations formelles peuvent servir de base à un travail sur le concept ou à des réalisations informatiques.

La formalisation suppose (et le plus souvent, impose) quant à elle une continuité entre le formalisme et l'objet représenté¹⁸, alors que la représentation évite soigneusement une telle hypothèse et utilise les formalismes uniquement pour leur aspects pratiques (ce que certains auteurs anglo-saxons dénomment utilisation *theory-free*). Dans la sémiotique du texte, et encore une fois en opposition avec la sémiotique du symbole, c'est l'approche représentationnelle qui domine.

4.5. Une approche nominaliste du contenu

La notion de primitive sémantique va de pair avec une théorie compositionnelle et non interprétative du sens ; elle suppose des mécanismes formels particuliers ainsi que l'existence d'universaux conceptuels (s'ils prétendent valoir pour toutes les langues) ou linguistiques (s'ils décrivent la sémantique d'une langue particulière). Cette approche se caractérise également par son analyticit , c'est-à-dire la conjonction de primitives stables et d'un formalisme assurant leur compositionnalité. De nombreux arguments réfutent l'approche analytique sur le plan théorique, mais il serait en principe possible d'approcher des notions synthétiques par un formalisme analytique si

¹⁸ Ce qui reviendrait à conférer une certaine réalité à l'inscription du contenu dans les symboles.

leur mise en œuvre ne se heurtait trop souvent à des limitations techniques comme la gestion d'héritages multiples ou de raisonnements terminologiques, incapables de produire une solution formelle au problème de la pertinence du niveau de description, ou même des formalismes qui soient tout simplement accessibles à une linguistique descriptive.

Supposons maintenant que les éléments sémantiques soient des traits synthétiques décrits et manipulés comme tels ; leur nombre va être considérablement plus élevé que celui des primitives, tout en restant fini à l'intérieur d'un corpus donné¹⁹. Le critère de minimalité est remplacé par un critère de cohérence interne qui est basé sur l'existence de régularités. Par exemple, des traits définis à l'intérieur de certaines classes de signification possèdent un pouvoir de description étendu valable pour d'autres classes éloignées (cf. le trait /extra-urbain/, (Cavazza, 1994)). Mais le caractère synthétique va s'opposer à des hiérarchisations ou une typologie (autre qu'empirique) des traits : c'est cette absence de discours sur des types et la prévalence des sèmes en tant qu'entités singulières qui introduit une approche « nominaliste » de la description des sèmes²⁰.

La dénomination des sèmes se trouve être un problème majeur pour toutes les théories, mais plus particulièrement dans l'approche synthétique. En effet, si le sème peut être décrit comme une « périphrase à vocation métalinguistique », on oscille entre deux problèmes : une exigence de lisibilité du sème (voir aussi la notion de *gensym fallacy* chez Wilks (1991)) et la nécessité d'un intitulé simple, qui facilite également la vérification de la cohérence d'une description manuelle sur des vocabulaires ou des corpus étendus. L'enjeu théorique de la dénomination des sèmes est celui de la *sémiosis illimitée* (cf. *supra*) ; mais l'enjeu pratique est celui de la description manuelle de lexiques sémantiques de taille significative, qui impose une bonne

¹⁹ La notion de corpus n'est pas à prendre ici au sens restrictif d'échantillon, mais peut être étendue au corpus d'un genre donné.

²⁰ De ce que le nominalisme et le constructivisme s'opposent tous deux au réalisme naïf dominant la sémantique des langues naturelles, il ne s'ensuit pas directement qu'on puisse considérer le nominalisme comme une sorte de constructivisme. Pourtant, nous ferons l'hypothèse, qui sera développée ailleurs, que, dans le contexte d'une sémiotique du texte, le nominalisme entretient des relations privilégiées avec le constructivisme et qu'il est même assimilable à un « constructivisme sémiotique ». D'autres indices du même ordre sont présents chez Largeault (1993), et, dans une certaine mesure, dans la lecture goodmanienne de Cassirer.

lisibilité tant pour des raisons de maintenance que d'échange entre différents intervenants.

4.6. Conclusion sur la notion de lecture

La compréhension automatique d'un texte a été assimilée à la construction d'une représentation de ce texte, et plus précisément des événements décrits dans le texte. Ces représentations, dites conceptuelles, sont obtenues précisément en « supprimant » le niveau linguistique. Or, bon nombre de problèmes pratiques posés à la manipulation automatique des textes sont imparfaitement décrits par cette approche. Une partie des lectures se limite pourtant à un niveau de description différent ; une impression référentielle, la sélection d'une thématique sur un fond donné, un style général, etc. Et nombre d'applications pratiques seraient mieux servies par ce type d'approche, que par la recherche exclusive de structures actanciennes et événementielles qui, sur le plan des traitements automatiques, nécessite des traitements complexes dont la performance est souvent médiocre. À propos de l'analyse textuelle, Barthes (1985, p. 332) écrivait :

Cette manière de procéder est théoriquement importante : elle signifie que nous ne visons pas à reconstituer la structure du texte, mais à suivre sa structuration, et que nous considérons la structuration de la lecture plus importante que celle de la composition.

On peut maintenant proposer une définition simplifiée de la notion de lecture dans la perspective des applications informatiques : une lecture se compose d'un ensemble de descriptions initiales (le lexique sémantique) et d'un ensemble d'interprétants (interprétants syntaxiques, normatifs, argumentatifs, etc., formulés comme des règles d'inférence sur le contenu des sémèmes). En tant qu'interprétation résultante, elle produit des structures actualisées utilisables pour la description, et dont la validité dépasse celle du texte étudié pour s'étendre à une classe de textes de même genre. La formalisation des lectures ne permet sans doute pas d'automatiser de nouvelles lectures, mais peut être susceptible de réutiliser partiellement des lectures, ce qui suffirait à justifier pleinement l'approche.

5. STRUCTURE DES SIGNES ET STRUCTURE DU SIGNIFIÉ

La sémiotique du symbole et la sémiotique textuelle diffèrent non seulement par la nature mais aussi par la structure des unités qu'elles décrivent, et notamment du signifié. La sémiotique du symbole impliquant une sémantique structurelle, elle postule des signifiés

élémentaires au niveau de ses signes de base²¹. La sémiotique textuelle est quant à elle appelée à manipuler des signifiés plus complexes ; il est de plus nécessaire de proposer un certain niveau d'automatisation de la construction de structures sémantiques de niveau supérieur (isotopies, molécules sémiques, etc.) à partir des structures sémantiques élémentaires (sémèmes). Mais comme nous le verrons plus loin, il ne s'agit pas simplement de reporter une partie des fonctions de la syntaxe dans le lexique.

La structure du sémème doit posséder des propriétés descriptives, qui conservent l'information acquise lors de la description du sémème, en particulier en ce qui concerne les classèmes et leur hiérarchie locale ; cette structure a donc des propriétés linguistiques. Une autre fonction de la structure du sémème est de servir de base à la constitution de structures sémantiques de niveau supérieur au sémème (les différentes sortes d'isotopies et les molécules sémiques), mais cette seconde fonction peut exiger également des interprétants appropriés (cf. *infra*).

5.1. Le problème de la structure du signifié

La description du signifié lexical en termes de traits sémantiques est vite limitée par l'absence de structure de représentation capable d'accueillir ces traits, et cela indépendamment de la qualité de description des traits. La structuration du sémème est un thème qui apparaît très tôt dans la littérature, puisqu'il est déjà présent chez Greimas, où la préoccupation des traitements automatiques est déjà rencontrée :

— La structuration *propositionnelle* du sémème est présente aussi bien chez Greimas (1966) que chez Le Ny (1979). Cette hypothèse est selon nous fortement reliée à la dénomination du sème lui-même, lorsque les sèmes sont interprétables dans le langage source. Il est tentant de considérer que de leur interprétation découlent des relations semblables à celles qui relient entre eux les mots d'une phrase.

— La structuration hiérarchique est une alternative proposée par Greimas (1966). Une hiérarchie simple des sèmes aurait l'avantage d'être proche de leur mode de description, elle permettrait une

²¹ La sémiotique du signe-symbole reste essentiellement une sémiotique du signe élémentaire (cf. les *objets simples* du *Tractatus Logico-Philosophicus* de Wittgenstein), qui confie la constitution de signes complexes (les signes propositionnels) à la syntaxe.

combinatoire à différents niveaux qui serait quand même sous le contrôle de la spécificité lexicale.

— La structuration logique et ses variantes : une des difficultés majeures dans la réalisation d'applications pratiques résulte du conflit qui existe entre la notion de prédicat et la description des contenus sémantiques spécifiques.

— La recherche de formalismes lexicaux spécifiques est quant à elle beaucoup plus récente, le développement de formalismes lexicaux se faisant en quelque sorte au détriment de l'approche traditionnelle de la syntaxe. En effet, dans une conception où la syntaxe domine, c'est-à-dire où analyser la phrase c'est retrouver sa structure prédicative (typiquement, la f-structure en terminologie LFG), le contenu lexical est essentiellement utilisé pour la restriction de sélection ou certaines combinaisons (comme les descriptions nominales), qui suivent l'ordre interprétatif imposé par la syntaxe.

5.2. L'évolution des formalismes lexicaux

Le rôle traditionnellement dévolu à la structure lexicale est celui d'assurer l'interface entre syntaxe et sémantique lexicale, dans la perspective où la syntaxe détermine à la fois la structure de la phrase et l'assemblage des éléments sémantiques. Cette conception restrictive ramène au niveau du lexique le problème de la frontière entre syntaxe et sémantique. Les évolutions récentes du domaine vont plutôt dans le sens d'une lexicalisation croissante, ce qui pose le problème des interactions syntaxe-sémantique dans le lexique en termes différents : on pourra citer les grammaires d'arbres adjoints (*Tree-Adjoining Grammars* ou TAG) (Abeillé, 1991), le lexique génératif (Pustejovsky, 1991), et les structures sémantiques (SemS) de Melcuk (1988).

Le *modèle sens-texte* de Melcuk appartient à la famille des grammaires de dépendance ; la notion de structure sémantique (SemS) y sert à structurer le *sémantème* (il s'agit de la terminologie de Melcuk ; nous considérerons ici qu'il est équivalent au sémème lexical en sémantique différentielle). Les foncteurs sont à la base de la structuration des SemS. Melcuk donne des exemples de ces foncteurs ; ainsi, pour décrire la phrase *Mary discovered the loss of her flask*, les fonctions 'discover₂', 'disappear₁', 'belong to₂', 'be the moment of₂' et 'be before₂' sont nécessaires. Une vision partielle de la structure obtenue donne des indications sur le mode de structuration :

'discover₂' ('Mary' ; 'disappear₁' ('flask')) ; 'belong to₂' ('flask' ; 'Mary')

On pourra utilement comparer le problème de la dénomination des fonctions à celui de la dénomination des sèmes abordé plus haut. Melcuk pose également le problème du niveau de décomposition, pour favoriser un niveau de décomposition minimal, en fonction d'un principe de « pertinence » qu'il trouve chez Miller et Johnson-Laird (1976) : « speakers use as little decomposition as they can in a particular situation ». Ce point mérite d'être souligné dans la mesure où la théorie de Melcuk postule l'existence de primitives sémantiques, bien qu'il en perçoive les difficultés pratiques : « one should not necessarily try to decompose all the semantemes into semes. Such a decomposition would make the semantic network *unreadable and unmanageable* ²² ».

Le Lexique Génératif (Pustejovsky, 1991) représente une tentative de créer un formalisme lexical spécifique qui soit à la fois expressif et puisse servir de base à des opérations « interprétatives²³ » (Pustejovsky & Boguraev, 1993). La structuration du contenu lexical est réalisée par la *qualia structure* ; celle-ci est principalement basée sur les propriétés du référent, et veut être une extension aux noms des structures d'arguments habituellement utilisées pour les verbes. Un exemple de qualia structure permet de décrire les différents champs :

< door(x,y)
 CONST = aperture(y)
 FORMAL = physobj(x)
 TELIC = walk-through(P,w,y)
 AGENTIVE = artifact(x) >

CONST décrit les relations méréologiques de l'objet, essentiellement pour les objets physiques. Le rôle FORMAL est censé distinguer l'objet d'autres objets du domaine ; les exemples donnés en font la plupart du temps une sous-classe d'objets. Le qualia TELIC représente la fonction de l'objet ; c'est sans doute le qualia le plus important, et celui dont la description, en terme de « trait » apparaît la plus discutable ; si l'on peut admettre que, au moins dans la plupart des cas, le TELIC d'un sandwich soit d'être mangé (**eat(P,w,x)**), le fait que celui d'une porte soit d'être traversée (**walk-through(P,w,y)**) est plus difficile à concevoir en toute généralité. Le rôle TELIC se voit confier

²² C'est nous qui soulignons : en effet, ces notions de lisibilité et de manipulation sont des problèmes auxquels aucune représentation lexicale n'échappe.

²³ Le terme *interprétatif* n'est pas à prendre au sens habituel en sémantique interprétative. Le lexique génératif repose au-delà de la *qualia structure* sur un certain nombre d'autres constructions que nous ne reprenons pas ici.

un certain nombre de fonctions importantes, comme la sélection du niveau de description pertinent ; cependant, la capacité de prise en compte du contexte se heurte aux limitations générales des approches hiérarchiques. Enfin, le rôle AGENTIVE contient l'histoire ou mode de création de l'objet. L'essentiel de la structure de qualia est donc plutôt destiné à représenter les artefacts. Le Lexique Génératif n'apporte pas de précisions spécifiques quant à la définition de l'intitulé des traits ; il fait en revanche une hypothèse d'analyticité sur les descriptions, utilisant le calcul des prédicats. Dans le même ordre d'idées, il propose des structures typées avec héritage multiple, mais sans aborder le problème de la pertinence de la description.

D'une façon plus générale, on peut considérer que la structure lexicale conditionne l'agrégation des éléments sémantiques, quelle que soit la stratégie adoptée. Sa relation avec la syntaxe n'est donc qu'un cas particulier des liens entre stratégie interprétative et formalisme lexical ; ce type de relation mérite donc d'être approfondi et généralisé, et pourrait fournir les bases d'un utile programme de recherche.

5.3. Les TAG et le distributionnalisme

Les TAG représentent une forme très poussée de lexicalisation de la syntaxe, processus déjà initié par les autres grammaires d'unification depuis le début des années 80²⁴. L'intégration syntaxe-sémantique y est naturellement représentée, au moins sous forme statique, par les relations de dépendance qui sous-tendent les arbres substituables. Si cette première forme d'intégration contribue à réduire l'ambiguïté de rattachement, celle-ci ne disparaît pas et des éléments sémantiques sont souvent nécessaires pour déterminer, par exemple, les rattachements prépositionnels corrects (notons aussi que les valences sémantiques ne sont pas spécifiées dans le formalisme de base).

²⁴ Bien que les TAG remontent au milieu des années 70, les variantes avec unification sont contemporaines des autres grammaires d'unification.

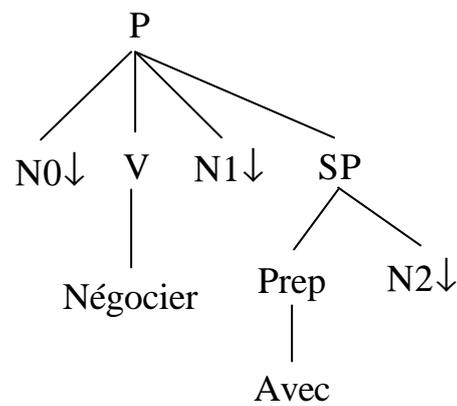


figure 1. Exemple d'arbre en TAG.

L'application des TAG au français est principalement l'œuvre d'Abeillé (1991) et l'un de ses apports essentiels, à notre avis insuffisamment souligné, a été de relier le formalisme des TAG à une certaine forme de distributionnalisme²⁵ (au sens de Gross (1975), voir par exemple Abeillé (1991, pp. 338-342)). Ce lien est plus qu'une simple communauté de notation, et fait des TAG des grammaires plus linguistiques que cognitives, en les reliant aux corpus et aux occurrences réelles qui, elles, ignorent la séparation entre phénomènes syntaxiques et sémantiques.

On pourrait donc proposer, mais ceci dépasse largement le cadre de cet article, une nouvelle forme d'extension sémantique des TAG, fondée sur l'analyse distributionnelle des classes sémantiques reliées par la composante de dépendance des TAG. Un exemple préliminaire sera proposé dans la suite de l'article.

5.4. Structure du sémème en sémantique textuelle

On peut considérer la structure lexicale sous un angle statique ou sous un angle dynamique. Dans le premier cas, la représentation lexicale doit être une structure *lisible* qui contienne les connaissances lexicologiques ; dans le second, elle doit servir de base à la formation de structures sémantiques de niveau supérieur, à partir de son propre contenu sémantique. Dans le cas de la sémiotique textuelle, la stratégie interprétative prime, et la structure lexicale devra permettre la construction de structures sémantiques à un niveau qui sera textuel et phrastique. Une spécificité supplémentaire de la sémiotique textuelle tient dans le rôle joué par les interprétants dans les transformations du

²⁵ Ces notions complexes mériteraient certainement une discussion plus rigoureuse que l'utilisation, largement heuristique, que nous en faisons ici.

contenu lexical, qui peut se substituer à certaines fonctions traditionnellement attribuées à la structure lexicale elle-même.

5.4.1. Description et représentation du sémème

Si l'on suit les concepts de la sémantique différentielle, qui définissent les traits sémantiques en fonction des différentes classes de signification, on obtient une description lexicale brute, sur le modèle ci-dessous, qui ne préjuge d'aucun format lexical :

```
< Entrée lexicale : “grève”>
  dimension : //actions collectives//
  domaine : //socio-économie//
  champ : //relations sociales//
  taxème : //actions revendicatives//
  spécificité : /conflit/
  afférences : { /conflit/ /échec politique/, etc. }
</ Entrée lexicale >
```

Le premier problème de ce type de présentation consiste dans le choix de l'intitulé des traits, qui va chercher à rendre la description simple et expressive, en compensant l'absence de mécanismes combinatoires. Il s'agit en fait de pouvoir manipuler explicitement des traits synthétiques à un niveau de généralité optimal, en remplacement de structures typées et/ou d'héritage multiple ; on choisit de confier cette tâche à la méthodologie de description différentielle elle-même, seule capable de traiter la multiplicité des descriptions par rapport aux données du texte. Ainsi, si dans le taxème des //actions revendicatives//, “grève” s'oppose à “négociation”, dans celui des //relations économiques//, “négociation” peut s'opposer à “OPA” ; ce type de phénomène est difficilement systématisable hors contexte. Pour des raisons similaires, les afférences sont mieux représentées sous forme explicite, par les traits constituant les extrémités de la relation d'afférence.

Cette description brute n'aborde pas le problème de l'établissement de la structure fonctionnelle (c'est-à-dire de l'actance) ; celui-ci se pose en des termes nouveaux quand les structures à construire sont d'une amplitude qui dépasse la phrase. La conception traditionnelle des relations syntaxe-sémantique dans le sémème est celle des *restrictions de sélection* ; elle consiste à superposer à une structure argumentale (du type négociateur(X, Y, Z), où X et Y seraient les acteurs de la négociation et Z son objet) des conditions de compatibilité sur les

classes sémantiques des sémèmes à relier. La structure argumentale est le plus souvent définie par des catégories syntaxiques (par exemple N0-négociateur-N1-avec-N2) ou des fonctions casuelles, qui représentent un degré d'intégration syntaxe-sémantique plus élevé et s'accommodent mieux des phénomènes textuels que ne le fait la syntaxe traditionnelle. Nous avons cité plus haut les solutions proposées par le lexique génératif et le modèle sens-texte, qui représentent une première évolution par rapport aux relations syntaxe-sémantique traditionnelles. Il semble cependant nécessaire de passer de la restriction de sélection à la *prescription de sélection*, mais ce passage s'accompagne souvent d'une spécification des classes sémantiques à prendre en compte. Nous pouvons maintenant revenir sur la notion de distributionnalisme sémantique proposée dans la section 5.3. Considérons par exemple le cas du sémème "négociation" : il est susceptible d'établir des relations avec les sémies correspondant à certaines classes sémantiques particulières, comme //paramètre// (qui contient, par exemple, "prix", "temps de travail", etc.), //actions collectives// ("reprise du travail", "levée de l'embargo") ou encore //objet contractuel// ("clause", "disposition", etc.). L'établissement de ce type de relation semble principalement du ressort des interprétants, mais peut-être une analyse distributionnelle sur des corpus étendus permettrait-elle de les inclure dans un formalisme comme les TAG ?

5.4.2. Les interprétants comme alternative à la structure lexicale

À partir de la définition originale des interprétants :

L'interprétant est une unité linguistique ou sémiotique qui permet de sélectionner la relation sémique pertinente entre les sémies reliées par le parcours interprétatif. Parmi les interprétants, il faut souligner l'importance des axiomes normatifs implicites (ou *topoi*) (Rastier *et al.*, 1994, p. 70-72).

il est possible de proposer une alternative à la structuration du sémème, en confiant l'établissement de relations entre les sémies aux interprétants eux-mêmes plutôt qu'à des relations pré-existant dans le sémème. L'utilisation des interprétants peut également éviter de contraindre la structure des représentations sémantiques textuelles par les choix faits pour la structure du sémème.

Les interprétants peuvent trouver leur origine, comme nous l'avons vu plus haut, dans des structures syntagmatiques (de la simple énumération à certaines structures syntaxiques remarquables), des structures argumentatives ou des axiomes normatifs. Les deux premiers effectuent des opérations formelles alors que le troisième type d'interprétant est fortement lié aux classes sémantiques mises en

œuvre. Les opérations formelles servent à déterminer les conditions d'activation ou de propagation du contenu (relations d'afférence, détermination des classes de généralité pertinentes) alors que les opérations relationnelles concourent à la formation de structures sémantiques au niveau textuel (molécules sémiques, isotopies, etc.).

Un interprétant peut être implémenté au moyen d'une règle figée décrivant des conditions de modification ou d'aggrégation du contenu lexical ; la même forme d'implémentation peut cependant servir les deux grandes catégories d'interprétants. D'après ce qui précède, et sans prétendre aucunement établir une classification des interprétants, nous pouvons déjà noter la diversité des interprétants décrits jusqu'ici :

— Interprétants *syntaxiques faibles*, comme ceux engendrés par la ponctuation, les énumérations, la parataxe.

— Interprétants *syntaxiques forts*, correspondant à des structures syntaxiques plus complexes ou aux structures syntactico-argumentatives discutées dans la section 4.1.

— Interprétants *topiques*, qui se fondent sur le contenu normatif ou *doxa*.

Enfin, si l'on confie aux interprétants un rôle majeur, permettant de passer de la restriction de sélection à la prescription de sélection, il faut alors admettre une spécification des interprétants, qui vont se voir confier l'établissement de relations très spécifiques, et dont le nombre va augmenter.

6. SEMIOTIQUE TEXTUELLE ET TRAITEMENTS AUTOMATIQUES DES TEXTES

L'interprétation n'est pas, en l'état actuel, pleinement accessible aux technologies de traitement automatique, qui reposent pour l'essentiel sur une combinaison d'éléments sémantiques dirigée par la syntaxe. Les limitations de cette approche sont aujourd'hui largement reconnues, et on peut les résumer ainsi :

— La couverture de la syntaxe dans les formalismes actuels demeure limitée et peu adaptée aux textes réels.

— La prescription des opérations sémantiques par la syntaxe n'est pas en mesure de réaliser des opérations sur l'ensemble d'un texte.

— L'analyse détaillée de l'ensemble du texte est à la fois trop puissante et trop peu performante lorsqu'il s'agit de retrouver les informations pertinentes.

Dans l'état actuel de la technologie, il y a en fait incompatibilité entre les structures d'analyse d'une phrase et les structures d'analyse d'un texte. Ces dernières n'ont été développées que pour l'intégration des récits à partir d'éléments narratifs élémentaires ; comme nous l'avons vu plus haut, l'usage réel fait des textes se rapprocher plus de la construction de structures sémantiques au niveau textuel qu'au niveau phrastique. Indépendamment de la sémiotique textuelle, la nécessité d'obtenir des résultats pratiques dans le domaine du traitement automatique des textes a conduit au développement de méthodes beaucoup plus utilisables dont nous allons donner un bref aperçu.

L'Extraction d'Informations se trouve être la forme la plus aboutie des systèmes d'analyse de textes fonctionnant sur des applications réelles ; elle s'est principalement développée depuis le début des années 1990. La recherche d'une analyse syntaxique exhaustive, qui était l'apanage d'une linguistique de la phrase, s'est progressivement effacée au profit d'analyses locales, d'analyses basées sur des relations sémantiques (Jacobs, 1990), mais surtout la problématique va s'inverser ; au lieu de tester l'applicabilité des théories les plus complexes, les systèmes vont chercher la meilleure efficacité au moindre coût théorique. Et cela, essentiellement par un rôle plus important du lexique et une simplification des techniques d'analyse. Si l'on trouve dans les premières évaluations « MUC » une supériorité des systèmes utilisant des analyseurs syntaxiques (Sundheim, 1991), cette supériorité va être progressivement remise en question, ce qui culminera en 1992-93 avec les deux systèmes concurrents du SRI que sont TACITUS et FASTUS. TACITUS (Hobbs *et al.*, 1992) représente la forme aboutie de l'ingénierie des « grands systèmes », utilisant un analyseur syntaxique complexe (comprenant un *chart-parser* sophistiqué) et une architecture qui intègre des modules dédiés aux différents traitements syntaxiques, sémantiques et discursifs.

FASTUS (Appelt *et al.*, 1993) est quant à lui un système basé sur une cascade d'automates à états finis, qui sont chacun spécialisés dans une tâche empirique prenant comme point de départ l'étude des textes à analyser. Les différents automates sont dédiés à l'identification des noms, des groupes nominaux et verbaux élémentaires, des structures d'actance, des principaux événements, ainsi qu'à la résolution des coréférences. La conjonction d'une approche empirique et d'une description fine a produit très rapidement des résultats plus que satisfaisants, supérieurs à ceux obtenus par les approches traditionnelles, confirmant la primauté d'une description fine sur une formalisation poussée (cf. également (Wilks, 1991)). Cependant, une description elle-même empirique peut conduire à des limitations, et la

sémiotique textuelle est susceptible de servir de cadre méthodologique à ces descriptions. Un des points fondamentaux de la théorie textuelle est que les structures peuvent être décrites par des formes similaires à différents niveaux (micro-, méso- et macrosémantique)²⁶ ; ceci peut servir de base à une implémentation à base de cascade d'automates à états finis, en tirant profit de cette propriété. Cette approche nécessite encore des recherches complémentaires dans la description d'un certain nombre d'interprétants, en particulier ceux reliant la *topique* à la construction de structures sémantiques.

CONCLUSION GENERALE

Nous avons choisi pour cet exposé un cadre théorique précis, et l'on pourra bien sûr nous objecter que les idées qui ont été avancées sont difficilement transposables à d'autres cadres, ou peuvent même être impossibles à mettre en œuvre en l'état actuel de nos connaissances. Si les directions proposées peuvent parfois sembler préliminaires, il importe cependant de souligner la convergence de recherches issues d'horizons variés, qui tendent à réhabiliter des problèmes encore considérés comme secondaires il y a quelques années.

Alors que la sémiotique, particulièrement dans le cadre interdisciplinaire des sciences cognitives, tend à être dominée par les approches symboliques, la sémiotique du texte propose une approche fondée sur le contenu, et néanmoins capable de servir de base à des réalisations informatiques. La sémiotique du texte étudiant les systèmes de signifiés, elle est autonome par rapport aux formalismes et peut ainsi fonder une extension transversale entre les modalités sémiotiques :

La sémiotique pourrait faire preuve d'une ambition plus haute : loin d'éliminer le contexte socio-historique, penser son rapport au texte comme un rapport d'intersémiotité (Rastier, 1987, p. 219)

Ce programme de recherche peut constituer une alternative partielle au cognitivisme symbolique qui a longtemps dominé les Sciences Cognitives, avec pour horizon le développement de technologies culturelles en lieu et place des technologies de l'intelligence.

²⁶ Voir par exemple la description des structures thématiques et dialectiques dans (Rastier, 1989).

Bibliographie

- Abeillé A. (1991). *Une grammaire lexicalisée d'Arbres Adjoints pour le français*. Thèse de Doctorat, Université Paris 7.
- Appelt D. E., Hobbs J. R., Bear J., Israel D. and Tyson M. (1993). FASTUS: A Finite-state Processor for Information Extraction from Real-World Text, *Proceedings of the IJCAI'93 Conference*.
- Ballón Aguirre E., Cerrón-Palomino R. & Chambi Papaza E (1992). *Vocabulario razonado de la actividad agraria andina*. Cusco : Monumenta Linguistica Andina.
- Barthes, R. (1967). *Système de la mode*. Paris, Éditions du Seuil.
- Barthes, R. (1985). *L'aventure sémiologique*. Paris, Éditions du Seuil.
- Boguraev, B. and Briscoe, T. (1989). *Computational Lexicography for Natural Language Processing*, New York, John Wiley & sons.
- Cavazza, M. (1994). La description du contenu lexical, In : Rastier, F., Cavazza, M. et Abeillé, A., *Sémantique pour l'analyse*, Paris, Masson.
- Church, K. and Hanks, P. (1989). Introduction to the special issue on computational linguistics using large corpora, *Computational linguistics*, 19, 1, pp. 1-24.
- DARPA, (1993). Defense Advanced Research Project Agency, *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, Morgan Kaufmann.
- de Libéra, A. (1995). *La querelle des universaux: de Platon à la fin du Moyen Age*, Paris, Éditions du Seuil.
- Eco, U. (1988). *Sémiotique et philosophie du langage*. Paris, Presses Universitaires de France.
- Eco, U. (1983). L'antiporfirio, in: *Il pensiero debole*, Vattimo G., Rovatti P.A. (Eds.), Milan, Feltrinelli.
- Granger, G.-G. (1979). *Langages et épistémologie*. Paris, Klincksieck.
- Greimas, A. J. (1966). *Sémantique Structurale*, Paris, Presses Universitaires de France.
- Gross, M. (1975). *Méthodes en syntaxe*, Paris, Hermann.
- Hjelmslev, L. (1968). *Prolégomènes à une théorie du langage*, Paris, Éditions de Minuit (première édition: 1943).
- Hobbs, J. R., Appelt, D. E., Bear, J. and Tyson, M. (1992). Robust Processing of Real-World Natural-Language Texts, *Proceedings of the third Applied Natural Language Processing Conference*.
- Jacobs, P. (1990). To Parse or not to Parse: Relation-Driven Text Skimming, *Proceedings of COLING'90 Conference*, Helsinki.
- Jayez, J. (1988). *L'inférence en langue naturelle*. Paris, Hermès.
- Largeault, J. (1993). *Intuition et intuitionisme*, Paris, Vrin.
- Le Ny, J.-F. (1979). *La sémantique psychologique*. Paris, Presses Universitaires de France.

- Martin, E. (1990). L'exploration textuelle assistée par ordinateur : l'interrogation thématique. *Colòquio de lexicologia e lexicografia*, Universidade Nova de Lisboa.
- Melcuk, A. (1988). *Dependency Syntax*, Albany, State University of New York Press.
- Mettinger, A. (1994). *Aspects of Semantic Opposition in English*, Oxford, Clarendon Press.
- Metz, C. (1977). *Essais sémiotiques*, Paris, Klincksieck.
- Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and Perception*, Cambridge, Cambridge University Press.
- Morris, J., and Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of a Text. *Computational Linguistics*, 17, 1.
- Pustejovsky, J. (1991). The Generative Lexicon, *Computational Linguistics*, 17, 4.
- Pustejovsky, J. and Boguraev, B. (1993). Lexical knowledge representation and natural language processing, *Artificial Intelligence*, 63, 193-223.
- Rastier, F. (1987). *Sémantique Interprétative*, Paris, Presses Universitaires de France.
- Rastier, F. (1989). *Sens et Textualité*, Paris, Hachette.
- Rastier, F. (1990). La triade sémiotique, le trivium et la sémantique linguistique. *Nouveaux actes sémiotiques*, 9.
- Rastier, F. (1991). *Sémantique et Recherches Cognitives*, Paris, Presses Universitaires de France.
- Rastier, F., Cavazza, M. et Abeillé, A. (1994). *Sémantique pour l'analyse*, Paris, Masson.
- Sowa, J. (1984). *Conceptual Structures: Information processing in mind and machines*, New York, Addison-Wesley.
- Sundheim, B. (1991). Third Message Understanding Evaluation and Conference (MUC-3): Methodology and Test Results, in: J.G. Neal and S.M. Walter (Eds.), *Natural Language Processing Systems Evaluation Workshop*, Rome Laboratory, Technical Report RL-TR-91-362.
- Wilks, Y. (1991). Form and content in semantics, in: *Computational linguistics and formal semantics*, Rosner M., Johnson, R., (Eds.), Cambridge, Cambridge University Press.