Chapitre 2

Le Langage : "Si loin, si proche"

Si loin de nous, en ce qu'il est difficile à appréhender.
Si proche, en ce qu'il est une des principales activités humaines.

Ce n'est pas uniquement parce que le langage est difficile à appréhender que nous avons du mal à l'aborder, c'est aussi parce qu'il est proche de nous que nous avons du mal à nous en détacher pour le prendre aux mots ou à la lettre.

2.1. De l'étude de faits linguistiques

2.1.1. Les faits linguistiques sont empiriques

2.1.1.1. Si la linguistique est une science, elle est une science empirique[1]

Le langage possède la propriété remarquable d'être quantifié : il met en jeu divers types d'unités discrètes, les phonèmes qui composent les mots, les mots qui composent les phrases... Il s'agit d'une des rares activités humaines qui puisse être ainsi discrétisée. A partir de cette propriété de quantification du langage, il a paru possible de transformer la linguistique en une véritable science.

2.1.1.2. Le langage est à la fois, support, instrument et fondation de nos activités de représentation

Le principal obstacle à l'étude scientifique du langage réside dans la difficulté d'isoler un domaine objectivable : le langage se confond avec son propre fonctionnement et y cache ses secrets. Le linguiste et son objet sont difficilement séparables, il lui est très difficile de prendre de la distance par rapport à son matériel d'étude. Le langage est *"un système ouvert et au delà du rapport désignatif à des* ***réalités****, il est l'objet même de ses propres représentations...il est à la fois, support, instrument et fondation de nos activités de représentation et par suite, de nos*

[1] *"Par empirique, on entend l'ensemble de ce qui est représentable dans l'espace et dans le temps"* (Milner 1989).

actions symboliques sur le monde; on ne peut ainsi analyser le linguistique sans prendre en compte cette capacité structurelle de ***manipulation des signes****, en conséquence d'abstraction générique vers la catégorisation permanente des choses et du rapport à ces choses"* (Vignaux 1986).

2.1.1.3. Du travail sur les faits de langue[2]

Etant donné une langue, le problème d'un linguiste consiste d'une part à déterminer ses unités discrètes minimales et d'autre part (et simultanément) à étudier les combinaisons de ces éléments. Les difficultés apparaissent très vite : le linguiste doit procéder à des regroupements et invoquer des argumentations fondées sur la répétition des unités discrètes et leur combinaison à l'échelle de la langue. Le linguiste forme des hypothèses de règles de syntaxe. Il construit alors les assemblages de mots prévus par les règles, il vérifie la validité de ses hypothèses en portant des jugements faisant appel à l'intuition sur l'acceptabilité de ses constructions. Certains faits peuvent ainsi être dégagés au moyen d'intuition d'acceptation et de refus de certaines formes. Ce sont ces jugements d'acceptabilité qui constituent la base expérimentale de la syntaxe et tout le progrès de la syntaxe provient de ce que ces intuitions ont une bonne reproductibilité.

2.1.2. De la finesse des questions linguistiques

Les problèmes traités par la science linguistique touchent en général à des phénomènes d'une extrême finesse[3] : quand on travaille sur la langue, on est contraint d'étudier des microphénomènes. Le travail du linguiste est une recherche d'équilibre entre différentes contraintes; le résultat de ce travail est toujours remis en question par la prise en compte de nouvelles informations. Il est donc condamné à rester provisoire.

2.1.3. De l'interaction de phénomènes divers dans les questions linguistiques

Les questions auxquelles sont confrontés les linguistes mettent en jeu de nombreux micro-phénomènes[4] de natures diverses qui interagissent de manière complexe. Quand on dépasse l'étude des mots isolés (ces derniers n'appartenant pas à l'objet empirique de la linguistique), on est confronté, face à un énoncé linguistique, à des phénomènes lexicaux, syntaxiques, sémantiques... Un énoncé s'inscrit dans une situation sociale et temporelle qui n'est généralement pas sans incidence sur cet énoncé.

2.1.4. Du non-achèvement des résultats dans la science linguistique

[2] *"La linguistique doit se donner comme des faits primitifs:*
- *le fait de la langue, le factum linguae, le fait que ce que parle un être parlant mérite le nom de langue, à bien distinguer du factum loquendi ou le simple fait qu'il ait des gens qui parlent;*
- *le fait des langues, ou factum linguarum, c'est-à-dire qu'elles soient diverses, tout en contituant une classe homogène; cette multiplicité-homogénéité est censément attestée par le fait de la traduction;*
- *le fait que les langues soient descriptibles en termes de propriétés. Ce fait peut recevoir un nom plus précis: c'est le fait de la grammaire, factum grammaticae. On peut le résumer ainsi: l'activité grammaticale existe dans la plupart des communautés linguistiques. Or cette activité a des caractères propres: elle suppose qu'on puisse attribuer des propriétés à une formulation langagière sans avoir aucun égard ni à celui qui la profère ni à son éventuel destinataire ni aux circontances de la profération. D'où il suit que certaines de ces propriétés seront hors circonstances, c'est-à-dire constantes.*

De là suit la disposition terminologique la plus courante. Tout de même que le terme langage, dans son usage courant, sténographie le factum loquendi, de même, dirait-on, le complexe des trois faits - factum linguae, factum linguarum, factum grammaticae - se trouve sténographié par le terme langue" (Milner 1989).

[3] *"Les questions que soulève la science du langage, dans toutes ses versions, sont des questions fines; dès qu'elle dépasse la banalité, une proposition de linguistique concerne peu de données à la fois et elle y fait apparaître généralement ce que l'opinion courante tiendrait pour du détail"* (Milner 1989).

[4] *"La donnée minimale de langue, qui est une phrase, est toujours trop complexe par rapport à la proposition minimale de linguistique.(...).Les langues naturelles sont organisées de telle sorte que jamais une propriété ne se trouve à l'état isolé"* (Milner 1989).

Quand on s'attaque à l'étude des activités langagières, le nombre de problèmes à résoudre est considérable et les résultats établis dans la science linguistique ne sont jamais définitifs. L'évolution perpétuelle des langues ne laisse pas intacte les connaissances que l'on peut avoir sur les langues.

On ne peut jamais en savoir assez sur le système osseux et sur le système nerveux de la langue

Vouloir construire des systèmes de représentation de la langue, c'est toujours (re)construire ces systèmes car travailler sur un fait linguistique, c'est toujours (re)travailler sur ce fait de langue.

2.1.5. La science linguistique résiste au formalisme

2.1.5.1. De la formalisation[5] du langage naturel

Un langage formel définit un ordre syntaxique en faisant opérer un ensemble de règles sur une suite de symboles. Le langage formel ne contient pas par définition les références qu'il est possible d'attacher à chacun des symboles qu'il manipule. Il faut pour cela opérer une mise en rapport entre le monde représenté et le système formel construit. Inversement, un code symbolique définit un rapport direct entre les symboles et les objets représentés, mais ne définit pas de lien syntaxique entre les symboles utilisés. Si l'on a pu penser qu'une formalisation des activités linguistiques passait par la construction d'un langage formel symbolisé, où le référentiel est soumis au syntaxique, il est apparu rapidement que la composante pragmatique du langage devait être intégrée à cet ensemble. En fait les langues mettent en jeu des ordres différents (Rastier 1994) :

- L'ordre syntagmatique qui traite de la linéarisation du langage dans l'espace spacio-temporel.
- L'ordre paradigmatique qui permet de donner une valeur à une unité relativement à d'autres qui sont commutables avec elle.
- L'ordre herméneutique qui traite des conditions de production et d'interprétation des textes.
- L'ordre référentiel qui traite des rapports entre les signes, les concepts et les choses, et des rapports entre les phrases, les propositions et les "états des choses".

2.1.5.2. De l'excès dans la langue à la formalisation

Si on considère la langue comme un système et si on veut le modéliser, il faut d'une part essayer de déterminer les connaissances à l'oeuvre dans les activités langagières et d'autre part il faut étudier la mise en oeuvre et la structuration de ces connaissances. Malheureusement, le problème de définir les modes de représentation et les interactions entre ces modèles de représentation n'est pas trivial. Celui de déterminer les différentes types de connaissances

[5] *"Un moment décisif de la science du langage devra consister en ceci qu'on raisonne sur des entités purement littérales, en ne tenant compte que des lois qui régissent les combinaisons de ces entités. A cet égard ce qui se passe du côté des formalismes logico-mathématiques importe au premier chef; ils constituent un réservoir naturel de contraintes aveugles . Or dans ce domaine, l'invention existe et un nouveau type de contrainte (ou une propriété encore mal dégagée d'un formalisme déjà connu) peut toujours se révéler. Comme on l'a souligné, c'est une grande question de savoir si les formalismes logico-mathématiques forment en dernière instance la seule et unique base de littéralisations. Le structuralisme ne le croyait pas. Il n'est pas sûr qu'il ait eu raison. Pour autant, la science du langage ne cesse d'être empirique; il faut donc que le maniement aveugle produise, en fin de parcours, des propositions empiriques réfutables. En retour , choisir tel type de contrainte plutôt que tel autre dépend de la fécondité empirique qu'on peut lui reconnaître. Le croisement entre la réfutabilité (fondée par un différentiel de langue) et la littéralisation (qu'elle soit explicitement empruntée aux formalismes mathématiques ou qu'elle ne le soit pas), voilà donc ce qui doit caractériser la science du langage dans son ensemble"* (Milner 1989).

nécessaires pour un traitement automatique du langage naturel (TALN) ne l'est pas non plus. Les connaissances nécessaires pour comprendre ou engendrer des textes sont considérables, surtout s'il s'agit de vouloir atteindre un niveau de résultat équivalent à celui produit par un être humain. On touche là un problème crucial pour la construction de modèles de représentation du langage naturel. Comment formaliser une telle complexité d'interactions de processus qui sont mis à l'oeuvre dans la production langagière? Faut-il construire des systèmes qui dépassent nos limitations dans notre intelligence de production et de compréhension du langage naturel ou faut-il suivre le comportement humain dans les choix fondamentaux à effectuer pour modéliser les activités langagières?

2.1.5.3. Le TALN ou le dur désir de formaliser les faits de langue

Les formalismes dont on doit pouvoir disposer pour le traitement du langage naturel (et notamment pour mettre en place des dispositifs de représentation de ces modèles) doivent prendre en compte les complexités inhérentes aux activités langagières. Celles-ci mettent en jeu dans un énoncé des processus complexes interagissants : la construction syntaxique d'un énoncé s'inscrit dans l'espace et le temps et cette structuration progressive et matérielle accumule tout au long de son énonciation des indices sémantiques qui se répartissent de manière non linéaire sur les éléments de l'énoncé. Ces derniers se répondent tout au long de la production de l'énoncé et grâce aux interactions de ces traces sémantiques disposées par les éléments constitutifs de la chaîne produite dans l'espace et le temps de l'énonciation, il est généralement possible de (re)construire un des sens associé à l'énoncé perçu. Certes, modéliser de tels mécanismes n'est pas simple. D'une part, les éléments en jeu dans ce travail interprétatif sont difficiles à appréhender, car il n'est pas simple de pouvoir isoler ces éléments. D'autre part, dans le cadre d'un dispositif informatique de représentation matérielle de ces modèles, les outils manquent encore pour représenter ces mêmes phénomènes : on pense en particulier aux mécanismes réflexifs, perpétuellement à l'oeuvre dans les activités langagières, mais dont l'utilisation qui en est faite dans les langages de programmation de haut niveau n'a pas encore atteint des résultats satisfaisants. Les activités langagières mettent en jeu des mécanismes de rétro-action, d'ajustement, de masquage...; on comprend qu'il soit problématique de représenter de tels processus : ils sont inhérents à notre activité de parole, nous les manipulons en permanence, mais on a beaucoup de mal à les isoler et donc à les représenter.

2.1.5.4. Formalisation et réalité linguistique

Ainsi, même si les langages de programmation, utilisés dans le cadre des dispositifs informatiques, mettent en place des mécanismes puissants (c'est le cas des processus réflexifs pour certains d'entre eux), il n'est pas sûr que ceux-ci soient adéquats pour représenter ces mêmes processus réflexifs à l'oeuvre dans le langage[6] : leur existence dans les langages informatiques n'est pas une modélisation définitive de ces phénomènes, nous ne sommes pas encore en mesure de la réaliser. Il faut donc garder en mémoire que *"toute conception rationaliste voire formaliste du linguistique se trouve en péril non seulement de **figer** la grammaire quand on sait que celle-ci n'est que le produit d'une longue activité phylogénétique mais encore d'ignorer que dès le syntaxique, se trouvent intriquées aux **formes-signes**, les traces des manipulations sémantiques c'est à dire encore pragmatiques de ces formes"* (Vignaux 1986).

2.1.6. La science linguistique et le dispositif

[6] *"Il convient toujours de poser la question lorsqu'on rencontre une théorie linguistique formalisée: les procédures formelles sont-elles dues à une exigence de rigueur, en sorte que l'on aura choisi telle procédure plutôt qu'une autre parce qu'elle est plus conforme à l'idéal métathéorique que l'on s'est proposé? Ou bien les procédures formelles constituent-elles des hypothèses sur les propriétés du langage?"* (Milner 1989)

"La science du langage n'a apparemment que deux possibilités : ou bien demeurer vague et renoncer à attribuer à son objet des propriétés spécifiques, ou bien rechercher la précision empirique et dépasser les données observables en proposant un dispositif" (Milner 1989).

Un dispositif[7] propose une représentation matérielle des phénomènes traités. Dans notre cadre de travail, les dispositifs, dont il est question, sont constitués par des outils informatiques développés dans le cadre de la programmation à Objets; il s'agira d'établir qu'on a bien une adéquation forte entre les opérations de connaissance que ces outils permettent et la langue. De plus, *"la relation d'une proposition de linguistique à une procédure technique ne saurait se faire directement et simplement."* (Milner 1989). L'activité langagière manipule des processus de contrôle, de rétro-action, de masquage,..., sur les formations qu'elle produit. Les dispositifs (et notamment les dispositifs informatiques que nous manipulons) doivent impérativement prendre en compte ces phénomènes à travers les systèmes de représentation qu'ils développent. C'est en ce sens qu'il convient de les interroger pour que les modèles qu'ils représentent puissent être considérés comme des représentations empiriquement correctes. Ces phénomènes sont au coeur de toutes les problématisations sur les systèmes de représentation des connaissances. En effet, les questions qui sont en jeu avec le traitement automatique du langage naturel dépassent le cadre de l'étude du langage naturel. Il ne faut pas oublier d'autre part que *"le recours à une technique ou à une science comme dispositif est éminemment contingent et déterminé par des valeurs de convenance : on choisit la référence qui, à un moment donné, fournit le dispositif le plus clair et le plus distinct pour la représentation, celui qui comporte le moins d'incontrôlable. Autrement dit ce recours est aussi précaire : une science peut être amenée à changer de dispositif à tout moment, en fonction de son propre progrès"* (Milner 1989).

2.1.7. Des industries de la langue ?

"On peut même noter, entre linguistes et techniciens de la langue, une indifférence et une ignorance réciproques plus accentuée qu'il n'est de mise dans d'autres domaines" (Milner 1989).

L'étude théorique des automates et des langages formels a pendant un temps affecté simultanément les recherches informatiques et linguistiques. Le travail fait par Chomsky sur les langages formels atteste de la relation privilégiée entre l'informatique et la linguistique. On a d'ailleurs très tôt (trop tôt) pensé que l'informatique pourrait constituer une technique pour la linguistique[8]. Encore aujourd'hui il est irréaliste de penser que l'informatique puisse servir de technique (au sens milnérien défini supra) à la linguistique, même si d'aucuns souhaiteraient voir se développer rapidement un marché des industries[9] de la langue et pouvoir ainsi disposer d'outils informatiques adaptés à tous les besoins possibles et imaginables pour résoudre toutes sortes de problèmes linguistiques. En effet, l'informatique est une discipline neuve en perpétuel développement. Or une industrie suppose préalablement une science achevée (provisoirement). Il ne saurait en être ainsi pour les langues naturelles : les descriptions sont toujours partielles, partiales et destinées à le rester : *"En tant que la science les saisit, les langues et le langage ne sont pas des matières réalisées; ce sont plutôt les lois qui régissent ces "matières"* (Milner 1989). Le TALN doit placer au centre de ses préoccupations la

[7] *"Dans le dispositif, il s'agit de proposer une représentation - et en fait une représentation matérielle - de ce qui a été découvert; si l'on adopte un dispositif informatique pour représenter le langage, on ne vise pas seulement des raisonnements, on ne se borne pas à soutenir que la méthode linguistique est analogue à la méthode de l'informatique, mais l'on affirme que, dans le cadre du dispositif et pour toutes les opérations de connaissance qu'il autorise, le langage est effectivement tel qu'on le représente en termes informatiques"* (Milner 1989).

[8] Le projet de la traduction automatique a vu le jour en 1960. L'idée de base consistait à effectuer une analyse syntaxique des phrases et à s'appuyer sur la structure générale obtenue pour traduire des textes (en l'occurrence des symboles) particuliers. Les algorithmes d'analyse étaient fondés sur les combinaisons deux à deux des mots et/ou des groupes de mots. Le nombre des combinaisons à vérifier dans un texte pouvant être élevé d'où le recours à la machine, de nombreuses difficultés sont très vite apparues et la traduction automatique a fait une faillite retentissante en 1970. Il ne suffisait pas comme cela était pensé à l'époque de penser que l'information fournie par les grammaires et les dictionnaires du commerce suffiraient à construire de façon non ambiguë les interprétations des phrases.

[9] *"Quand on parle d'application industrielle, on songe trop aisément à l'industrie lourde. Or les questions que soulèvent la science du langage, dans toutes ses versions, sont des questions fines; dès qu'elle dépasse la banalité, une proposition de linguistique concerne peu de données à la fois et elle y fait apparaître généralement ce que l'opinion courante tiendrait pour des détails. En bref, il paraît peu vraisemblable qu'à partir de la science du langage, on puisse développer l'équivalent d'une sidérurgie ou d'une aéronautique"* (Milner 1989).

nécessité d'outils, de dispositifs permettant de mettre au point les connaissances à utiliser, de tenir compte de leur caractère avant tout provisoire, daté et évolutif.

Et, même si le TALN a pendant un temps bénéficié des résultats acquis par le développement de l'informatique, il reste à savoir si l'absence de solution de la continuité entre langages formels, langage naturel et langages de programmation n'est pas un postulat précipité, les points de contacts et de divergences n'ayant pas été explorés. De plus les formalisations disponibles ne permettent pas encore d'exprimer correctement les relations complexes existant, entre autres, dans la représentation des connaissances lexicales et syntaxiques d'une langue. En outre, elles ne disposent pas encore d'outils adéquats pour manipuler les représentations initiales en spécifiant des questionnements précis ou pour examiner les résultats produits grâce à des mécanismes associés à ces formalismes. Il s'agit encore pour le TALN d'utiliser l'informatique comme d'un outil de vérification (d'utilisation) des résultats connus d'une théorie linguistique.

2.1.8. La langue comme modèle de système de représentation des connaissances

"Les langues, utilisées constamment par les êtres humains et continuellement re-modelées par eux en fonction de leurs besoins, constituent des systèmes exemplaires de gestion des connaissances" (Pierre-Yves Raccah in (Dubois 1986)).

Le langage usuel est un moyen de représenter les choses, doté d'une capacité de description et d'explication considérable. L'être humain a créé des langues en fonction de ses besoins de communication et de ses possibilités de traitement de l'information. Nous sommes capables de parler et de comprendre un interlocuteur en temps réel bien que le temps de traitement de nos neurones se mesure en millisecondes. De plus l'être humain parlant structure ses connaissances sur la langue qu'il manipule de telle sorte qu'elles soient accessibles au moment où celles-ci s'avèrent nécessaires. Nous avons accès immédiatement à ce qui nous intéresse. De plus, quand nous rencontrons une ambiguïté, nous choisissons en général la solution la moins coûteuse qui consiste à ne pas retenir plusieurs interprétations et donc à ne pas maintenir plusieurs interprétations et ainsi à ne pas retarder la construction d'une structure (et cela à cause en partie de nos limitations de mémorisation). On garde malgré tout la possibilité de revenir en arrière quand les choix initiaux conduisent à une impasse en effectuant le retour le moins coûteux qui conduit à une solution encore une fois la moins coûteuse. Puisque le langage est adapté à nos besoins et à nos possibilités (ce qui prend en compte les défauts et les qualités de compréhension et de production) on peut peut-être souhaiter que les dispositifs mis en place pour le traitement automatique soient à l'image de ce que nous faisons, c'est-à-dire qu'ils tiennent compte du fait que dans notre activité langagière nous commettons des erreurs (que nous savons gérer dans le cadre de cette activité), ou encore que nous ne comprenons pas toujours tout ce que nous entendons, en précisant que toutes les opérations que nous effectuons dans notre langagière sont faites en temps réel! L'examen de diverses langues semble indiquer que certains principes combinatoires qui régissent l'organisation des unités discrètes sont généraux. D'où un second niveau niveau de problème : comment caractériser le potentiel linguistique d'un être humain? Quels sont les mécanismes qui permettent à un enfant d'apprendre une langue quelconque? Si nous savions quelles sont les méta-connaissances innées permettent à l'enfant d'apprendre à parler et à comprendre, il serait envisageable de disposer de systèmes capables d'agir de manière identique, on en saurait beaucoup plus sur la langue.

2.2. De la construction du sens

> Le destinataire d'un énoncé est un interprète. Au destinataire attentif, nul doute que le sens apparaît, même si les sons sont parfois troublés ou confus. (Une allusion à l'azur, M.Butel, l'Azur n°14)

Le sens : un papillotement de présences et d'absences.

"Ces mots poreux, chargés des discours qu'ils ont incorporés et dont ils restituent au coeur du discours en train de se faire, la charge nourricière et dépossédante, ces mots gigognes, qui se scindent, se déboîtent en d'autres mots, mots kaléidoscopiques dans lesquels le sens, multiplié à leurs facettes imprévisibles, s'échappe, en même temps, et peut, dans le vertige, se perdre, ces mots qui manquent, manquant à dire, manquant pour dire- défectueux ou absents - cela même qu'ils permettent de nommer, ces mots qui séparent, ceux-là même entre qui ils établissent le lien d'une communication, c'est dans le réel des non-coïncidences foncières, irréductibles, permanentes, dont ils affectent le dire, que se produit le sens. Ainsi, est-ce fondamentalement que les mots que l'on dit ***ne vont pas de soi****, mais si l'on veut ...* ***de l'Autre*** *: de l'Autre ouvrant le discours sur son extériorité inter discursive interne, la nomination sur sa perte relativement à la chose, la chaîne sur l'excès de sa* ***signifiance****, la communication sur sa béance intersubjective, et au total l'énonciation, sur la non-coïncidence à lui-même du sujet, divisé, de cette énonciation. Cet espace de non-coïncidences où se fait le sens, nourri de ces hétérogénéités qui le distinguent de la fixité une du signal, est aussi, indissociablement celui dans lequel il pourrait se défaire, si ne le protégeait, s'opposant à sa dispersion, une force de lien, de cohésion, de UN qui fait tenir une parole, qui fait que* ***tenir*** *une parole c'est, entre autres, faire* ***tenir ensemble*** *ce qui ne fait sens que de n'être pas un"* (Authiez-Revuz 1992).

De même que la production langagière est une construction permanente pour généralement dire quelque chose, la compréhension d'un énoncé est parallèlement une construction pour comprendre ce qui est dit. Et ce qui est dit est souvent, volontairement ou non, porteur d'éléments implicites, ambigus, ambivalents,..., auxquels le destinataire du message n'a pas accès ou qu'il peut construire (directement ou par inférence).

L'interprétation oblige souvent à résoudre dans le dire ce que le trajet du dire a produit comme masquages, brouillages, télescopages, etc., dans le vouloir dire.

Malgré tout, il existe un contraste entre l'omniprésence de la polysémie des diverses marques de surfaces dans les activités langagières et le petit nombre d'énoncés réellement ambigus. Quoi qu'il arrive en général dans une activité langagière, nous parvenons à récupérer les manques apparents, à rétablir les liens entre ce qui est perçu et une représentation de ce qui a voulu être dit, à identifier les flous du dire... De plus, *"si les êtres humains transmettent et stockent la plupart de leurs connaissances sous forme de textes en langue naturelle, c'est peut-être parce que l'ambiguïté est une qualité nécessaire à la représentation des significations, et non pas une tare léguée par l'histoire des langues, dont il convient de se débarrasser"* (Kayser 1989). Là encore, le langage manifestant une grande adaptation à nos besoins et à nos possibilités, il est souhaitable que les modèles de représentation du langage utilisés tiennent compte du fait que dans notre activité langagière nous commettons des erreurs (que nous savons gérer dans le cadre de cette activité), ou encore que nous ne comprenons pas toujours tout ce que nous entendons et que malgré tout nous réussissons généralement à transmettre/reconstruire du sens : **La représentation sémantique doit préserver l'ambiguïté et l'ambivalence inhérente au langage naturel.**

2.2.1. Représentation(s) du sens

"Trois problématiques de la signification, centrée sur le signe, dominent l'histoire des idées linguistiques occidentales :

a) La problématique de la ***référence****, de tradition aristotélicienne, définit la signification comme une représentation mentale, précisément un concept. Elle est reprise diversement aujourd'hui par la sémantique véric onditionnelle et la sémantique cognitive.*

*b) La problématique de l'****inférence****, d'origine rhétorique et de tradition augustinienne, définit la signification comme une action intentionnelle de l'esprit, mettant en relation deux signes ou deux objets. Elle est développée aujourd'hui par la pragmatique.*

c) La problématique de la ***différence****, d'origine sophistique, développée par les synonymistes des Lumières, puis par la sémantique dite structurale, définit la signification comme le résultat d'une catégorisation.*

Enfin la problématique du ***sens*** *prend pour objet le* ***texte****, plutôt que le signe, et définit le sens comme interprétation, passive ou active.*

/./ Dès que la problématique de la différence tient compte du contexte, elle dépasse le problème de la signification et s'ouvre à la question du sens...La prééminence du paradigme du sens se marque dans le fait que les sèmes inhérents ne sont actualisés qu'en fonction de licences ou prescriptions contextuelles, ce qui place en somme la signification sous le contrôle du sens" (Rastier in (Rastier & al. 1994)).

L'unification des paradigmatiques passe donc par une unification du *"paradigme lexical de la différence et le paradigme du sens, le sémiotique et le textuel, pour rendre compte de l'inférence et de la référence"* (Rastier in (Rastier & al. 1994))

Pour représenter la complexité de phénomènes à l'oeuvre dans la construction du sens d'un énoncé il ne suffit pas d'empiler des couches.

Les langues naturelles constituent un moyen irremplaçable pour communiquer et pour représenter des connaissances et toute activité langagière opère une recherche de marques de repérages que le langage construit afin que la compréhension se fasse. Ce constat ne nous est malheureusement pas suffisant pour déterminer les processus qui permettent d'identifier les connaissances présentes dans un fait de langue. La nature même de ces connaissances (et leurs rapports) restent mal connues.

Aucune énumération, aucune décomposition des unités qui composent un énoncé ne peut donner une somme de sens correspondante.

L'étude d'un énoncé consiste en un examen des articulations entre les différents niveaux de représentations des connaissances que le langage met en place : niveaux morphologiques, lexicaux, syntaxiques, sémantiques, pragmatiques. Si les linguistes ont pu établir de nombreuses théories sur chacun de ces champs particuliers de l'activité langagière, une théorie globale est loin d'être envisageable, d'autant plus que certains de ces champs d'étude sont encore loin d'être couverts. Il va ainsi pour les questions sémantiques.

En effet, l'analyse du langage pose (au moins) deux difficultés importantes quand il s'agit d'étudier les questions sémantiques. La première difficulté résulte de l'ambiguïté inhérente à toute activité langagière. Si nous avons du mal à représenter/formaliser les processus en jeu dans un énoncé, il ne faut pas oublier que l'élaboration de cet énoncé porte elle aussi souvent les traces de notre difficulté à coder sous une forme linguistique logique, rigoureuse et unique le sens d'un énoncé. C'est grâce à cette potentialité du langage de laisser les énoncés ouverts à différentes interprétations que le langage permet une grande richesse d'expressivité; et c'est

d'ailleurs, en partie, en raison de cette potentialité de produire des énoncés porteurs de sens flous qu'il est difficile de formaliser les activités de langue. La seconde difficulté résulte de l'immense quantité de connaissances nécessaires pour traiter des questions touchant au langage naturel. On l'a déjà dit précédemment, cette interaction complexe de phénomènes divers pose des problèmes non triviaux. D'autant plus quand on aborde les phénomènes sémantiques qui mettent en place des mécanismes d'analyse à profondeur variable. Une des grandes difficultés de la représentation des phénomènes linguistiques "tourne autour" des questions du sens : la formalisation et donc le codage des processus sémantiques est un tâche d'autant plus ardue que ces processus manipulent des indices d'une portée générale touchant aux activités fondamentales de l'intelligence humaine. On perçoit d'ailleurs très bien qu'une grande partie de la qualité du TALN est conditionnée par un résultat "global[10]" de cette formalisation. Cela est d'autant plus important que les questions sémantiques ont longtemps été négligées dans les travaux de réflexion sur le langage.

Une phrase signifie toujours plus. Même un mot, pris dans la texture de ses innombrables connotations est généralement irréductible.

2.2.2. De la construction du sens comme un travail

2.2.2.1. L'interprétation : une mise en action

L'interprétation consiste à proposer une cohérence sémantique du dit qui n'est pas directement le reflet d'un sens visé par le vouloir dire.

La piste sonore fournit des pistes de sens. Il s'agit ensuite de choisir parmi les chemins de sens possibles, ceux qui permettent d'accéder à un espace de sens cohérent. La construction du sens n'est pas assimilable à un processus linéaire de lecture d'étiquettes sémantiques. Il ne suffit pas en général de considérer un énoncé comme une entité dont le sens se construirait par une lecture séquentielle des indices sémantiques de ses composants. Les énoncés de langue produits sont porteurs de non-dits, de flou, de polysémie. *"Les expressions de langue ne sont plus formes désincarnées dotées d'un sens littéral, et ensuite enfilées bout à bout pour fabriquer un discours"*, elles représentent *"des indices relatifs à la construction du discours, généralement vu comme une mise en place de domaines liés entre eux, et structurés de façon interne par des modèles cognitifs"* (Fauconnier 1992). Le sens se nourrit de l'espace-entre, entre les fixités des sons produits, entre les définitions, entre les bruits; le sens se nourrit de chuchotements, mais aussi de gonflements. Le sens se négocie. Une conversation tient souvent lieu de négociations sur le sens des mots utilisés.

L'interprétation est une mise en action des matériaux mis en place par le dire.

L'analyse du sens d'un énoncé n'est donc généralement pas déterminée uniquement par la présence et la nature des formes linguistiques. Le sens n'est pas une donnée explicite des énoncés produits : le sens est une construction, le sens est un processus, un travail interprétatif. La production langagière met en place tout un réseau *ouvert* de correspondances entre les indices sémantiques présents dans un énoncé et le calcul interprétatif permet généralement de construire un (des) sens attaché(s) à l'énoncé produit. Chaque texte donne lieu à la construction d'un ensemble de mondes. L'interprétation consiste généralement à décrire les inférences qui permettent de passer de la linéarité du texte à une compréhension plus précise : le découpage du texte, l'inventaire et l'articulation des relations entre mondes jouent un rôle essentiel pour exprimer un des sens possibles du texte visé.

L'interprétation, c'est aussi s'investir dans ce processus de mise en action.

[10] Ce terme doit être compris en tenant compte du non-achèvement des résultats linguistiques et sur leurs remises en cause continues.

Le raisonnement au cours d'une interprétation est généralement non-monotone. On ne raisonne pas sur la seule base des informations explicitement accessibles. On est souvent amené lors d'un processus d'interprétation à faire des hypothèses, des choix... Un énoncé peut en effet porter et produire, volontairement ou involontairement, plusieurs interprétations, et c'est le résultat de cette confrontation qui fait le sens : l'énoncé `"Le coup d'état de grâce"` (Habert & Fiala 1989) émis au moment de la réélection de François Mitterrand est volontairement porteur de plusieurs sens, et c'est la confrontation de ces sens visés qui produit l'interprétation. De même, dire *"nous tripions"* (Jean Tardieu, entretien avec Pierre Dumayet) transforme la conversation en *"une action qui nous prend aux tripes"*, ce qu'elle est souvent d'ailleurs.

Les souvenirs, la mémoire peuvent aussi se faire reconnaissance.

Les unités linguistiques peuvent être les dépositaires d'informations qui dépassent le cadre de leur réalisation. Un mot peut être le "déclencheur" de mécanismes de rappel de situations spatio-temporels antérieures (enregistrées sous forme d'images mentales ou autres). Comment évaluer la portée de tels mécanismes dans les processus d'interprétation?

2.2.2.2. La construction du sens : un télescopage de savoirs

La construction du sens nécessite l'intervention de compétences de natures diverses qui fonctionnent en interaction. Le sens global d'un énoncé peut être assimilé à une espèce de feuilleté, un ensemble de couches articulées les unes aux autres, et dont l'organisation est à déterminer et à construire pour déterminer le sens[11] de cet énoncé. La question de savoir s'il est possible d'organiser (de manière hiérarchique ou autre) ces informations sémantiques se pose d'ailleurs de manière fondamentale dans la représentation des savoirs linguistiques. Une réelle interprétation de la polysémie, du flou, du vague, etc., nécessite la prise en compte de phénomènes qui dépassent parfois le cadre de l'énoncé produit. Aux connaissances statiques dont l'énoncé est porteur, il faut adjoindre des connaissances dynamiques qui éliminent ou atténuent les indéterminations, les imprécisions que l'énoncé révèle. Dans cet exemple, cité par (Kleiber 1994) :

```
"Je suis passé de la salle de bain dans la cuisine. Le plafond était
                                très haut."
                            (plafond[cuisine])

"Je suis passé de la salle de bain dans la cuisine. Le plafond était
                                trop haut."
                        (plafond[salle de bain])
```

Si l'interprétation contextuelle des deux phrases `"Le plafond était très haut"`, bien que précédées par la même phrase, conduit à un résultat différent, c'est parce que c'est la phrase qui construit le contexte pertinent. Le contexte se trouve construit pendant l'interprétation de la phrase elle-même. De même, si le sens d'un mot hors contexte nous échappe, il peut être possible de retrouver une des significations de celui-ci dans un contexte donné. Ainsi si le verbe `"stipendier"` n'a pas de signification précise pour bon nombre de locuteurs du français, il ne fait aucun doute que cette signification sera éclairée par la mise en contexte suivante :

```
"Les jockeys du Prix de l'Arc de Triomphe ont été stipendiés pour
 fausser la course, ils font l'objet de poursuites judiciaires."
```

Le contexte est infini : aucune saturation du contexte n'est possible, on peut toujours en dire plus.

[11] Quand on utilise l'expression "le sens d'un énoncé", il faut entendre "un des sens possibles que l'on peut attacher à cet énoncé", en particulier un de ceux que le destinataire peut construire quand il interprète cet énoncé.

La construction du sens d'un énoncé peut s'accompagner d'une instruction à mettre en place un contexte approprié pour pallier les problèmes posés par l'incomplétude de certaines expressions (Kleiber 1994). Ainsi si l'on considère l'énoncé (J. Moeschler 1989 repris dans (Kleiber 1994)) :

```
"Il fait beau, mais le pape est célibataire"
```

si l'on nous interroge sur la "validité linguistique" de celui-ci, et si notre intuition nous conduit à le rejeter, c'est peut-être dû au fait qu'il ne nous a pas été possible, *"en temps limité et dans les conditions particulières et artificielles de la demande, de construire une situation appropriée"* (Fauconnier 1992).

2.3. Modèles pour la construction du sens

2.3.1. Connexionnisme et Composés

Le traitement des composés peut être considéré comme un domaine d'études intéressant pour l'utilisation de réseaux de neurones appliqué au langage naturel. Si le connexionnisme traite du rapport forme/sens quand il travaille sur les données de langue, les composés constituent un terrain particulièrement propice pour rendre compte des phénomènes complexes qui interagissent dans ce rapport. De plus, l'utilisation de tels réseaux qui mettent en place des mécanismes interprétatifs qui adaptent les paramètres de résolution en fonction des données traitées, peut se révéler pertinente pour traiter certaines configurations nominales dont la construction peut s'accompagner de mécanismes mettant en jeu un affinage progressif de la représentation.

2.3.1.1. Apprentissage et réseaux connexionnistes

Les réseaux de neurones sont des modèles inspirés du fonctionnement des neurones biologiques. Au cours des années 40, des chercheurs ont commencé à travailler sur la modélisation du système nerveux. Les neurones étaient simulés par automates élémentaires reliés entre eux par des liaisons ou connexions (les synapses) qui permettaient de faire passer un influx plus ou moins intense. Chaque neurone réémet des influx en fonction de la force des potentiels reçus. L'information dans un réseau de neurones se situe à deux niveaux :

- Dynamiquement, les influx circulant d'unité en unité représentent la réaction du réseau à des stimuli reçus;
- Statiquement, la force des connexions et la sensibilité des neurones caractérisent la connaissance emmagasinée dans le réseau, sa mémoire à long terme, ils modulent les réactions du réseau.

Mimant les phénomènes de plasticité synaptique observées par les neurobiologistes, les liaisons entre automates sont capables de se modifier dynamiquement, simulant ainsi des phénomènes d'apprentissage. L'apprentissage (avec ou sans modèle correct) consiste à modifier les paramètres du réseau afin d'induire un comportement ultérieur de celui-ci qui soit mieux adapté aux données reçues. L'idée de reproduire des architectures neuronales par des réseaux de neurones artificiels a été essentiellement motivée par des problèmes de reconnaissance de formes ou de classification.

2.3.1.2. Réseau de neurones et TALN

Dans le cadre du traitement des données symboliques telles que le langage naturel, on distingue deux types de réseaux : les réseaux localistes et les réseaux subsymboliques.

- **Les réseaux localistes :**

Ce sont les héritiers directs des réseaux sémantiques. Dans un tel réseau, chaque unité est associée à un des symboles du domaine (symbole de grammaire, mot du lexique ou trait sémantique). Les connexions y ont également une valeur symbolique telle que la compatibilité sémantique ou syntaxique entre des mots lorsqu'elles sont positives ou leur antagonisme dans le cas contraire. Seule l'activité propagée entre les unités est indifférenciée.

- **Les réseaux subsymboliques :**

Aucune valeur interprétative n'est assignée a priori aux composants du réseaux ou aux activités qui s'y propagent. Les stimuli appliqués sur le réseau sont encodés de façon opaque (codage aléatoire) ou parcellaire (utilisation de micro-traits). S'il émerge une organisation interne qui n'a pas été explicitement introduite initialement, elle est révélée par une analyse des modifications après apprentissage. L'utilisation d'un réseau se passe en deux temps :

> Phase d'acquisition : une procédure d'apprentissage et un corpus d'entraînement servent à l'ajustement automatique des paramètres. Dans les réseaux localistes il est possible d'attribuer à la main des valeurs aux paramètres du réseau.
>
> Phase d'exploitation du réseau : on observe les capacités de classification et de prédictivité sur des données nouvelles inconnues.

Les réseaux sont utilisés, par exemple, pour apparier sens et forme et rendre compte de la polysémie et de compositionalités complexes (Fuchs &Victorri 1990).

2.3.1.3. Réseau de neurones et "noms composés"

(Habert & Jacquemin 1993) présentent un panorama des traitements automatiques mis en place pour l'analyse des noms composés. Dans le cadre des méthodes connexionnistes, nous reprenons leur présentation du travail de C. Jacquemin.

- (Jacquemin 1992) met en place un réseau localiste pour la reconnaissance de noms composés subissant des variations tels que `câble à haute tension`. A chaque entrée lexicale est associé un lemme. Le réseau est divisé en sous-réseaux : un sous-réseau a pour tâche de reconnaître un nom composé particulier. *"Par baisse de seuil, le réseau révèle les variations inattendues des composés et un apprentissage supervisé permet d'intégrer cette connaissance dans le réseau par une modification de poids. La reconnaissance ultérieure de la modification ne nécessitera plus alors de baisse de seuil, mais sera acceptée par la nouvelle configuration de poids obtenue par apprentissage"* (Habert & Jacquemin 1993).

- (Jacquemin 1993) met en place un réseau qui s'inspire de la neurophysiologie pour la mise en oeuvre des mécanismes d'apprentissage et de propagation sur deux tâches liées à la composition nominale. *"Le premier ensemble de données introduit dans le réseau est composé d'interprétations sémantiques de composés en `N à N` et des acceptabilités de ce composé pour certaines variations syntaxiques et certains tests interprétatifs. Par exemple `verre à pied` a pour sémantique `AVEC (partie-de)` et accepte des variations telles que la modification adjectivale du deuxième nom dans `verre à pied en métal`. L'observation du réseau met en évidence des connexions renforcées pour certaines associations statistiquement pertinentes"* (Habert & Jacquemin 1993). La deuxième tâche doit rendre compte *"de la compositionnalité dans les composés en `cuisson Prép N` en fonction de la préposition et du nom en jeu /./. Le renforcement de certaines connexions distingue le rôle inférentiel des prépositions incolores `(à, de)` du codage sémantique des prépositions colorées `(sur, nous)` qui imposent l'interprétation finale"* (Habert & Jacquemin 1993).

2.3.2. Représentation hiérarchisées : mots ou choses

2.3.2.1. Hiérarchies

Ontologie vs. Classifications linguistiques

- Les choses telles qu'elles sont. Et les classes d'objets qu'une langue donne à voir.

La construction du sens met en oeuvre une confrontation permanente entre les faits de langue et les existants du monde. Le traitement de la polysémie pose d'ailleurs clairement le problème de *"la conceptualisation des entités de base supposées"* (Habert & Cadiot 1997).

- L'exemple de `"livre"`.

> Un livre peut être considéré comme un `Objet-Physique`.
>
> Un livre peut aussi être appréhendé via le contenu qui le compose.
>
> Il est aussi un élément constitutif du processus de publication ou de commercialisation attaché à la réalisation de ce livre.

Savoir sur le monde et savoir sur la langue

a. Articulation Langagier-Concepts

Définir le sens est une tâche délicate. Si comme Mallarmé, on n'accorde au mot `"fleur"` que la seule légitimité d'être `"l'absente de tous bouquets"`, on aboutit à une sémantique de l'absence qui interdit toute définition sémantique formelle. Ce qui est en cause fondamentalement dans les problèmes de construction du sens, et plus généralement dans l'étude de modèle de représentation du langage naturel tourne autour des notions de concept, de signe et de référent. Dans le système de la langue, le signe `"fleur"` est un assemblage arbitraire de deux voyelles et trois consonnes. Il est donc à coup sûr porteur d'une forme. Le passage de ce signe dans le discours va permettre de référer à un objet unique : `"cette fleur est bleue"`. Le sujet de cette phrase désigne un objet du monde identifiable (le référent). Le signe est donc un simple élément capable d'éveiller dans l'esprit une idée et qui n'est pas sa représentation : le signe apparaît comme un déclencheur. Les mots et les phrases ne présentent aucune affinité préétablie avec les objets. *"La problématique de la référence, de tradition aristotélicienne, définit la signification comme une représentation mentale, précisément un concept. Elle est reprise diversement aujourd'hui par la sémantique vériconditionnelle et la sémantique cognitive"* (Rastier & al. 1994). Le concept est le fruit de l'activité de l'intelligence (conception). *"Depuis G.Frege (Frege 1893), un concept peut être vu comme une fonction définie sur un certain domaine et à valeur dans l'ensemble de valeurs de vérité {vrai faux}. Cette fonction a pour but de discriminer les objets auxquels s'appliquent le concept..."* (Haton & al 1991). Selon Lyons, le concept *"exprime toute idée, toute pensée ou toute construction mentale au moyen de laquelle l'esprit appréhende les choses ou parvient à les reconnaître"* (in Haton & al 1991), selon Kayser, un concept est une *"entité symbolique recevant des propriétés inférentielles"* (in Haton & al 1991).

1. Inutilité de définir le sens d'un mot

L'approche de D. Kayser (Kayser 1987,1992) vise à modéliser non pas le sens, mais les processus de compréhension des phrases en leur associant des ensembles de transformations gouvernées par des règles d'inférence plausible. De plus, cette approche vise à redéfinir la notion de concept traditionnellement associée aux mots. `"Livre"` selon D. Kayser renvoit à une multiplicité de types de référents possibles dont on peut rendre compte en se passant de la notion de sens.

`"Livre"` peut renvoyer à :

- un élément indéterminé d'une classe d'équivalence

(1) Ce livre se trouve dans toutes les bonnes librairies

- une entité à partir de laquelle les éléments de cette classe ont été créés

(2) Ce livre a été tiré à 150000 exemplaires

- un objet manuscrit

(3) Jean est parti à la campagne pour écrire un livre

- aux idées contenues dans ce livre

(4) Ce livre a fortement influencé les révolutionnaires de 1789

- à la commercialisation de ce livre

(5) Ce livre a été un fiasco pour son éditeur

- à la lecture de ce livre

(6) Ce livre est ennuyeux

La variation de prédicat entraîne également un référent différent et donc un sens différent pour "livre" dans (3), (4), (5).

D. Kayser émet l'hypothèse du mot générateur d'entités dans un réseau sémantique :

"Lorsque l'interprétation d'un mot est souhaitée, on fait fonctionner le générateur au premier niveau. On obtient ainsi un ou plusieurs noeuds (au groupe nominal "un livre" correspond un noeud situé dans la branche hiérarchique dominée par le noeud "Objet-Physique"). Si ce nouveau noeud remplit les conditions souhaitées, le processus peut s'arrêter là. Sinon on demande au générateur de fabriquer un ou plusieurs autres noeuds ("représentant indéterminé d'une classe", actions acceptant "livre" comme objet : la lecture, l'écriture, la publication...). On réitère ce processus jusqu'à trouver un (ou plusieurs) noeuds satisfaisants, ou jusqu'à avoir épuisé les ressources mises à sa disposition" (Kayser 1987).

2. Utilité de définir le sens d'un mot

Pour Kleiber et Riegel (Kleiber & al. 1989), il convient de maintenir pour "livre" l'hypothèse d'un sens unique, capable d'expliquer à la fois :

- la variation référentielle des SN comportant "livre" dans les énoncés (1) et (7);
- la possibilité d'avoir avec un référent stable des effets interprétatifs différents dus à ce qu'ils appellent le principe de métonymie intégrée (certaines caractéristiques de certaines parties peuvent caractériser le tout).

1. Pour Kleiber et Riegel, il n'est guère besoin de procéder à un éclatement sémantique de "livre" pour rendre compte du changement de référent des SN des énoncés suivants :

(7) Ce livre est sale est déchiré

Ce livre se trouve dans toutes les bonnes librairies

Il s'agit de mécanismes référentiels généraux liés à la distinction type/token. La structure conceptuelle du mot "livre" (i.e. son sens) révèle l'existence d'entités individuelles regroupant elles-mêmes des occurrences.

2. On peut résoudre les interprétations (3), (4) et (5) en optant pour un référent qui reste le même et en expliquant les phénomènes d'interprétation globale au moyen du principe de la métonymie intégrée.

b. Représentation des concepts

"Un des grands problèmes relevés dans la modélisation des connaissances en Langage Naturel est celui du principe d'une relation "simple" entre le mot (réalisation lexicale) et le concept (conception cognitive)" (Kayser 1991).

Traditionnellement, on considère le concept comme un atome de signification que l'on traduit en logique par des prédicats ou par des constantes. On peut aussi le représenter sous forme de molécule qui se décompose en primitives atomiques. Dans les deux cas, un concept se traduit par un objet unique. Dans l'approche développée par Kayser, un concept C est représenté par une famille ouverte d'objets : l'ensemble des x pour lesquels C est un NOM-VALIDE dans un certain contexte Contexte.

Représentation de C : = { x tels que NOM-VALIDE(C, x, Contexte)}

D. Kayser (Kayser 1988a) décrit *"un certain nombre de caractéristiques que possèdent les concepts utilisés dans la vie courante"* et montre que *"ces caractéristiques ne sont pas des traits superficiels, liés au langage au niveau lexical mais des traits profonds, liés à la nature même de nos connaissances"*.

"Il existe un contraste frappant entre le nombre de situations que nous pouvons discerner et le nombre d'éléments de description que notre appareil cognitif nous permet de manipuler. En effet, la finesse de nos perceptions nous apporte une quantité d'informations sans commune mesure avec nos facultés de manipulation et de mémorisation; faute d'être capables de gérer des assemblages de concepts distincts pour représenter des situations distinctes, nous réutilisons les mêmes concepts, mais nous savons que ces concepts n'ont plus alors de définition précise, et des mécanismes d'analyse à profondeur variable nous permettent à chaque occasion de reconstituer l'interprétation la plus adéquate du concept considéré" (Kayser 1992).

c. Le Concept : un point d'entrée pour des inférences

Kayser avance que *"les suggestions habituellement proposées (logiques multi-valuées et (ou) non monotones) pour tenir compte de la "typicalité" ne cernent pas parfaitement le phénomène. L'idée force consiste à séparer la notion de concept de celle d'élément de base (prédicat, noeud) et de représenter un concept par une famille ouverte d'entités du système. Chaque entité dénote une interprétation possible du concept, et ces interprétations sont ordonnées par profondeur croissante"* (Kayser 1988a). Les concepts sont donc des points d'entrée qui permettent de déclencher des inférences, de plus ils dominent des graphes ouverts et évolutifs car *"il importe de pouvoir modifier le grain de représentation au cours d'un raisonnement"* (Kayser 1992).

Ainsi à propos du concept de "lion" : on peut dire que *"ce concept ne se traduit pas par une seule entité dans le système de représentation, mais par une famille ouverte, et partiellement ordonnée, d'interprétations de ce concept."*

Kayser pose les définitions suivantes :

notion : ce que l'on veut représenter,
entité : élément du système de représentation,
concept : cas où l'on postule une correspondance biunivoque entre les deux.

"Au lieu de représenter une notion par une seule entité du système de représentation, nous décidons de lui faire correspondre une famille ouverte d'entités. On accède à cette famille par un point d'entrée (un mot du langage indexe un point d'entrée, sauf dans les cas d'homonymies "accidentelles"); on ne substitue à ce point d'entrée d'autres éléments de la famille que si les inférences en cours l'exigent" (Kayser 1992).

"Cette approche permet :

- *de réduire la distance entre mot et concept : `C` est le nom d'un concept; c'est aussi un mot de la langue : un mot ne correspond à une multiplicité de concepts qu'en cas d'homonymie `accidentelle`;*
- *de rendre compte de certains mécanismes de coréférence : le principe circonscriptif tend à maximiser la coréférence, et la logique des défauts permet de formaliser le choix de l'antécédent contextuellement le plus plausible;*
- *mais surtout de traiter certaines régularités des phénomènes d'extension de sens et de glissement de sens : une même expression sert à désigner différentes notions qui entretiennent entre elles des rapports constants; /./ De tels ajouts à l'ensemble des significations possibles d'un mot sont représentables par une règle générale d'extension de sens (règle de type (RD) pour prendre en compte les exceptions locales); de plus, un tel ensemble se construit dynamiquement, au fur et à mesure que la profondeur d'analyse augmente"* (Kayser 1988b).

On s'écarte donc de la notion traditionnelle du sens[12] dans la mesure où il n'existe pas d'espace de sens et qu'il est donc impossible d'exhiber un objet (formule, graphe, point, ...) en affirmant : voici le sens de cette expression (Kayser 1992).

d. `Moulin` et la profondeur variable

L'approche suivie par D. Kayser vise à établir un équilibre entre des savoirs établis (non modifiables par inférence) et le raisonnement (les processus qui affinent par inférence les savoirs établis). La profondeur variable suppose en effet une hiérarchie de concepts aux noeuds de laquelle sont attachées des connaissances, des rôles qui permettent des inférences. Cela structure le savoir sur le monde effectivement mis en oeuvre. Cette hiérarchie est le support de l'inférence. Le savoir sur le monde est structuré et cette organisation guide le travail inférentiel.

Ainsi dans le cadre des séquences `moulin à N2`, on peut mettre en oeuvre la profondeur variable défendue par D. Kayser : on construit au départ une représentation sémantique initiale peu déterminée (uniquement le prédicat pour `moulin`) qui peut ensuite être affinée en tenant compte des éléments d'informations complémentaires permettant d'étendre la représentation initiale. On retrouve là la notion du sens comme processus : le sens n'est pas une donnée, il est reconstruit par affinements successifs. Ce type d'approche permet tous les remodelages possibles. **Il reste à pouvoir contrôler ces ajustements ou ces remodelages** (comment éviter de surgénérer ?) et donc à fournir des processus de contrôle sur les entités visées si l'on veut évaluer les modifications réalisées. Cette approche met aussi clairement en avant la difficulté qu'il y a à articuler les différents niveaux de savoirs manipulés. La mise en place d'articulations dynamiques entre les savoirs décrits n'est pas un problème qui se laisse résoudre

[12] *"Ce que nous avons dit des concepts et des mots correspond à l'idée qu'ils n'ont pas de signification propre, et que cette signification peut se modifier au cours des processus inférentiels. On arrive ainsi à des notions d'ambiguïté conceptuelle, voire de pensée ambiguë. De telles notions sont incompatibles avec des modèles vériconditionnels du sens. Elles pourraient donc contribuer à orienter la sémantique formelle dans des directions fort différentes de celles où elle s'est dirigée jusqu'ici"* (Kayser 1992).

facilement surtout si les structures de représentation utilisées ne permettent pas de tenir compte de tels ajustements toujours possibles.

Une approche conceptuelle à la Kayser met aussi en lumière les difficultés de penser et de représenter une articulation cohérente des connaissances conceptuelles et des connaissances lexicales habituellement manipulées dans les modèles de représentation du langage naturel : de bons outils de ce type ne sont pas encore disponibles en raison justement de cette difficulté à mettre en place une telle articulation. D'autant plus qu'il ne s'agit pas de privilégier une pratique par rapport une autre : à l'inverse d'une approche lexicaliste, l'approche conceptuelle a ainsi du mal à rendre compte de certains phénomènes comme les régularités sémantiques ou morphologiques. Cette approche met aussi en lumière les limites des formalismes d'unification : comment, dans un cadre symbolique de représentation, peut-on rendre compte des probabilités différentes des interprétations ?

2.3.2.2. Accès à des scénarios (Pustejovsky & Boguraev 1991)

Deux approches pour traiter la polysémie

a. Approche lexicographique traditionnelle

Cette approche consiste à créer autant de d'entrées qu'un mot a de types sémantiques. On dira qu'il y a deux entrées `"livre"`, une pour l'objet physique et l'autre pour son contenu.
Cette approche pose les problèmes suivants :

- Elle postule qu'il est possible de lister tous les sens et ne rend pas compte de la possibilité de créer des sens en contexte.
- Elle génère aussi de l'ambiguïté lexicale et pose donc le problème de la désambiguïsation.
- Elle n'explique pas non plus le lien qui existe entre les différents sens (le lien entre l'objet physique et son contenu pour `"livre"`).
- Elle traite enfin toutes les ambiguïtés de la même manière.

b. Approche semi-polymorphique du lexique génératif de Pustejovsky

Le travail de Pustejovsky vise à une prise en compte de l'évolution permanente du lexique et à intégrer ces phénomènes dans le travail de représentation de celui-ci. Cette évolution permanente est le résultat de la création perpétuelle de mots nouveaux mais aussi le résultat d'une mise en contexte des éléments lexicaux capable de produire des transferts, des élargissements, des glissements de sens. Une représentation statique du lexique n'est évidemment pas apte à rendre compte des mécanismes de sélection du sens d'un mot dans la mesure où elle ne met pas en place les liens qui permettraient une discrimination entre les différents sens possibles. De plus les facteurs contextuels ne sont pas toujours suffisants pour préciser le sens d'un mot.

Repérer les régularités lexicales

Il ne s'agit pas de créer autant de représentations du sens d'un mot qu'il y a de réalisations syntaxiques de celui-ci produisant une interprétation différente mais au contraire de capter les généralisations pertinentes entre les différents sens d'un mot (Pustejovsky 1991) tout en définissant la notion d'inférence lexicale. Le but de cette approche est de ne pas avoir un type sémantique par mot, mais de donner aux mots une riche représentation typée qui peut être

considérée comme une réserve de types, en indiquant les différentes extensions sémantiques possibles d'un mot. Chaque mot est donc représenté par plusieurs niveaux de représentation qui sont connectés entre eux.

- La structure argumentale spécifie le nombre et le type des arguments d'un item.
- La structure événementielle permet de définir le type et la structure événementielle.
- La structure des `qualia` indique les caractéristiques sémantiques essentielles de l'item lexical, c'est-à-dire les propriétés et les événements qui lui sont liés, et ce dans quatre rôles :
    - Un **rôle constitutif** qui indique la relation entre un objet et ses constituants ou parties : matériau, poids, parties.
    - Un **rôle formel** qui distingue l'objet à l'intérieur d'un domaine plus large : orientation, grandeur, forme, dimension, couleur, position.
    - Un **rôle télique** qui exprime la fonction et la finalité de ce que décrit l'entrée lexicale visée : but, fonctions constitutives.
    - Un **rôle agentif** qui spécifie ce qui se rapporte à l'objet en général, en particulier à son origine : créateur, artefact, espèce naturelle.

Une entrée lexicale est donc associée à un prédicat au sein de la structure que constitue le `qualia`. La structure de `qualia` est décomposée sous forme de champs descriptifs qui permettent la composition sémantique.

```
livre
CONSTITUTIF = pages
TELIQUE = lire
AGENTIF = écrire
FORMEL = objet physique
```

Les différentes facettes de signification d'une entrée lexicale sont donc réunies dans une structure `qualia` et les relations d'héritage permettent d'infléchir l'interprétation des entrées lexicales. Cette hiérarchie doit traduire les liens transversaux qui existent entre les entrées lexicales et les liens orthogonaux que l'on peut définir entre la représentation lexicale et une représentation des concepts organisée hiérarchiquement. La signification d'une entrée lexicale est saisie au travers d'une série d'oppositions plutôt qu'en décomposant cette entrée en une suite de primitives sémantiques.

Mais cette structure traduit aussi un certain point de vue sur le mot `"livre"` et sur son rapport au monde. On peut modifier les valeurs associées aux rôles associés à la structure `qualia` si on adopte un autre point de vue : on peut en effet considérer que le livre est constitué de parties ou de chapitres différents (`CONSTITUTIF = chapitres`), qu'il est porteur d'informations (`FORMEL = information`), qu'il est conçu pour être acheté (`TELIQUE = acheter`) et enfin qu'il est produit par un imprimeur (`AGENTIF = imprimer`). La question de savoir quelle attitude adoptée pour appréhender ce mot reste donc ouverte et la structure `qualia` ne propose en fait que la mise en place dans la définition du mot que d'une description particulière.

L'opération de coercion

La coercion est une opération sémantique qui convertit l'argument d'une fonction en un type attendu par celle-ci :

> *"if α is of type <b, a>, and β is of type b, then $\alpha(\beta)$ is of type a"* (Pustejovsky 1991)

Si on suppose que chaque expression a à sa disposition un ensemble d'opérateurs de conversion, Σ_α, qui peuvent agir sur elle (en changeant son type et sa signification), on peut s'y référer comme à des alias de cette expression. De plus, il est possible de faire appel directement à l'un de ces alias dans les règles définies pour la description des items lexicaux.

> *"if α is of type <b, a>, and β is of type c, then :*
> *(i) if type c = b, then $\alpha(\beta)$ is of type a;*
> *(ii) if there is $\sigma \in \Sigma_\beta$, such that $\alpha(\beta)$ results in an expression of type b, then $\alpha(\sigma((\beta))$ is of type a;*
> *(iii) otherwise a type error is produced"* (Pustejovsky 1991)

Cette opération permet la prise en compte des phénomènes de coercion de sens sur l'énoncé :

```
"Cette 2CV conduit dangereusement"
```

Comme il n'est pas souhaitable de classer `"2CV"` dans les `Nhumains`, cette opération permet justement, dans ce contexte particulier et temporairement, de modifier le type initialement associé à `"2CV"`. Cette opération permet aussi de prendre en compte l'interprétation des configurations nominales du type `moulin à paroles` qu'il est possible de considérer comme une coercion de type.

L'exemple de `"fast"`

Pustejovsky (Pustejovsky & al. 1993) illustre son travail de représentation de la polysémie lexicale sur l'entrée lexicale "fast" à laquelle on peut attribuer au moins les différents sens suivants :

- fast1 : se déplacer rapidement

```
"The island authorities sent out a fast little government boat, the
                    Culpeper, to welcome us"
```

- fast2 : réaliser une action rapidement;

```
"a fast typist"
```

- fast3 : disposer d'un court laps de temps pour réaliser quelque chose;

```
"a fast book"
```

- fast4 : prendre rapidement des initiatives.

```
"You may decide that a man will be abble to make the fast, difficult
                        decisions"
```

a. Analyse de `"fast"` comme s'appliquant au rôle télique du nom qu'il modifie

Comme on l'a vu précédemment, il ne s'agit pas de proposer une représentation qui énumère les différents sens recensés. Pustejovsky propose au contraire de construire une représentation

lexico-sémantique qui indiquera que l'adjectif agit sur le prédicat de type "event" (en modifiant le type d'activité) contenue dans le rôle télique du nom qui lui est associé : "fast" peut différencier le prédicat attaché au nom tête qu'il modifie.

```
car (*x*
Const : {body, engine,...}
Form : physobj(*x*)
Telic : drive(T,y,*x*)
Agentive : artifact(*x*)
```

Considérons "fast" comme partie intégrante du prédicat associé au rôle télique du nom qu'il modifie. Ici "car" dans "a fast car". Dans cette expression, c'est l'activité de conduire qui est modifiée par l'adjectif. Les mêmes remarques s'appliquent à (nom tête:"typist", rôle télique : type(T,y,*x*)) dans "a fast typist" ou à (nom tête:"reader", rôle télique : read(T,y,*x*)) dans "a fast reader".

La structure qualia et l'opération de coercion définies par Pustejovsky donnent la possibilité de générer différentes interprétations d'une entrée lexicale sans avoir à les lister toutes de manière rigide.

Les opérations génératives qui opèrent sur cette représentation sont :

- la possibilité de générer, à partir de la structure des qualia, les différents types sémantiques du mot;
- la coercion de types qui permet à un prédicat de changer le type de son argument vers celui qui est requis par le prédicat;
- le liage sélectif qui permet à un prédicat de ne modifier qu'une partie de la structure des qualia;
- la cocomposition, qui permet à un argument de changer la sémantique du prédicat en composition.

Représentation de la polysémie nominale

Cette approche distingue deux types de nom : les noms non-polysémiques qui ont des types sémantiques simples et les noms polysémiques qui ont des types sémantiques complexes. Un nom a un type simple s'il réfère à un seul type. Dans ce cas, il a un seul argument, qui correspond directement au type du rôle formel. Un nom a un type complexe quand il réfère à un type pointé[13] (t1.t2). Ce type correspond au produit cartésien des types définis dans la structure argumentale, c'est-à-dire t1 et t2. Si le nom a deux arguments : le rôle formel définit la relation entre les deux arguments; le nom dénote trois types différents t2.t3, t2, t3. L'ensemble des types dénotés par un item constitue son "paradigme conceptuel lexical". "Livre" par exemple a un type complexe : il a deux arguments dont le rôle formel indique la relation : y contient x. Sa fonction est d'être lu et il est créé par l'acte d'écriture. Le nom "livre" peut donc projeter séparément les trois types de son "paradigme conceptuel lexical" : le type pointé, informationnel et phys-objet. Les types événements seront aussi accessibles par coercion ou liage sélectif.

Evaluation paresseuse par nécessité (Copestake & Briscoe 1994)

[13] L'introduction du type pointé permet ainsi de lever l'objection soulignée supra sur le choix du rôle formel associé à livre.

A la manière de Pustejovsky, (Briscoe & Copestake 1993) étudient les mécanismes d'inférence sur le lexique qui permettent d'articuler analyse syntaxique et interprétation en proposant une décomposition lexicale du même type. Là encore leur approche met en avant la nécessité de définir une hiérarchie des concepts en parallèle à la hiérarchie lexicale. Leur approche utilise un langage de représentation lexicale qui manipule des attributs typés pour définir les entrées lexicales : les types permettent une structuration des entrées lexicales qui sont associées à des structures d'attributs typés. Là encore, on est confronté à une difficulté liée à la mise en place de hiérarchies rigides qu'il est souvent difficile de faire évoluer.

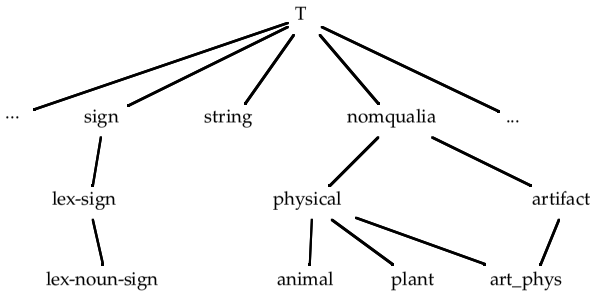

Figure 2.1 : Fragment de la hiérarchie de types (Briscoe & Copestake 1993)

Une description lexicale est par exemple la structure suivante :

```
book 1
< > = lex-noun-sign
< QUALIA > = art_phys
< QUALIA TELIC PRED > = read
< QUALIA FORM > = indiv
```

Cette utilisation de la structure de qualia (qui utilise une représentation hiérarchisée de concepts) permet de décrire des entrées lexicales par héritage de structure qualia. C'est le cas pour un des sens associé à l'entrée "novel" :

```
novel 1
< QUALIA > < < book_1 < QUALIA > .
```

qui spécifie que cette entrée hérite d'une structure qualia (via le symbole <) construite pour "book". Il est aussi possible d'affiner une relation d'héritage en rendant modifiable une partie des attributs pour rendre possible la redéfinition de certaines entrées lexicales dans la structure hiérarchique. Certaines valeurs d'attributs peuvent ainsi être marquées comme modifiables (via un "/" à la droite de la valeur), ou non-modifiables (via un "/" à la gauche de la valeur). Pour inscrire "dictionary" dans la structure hiérarchique initiée la représentation de "book", il est nécessaire de modifier le rôle télique de "book" et le rendre modifiable dans la mesure où si "dictionary" peut hériter de la structure qualia définie par "book", il est nécessaire d'ajuster le rôle télique en "refer-to".

```
book 1
< > = lex-noun-sign
< QUALIA > = art_phys
< QUALIA TELIC PRED > = /read
< QUALIA FORM > = indiv

dictionary 1
< QUALIA > < < book_1 < QUALIA > .
< QUALIA TELIC PRED > = /refer_to
```

a. Traitements de la polysémie

- Polysémie de construction (Constructional Polysemy)

"There are many cases of apparent polysemy which we would argue are better treated as 'constructional' polysemy, in that the lexical item is assigned one (often more abstract) sense and processes of syntagmatic combination or 'co-composition' are utilised to specialise this sense appropriately. We treat this as a process of sense modulation, represented by specialisation in the Lexical Knowledge Base, in contrast to the process of sense extension /./, which we represent using lexical rules" (Briscoe & Copestake 1993).

Cette notion correspond à la nécessité de choisir l'aspect de sens approprié (dans le possible défini) à une situation donnée plutôt qu'un changement de sens (exemple de `"fast"` in (Briscoe & Copestake 1993) p20).

- Extension de sens (Sense extension)

"By contrast with constructional polysemy, we argue that there are systematic polysemies which are best represented as lexical rules, which we refer to as sense extensions; that is predictable creation of different but related senses" (Briscoe & Copestake 1993).

Au lieu de lister toutes les extensions de sens dans le lexique, ce qui impliquerait d'ailleurs une remise à jour permanente du lexique, il s'agit de représenter via des règles lexicales les mécanismes qui réalisent ces extensions de sens. C'est le cas des emplois métaphoriques des noms d'animaux appliqués pour décrire des êtres humains.

b. Les règles lexicales

"Lexical rules are used to create derived sense. /./ We use the lexical rules to dynamically generate alternatives during constraint resolution of nodes with lexical types" (Briscoe & Copestake 1993).

Les règles lexicales peuvent être vues comme des processus de conversion. Dans l'emploi métaphorique des noms d'animaux pour qualifier les êtres humains, il suffit d'ajuster la représentation sémantique initiale, en ajoutant un type `animé` et deux sous-types `humain` et `animal`, puis construire un mécanisme de conversion qui réalise l'interprétation de ces emplois métaphoriques.

Telle entrée lexicale : tels scénarios?

a. Tous les noms sont-ils descriptibles par une structure prédicative ? Tous les noms ont-ils des rôles téliques ?

"Pustejovsky ne cherche pas fonder linguistiquement les distinctions qu'il établit (sauf pour les types P, S et T). Celles-ci restent avant tout cognitives même si les sous-champs discriminés diffèrent dans l'utilisation linguistique qui en sera faite. Il est même plausible de soutenir que les deux ordres de discriminations (cognitif et linguistique) doivent être maintenus et ne sont pas forcément isomorphes. Le point important, sur lequel l'accord paraît établi, est qu'il convient de circonscrire l'accès à la Représentation Sémantique lexicale" (Fradin 1993). Les champs définis dans la structure `qualia` peuvent-ils s'appliquer à toutes les entrées lexicales? Fradin (Fradin 1993) pense que non, notamment pour ce qui concerne la distinction rôle formel et rôle constitutif. De plus, *"ces distinctions ont-elles une pertinence linguistique. Rien n'est moins sûr, car la frontière de l'inaliénabilité, par exemple, passe à l'intérieur de chacune"* (Fradin

1993). *"Conceptuellement, ce qui ressortit au rôle télique est plus clair. D'un point de vue formel, il devrait regrouper des formules avec plusieurs arguments. D'autres cependant resteraient sous le rôle agentif, notamment toutes celles exprimant l'origine de l'objet référé. tel que nous le comprenons, le rôle agentif a trait à la catégorisation de l'objet (ou de l'évenance) et à sa place dans un chaînage de causalité"* (Fradin 1993). Fradin propose d'ailleurs de réorganiser la structure `qualia` en la réduisant sous la forme suivante :

- Un rôle aspectal qui *"concerne les propriétés aspectales et constitutives : rapports tout/partie, figure/site et les propriétés servant à identifier l'objet en question"* (Fradin 1993).
- Un rôle télique qui indique les propriétés téliques au sens large : *"la fonctionnalité de l'objet aussi bien que son origine"* (Fradin 1993).
- Un rôle qui traduit le mode d'existence de l'objet, *"essentiellement sa catégorisation par rapport aux enchaînements de causalité"* (Fradin 1993).

b. Aspect fini de la représentation : on fixe le nombre de scénarios

Ce type de représentation reste en effet en deçà des possibles de réalisation d'une entité donnée. Puisque les faits de langue sont soumis à des évolutions permanentes, leur représentation se trouve en permanence remise en question. Il faut pouvoir les ajuster en tenant compte des incidences potentielles que les conditions de leur réalisation révèleront et il est difficile de prévoir tous les scénarios possibles à attacher aux mots.

2.3.2.3. Les classes de mots

G. Gross (Gross 1994) travaille dans le cadre théorique d'un lexique composé de phrases et non de mots isolés et propose un modèle pour le traitement informatisé du lexique à des fins notamment de traduction automatique. Il construit pour cela des classes de noms clairement spécifiées. Les classes de noms sont définies non pas "dans l'absolu", mais toujours en fonction d'un opérateur approprié (plus précisément d'une construction syntaxique). Les classes d'objets représentent en fait le domaine d'arguments de cet opérateur. Les classes d'objets sont donc définies syntaxiquement et elles sont destinées à classer les noms et à décrire le domaine d'arguments des prédicats verbaux. Cette approche vise à établir une véritable approche sémantique via une méthode de classification par opérateurs syntaxiques. D. Le Pesant (Le Pesant 1994) en donne une illustration sur les compléments du verbe `"lire"`. Cette étude porte sur 1500 noms environ, classés en fonction de leurs opérateurs appropriés. Ce classement formel constitue aussi un classement sémantique qui distingue les "textes" des "supports de texte" et établit des sous- classes sémantiques homogènes.

(1)
N0+h lire que P
N0+h lire du N1 nph
N0+h lire N1(-/+)h

N0+h : groupe nominal sujet humain
que P : proposition complétive
N1 : groupe nominal complément d'objet
nph : nom propre humain
(-/+)h : humain ou non humain

```
"Je suis en train de lire que la critique ne considère
pas Pynchon comme un romancier américain"
"Je suis en train de lire du Pynchon"
"Je suis en train de lire un romancier américain"
```

(2) **N0+h lire N1-h**

"Je suis en train de lire un poème"

N0+h V N1-h (dans + sur) N2 inc

(3) V : déchiffrer, décrypter, lire, relire
Restriction : (dans + sur) N2 est une détermination
non de N0 mais de N1
inc : inanimé concret

"Je lis un poème sur un radiateur"

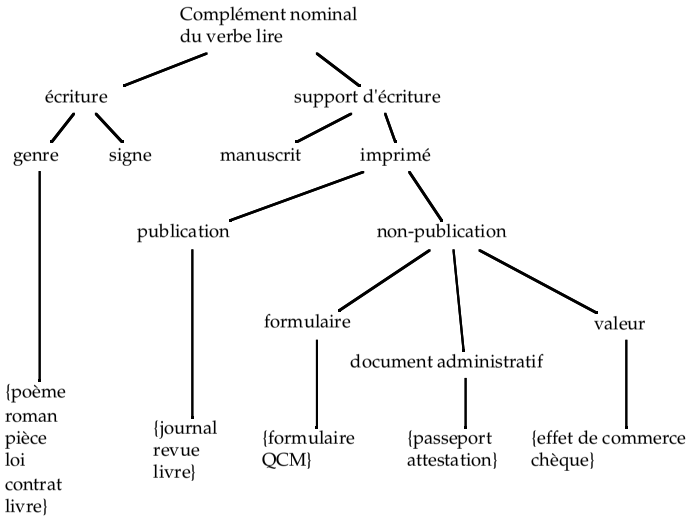

Figure 2.2 : Classement des compléments nominaux du verbe lire.

Ce type d'approche pose des problèmes cruciaux pour la représentation des connaissances des faits de langue. On décrit un certain état de la langue, à un moment donné. Et le classement produit vise à stabiliser une somme de savoirs en construisant une hiérarchie qui subdivisionne à l'infini les savoirs classés pour prendre en compte leur extrême finesse. Ce classement doit aussi traduire le fait qu'un mot peut appartenir à plusieurs classes :

livre : `<support de l'écriture> <texte>`

Le point important à souligner est qu'on aboutit à une représentation hiérarchisée qui tend à "clore" ou à fermer le domaine représenté (aspect fini de la représentation) : on fixe le nombre de classes, et il est difficile de revenir sur cet acquis. Ce type de représentation reste donc en deçà des possibles de réalisation d'une entité donnée. Puisque les faits de langue sont soumis à des évolutions permanentes, leur représentation se trouve en permanence remise en question. Il faut pouvoir les ajuster en tenant compte des incidences potentielles que les conditions de leur réalisation révèleront. On peut penser à une mise à jour de la classification par spécialisation. Mais il est souvent insuffisant de créer des sous-types pour traduire les extensions de sens potentielles sur les mots. A contrario, créer des sur-types peut conduire à des conflits d'héritage. Mais surtout, il reste délicat voire impossible de changer dynamiquement un

graphe d'héritage construit. Surtout quand on sait que généralement, les représentations des savoirs linguistiques peuvent construire des hiérarchies de savoirs parallèles qui articulent les différents types de savoirs à l'oeuvre dans le langage naturel.

De plus un examen de réalisations textuelles révèle le plus souvent que les comportements des mots sont soumis à des variations quand on passe d'un corpus à un autre.

"Two findings repeatedly come out of corpus-based studies : first, individual linguistic features are distributed differently across registers, and second, the same (or simular) linguistic features can have different functions in different register" (Biber 1993).

En effet, le travail sur les sous-langages révèle que les classements généraux que l'on peut établir sur la langue ne sont pas toujours pertinents. L'utilisation des mots dans des domaines de spécialités particuliers peut induire des comportements linguistiques nouveaux qui mettent justement en lumière un nouveau type de rapport avec l'objet visé dans ce domaine particulier (Cadiot & Nemo 1997a).

"Corpus-based research shows that our intuitions[14] *about lexicals patterns for grammatical structures, for many words there is no general pattern of use that holds across the whole language; rather, different word senses and collocational patterns are strongly preferred in different registers"* (Biber 1993).

Si notre travail s'éloigne sur ce point de l'approche suivie par Gross, il en retient la reprise de l'approche harissienne (Daladier 1990, Dachelet 1994, Sager & al. 1987, Habert & al. 1996) : en fait la détermination de classes d'opérateurs et d'opérandes par le fonctionnement linguistique. Il ne s'agit donc pas d'une sémantique ontologique i.e. ayant des existants hors de toute réalisation linguistique.

2.3.2.4. Types et distributions (Godard & Jayez 1995)

(Godard & Jayez 1995) propose le multitypage des entrées lexicales (livre : `objet matériel & objet informationnel`) pour prendre en compte les différentes réalisations sémantiques d'une entrée lexicale. Ce typage (structure de type simple ou complexe) est réalisé dans une approche linguistique avec le dessein d'obtenir des classes en nombre limité (à assez gros grain).

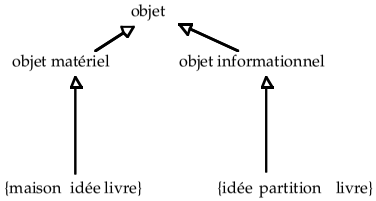

Figure 2.3 : Multitypage des entrées lexicales.

Là encore, on retrouvre le problème d'une approche de représentation qui tend à figer les savoirs et qui limite leurs évolutions potentielles. Si le type d'un mot préexiste aux occurrences de ce mot dans des textes, on limite l'interprétation de ces occurrences au types prédéfinis.

14 *"The collocational patterns (confirmed by concordance listings) identify two surprising fact about (these) words : 1) 'certain' is commonly used to mark uncertainty rather than certainty; and, 2) certainty is rarely expressed in social science at all. Thus, the most common collocations for 'certain' in social science reflect a kind of vagueness, marking a referent as possibly being known (by someone) but not specified in the text (e.g., a certain kind of..., in certain cases..., there certain indications that...). These collocations are relatively rare in fiction. In contrast, those collocational patterns for 'certain' that directly state that someone or something is certain - he/she/they/you/it + BE + certain and i/we + BE + certain - are extremely rare in social science but relatively common in fiction"* (Biber 1993).

2.3.3. Polysémie et changements de sens

2.3.3.1. Changements réguliers de sens (CRS) et hiérarchies

Multitypage et choix en contexte : (Godard & Jayez 1995)

Le multitypage permet un regard en avant sur des possibles que les énoncés en situation doivent ajuster comme dans l'exemple suivant dans lequel le mot `livre` est associé à deux types différents :

Paul écrit un livre qui pèse déjà un kilo et qui fera du bruit.

livre : `objet matériel & objet informationnel`

On est donc confronté ici à une difficulté liée à la mise en place de hiérarchies rigides qu'il est souvent difficile de faire évoluer dans la mesure où le multitypage s'accommode d'un graphe des types figés.

Les règles lexicales (Briscoe & Corperstake 1993)

Les règles lexicales peuvent être vues comme des processus de conversion : elles sont utilisées pour créer les sens dérivés. Elles permettent ainsi de générer automatiquement des alternatives possibles de sens lorsque la résolution de contraintes l'exige. L'utilisation de règles lexicales repose donc de manière différente le problème de la fermeture de la représentation du domaine étudié. Si une approche hiérarchique tend à "clore" la représentation du domaine et à limiter les évolutions potentielles des savoirs ainsi décrits, les règles lexicales permettent d'agir sur ce type de représentation statique en modifiant temporairement les valeurs attachées à tel noeud de la hiérarchie (quand cela est nécessaire). Ce type de clôture reste malgré tout plus proche d'un artifice calculatoire. Si la volonté initiale de ne pas lister tous les sens d'un mot se traduit par la mise en place d'une liste de règles lexicales (que l'on peut étendre à l'infini), le modèle de représentation ne traduit pas adéquatement ce qui est réellement en jeu.

Coercion de type (Pustejovsky 1991)

Dans l'énoncé qui suit, la coercion permet au prédicat de changer le type de son argument vers celui qui est attendu par le prédicat:

`"Il a commencé le livre"`

`commencer` attend un `événement`, le prédicat ajuste donc le type de livre en fonction des rôles associés à l'argument.

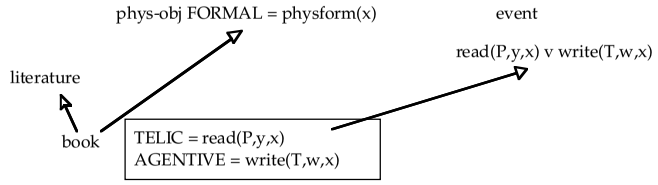

Figure 2.4 : Coercion de type sur `livre`.

On peut aussi penser à la métonymie intégrée de Kleiber. Et là encore, on est confronté à une difficulté liée à la mise en place de hiérarchies rigides qui s'accommodent d'un graphe des types figé.

2.3.3.2. Contraintes

Continuité du passage d'un type à un autre

a. Un noyau de sens pour conduire l'interprétation

1. Du discret au continu

La problématique développée par C. Fuchs met en avant la difficulté à définir les modes d'application des mécanismes qui réalisent la construction du sens d'un énoncé. Comment attribuer des valeurs aux unités linguistiques à représenter, et quelles valeurs retenir? Si l'on fige la représentation sémantique d'une unité de langue, celle-ci peut être modifiée intégralement dans une interprétation donnée. Sachant que l'interprétation peut conduire à une utilisation de valeurs sémantiques éloignées des valeurs habituellement attachées aux entités manipulées. Il convient donc d'exprimer la notion de variation graduelle, c'est-à-dire *"se représenter la signification en termes non plus en termes de valeurs discrètes et disjointes mais en termes d'un continuum de valeurs"* (Fuchs & Le Goffic 1985). La formalisation mise en place pour représenter ce continuum de valeurs se situe dans le cadre d'une modélisation utilisant les mathématiques du continu.

1.1. Modèle mathématique de la polysémie

A toute expression polysémique sont associés deux espaces :

- **un espace des indices co-textuels** : il *"permet de représenter les données susceptibles par leur présence dans l'énoncé d'influer sur l'interprétation de l'expression étudiée"* (Fuchs & Victorri 1990),

- **un espace sémantique** : dans celui-ci, *"les diverses significations de l'expression seront représentées comme des régions plus ou moins vastes"* (Fuchs & Victorri 1990).

De plus, *"le processus d'interprétation est modélisé par une fonction qui associe à chaque point de l'espace des indices une dynamique sur l'espace sémantique"* (Fuchs & Victorri 1990) : à chaque énoncé correspond un relief sur l'espace sémantique et *"les points de basse altitude sur ce relief correspond(e)nt aux significations de l'expression dans l'énoncé considéré"* (Fuchs & Victorri 1990). Dans ce modèle, on associe à chaque énoncé une dynamique sur l'espace sémantique, et la signification de l'expression dans un contexte ne se réduit pas à un ou plusieurs points, mais est matérialisée par des régions (les bassins attracteurs de la dynamique). *"L'espace sémantique associé à une expression polysémique est un espace géométrique continu, structuré de telle manière que des significations proches soient représentées par des régions voisines"* (Fuchs & Victorri 1990).

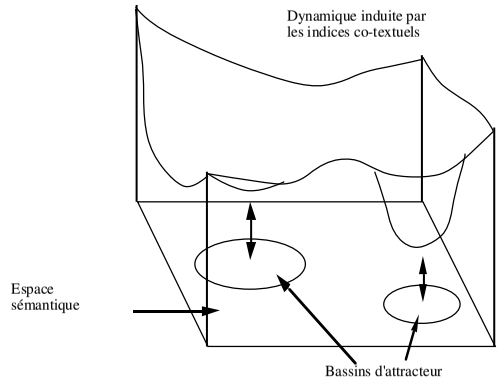

Figure 2.5 : Dynamique associée à un énoncé.

Une ambiguïté est alors représentée par l'existence de deux points d'équilibre stable dont les bassins d'attraction ont des formes traduisant le type d'ambiguïté (alternative, ambivalence, indétermination, etc.).

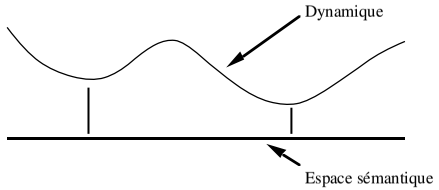

Figure 2.6 : Interprétation avec ambiguïté-alternative.

Un bassin unique correspond à un interprétation univoque :

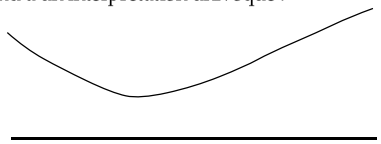

Figure 2.7 : Interprétation univoque.

Un bassin très large et peu profond traduit une signification peu déterminée :

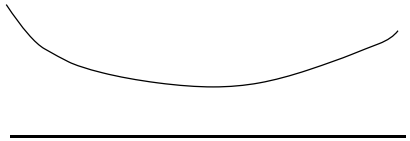

Figure 2.8 : Interprétation avec indétermination.

Ce modèle, implémenté à l'aide de réseaux connexionnistes qui fonctionnent par apprentissage, permet le traitement du sémantisme de certaines unités grammaticales. Le point fondamental dans cette étude est la construction de l'espace associé à l'expression étudiée : *"loin de ne*

constituer qu'un codage arbitraire et sans conséquence, cette construction implique de définir une topologie sur ces espaces qui va conditionner les capacités de discrimination et de généralisation du système" (Fuchs & Victorri 1990).

1.2. Hypothèse méthodologique pour le classement des significations

Pour une expression polysémique, *"on postule l'existence de significations discriminales, et à propos desquelles il est possible d'émettre des jugements de proximité"* (Fuchs & Victorri 1990). Aussi, il s'agit d'établir *"des jugements qualitatifs relatifs assez spontanément mis en oeuvre dans la pratique langagière quand on cherche par exemple à expliciter le sens précis d'une expression de la langue dans un contexte quelque peu inhabituel"* (Fuchs & Victorri 1990).

1.3. Accès aux valeurs sémantiques

Pour cela, on utilise des jugements de recoupement du type :

"La signification de l'expression dans l'énoncé E_2 recoupe la signification que l'on observe pour cette expression dans l'énoncé E_1" (Fuchs & Victorri 1990),

et la variante plus forte :

"La signification de l'expression dans l'énoncé E_2 inclut la signification que l'on observe pour cette expression dans l'énoncé E_1" (Fuchs & Victorri 1990).

Ce type de jugement implique *"que toute signification possède une certaine extensionnalité dans l'espace sémantique : les significations les plus indéterminées seront celles de plus grande extension, et inversement les plus précises correspondront à de très petites régions de l'espace sémantique"* (Fuchs & Victorri 1990). La notion de valeur peut être définie comme étant *"une signification d'extension minimale (c'est-à-dire telle que toute signification qui la recoupe la contienne entièrement)"* (Fuchs & Victorri 1990). A la différence des approches componentielles, ce modèle pose *"l'existence de ce niveau de valeurs, qui ne sont pas des éléments de sens (sèmes, sémèmes, traits, etc.), mais des sens purs, d'extension minimale ou plutôt d'extension nulle, obtenus en réduisant progressivement l'extension des sens attestés jusqu'à zéro par passage à la limite"* (Fuchs & Victorri 1990).

2. Le noyau de sens

Le noyau de sens *"ne serait pas à proprement parler une valeur, mais une description minimale valable"* (Fuchs & Victorri 1990) pour l'entrée lexicale considérée, *"à partir de laquelle on pourrait définir des dimensions canoniques qui caractérisent chacune des valeurs"* . Le noyau de sens doit permettre *"de choisir un bon repère de l'espace sémantique associée à l'expression"* (Fuchs & Victorri 1990). Il est "*une abstraction sous-déterminée, que l'on peut se représenter comme un opérateur abstrait : ce qui est défini en effet c'est l'opération, unique, qui doit s'effectuer chaque fois que l'on se donne des instanciations concrètes"* (Fuchs & Victorri 1990) des paramètres décrivant ce noyau de sens. A chaque expression polysémique, par exemple, est associé un noyau de sens sous-déterminé et des directions dans lesquelles peuvent se déployer des significations plus déterminées en contexte. Si l'analyse n'exige pas de précisions, on en reste au noyau de sens. Dans le cas contraire, *"le calcul de significations plus déterminées fait appel à des règles qui selon la complexité de l'expression, sont soit des règles symboliques, soit des règles implicites apprises par un réseau connexionniste à partir d'exemples"* (Fuchs & Victorri 1990).

3. Définir un noyau de sens comme un opérateur abstrait auquel on peut faire correspondre une opération cognitive

Le noyau de sens ouvre la voie à une représentation cognitive : il peut être vu comme *"un schéma conceptuel dont les ingrédients (un domaine, le parcours d'une trajectoire dans ce domaine, une proposition à valeurs de vérités définie sur ce domaine) peuvent se retrouver dans de nombreuses opérations de pensée"* (Fuchs & Victorri 1990).

b. Interpréter, c'est opérer une confrontation frontale de tous les indices sémantiques présents dans un énoncé

En plus de la construction d'un espace sémantique (décrit par le noyau de sens), il faut définir l'espace des indices co-textuels qui influent sur l'interprétation. Puis évaluer les relations entre les présences/absences de ces indices sur le cas de figure interprétatif visé.

- De la tension interaction vs calcul ordonné (hiérarchisé)

"Ce qui est en question, c'est la possibilité d'un calcul hiérarchiquement organisé, où la valeur sémantique de chaque composant ne dépend que de la valeur de ceux qui, du fait de leur position dans l'arbre, ont été calculés avant lui. L'ordre de prise en compte des constituants dépend des formes présentes dans l'énoncé et doit donc être calculé d'abord lui-même avant d'être utilisé. Cela enlève déjà en grande partie l'intérêt de mener un calcul hiérarchique. Et ce n'est pas tout. Il nous faut prendre en compte un autre phénomène qui rend encore plus problématique l'application d'un tel calcul : les indices présents dans l'énoncé interviennent avec une certaine "force" en faveur de telle ou telle valeur et ces forces peuvent se combiner de différentes manières. En particulier, plusieurs indices faibles peuvent concourir à imposer une valeur, malgré la présence d'un indice "fort" en faveur de la valeur opposée" (Fuchs & al. 1991).

C'est dans son processus de création, inscrit dans l'espace et le temps de son énonciation, que le sens d'un énoncé se dévoile : il y a obligation de réaliser un calcul dynamique du sens en contexte.

"On est alors conduit à une vision plus globale du problème : tous les indices contribuent à construire la signification de l'énoncé pris comme un tout, et c'est en fonction de ce tout ce que l'on peut assigner à chaque constituant une valeur sémantique. Le calcul des valeurs des différentes caractéristiques sémantiques doit donc se mener "de front", à partir de toutes les formes présentes dans l'énoncé. /.../. La représentation du calcul sémantique doit prendre en compte :

- *les relations exprimant l'influence des formes présentes dans l'énoncé sur la valeur des différentes caractéristiques sémantiques;*
- *les contraintes (plus ou moins fortes) entre les caractéristiques sémantiques"* (Fuchs & al 1991).

Modèle discret vs. continu : notre problème est ailleurs

La notion de continuité pour décrire les processus sémantiques n'est pas une nécessité absolue : (Kayser 1994) montre qu'il est possible avec des modèles discrets de représenter une continuité conceptuelle. Mais l'utilisation de la continuité pour la modélisation des processus sémantiques (Fuchs & Victorri 1994) met surtout l'accent sur les difficultés rencontrées pour mettre en oeuvre (dans un dispositif) les phénomènes linguistiques ainsi décrits. En effet le point important soulevé par ce type d'approche est clairement énoncé dans (Victorri 1994) : *"/./ representing continuity on a machine is all but a simple problem"*, et cette question ne se

laisse pas résoudre facilement. La mise en oeuvre d'un dispositif informatique pour représenter ce type de phénomène rejoint d'ailleurs un point important de notre travail : la notion de contrôle et la méta-représentation (-> partie 3 chapitre 9). Si le connexionnisme par exemple est capable de rendre compte des phénomènes de la continuité, *"it has a drawback that prevents it from playing in continuous modeling the same role as classical IA tools play in discrete modeling. This flaw is related to an important notion developped in IA : the notion of control. A connectionist network remains 'a black box' which does not allow much reasoning about its functioning"* (Victorri 1994). Or, notre travail vise aussi à mettre en avant la nécessité de disposer de processus de contrôle pour le TALN. Si les descriptions/processus linguistiques à représenter (de manière discrète ou continue) sont potentiellement évolutifs, il est impératif de pouvoir évaluer les évolutions réalisées et donc de disposer d'outils de contrôle.

2.3.3.3. Statut des CRS

Les extensions de sens, dans quelle mesure relèvent-elles :

- d'une hiérarchie de concepts (Kayser 1992), (Pustejovsky 1991),
- de mécanismes pragmatiques très généraux (Kleiber & al. 1991).

Les entrées lexicales inversement se réduisent-elles à l'addition d'informations héritées d'une hiérarchie conceptuelle ?

Mécanismes pragmatiques très généraux (Kleiber & al. 1991)

- Le principe de métonymie intégrée

Certaines caractéristiques de certaines parties peuvent caractériser le tout.

`"Paul a le dos bronzé"`

caractérise le tout (Paul) au moyen d'une propriété assignée à l'une de ses parties.

Ce principe permet ainsi d'assimiler la commercialisation à une "partie du livre" dans la mesure où on peut considérer d'une part que la commercialisation fait partie du concept `"livre"` : un livre est un objet destiné par définition à la commercialisation. D'autre part, on peut dire qu'un livre est destiné à être lu par le plus grand nombre possible de lecteurs : là aussi, la commercialisation et la publication constituent à cet effet des facteurs importants. Le principe de métonymie exige en fait plus. Il ne suffit pas que la partie soit jugée suffisamment importante pour le tout, il faut encore que la caractéristique attribuée à cette partie soit caractéristique du tout. S'il n'en va pas ainsi, l'attribution au référent entier risque d'être bancale. Ainsi le prédicat `"être un fiasco pour son éditeur"` doit-il être pertinent non seulement pour la commercialisation, mais aussi pour le livre : il faut qu'il dise quelque chose en plus sur le livre en question. En disant qu'un livre a connu un fiasco éditorial, on n'indique pas seulement que la commercialisation ou publication du livre a échoué, mais on révèle en même temps indirectement quelque chose sur le livre lui-même : l'échec commercial est en même temps l'échec du livre en tant que livre.

Liés à une hiérarchie de concepts (Kayser 1992), (Pustejovsky 1991)

La profondeur variable de Kayser, comme le lexique génératif de Pustejovsky, suppose une hiérarchie de concepts aux noeuds de laquelle sont attachées des connaissances, des rôles qui

permettent des inférences. L'interprétation de `moulin à paroles` et des expressions similaires peut relever des mécanismes d'extension de sens à la Kayser, ou encore des coercions de type à la Pustejovsky. On a même intérêt à l'expliquer ainsi, dans la mesure où il faut rendre compte du sentiment qu'il s'agit d'un sens "figuré". Par exemple pour `moulin`, on peut associer le rôle `partie-de` à `objet-concret`, `énergie` à `instrument` et `entrée/sortie` à `instrument de transformation`. Cela structure ainsi le savoir sur le monde effectivement mis en oeuvre. Cette hiérarchie est le support de l'inférence. Le savoir sur le monde est structuré, et cette organisation guide le travail inférentiel (d'où l'appellation paradoxale de Pustejovsky "héritage projectif", où la création dynamique s'appuie sur une structure statique). Pour suivre Kayser, on aboutit à des représentations sémantiques très abstraites, très peu déterminées (le prédicat pour `moulin`), qui sont précisées si besoin est. Et qui sont très éloignées des référents de même nom. Les idées de coercion, ou de dérivation de sens, invalident pour partie le filtrage "statique", aussi fin soit-il, puisqu'aussi bien l'élément en position d'argument précisément ne présente pas les traits nécessaires et que ceux-ci doivent lui être rajoutés, au besoin en supprimant des traits présents et contradictoires (`moulin` devient `+humain` dans `moulin à sottises`).

2.3.4. Evolution des classifications

2.3.4.1. Hiérarchies rigides

S'il apparaît très clairement qu'un choix de représentation qui manipule la notion d'objet dans une taxonomie de classes est adéquate pour fixer un domaine de savoir établi dans une structure hiérarchisée[15], il s'avère parfois délicat de construire puis de prendre en compte les évolutions de telles représentations hiérarchiques quand celles-ci décrivent des domaines de savoir non stabilisés. Dans une telle approche, la mise en contexte des instances de classe peut révéler la nécessité d'une mise à jour de la structure initialement adoptée pour la définition des classes. Dans le cas de la construction d'une hiérarchie lexicale dans laquelle on souhaite inscrire dans la définition des entrées lexicales une articulation forme-sens, une telle représentation doit affronter le problème complexe de la confrontation d'un savoir établi et d'un savoir à venir (non pris en compte par le précédent). La prise en compte de nouvelles informations, de résultats issus de l'analyse, de configurations énonciatives particulières..., peut conduire à modifier, à affiner la représentation initiale. Dans le cadre de notre travail sur la modélisation des configurations nominales du type `moulin à N2`, une telle évolution sur le savoir initial est à l'oeuvre sur, par exemple, des formes du type `moulin à paroles` pour laquelle `moulin` devient `+humain`. Et c'est justement la travail de mise à jour d'une hiérarchie qui pose problème. Plus on structure un domaine de savoir, plus les liens de dépendance entre les noeuds du graphe d'héritage se complexifient, et il devient ainsi délicat de modifier cette structure hiérarchique.

2.3.4.2. Changer la hiérarchie

Vouloir classer les choses de manière figée pose donc des problèmes diffilement compatibles avec une nécessaire évolution dans les représentations construites des savoirs linguistiques à un moment donné. On sait bien que les classifications évoluent : la connaissance du monde bouge, le classement des mots change, un même domaine est aussi sujet à des classifications distinctes. Aussi, on pourrait vouloir changer les hiérarchies construites en spécialisant (créer des sous-types) ou en généralisant (créer des sur-types : le livre électronique comme objet matériel informatique (avec fichier, programme...), mais quel sur-type doit on considérer pour `"un livre est un livre"` ?). Il reste délicat voire impossible de changer dynamiquement un graphe d'héritage construit (comment remplacer (dynamiquement) un graphe de types par un autre?). Surtout quand on sait que généralement, les représentations des savoirs linguistiques

[15] Le multitypage ou la coercion de type ou les règles lexicales s'accommodent d'un graphe des types figé.

peuvent construire des hiérarchies de savoirs parallèles qui articulent les différents types de savoirs à l'oeuvre dans le langage naturel.

2.4. Conclusion

Notre travail s'inscrit fondamentalement dans la lignée des approches présentées supra, même si nous nous sommes montré critique sur certains aspects de ces modèles. Ces approches visent toutes à concevoir la construction du sens comme la mise en action de mécanismes. Ces approches s'écartent d'un appariement de termes et de sens. Elles ne consistent pas non plus à recourir uniquement à de l'induction ou à du raisonnement. Il s'agit plutôt d'établir un compromis entre des informations figées et des heuristiques qui peuvent enrichir ces informations. Ce qui est en jeu, c'est la confrontation entre la représentation de savoirs minimaux et les mécanismes qui permettent d'étendre ces savoirs en fonction des réalités contextuelles. Il s'agit pour certains d'organiser les savoirs de manière hiérarchique de telle sorte que des processus attachés à ce type de représentation permettent de reconstruire le sens; pour les autres, il s'agit au contraire de décrire plus précisément les mécanismes qui réalisent la construction du sens. Le problème fondamental pointé ici tourne donc autour de la difficulté à représenter les savoirs à l'oeuvre dans les descriptions linguistiques dans la mesure où ces derniers ne sont pas donnés une fois pour toutes.

2.4.1. Représenter, c'est figer un peu.

- Comment représenter la base de savoirs initiaux pour tenir compte du fait que l'on a affaire à des informations évolutives et non figées ? Comment enrichir ces savoirs établis quand de nouveaux types de descriptions linguistiques révèlent une certaine stabilité ?

Si les descriptions linguistiques se bornent à refléter un savoir partagé sur le monde, en particulier si le lexique est organisé (de manière hiérarchique ou non) sur la base de connaissances (le monde objectif), on perd une grande partie des mécanismes à l'oeuvre dans le langage naturel. L'interprétation permet en effet la mise en oeuvre de processus sémantiques qui créent de nouvelles acceptations et de nouveaux emplois non dérivables de la connaissance encyclopédique. Dans ce cas les parcours (arborescents ou non) n'apprennent rien que l'on ne sache déjà (Rastier 1995). Figer les savoirs linguistiques dans des taxonomies figées et non évolutives ne convient guère à la structure sémantique des langues et assez mal à celle des discours de spécialité (Biber 1993) : il n'est pas certain que les différents niveaux de l'arbre appartiennent à une même pratique (Rastier 1995). Un travail de représentation des faits de langue doit donc s'attacher à rendre les savoirs de base évolutifs en sachant qu'un ajustement est toujours potentiellement à venir sur ces savoirs.

2.4.2. Ajuster les processus d'ajustements : approches de la réflexivité

L'utilisation d'heuristiques pour étendre les savoirs initiaux représentés dans des configurations particulières pose quant à elle au moins deux types de problèmes.

- Comment représenter le fonctionnement des heuristiques pour rendre compte à la fois de leurs évolutions potentielles et de leurs applications sélectives (éviter de surgénérer) ?

Un dispositif pour le TALN est contraint d'enrichir ou affiner les processus inductifs mis en place (il en va de même pour les savoirs représentés). Dans la mesure où il est impossible de lister et de prévoir tous les scénarios possibles à attacher aux mots, il est nécessaire de pouvoir ajuster les processus définis au cas où telle heuristique doit justement être remodelée. On est donc confronté au problème fondamental qui consiste à travailler sur un nombre important de

connaissances interconnectées, toujours en mouvement, remodelées par des processus évolutifs et capables de s'enrichir.