Chapitre 1

Enjeux linguistiques et informatiques du traitement automatique de la construction du sens

1.1. Comment représenter la mouvance ?

1.1.1. Les noms composés : un micro-phénomène pour l'étude de la construction du sens

> *"Il est illusoire de songer à un traitement automatique avant de disposer d'un taux raisonnable de couverture des structures composées."*
> (Gross 1988)

> *"Les noms composés sont répartis sur un continuum qui va de l'opacité totale à la compositionalité limpide. La question est alors de répartir de manière cohérente les séquences considérées entre les deux pôles. L'autre difficulté étant, pour les séquences analysées de manière compositionnelle, de faire la part dans leur interprétation entre le savoir proprement linguistique et les connaissances pragmatiques."*
> (Habert & Jacquemin 1993)

1.1.1.1. Préambule

Il s'agit ici d'introduire[1] ce qui sera notre premier point d'ancrage linguistique, à savoir les problèmes de reconnaissance et de la construction du sens de certaines structures nominales au travers des modèles de représentation que l'on met en place pour le TALN. L'étude de la composition nominale est significative de la complexité du travail à accomplir pour modéliser les phénomènes qu'elle met en jeu dans un traitement automatique. La composition nominale

[1] Il ne s'agit pas ici de proposer une étude détaillée sur les problèmes des composés (le travail de (Habert & Jacquemin 1993) propose une réflexion complète sur les enjeux de cette étude).

illustre de manière fondamentale la difficulté à intégrer dans une modélisation les aspects lexicaux, syntaxiques, sémantiques, idiosyncratiques ou pragmatiques qui sont à l'oeuvre dans cette opération. Si cette opération met en jeu des mécanismes linguistiques réguliers pour la construction de la forme et du sens, elle s'inscrit aussi dans un cadre plus large qui nécessite la prise en compte de savoirs pragmatiques dans les raisonnements qui complètent ou affinent la représentation linguistique.

1.1.1.2. Une première tâche : étudier les formes nominales `moulin à N2`

Notre travail vise à repérer puis à classer les régularités qui se dessinent dans la construction du sens des configurations nominales du type `moulin à N2`. Il s'agira ensuite d'évaluer les affinements qu'il est possible d'apporter aux modèles de représentations adoptés. Outre les problèmes de classification que posent ces configurations nominales, une réflexion sur la construction du sens de ces configurations met en lumière la complexité des interactions entre les savoirs qui participent à cette construction. La formalisation puis la représentation "matérielle" de ces phénomènes entraîne une réflexion générale sur la mise en place des différents types de savoirs qui permette d'organiser au mieux les dépendances que ces savoirs établissent dans l'interprétation des faits particuliers de langue. Il s'agit donc d'étudier les problèmes de représentation de plusieurs de domaines de savoir qui participent ensemble à la résolution du codage ou du décodage de certains faits de langue. Et parallèlement, il faut analyser les mécanismes qui réalisent cette construction du sens en confrontant ces domaines de savoir.

1.1.1.3. Examen des évolutions de `moulin` sur corpus

Le corpus que nous avons utilisé pour l'étude des configurations nominales en `moulin` a été constitué en exécutant une requête sur `moulin` dans FRANTEXT. Les numéros donnés sont les références à ce corpus (-> annexe corpus `moulin` : extraits du corpus constitué). Dans ce corpus, on trouve `moulin` dans des configurations avec ou sans structure argumentale. Examinons tout d'abord les configurations avec structure argumentale.

- `Moulin à Vinf`[2]

```
moulin à scier: 71
moulin à frapper (les monnaies): 135
moulin à s'agiter (sans résultats):104
moulin à broyer: 167
moulin à battre: 227
moulin à distribuer (les cartes à jouer): 284
```

- `Moulin à N2`

Emplois non métaphoriques

```
moulin à vent: 1, 3, 4, 5, 6, 7,  9, 14, 15, 21, 22, 23, 33, 34, 35,
40, 42, 46, 47, 49, 50, 51, 54, 55, 58, 59, 60, 65, 68, 70, 91, 92,
95, 96, 97, 98, 103, 105, 106, 107, 113, 114, 121, 124, 128, 133, 139,
144, 145, 146, 149, 150, 153, 154, 155, 156, 157, 159, 160, 161, 163,
171, 172, 174, 175, 178, 179, 181, 184, 186, 188, 190, 197, 198, 199,
200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 213, 218,
221, 226, 228 (cité par Saussure dans le CLG!), 232, 233,  238, 239,
246, 247, 248, 251, 252, 254, 259, 260, 262, 263, 265, 266, 269, 270,
```

[2] La rareté des `Vinf` serait-elle un indice de la nature prédicative de "moulin" ?

271, 272, 273, 275, 277, 278, 283, 289, 291, 296, 298, 304, 307, 309, 316, 318, 322, 326, 331, 332, 333, 338, 339, 345
moulin à tan: 2, 118, 119, 120, 191
moulin à papier:8, 10, 12 (?), 13, 30, 32, 39, 43, 45, 212, 250, 286
moulin à bled: 28, 31, 32, 72, 77, 88, 90
moulin à blé:69, 93, 116, 191, 315
moulin à l'huile: 29 (archaisme ?)
moulin à huile:32, 90, 258, 287
moulin à eau: 35, 38, 42, 47, 48, 65, 91, 94, 112, 114, 122, 123, 126, 129, 130, 131, 132, 134, 136, 138, 141, 154, 183, 185, 187, 229, 231, 261, 308, 312, 313, 314, 322, 323
moulin à bras: 36, 37, 41, 61, 62, 63, 64, 101, 140, 169, 268, 297, 305, 334
moulin à poudre: 44, 45, 52, 66, 143, 170
moulin à sucre: 45, 109, 244
moulin à soie: 45
moulin à scie: 52, 72, 73, 74, 75, 76, 78, 80, 82, 83, 84, 85, 86, 87, 90, 115
moulin à coton:53
moulin à farine: 56, 79, 89, 240, 241, 242, 321
moulin à écorce : 56
moulin à ventilateur: 81
moulin à charbon: 125
moulin à café:147, 148, 151, 152, 173, 189, 196, 219, 220, 224, 225, , 243, 245, 255, 274, 279, 282, 303, 311, 317, 324, 327, 328, 329, 330, 335, 336, 340, 341, 346, 347, 349, 350
moulin à foulon: 90, 176, 177, 180, 235
moulin à poivre:192, 299, 300, 325, 342
moulin à vapeur: 194, 195, 298
moulin à sel : 214, 344
moulin à lin : 267
moulin à cacao : 295
moulin à marée : 301, 302
moulin à grains : 52
moulin à scies : 90
moulin à prières: 99, 236, 256, 264, 280, 320
moulin à planches : 142
moulin à barres: 276
moulin à légumes:310
moulin à bactéries : 319

Emplois métaphoriques[3]

moulin à paroles: 19, 20, 25, 162, 193, 290
moulin à papier:12 (des copistes?)
moulin à sophismes: 57
moulin à thèmes anglais: 102
moulin à systèmes: 108
moulin à ennui: 165
moulin à café à vent:217
moulin à café à monnaie: 253

[3] *"La métaphore réside en un transfert de dénomination fondé sur la représentation du partage d'une ou de plusieurs propriétés par les occurrences de deux classes référentielles. Une autre facon d'aborder la métaphore, non nécessairement contradictoire avec la première, consiste à poser qu'il s'agit en fait d'une restriction d'intension: le mot x servant à dénommer une categorie R voit son intension réduite à une ou plusieurs propriétés P. Une seconde catégorie R1 pourra être dénommée x si ces occurrences sont vues comme présentant P de façon saillante. Que l'on définisse la métaphore comme fondée sur la représentation d'un partage de propriétés(s) ou comme une identification simulée de deux classes, on laisse entendre que la focale peut être mise indifferemment sur toute propriété. Or il semblerait que certaines soient pointables de façon privilegiée: 1. Nombre de métaphores trouvent leur motivation dans des proprietés expérientielles /./ encore faut-il que la propriété focalisée soit susceptible de se réaliser dans divers particuliers, et que la catégorie-source de la metaphore présente un haut degré de typicité au regard de cette propriété /./. 2. /./ Les propriétés stéréotypiques sont d'autres propriétés privilegiés par la métaphore. La prégnance de la lune dans la vie quotidienne ou dans l'imaginaire des humains expliquerait ainsi cognitivement la multiplicité des classes référentielles dénommables à lunette(s)"* (Corbin & al. 1993).

moulin à images: 257
moulin à coups: 285
moulin à lunettes: 288
moulin à soleil: 292
moulin à galette: 337

Examinons maintenant les configurations sans structure argumentale :

- Le moulin comme moteur, machine, ou instrument

un cheval de moulin à faire un certains nombres de tours:67
les roues des moulins à cinq cents pas au dessous: 168
rendit le moulin à Marguerite:343
le moulin à mi-régime:348
le petit moulin : 353

- Le moulin comme bâtiment

deux moulins à chaque lieue quarrée: 27
fermé à l'un de ses pignons, moulin à l'autre: 137
le garçon de moulin à la balafre: 166
le moulin à qui les arbres faisaient place en s'écartant: 222
pas de moulin à l'horizon: 230
moulins à faible distance: 234
moulin à notre droite: 250
maître de moulins à Manet et à Cézanne:281
moulin à la même place: 349
les ruines d'un moulin : 351
qui traînent d'un moulin à l'autre : 352

- `Moulin` comme nom propre

Moulins ou Moulin: 11, 100, 110, 111, 215, 223

1.1.1.4. `Moulin` : Reconstruire une Représentation Sémantique

Les différentes réalisations du corpus permettent de tracer une première piste de description pour `moulin`. U ne forme initiale de description de `moulin` adoptée peut être la suivante :

`moulin` : nom masculin singulier

Dans le Petit Robert, on trouve associées à `moulin` les définitions suivantes :

> Appareil servant à broyer, à moudre le grain des céréales.
> Par extension, établissement qui utilise cet appareil.
> Bâtiment où ce type de machine est installé.
> Moulin servant à battre, à piler, à extraire le suc par pression.
> Sens figuré : moulin à paroles.
> Moulin à prières : cylindre renfermant des bandes de papier recouvertes de formules sacrées pour acquérir les mérites attachés à la répétition de cette formule.
> Sens familier actuel, désigne un moteur de voiture ou d'avion, "`faire tourner son moulin`","`emballer son moulin`".

- `Moulin` : le moteur avant le moteur

Moulin est donc principalement une machine. Si originellement moulin peut être considéré comme une machine destinée à moudre les céréales, on peut tout de suite remarquer que moulin n'existe seul que pour désigner un moteur, une machine ou un bâtiment où un certain type de moulin est installé ou l'a été.

"Un lecteur ouvre un volume de planches, il apperçoit une machine qui pique sa curiosité : c'est, si l'on veut, un moulin à poudre" (D'Alembert, Disc. Prélim. de l'Enclyclop. 1751)

"Qu'un homme /./ invente une machine, telle qu'un moulin à vent, ces productions de son esprit peuvent en faire un bienfaiteur du monde." (Helvetius, De l'esprit, 1758)

"Le célèbre mathématicien De la Hire ne passait jamais devant un moulin à vent sans ôter son chapeau, par respect, disait-il, pour la mémoire de celui qui l'avait inventé." (Bernardin de St Pierre, Harmonies de la nature, 1814)

"Le moulin à vent, père de tant de machines nouvelles et à venir." (Alain, Système des beaux-arts, 1920)

"Le monastère et le chapelet son universels comme l'arc et le moulin à vent..." (Alain, Propos, 1936 p1080)

Le "moulin" est le prototype même de la machine. Il sert d'ailleurs de référence pour désigner d'autres types de machines :

"Telles sont, dans la période de transition que traverse le langage des images parlantes et sonores, les relations commerciales, intellectuelles et morales qui se sont établies entre nos moulins à images et la collectivité." (Arts et littérature dans la société contemporaine, 1935 p7808)

Son empreinte sur la société évolue en même temps que celle-ci :

"Le moulin à bras nous donnera la société avec le suzerain, le moulin à vapeur la société avec le capitalisme industriel." (Vedel G., Manuel élémentaire de droit constitutionnel 1949 p204)

"La société est donc l'expression d'un certain type de technique : suivant une formule connue, "le moulin à eau a produit la société féodale, la machine à vapeur a donné naissance à la société bourgeoise." (Lesourd-Gérard, Histoire économique 19ème et 20ème siècle, 1968 p164)

Moulin fonctionne aussi comme un opérateur syntaxique qui prend en arguments une préposition et un nom. Une trace de l'emploi originel de moulin est peut-être à l'origine de la prédominance[4] de moulin à blé sur moulin à farine, le produit naturel pouvant potentiellement donner lieu à plusieurs transformations. Dans ce cas moulin traduit l'opération qui consiste à transformer un produit naturel :

un moulin à blé transforme du blé en farine
un moulin à café transforme du café en grains en café en poudre

On trouve aussi des réalisations de moulin à N2 qui traduisent le point de vue inverse. Un moulin est alors le créateur de produit issu d'une transformation :

[4] Dans le corpus donné en annexe, on trouve 7 réalisations en "moulin à bled" et 5 en "moulin à blé" contre sept réalisations en "moulin à farine".

un moulin à farine est un moulin qui crée de la farine

on a aussi :

un moulin à beurre est une machine qui crée du beurre
un moulin à papier est une machine qui crée du papier

Si les moulins à vent n'existent pratiquement plus en tant qu'instrument de production, la séquence "moulin à vent" est à l'origine des nombreux emplois métaphoriques :

"Mais cette dernière, profitant de la mise en accusation, se réforme, emprunte à la sociologie ses méthodes, de sorte que celle-ci ne se bat en réalité que contre des moulins à vent, contre des attitudes de pensées dépassées par la psychologie elle-même." (Vuillemin J., L'être et le travail, 1949 p115)

- Des moulins à paroles et des emplois métaphoriques

"Il se mit à détester cette prose de moulin à vent, avec son éloquence gesticulaire." (Sabatié R., Les chinois d'Afrique, 1966 p253)

Une "prose de moulin à vent" peut donc être assumée par un "moulin à paroles", qui outre la quantité de "bruit" générée, ajoute à cette production une gesticulation physique digne d'un moulin à vent. Dans la série des "moulins à paroles" rencontrés, l'expression désigne tantôt la langue en tant que telle, tantôt la personne bavarde. Le "moulin à vent" et le "moulin à café" sont les principaux fournisseurs de métaphores, l'un peut-être par sa forme caractéristique et par sa diffusion, sans compter l'emprunt net à Don Quichotte, l'autre peut-être comme premier instrument ménager. L'assimilation du mouvement de l'instrument "moulin" à un comportement physique dénotant une gesticulation est souvent de mise:

92 : "Il se moque de ces bras en moulin à vent..."
181 : "On dirait que ces bras sont des ailes de moulin à vent..."
269 : "Un jeune type agité, les bras en moulin à vent..."
283 : "/./ sa coiffe noire, vue de dos, se silhouettait comme un étrange moulin à vent."
338 : "On dirait un moulin à vent quand il se fraie un chemin parmi les infirmiers..."

De même pour le bruit provoqué par l'instrument "moulin" :

327 : "Un moulin à café ronronnait dans l'appartement voisin."
puis
328 : "Le moulin à café s'arrête et mon père ronronne à son tour..."

Un être humain peut aussi être assimilé à un automate qui se comporte comme un cheval de moulin :

67 : "Au nombre des meilleurs sont Laplace, Monge et Cousin, espèces d'automates, habitués à suivre certaines formules, et à les appliquer à l'aveugle, comme un cheval de moulin à faire un certain nombre de tours avant de s'arrêter."

Les emplois métaphoriques permettent d'élargir la prise en compte des N2 de type Nparole. En effet, on rencontre les séquences telles moulin à coups, moulin à systèmes, moulin à thèmes anglais, moulin à images, où N2 est un pluriel sauf pour

"moulins à papier" dans 12; on trouve aussi moulin à ennui où N2 est massif singulier. Dans tous les cas, il y a production[5] massive.

- Moulin et les déverbaux

Certains exemplaires de moulin à N2 ne décrivent pas ce que produit le moulin ni ce qu'il transforme, mais ils représentent un instrument constitué par le moulin. Dans moulin à scie, moulin à foulon, moulin à ventilateur, l'énergie du moulin sert respectivement à scier, à fouler, à ventiler. Ces trois mots sont des déverbaux et ce fait permet de sélectionner une interprétation : instrument dont l'énergie est apportée par le moulin.

La description initiale de moulin et l'examen de ses emplois mettent en évidence les mécanismes évolutifs à l'oeuvre dans les faits de langue. Un traitement automatique de ces phénomènes est contraint par la prise en compte impérative de cette mouvance illustrée sur moulin : **les emplois précités ne peuvent être considérés comme figés et non évolutifs**. Leurs représentations dans un dispositif de traitement automatique doivent donc pouvoir être ajustées voire remodelées pour rendre compte de ces évolutions.

1.1.1.3. De l'étude de certains faits empiriques à la construction d'un dispositif : problèmes de représentation

On perçoit déjà la grande difficulté à identifier dans les représentations à adopter ce qui relève, pour simplifier, du général (le "conceptuel") et du particulier (le "lexical"). Dans quelles mesures, les informations sémantiques doivent-elles être représentées dans une hiérarchie de concepts ou dans une hiérarchie lexicale? Et comment articuler ces environnements d'informations? Comment évaluer les (in)dépendances entre ce qui dû aux informations conceptuelles et ce qui est dû au langage lui-même? Il faut donc envisager la mise en place d'un système de représentation sous forme de couches (de feuilleté) en tenant compte que ces couches mettent en jeu des mécanismes ou tâches qui ne sont pas à évaluer sur un même plan; une activation de ces tâches en parallèle ne semble donc pas une solution cohérente. Là encore, ce qui est en question, c'est l'existence d'une organisation (hiérarchique ou autre) de ces tâches. Ce qui est sûr, par contre, **c'est la nécessité de disposer d'éléments de contrôle qui permettent une gestion adéquate de ces processus**. On retrouve là un des grands problèmes de la représentation des connaissances qui est celui de la définition de connaissances de contrôle pour gérer des telles complexités d'interaction :

"Les connaissances de contrôle sont les connaissances sur le système et sur les représentations des connaissances des domaines et portent sur la signification de ces représentations dans le système. Cela provient de l'écart irréductible entre connaissance et représentation. La représentation extraite de la connaissance reçoit un sens informatique dans le contexte du système. Les connaissances de contrôle précisent ce sens. Il y a deux étapes essentielles dans la représentation des connaissances : représenter les connaissances du domaine dans un formalisme de référence indépendamment du système final, représenter les connaissances de contrôle dans un formalisme tiré du système pour concilier représentations du domaine et système final" (Bachimont 1992).

1.1.2. Moulin à N2 : une approche hiérarchique de la construction du sens (Habert & Fabre 1993)

Le travail de (Habert & Fabre 1993) vise à présenter un modèle de la construction du sens pour l'acquisition de dénominations nouvelles:*"Dans l'ensemble des dénominations complexes, nous nous intéressons aux séquences nouvelles (par opposition aux termes attestés) et interprétables,*

[5] Sauf pour "moulin à soleil" (solution 292 du corpus) dont l'interprétation est d'ailleurs difficile.

potentiellement construites selon des schémas syntaxiques et sémantiques réguliers". Au sein du patron N1 à N2, il s'agit d'examiner les contraintes qui régissent les séquences moulin à N2. Pour cela, leur travail s'appuie sur un outil de représentation hiérarchisée du lexique, via le langage DATR (Gazdar & Evans 1990), et sur des règles à la PATR-II (Shieber 1986) pour montrer la mise en oeuvre de ces contraintes. *"L'examen critique de cette démarche débouche sur la proposition d'utiliser une organisation hiérarchisée des règles de construction du sens comme principe directeur de l'analyse"* (Habert & Fabre 1993).

1.1.2.1. Objectif : un modèle des règles de production de Dénomination Complexe

L'approche suivie s'inscrit dans un cadre de traitement automatique des dénominations complexes non attendues (non présentes dans les bases de connaissances ou les dictionnaires utilisés) pour y acquérir de nouvelles dénominations bien formées. Cette approche repose *"sur le calcul de l'interprétation sémantique de la séquence à partir des sens des composants, et par la mise en évidence de certains schémas de sens productifs"* (Habert & Fabre 1993). Quelles Dénominations Complexes Motivées (c'est-à-dire où le sens global est motivé, dans une proportion plus ou moins grande, par le sens des composants) peuvent être produites? Dans une telle perspective, quelles sont les règles à l'oeuvre?

1.1.2.2. Les règles à l'oeuvre dans le patron moulin à N2

Un premier classement

L'examen des expressions moulin à N2 prend appui sur le travail de (Poncet 1991)[6]. Dans ce dernier, la série des séquences moulin à N2 entre dans les séquences N à Ni désignant des noms d'instruments ou de machines, à l'exception de certains emplois figurés tels que moulin à paroles qui sont regroupés dans une classe résiduelle. La répartition des séquences moulin à N2 est représentée dans la figure qui suit.

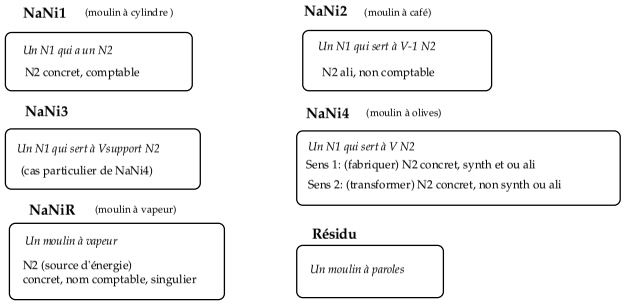

Figure 1.1 : moulin à N2 : un premier classement.

Le principe de classification sémantique présenté ci dessus utilise des classes sémantiques telles que aliment, noté ali, synthétique (pour des référents synthétiques, non naturels), noté synth, permettant de résumer un ensemble de traits élémentaires tels que concret, comptable :

[6] La perspective de (Poncet 1991) est de déterminer les propriétés qui amènent à classer un Na à Nb dans la catégorie Ni.

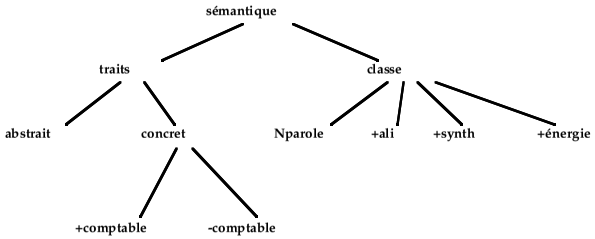

Figure 1.2 : Classification sémantique.

Réexamen de la classe résiduelle pour une analyse unifiante du sémantisme de `moulin`

"Considérer `moulin à paroles` comme figuré, figé, n'est pourtant pas vraiment satisfaisant. Lister de telles séquences et leur associer un sens unique (langue, bavard) ne rend pas compte en effet des possibilités de création sur ce schéma" (Habert & Fabre 1993). On trouve en effet une série de réalisations qui entrent dans la classe résiduelle de la figure précédente : `moulin à sophismes`, `moulin à conversation` (-> annexe corpus moulin). Ces séquences fonctionnent suivant la contrainte suivante : elles attendent un `N2` de type `NParole`, au pluriel ou de sens pluriel, et le `N` résultant est `+humain`. (Habert & Fabre 1993) proposent[7] donc d'associer à `moulin` un prédicat à quatre arguments de la forme suivante[8] :

`PRODUIRE(Entrée,Sortie,Energie,Partie-De)`

Le modèle sémantique proposé permet d'affiner les résultats existants sur la classification des `N1 à N2` (Poncet 1991) en intégrant notamment la classe jusqu'alors résiduelle de `moulin à paroles`, celle-ci s'intégrant au modèle proposé suivant les contraintes sémantiques précisées dans la figure qui suit.

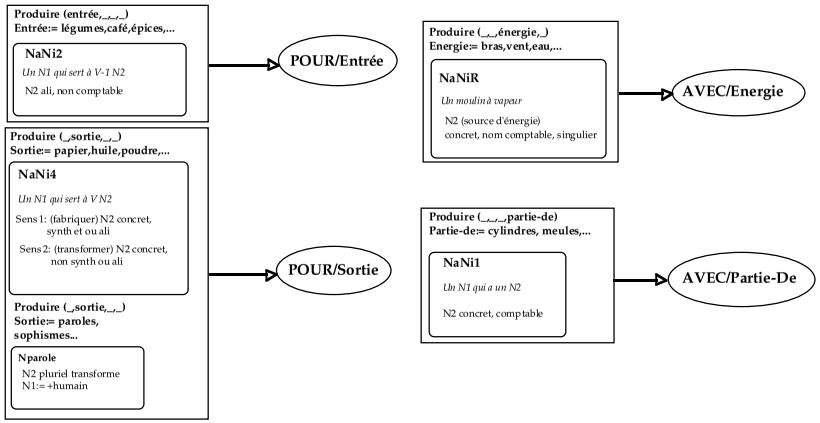

Figure 1.3 : Nouvelle répartition sémantique des `moulin à N2`.

[7] *"La productivité, restreinte, de ce schéma invite à réexaminer son rapport avec le sémantisme de `moulin`, et à donner de celui-ci une analyse unifiante"* (Habert & Fabre 1993).

[8] Pour reprendre ce qui est dit dans la note précédente, l'approche choisie ici vise donc plus à chercher les propriétés qui font de `Nb` un bon candidat pour être apparié à un `Na` dans un `Ni`.

Dans ce modèle, *"les sens dits figurés apparaissent alors comme une spécialisation d'une configuration citée. /./ Un `moulin à paroles` produit des paroles, c'est cette inscription dans l'une des configurations possibles pour `moulin` qui explique la productivité et la transparence relatives de ce patron. L'utilisation d'un N2 pluriel, `NParole` comme deuxième argument, entraîne une transformation régulière de N1 (`+humain`). En revanche le trait `excès` qui apparaît dans ces dénominations relève probablement d'un autre type de mécanismes : par exemple de maximes à la Grice. Si l'on caractérise une personne par la production de paroles, c'est que celle-ci est excessive. A une métaphorisation qui s'appuie sur les caractéristiques fondamentales du prédicat associé à `moulin` s'ajoute un nouveau trait, de manière là encore régulière et reproductible"* (Habert & Fabre 1993). Ce modèle permet aussi de préciser les propositions de (Cadiot 1992). `POUR` se subdivise en `entrée` et `sortie`, `AVEC` en `énergie` et `partie-de`. *"Les relations `POUR` correspondent aux arguments essentiels du prédicat, les arguments `AVEC` aux arguments secondaires"* (Habert & Fabre 1993).

Le traitement proposé utilise *"des règles de réécriture à base hors contexte où chaque élément de la règle est remplacé par un graphe de contraintes sur l'élément de la phrase"* (Habert & Fabre 1993). Une représentation graphique de cette vision unifiante de la construction du sens est donnée dans la figure qui suit.

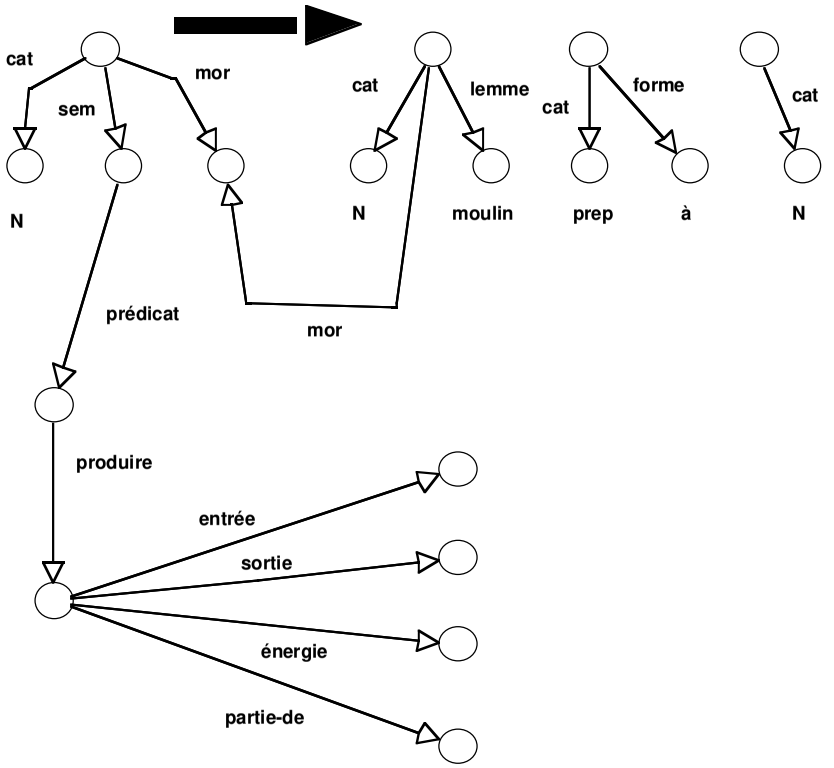

Figure 1.4 : Une vision unifiante de la construction du sens des séquences `moulin à N2`. (1)

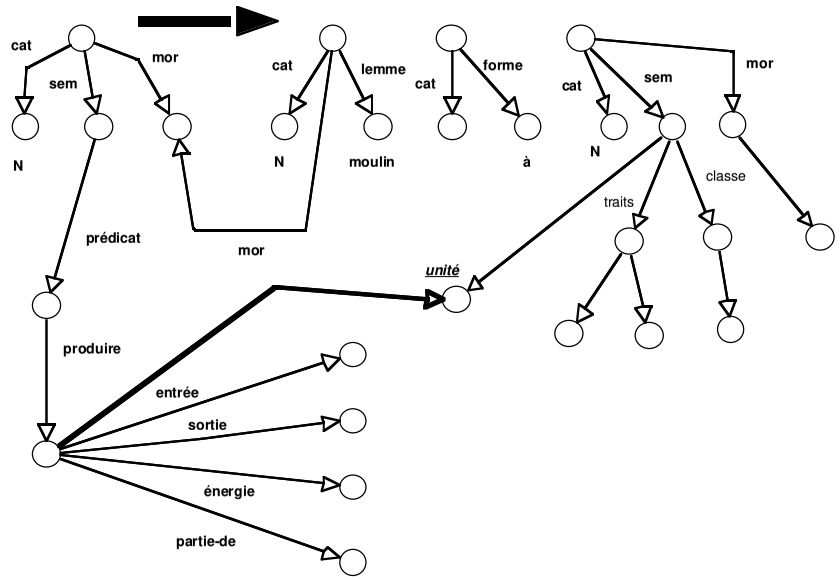

Figure 1.4 : Une vision unifiante de la construction du sens des séquences `moulin à N2`. (2)

Construction du sens des séquences `moulin à N2`

L'approche retenue par (Habert & Fabre 1993) s'inscrit dans un cadre qui privilégie la construction du sens non par juxtaposition et combinaison des sens des entrées lexicales, mais par un affinement d'hypothèses initiales *"quitte à rester dans l'indétermination quand les informations disponibles ne permettent pas d'aller plus avant"* (Habert & Fabre 1993). Le choix d'une telle approche vise à :

- Construire une représentation sémantique très abstraite, très peu déterminée qui est précisée si besoin est;
- Inscrire dans la représentation les scénarios, les processus d'interprétation.

L'analyse à mettre en oeuvre *"est vue comme un affinement progressif d'hypothèses dirigé par la structure hiérarchique des entrées lexicales et par le contexte des têtes lexicales"* (Habert & Fabre 1993). Le point fondamental dans cette approche est de maintenir une continuité entre une sous-détermination initiale et une spécialisation à venir.

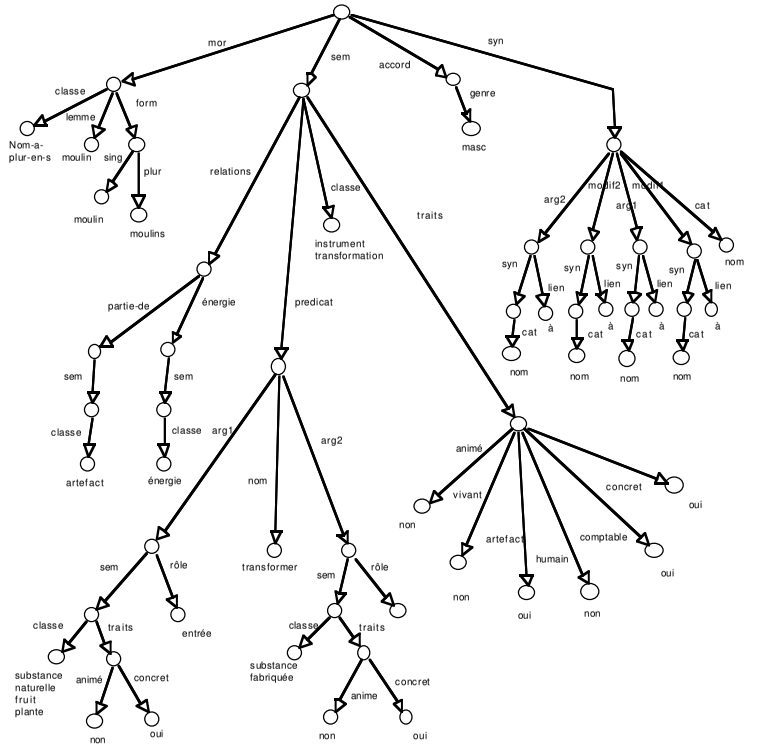

Figure 1.5 : Représentation scénarisée de `moulin`.

Cette approche de la construction du sens basée sur la polysémie et la sous-détermination se conjugue avec une approche de représentations hiérarchisées des connaissances lexicales qui traduit la volonté de limiter les effets d'un traitement homonymique du sens. La représentation lexicale doit prendre en compte la nécessité de disposer de mécanismes exprimant le passage du général au particulier. Il s'agit en particulier de mettre en avant les régularités lexicales qu'il est ensuite possible de préciser pour tel item donné. Pour cela, (Habert & Fabre 1993) utilisent le langage de représentation lexicale DATR (Gazdar & Evans 1990) qui permet un traitement des généralisations et des exceptions basé sur l'héritage multiple et la non-monotonie. C'est aussi une vision où il ne peut y avoir de violation de ce qui est acquis à un niveau supérieur. Il n'est pas sûr que cela ne soit jamais la cas.

Affiner le modèle sémantique proposé

Certains exemplaires de `moulin à N2` ne rentrent pas dans le cadre du prédicat à quatre places retenu : `scie, foulon, ventilateur.` Il s'agit pour ces derniers, non pas de ce que produit le moulin, mais de l'instrument constitué par le moulin : l'énergie du moulin sert à scier, à fouler, à ventiler. Ces trois mots sont des déverbaux et ce fait permet de sélectionner une interprétation : instrument dont l'énergie est apportée par le moulin et non `partie-de`. De plus, l'étude des emplois métaphoriques peut élargir la limitation faite ici au `NParole`. On peut ainsi avoir des séquences `moulin à N2` métaphoriques avec (`N2 pluriel : coups, thèmes, images, sophismes` ou avec (`N2 massif singulier : ennui`), dans les

deux cas, on a une production massive. La présentation du modèle faite précédemment doit donc prendre en compte ces exemplaires de `moulin à N2` qui apportent des précisions au modèle initial.

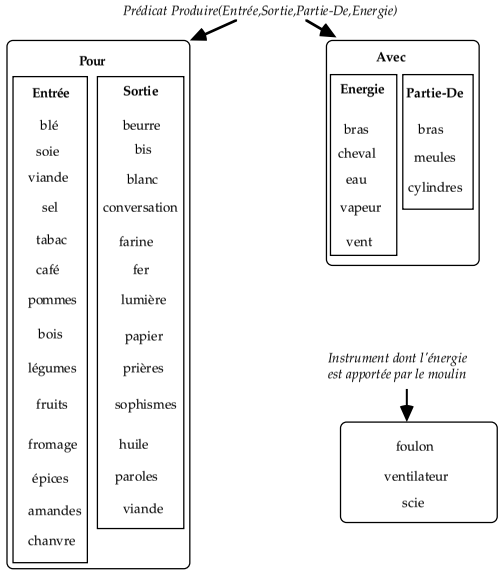

Figure 1.6 : Répartition lexicale dans le modèle sémantique construit.

1.1.3. Changements Réguliers de sens et organisation des connaissances

1.1.3.1. Limites du filtrage des arguments acceptables

Certains exemples mettent en avant les limites d'un simple filtrage des arguments acceptables. Les dérivations de sens invalident ce type de filtrage puisqu'aussi bien l'élément en position d'argument ne présente pas les traits nécessaires et que ceux-ci doivent lui être ajoutés (`moulin` devient `+humain` dans `moulin à paroles`). Il en va de même pour `bras` ou `cheval` dans `moulin à bras` ou `moulin à cheval`. Il n'y a en effet aucun intérêt à adapter la description de `bras` ou de `cheval` en ajoutant un sens `Nénergie`. Il s'agit donc de mettre en place des strates qui permettent :

- un filtrage fin en fonction des scénarios associés au nom tête,
- une extension éventuelle en fonction de rôles hérités,
- une inférence plus grande à partir d'informations conceptuelles.

Il est important de souligner malgré tout qu'il semble difficile de prendre en compte des interprétations qui ne seraient pas préalablement inscrites dans les mécanismes prédéfinis.

1.1.3.2. Structurer les connaissances syntaxiques

Les interprétations possibles dans la cadre de la construction du sens des formes `moulin à N2` sont toutes placées au même plan, alors qu'un poids différent leur est attaché. On aimerait pouvoir dire qu'une interprétation est plus centrale qu'une autre dans la manière même de représenter les connaissances. Les modèles de représentation pour les composés doivent préserver l'ajustement et l'ambivalence dans la construction du sens de ces composés : il faut permettre la construction du sens, non par juxtaposition et combinaison des sens des termes de la structure complexe en cause, mais par l'affinage progressif des hypothèses initiales (qu'il est préférable de laisser floues) quitte à rester incertain quant à l'issue du calcul en cours pour les enrichir quand de nouvelles informations pertinentes le permettent. Il faudrait ainsi pouvoir disposer pour représenter la construction de sens de dénominations complexes de mécanismes qui permettent :

- d'ordonnancer des règles : une règle doit pouvoir n'être appliquée qu'après l'échec des règles précédentes. cela devant s'accompagner d'une plausibilité de l'interprétation;
- de donner aux règles la possibilité de remodeler tout ou partie des entrées lexicales qu'elles prennent en argument (comme cela se passe pour `bras = énergie` dans `moulin à bras`);
- de respecter des mécanismes transversaux (opérations régulières).

En cela, on prend le contre-pied de deux choix habituels dans les formalismes d'unification (-> chapitre 3) : la construction de tous les possibles (qui sont mis sur le même plan) et la non-monotonie. Par ailleurs, c'est une position sinon opposée au lexicalisme, du moins limitant son champ d'action. Cela consiste à dire en effet qu'après chaque passage de règles normales de construction du sens, il y a des règles de récupération qui essaient, par coercion[9] en particulier, de produire tout de même une interprétation. Enfin cela pousse à chercher des modes de structuration du lexique qui fassent la part moins belle à l'homonymie, et qui recherchent un partage réel des informations et non une factorisation statique.

[9] (Pustejowsky 1991).

1.2. Quels savoirs pour les mots ?

Les développements récents des travaux en linguistique manifestent un intérêt grandissant pour la constitution de gros corpus. Il ne s'agit plus d'analyser en profondeur des domaines restreints, mais plutôt d'évaluer quantitativement des performances de systèmes sur des données de plus en plus importantes. Même si les travaux quantitatifs sur des corpus de plus en plus importants permettent d'élargir la couverture des problèmes traités, **les savoirs linguistiques établis sur des corpus ne peuvent être considérés comme figés et non évolutifs**. Le travail d'extraction de savoirs à partir de corpus montre en effet qu'il est difficile de traiter les savoirs linguistiques comme des généralités.

- En dehors de la connaissance précise d'un domaine de spécialité, il est souvent impossible de savoir si une construction est "acceptable" ou non.

A. Daladier (Daladier 1990) indique : *"en biologie, la phrase `des antigènes ont été injectés dans les anticorps d'un lapin` est dénuée de valeur informative. Il ne serait pas suffisant de dire que cette phrase représente une information fausse parce qu'il n'y a aucune façon d'obtenir une signification, même fausse, à partir des combinaisons de mots qui la constituent, pour un biologiste. Dans un sous-langage, une phrase est interprétable si elle est acceptable"*.

- Les savoirs rencontrés sur un corpus donné peuvent remettre en question des résultats généraux établis par ailleurs.

  En médecine coronarienne, `longue` fonctionne comme un adjectif qui marque le degré quand il est utilisé comme modifieur de `occlusion`. C'est le cas dans la séquence nominale `"une occlusion longue"`. Dans d'autres sous-langages, cet adjectif n'est pas typé pour marquer le degré.

- Les savoirs généraux que l'on peut associer aux mots ne sont pas toujours pertinents :*"La plupart du temps, les généralisations globales ne sont absolument pas fiables, parce qu'il n'existe pas de caractérisation générale d'une langue entière; au contraire, il y a des différences considérables entre registres (ou sous-langage). La description complète d'une langue entraîne par conséquent une analyse composite de traits tels qu'ils fonctionnent dans des registres divers"* (McNaught 1993). Il n'est donc pas satisfaisant de se contenter d'un modèle apriorique pour construire une représentation des comportements des unités lexicales.

- On ne connaît pas toutes les caractéristiques[10] des combinaisons des mots sur un sous-langage donné

  `coronaire` et `coronarien` partagent des contextes[11] :

  ```
  coronaire {artere ~, chirurgie ~, heredite ~, lesion ~, maladie ~,
  reseau ~, spasme ~}, coronarien{artere ~, chirurgie ~, heredite ~,
  lesion ~, maladie ~, reseau ~, spasme ~}.
  ```

  Mais les modes de combinaison de ces deux mots sont divergents :

[10] *"En ce qui concerne la variabilité liée au domaine, J.McNaught (1993) met l'accent sur ce qu'il perçoit comme une des caractéristiques les plus marquées du fonctionnement lexico-syntaxique des sous-langages: des verbes qui sont des homographes de verbes en langue générale ont fréquemment dans un sous-langage donné des structures prédicat-arguments complètements différentes de celles qui leur correspondent en langue générale ou dans d'autres sous-langages. Ceci a trait à la relation extrêmement serrée entre syntaxe, sémantique et domaine conceptuel qui caractérise les sous-langages, et qui a l'attrait de ne permettre en général qu'une seule analyse même dans le cas de séquences longues de syntagmes prépositionnels"* (Péry-Woodley 1995).

[11] Les exemples donnés ici sont extraits du corpus MENELAS (Zweigenbaum 1994).

`coronarien` est associé à des adjectifs évaluatifs `{severe, significatif, important}` et `coronaire` ne l'est pas.

- Les comportements lexicaux ne suivent pas des parcours uniformes sur des entités de même catégorie et surtout il est difficile de les prévoir. Les mêmes comportements lexicaux ne se réalisent pas sur tous les mots d'une même catégorie, ils se distribuent sur des sous-familles particulières. De plus ces comportements ne se réalisent pas de manière uniforme sur les différentes familles de mots d'une même catégorie.

  `artère` et `infarctus` entrent dans des relations de localisation qui se réalisent de manières distinctes:

  ```
  artere {coronaire circonflexe diagonal...}, infarctus {anterieur
  inferieur apical}
  ```

- De plus, *"bon nombre des problèmes de traitement des langues sont en fait limités à la langue utilisée dans un domaine particulier (...). La variété de langue utilisée dans un domaine scientifique ou technologique donné est non seulement beaucoup plus restreinte que la langue dans son ensemble, elle est aussi plus systématique sur le plan de la structure comme sur celui du sens"* (Grishman & Kittredge 1986).

Il y a peu de sens à vouloir faire de l'acquisition sémantique en dehors d'un sous-langage

En suivant ce constat, notre travail s'inscrit dans une approche qui vise à ne pas présumer complètement des choses à représenter. Cette démarche met en place un processus de représentation évolutif : les structures de représentation construites devront pouvoir être affinées dès que de nouveaux savoirs seront mis à jour. On peut en effet considérer que les comportements des mots ne sont pas tous prédéfinis mais que ceux-ci "émergent" dans le contexte dans lequel ces mots "agissent". Il n'est donc pas raisonnable de les considérer comme acquis par définition; il s'agit au contraire de les mettre en lumière ainsi que les corrélations multiples qui existent entre les mots dans un flot continu de discours. Notre démarche consiste en quelque sorte à "faire émerger" les comportements des mots puis à les représenter ou à affiner les représentations existantes.

Pour autant il ne s'agit pas pour nous de nier en bloc toute forme de savoirs stabilisés dans ces processus de représentation automatiques des descriptions linguistiques. Les savoirs généraux en langue ne manquent pas. Les dictionnaires constituent d'ailleurs une trace de la fixation sur le moyen terme de ces savoirs. Il ne s'agit donc pas de tout réinventer. Notre travail vise au contraire à prendre appui sur des connaissances très générales (on utilise au départ un cadre de sous catégorisation approximatif qui sera présenté infra) qui sont affinées voire remodelées et changées au gré des observations rencontrées de la même manière que les savoirs linguistiques évoluent (les dictionnaires sont mis à jour périodiquement...). Il s'agit ainsi pour nous de tendre vers un modèle cohérent qui ajuste les descriptions linguistiques dès qu'une certaine stabilité apparait. De même, dès qu'un certain type de comportement se révèle pertinent, il convient d'ajuster les savoirs définis pour en tenir compte.

1.3. Approche d'une grammaire de sous-langage

1.3.1. Notion de sous-langage

Les caractéristiques des sous-langages[12] (Dachelet 1994) - dimensions réduites, systématicité structurelle et sémantique - ont trait au modèle linguistique dans lequel prend la notion de sous-langage : ce modèle a été développé par Z. Harris à partir des années 60. La thèse de R. Dachelet donne une présentation complète de ce modèle.

1.3.2. L'analyse des sous-langages : la démarche harrissienne

> *"In the absence of an external metalanguage, the entities of each language can be identified only if the sounds, markers, or words of which they are composed do not occur randomly in utterances of the language. That is, the entities can be recognized only if not all combinations occur, or are equally probable. This condition is indeed satisfied by languages. A necessary step, then, toward understanding language structure is to distinguish the combinations of elements that occur in the utterances of a language from those that do not : that is, to characterize their departures from randomness."*
> (Harris 1988)

Harris met en oeuvre une conception "distributionnaliste" de la sémantique : le sens des mots[13] se déduit des constructions[14] dans lesquelles ils figurent : *"In fact we find in language that each word has a particular and roughly stable likelihood of occuring as argument, or operator, with a given other word, though there are many cases of uncertainty, disagreement among speakers, and change through time. These roughly stable likelihoods, and especially the selection frequency, conform to and fix the meanings of words "* (Harris 1988). Harris part de la constatation empirique que les éléments ou les suites d'éléments ne se rencontrent pas arbitrairement les uns par rapport aux autres et les arrangements ou environnements expriment une certaine régularité qui dévoile la structure du texte : *"les parties d'une langue ne se rencontrent pas de façon arbitraire les unes par rapport aux autres; chaque élément se rencontre dans certaines positions par rapport à certains autres éléments"* (Harris 1970). La reconstitution des phrases élémentaires d'un corpus de langue de spécialité débouche sur la mise en évidence de classes d'opérateurs[15] et de classes d'arguments, qui sont supposées fortement liées aux notions et aux relations du domaine considéré. La méthode consiste donc à construire des classes définies distributionnellement. On établit ainsi les occurrences et les co-occurrences d'un élément ou la suite d'éléments, puis on établit les équivalences[16]. C'est l'examen des classes d'équivalence dans la relation opérateur-argument[17] qui permet d'obtenir

12 *"Nous utilisons indistinctement sous-languages et sous-langues. Tous deux traduisent le mot anglais 'sublanguage'. C'est un problème conceptuel récurrent que les mots 'langue' et langage' sont tous deux des traductions possibles de l'anglais 'language'"* (Dachelet 1994).

13 *"At what point do words get meanings? One should first note something that may not be immediately obvious, and that is that meanings do not suffice to identify words. They can give a property to words that are already identified, but they don't identify words. Another way of saying this is that, as everybody who has used Roget's Thesaurus knows,* ***there is no usable classification and structure of meaning per se, such that we could assign the words of a given language to an a priori organization of meanings. Meanings over the whole scope of language cannot be arranged independently of the stock of words and their sentencial relations.*** *They can be set up independently only for kinship relations, for numbers, and for some other strictly organized parts of the perceived world"* (Harris 1988).

14 *"The components that go into the making of the structure are the components that go into the making of the information"* (Harris 1988).

15 *"In 'The man asserts that a tree sleeps', the 'asserts' is the operator on the argument pair 'man', 'sleep' (forming 'man asserts sleep'), while 'sleep' is the operator on its argument 'tree'. This is a dependence, , with each word in the sentence requiring particular other words, more exactly any one word of particular classes: 'sleep' requires the classes of 'man, tree'. This dependence is a partial order"* (Harris 1988).

16 *"On dira donc que des éléments (parties du texte, morphèmes ou séquences de morphèmes) sont équivalents entre eux s'ils se présentent dans l'environnement d'autres éléments identiques ou équivalents. Tout ensemble d'éléments équivalents entre eux s'appelle une classe d'équivalence"* (Harris 1969).

17 *"In language as a whole almost every word has a unique subset of operators on it (or arguments under it) in respect to which it has 'selectional' (higher than average) frequency. Over its whole range of operators (or arguments) this selectional subset characterizes the central meaning of the word. In respect to each operator many different arguments have different likelihoods"* (Harris 1988).

et les classes de mots et leurs relations de co-occurrence. Cet examen se révèle mettre à jour un petit nombre de phrases-type constituées de combinaisons spécifiques de sous-classes de mots.

Propriétés des sous-langages

- En général, le lexique est restreint.
- On y trouve peu de types de phrases différents. Chaque type est caractérisable comme une combinaison particulière de sous-classes de mots.

"Le trait distinctif d'un sous-langage est que pour certains sous-ensembles de phrases du langage, le phénomène de sélection, pour lequel on ne peut formuler de règles pour le langage dans son ensemble, fait partie de la définition de la grammaire. Dans un sous-langage, les classes de mots correspondant aux restrictions de sélection ont des frontières relativement nettes, ce qui reflète la division des objets du monde réel en classes qui sont clairement différenciées dans le domaine" (Sager 1986).

- La contrainte de probabilité y est neutralisée : dans la relation opérateur-argument, il n'existe pas de gradation due aux restrictions de sélection.
- Alors que les sous-langages sont des sous-ensembles de la langue, la grammaire d'un sous-langage n'est pas un sous-ensemble de la grammaire de la langue générale. Les grammaires sont plutôt en intersection.

"The sublanguage has important constraints which are not in the language : the particular word classes, and the particular sentences types made these. And the language has important constraints which are not followed in the sublanguage. Of course, since the sentences of the sublanguage are also sentences of the language, they cannot violate the constraints of the language, but they can avoid the conditions that require those constraints. Such are the likelihood differences among arguments in respect to operators; these likelihoods may be largely or totally disregarded in sublanguages. Such also is the internal structure of phrases, which is irrelevant to their membership in a particular word class of a sublanguage" (Harris 1988).

Obtention de classes d'opérateurs et d'arguments à partir de phrases élémentaires

- Dans la lignée des travaux de Harris, l'équipe de Naomi Sager a développé une méthodologie (Sager & al. 1987) de mise à jour, à partir d'une analyse distributionnelle en phrases élémentaires, des classes de concept et de relations d'un sous-langage lié à un domaine d'activité donné (Habert & Nazarenko 1996).
- (Habert & Nazarenko 1996) présentent une démarche qui reprend partiellement l'approche harrissienne : étude des fonctionnements syntaxiques qui sont propres à un corpus donné, analyse de groupes nominaux, automatisation du traitement.

La démarche harrissienne de mise au point de grammaires de sous-langages est moins connue en France que l'utilisation de sa théorie de la langue générale. Seules les études de textes politiques ont exploré cette voie : l'analyse automatique de discours politique de M. Pêcheux (Pêcheux 1969). M. Pêcheux *"a developpé une entreprise qui, en se fondant sur la scienticificité de la linguistique, visait à élaborer des outils d'analyse capables de révéler les processus idéologiques à l'oeuvre dans le discours /./ Le cadre épistemologique de l'A.A.D se définit comme l'articulation de trois régions de connaissances scientifiques : le matérialisme historique, la linguistique (théorie des mécanismes syntaxiques et des processus d'énonciation), une théorie du discours comme théorie de la détermination historique des*

processus sémantiques, ces trois régions étant traversées par une référence à une théorie psychanalitique du sujet" (Maingueneau 1991).

1.4. Conclusion

> Puisqu'il est difficile d'appréhender globalement les faits de langue examinons au plus près leurs composants.

La langue présente trop de variations pour pouvoir être appréhendée globalement. Ces variations se retrouvent le plus souvent au niveau des mots où elles sont plus directement accessibles. Dans les systèmes actuels de TALN, le lexique joue un rôle actif à tous les niveaux d'analyse. Notre travail s'inscrit dans une approche lexicaliste et vise à associer aux mots les savoirs minimaux de base qu'il est possible de leur associer et qui permettent des ajustements ou des affinements potentiels à venir. Les savoirs associés aux mots ne sont pas des valeurs définitives ou prédéfinies de manière absolue : nous considérons en effet que ces savoirs ne sont pas prédictibles et qu'ils peuvent évoluer de telle sorte qu'aucune représentation ne soit capable de fixer de prime abord des valeurs définitives pour ces savoirs. De même les savoirs sémantiques que l'on peut associer aux mots constituent des noyaux de base pour la représentation sémantique. Il ne s'agit pas pour autant de valeurs absolues qui exprimeraient un noyau irréductible de signification, aucune prescription ontologique ne guide la définition de ces noyaux de base associés aux mots. Ce noyau de base doit contenir ou indiquer les pistes de sens possibles à partir d'une valeur de base. Notre tâche n'est pas de prescrire le ou les sens mais de permettre de reconstruire ce ou ces sens via le noyau de base i.e. le support pour dériver les différentes pistes de sens possibles.