DEMANDE D'HABILITATION À DIRIGER DES RECHERCHES

Intégration structurale des points de vue componentiels et compositionnels : pourquoi et comment

Dominique Dutoit

Jury :

Stefan Darmoni	Professeur, Univ. de Rouen	Rapporteur
Anne Nicolle	Professeur, Univ. de Caen	Rapporteur
Max Silbertzein	Professeur, Univ. de Franche-Comté	Rapporteur
Jacques François	Professeur, Univ. de Caen	Membre du Jury
Thierry Lecrocq	Professeur, Univ. de Rouen	Membre du Jury
Pierre Zweigenbaum	Directeur de Recherche, Limsi	Membre du Jury

Soutenance le mardi 16 juin 2009, au CHU Hôpital CHarles Nicolle, Cour Leschevin, porte 21, 3ème étage

Remerciements

A tous ceux et celles qui m'ont supporté jusqu'à maintenant, le poids de la HDR s'ajoutant un peu aux autres poids considérables pour moi de mes entreprises.

Je remercie en premier lieu Julie qui a connu la solitude et ne m'en tient pas rigueur. Je remercie mes collègues qui ont supporté mon humeur quand je suis au-delà de la surcharge.

Et je remercie bien sûr nombre d'enseignants qui m'ont offert la possibilité de cette soutenance. Il s'agit bien sûr, déjà, de Pierre Nugues, mon directeur de thèse, aujourd'hui Professeur en Suède. De Patrice, qui me fit d'abord confiance, tout en me souhaitant d'abandonner la complexité. C'était le conseil d'un ami qui souhaite à un ami une vie sereine. Mais on ne se change pas et je n'ai pas pu l'écouter. Je souhaite maintenant remercier ceux qui m'ont accompagné jusqu'à ce jour. Il s'agit de Nadine, de Stefan, de Max, d'Anne, de Pierre ou de Jacques qui me consacrèrent aussi de ce temps dont il me faudra reconnaître enfin toute la valeur pour ne pas déméríter de leur générosité.

A vous, je fais la promesse de prendre davantage soin de mon propre temps.

Ce document comprend trois parties :

Première Partie : Fiche résumé

Deuxième Partie : Travaux

Troisième Partie : Liste des contributions

Première Partie

Fiche Résumé

Titres

Docteur Université de Caen en 2000.

Félicitations du Jury.

Quelques opérations sens->texte et texte->sens->texte utilisant une sémantique universaliste apriorique

Mémoire de Dea économie 1988 : Système stochastique à génération de capital

Fonction actuelle

Gérant de la Société Memodata, directeur de la recherche dans l'entreprise.

Domaine d'activité

Informatique linguistique, sémantique computationnelle.

Mots-clés

Isotopie, sème, sémantique lexicale, sémantique dérivationnelle, ontologie, paraphrase, agent, complexité, structuralisme.

Travaux de recherche

4 objets – instruments :

Dictionnaire Intégral	comment représenter le dictionnaire de langue pour le rendre accessible à une utilisation automatique sémantique?
Lexidiom	outil de gestion de réseaux lexico-sémantiques (4 millions de feuilles, 5 millions de relations)
Sémiographe	outils d'analyses linguistiques et d'inférences
Sémiographe I	instrument ensembliste de mesure d'isotopie componentielle
BabySemio II	instrument méréologique de mesure d'une intégration compositionnelle et componentielle

1 objet de dissémination

Alexandria	outils de diffusion et de promotion des résultats obtenus (web)

Travaux de recherche en projet

Fondements méréologiques d'une sémantique componentielle, modélisation cinétique de la sémantique lexicale.

Principales publications et disséminations

Publications

21 publications avec comité de lecture (dont 16 en première place) dans des revues à comité de lecture.
3 organisations de colloques
2 conférences invité

Expertise

Expert Technolangue
Membre de comités de lecture (6 fois).

Contrat de recherche :

14 contrats.

Projets nationaux

Projet 1 Dicologique, (Min. de la recherche), page 29 (coordinateur)
Projet 2 Amélioration de Dicologique, (MENRT), page 29 (coordinateur)
Projet 5 : AGIR, (Min de l'industrie), page 48, (partenaire)
Projet 7 : IVOMOB, (Min. de la recherche), page 48, (partenaire)
Projet 12 OSEO ANVAR ALEXANDRIA, page 66 (2004) (coordinateur)
Projet 13 VODEL, (ANR 2005), page 66 (coordinateur)
Projet 14 INTERSTIS, (ANR TECSAN) page 68 (partenaire)

Projets européens

Projet 3 CRISTAL, (DGXIII-CEE), page 29, (coordinateur scientifique)
Projet 4 : MARLEN, (LEONARDO-CEE), page 47 (partenaire)
Projet 6 : EuroWordnet., (E-content), page 48, (1998) (partenaire)
Projet 8 VIVIAN (ITEA), page 66 (1999)
Projet 9 Balkanet, (E-content) page 66 (2000) (coordinateur scientifique)
Projet 10 Ambience (ITEA), page 66 (2002) (partenaire)

Enseignements et direction scientifique

Enseignements

Enseignement universitaire
1988-92, Université de Caen
TD statistiques (niveau DEUG)
Cours magistral Méthodologie de conception des Systèmes d'information
MERISE (UV de 25 H, Maîtrise Economie et MSTCF)
Théorie des systèmes, théorie des Jeux

Formation continue

Cessions de 5 journées en informatique linguistique et sémantique lexicale

Centre de recherches
CAP GEMINI INNOVATION
ALCATEL
Paribas
THALLES
Thomson Multimedia Rennes
C.E.A.

Entreprises innovantes
Diverses PME.

Etablissements publics, collectivités territoriales :
Préparation au concours d'attaché territorial (Organisation)
Membre du jury de rédacteur (finances)
Formation initiale après concours (rédacteurs : système d'information, organisation).

Remarque : plusieurs ex-étudiants ou ex-stagiaires sont aujourd'hui maître de conférences ou chargé de recherche.

Direction scientifique

Thèses
3 thèses soutenues ont fait emplois et références directs au Dictionnaire Intégral ou au Sémiographe.

DEA – Master 2 recherche

6 mémoires de DEA (GREYC et PARIS VII)

Autres encadrements

douze DESS ou ingénieurs de 3°année
une cinquantaine d'étudiants de Licence-Maîtrise-2°année d'ingénieur

Jury de thèse
Marianne Dabbadie, Université Charles-de-Gaulle, Lille 3 (2007)

Deuxième Partie

Intégration structurale des points de vue componentiels et compositionnels : pourquoi et comment

Il faut toujours avoir deux idées : l'une pour tuer l'autre.
Georges Braque.

Table des matières

1 INTRODUCTION ET PLAN

Les travaux que nous décrivons dans ce mémoire servent à justifier la soutenance d'une habilitation à diriger des recherches. La soutenance repose sur :
- la fiche résumé et le CV
- les travaux de recherche réalisés et les publications.

En tant que directeur d'une société privée dont le noyau comprend trois personnes, depuis 1989, j'ai été responsable et/ou à l'initiative de quatorze projets de recherche soutenus par des institutions de recherche, en réponse à des appels d'offre (Framework projects de la CE, Ministère de la recherche, Ministère de l'Industrie, Agence Nationale de la Recherche). En tant que chercheur-directeur de société ou Directeur de Recherche Associé au CNRS, j'ai été rédacteur unique, principal rédacteur ou corédacteur de vingt et une publications avec comité de sélection. Depuis 1989, je ne me suis posé qu'une même question qui peu à peu s'est organisée dans une dualité dynamique : une dialectique qui parle du Signe et qui fournit une réponse complexe à la complexité du Signe.

Nous fournissons un résumé de ces travaux d'enquête et d'investigation.
Ce résumé est séparé en deux parties :
- une première partie, allant des chapitres 1 à 5, suit un plan chronologique organisé par grandes périodes. Les chapitres de cette partie présentent des considérations théoriques, des difficultés et des réalisations pratiques. Les interrogations que ces artefacts que sont les réalisations ont convoquées en nous sont souvent formulées avec le vocabulaire de l'époque considérée. Nous espérons que cette façon de procéder facilitera la transmission des problématiques que nous avons définies.
- une deuxième partie tenant en le seul chapitre 6 reprend l'ensemble des problématiques et décrit une sorte de méta-modèle qui transforme la diversité des problématiques en une problématique unifiée.

2 UN RESEAU COMPONENTIEL (1989-1991)

Après une proposition de positionnement initial de nos travaux (2.1), nous présentons les concepts que nous avons développés et qui sont encore aujourd'hui utilisés ne serait-ce que pour des raisons historiques (2.2), nous présentons sommairement les projets et documents de l'époque (2.3) et nous concluons par les perspectives de recherche à l'issue de cette première période (2.4).

2.1 Positionnement de nos travaux

Considérant le mot - *Son monosyllabique ou polysyllabique, composé de plusieurs articulations, qui a un sens* (Littré) - et le sens - *Idée ou ensemble d'idées intelligible que représente un signe ou un ensemble de signes* (petit Robert) – nous supposons qu'une caractérisation d'un même mot dans un même sens supporte **plusieurs** localisations dans un système semi-formel quelconque. Qu'entendons-nous ici par *plusieurs localisations?*

D'une manière naïve, en 1989, nous avions examiné la transitivité des deux définitions proposées, et considéré la pluralité suivante de la relation entre *mot* et *idée* :

mot --> sens --> *idée ou* ***ensemble d'idées***.

Cet *ensemble d'idées* lié au signe, que nous prendrons comme *signe linguistique* dans la définition de Saussure (*entité double, faite du rapprochement des deux termes* [signifié et signifiant], *tous deux* ***psychiques*** *et unis par le lien de l'association*[1]) conduit par transitivité à :

---

[1] Définition de Jean Dubois, "*Dictionnaire de linguistique, Librairie Larousse, édition de 1973,* page 439). Nous ne commenterons pas ici cette définition qui nous sert uniquement à la définition de notre domaine. Notons qu'en linguistique des auteurs ont des visions plus amples, comme par exemple J.J. Franckel et D. Paillard qui, tous deux inspirés par la théorie des repérages énonciatifs et notionnels d'A. Culioli, ont introduit le concept de *forme schématique.* De notre côté, notre article publié dans CIDE 7 [Dutoit, 2004] a visé principalement à réfuter cette séparation bipartite à partir de l'étude sémasiologique du signe le plus simple qui soit, une simple lettre, la lettre *i*, et donc pour appeler à l'étude de représentations informatisées du sens plus complexes que celles sous-jacentes à cette bipartition.

mot --> sens --> *idée ou* ***ensemble d'idées*** --> signe linguistique --> signifiant/signifié

et, si nous l'acceptons à :

signifiant/signifié --> signifiant --> mot

Entre un mot et lui-même, par exemple, il peut exister un ensemble d'idées qui appartiennent à la durée[2] durant laquelle nous avons réfléchi ce mot. Le parcours que nous venons de réaliser définit finalement bien notre domaine tel que nous l'avons conçu à l'époque. Il s'agit de passer du mot aux idées et réciproquement des idées aux mots d'autant de façons qu'il est possible. Le passage du mot aux idées se fait au moyen de plusieurs localisations, et ces différentes localisations peuvent être nécessaires également pour un mot monosémique. Par exemple :

- une reconstruction naïve et minimale du signe *samouraï* implique l'évocation des concepts de [guerre] (il est un guerrier), de [Japon] (il est un japonais), de [noblesse] (il est un noble).
- de l'autre côté, le passage des idées aux mots, c'est à dire un parcours allant des concepts aux mots, conduit à ce que les idées de [guerre], de [Japon] ou de [noblesse], prises ensemble ou séparément peuvent amener au mot "monosémique"[3] *samouraï.*

C'est le 1° octobre 1989 que nous créons, mon collègue Patrick de Torcy et moi-même la société MEMODATA. À cette date, nous avions achevé un éditeur nommé Lexilog qui était capable de gérer dans une interface efficace pour le genre de *lexicographie* envisagé un graphe orienté acyclique[4] (*Directed_acyclic_graph, DAG*[5]) à base de concepts et dans lequel les mots sont des feuilles[6]. On pouvait y faire des recherches, se déplacer, créer, supprimer, corriger, effectuer des contrôles d'intégrité, réaliser des suppressions logiques etc. Afin de préciser le contexte de ce travail de l'époque, situons le développement selon les :

- points de vue linguistiques et relations avec d'autres travaux concernant des sémantiques lexicales ou des dictionnaires
- relations avec les travaux formels de description

Avant d'aborder ces questions, nous présentons d'abord notre vision linguistique.
A cette époque, notre travail était uniquement empirique et exploratoire. Empirique, puisqu'il s'agissait de voir ce que nous pourrions apprendre de la pratique régulière d'une même activité, ici la lexicographie appliquée, à l'échelle d'une langue, avec le double regard de la sémasiologie et de l'onomasiologie. Exploratoire, puisqu'il s'agissait, du fait que notre travail était informatisé, d'être attentif à l'existence éventuelle d'usages propres au support (l'ordinateur).

---

[2] Dans un sens bergsonien "*Dans la durée envisagée comme une évolution créatrice, il y a création perpétuelle de possibilité, et non pas seulement de réalité*". Voir Bergson [1907].

[3] Une polysémie de ce mot pouvant toutefois apparaître dans la durée de nos représentations de *samouraï.*

[4] Nous remercions Pierre Zweigenbaum pour cette expression. Voir http://fr.wikipedia.org/wiki/Graphe_acyclique_orienté. En particulier, le terme "treillis" ne s'applique pas à notre travail puisqu'un treillis est un ensemble ordonné où toute paire d'éléments a une borne supérieure et une borne inférieure uniques.

[5] Nous utiliserons cet abréviation par la suite.

[6] A partir de l'introduction en 2003 de fonctions lexico-sémantiques proches de celles de Mel'çuk [1986], l'outil ne considérera plus les mots comme des feuilles (voir 3.1.2 L'enrichissement du modèle : le Dictionnaire Intégral (LDI), page 24).

L'approche a été aussi celle d'un sceptique. En effet, nombre de positions théoriques comme celle de Wittgenstein [1961]

> *Un item lexical n'a pas de "sens" en soi, les différentes lectures résultent entièrement de la variété des environnements contextuels* (citation de Desclés [2005], prise dans J. François [2007])

nous laissaient un peu étonné puisque après tout nous ne pouvons en général obtenir n'importe quelle lecture concernant un mot donné depuis n'importe quelle variation de son environnement contextuel sauf à redéfinir totalement ledit mot dans un contexte.
Dans cet ordre d'idée qui consiste à défendre au contraire de Wittgenstein qu'un mot a un sens, l'usage fréquent de la définition de la synonymie (*la possibilité de se substituer l'un à l'autre dans un seul énoncé isolé*[7]*)* pour montrer ce sens en contexte ne cessait pas de nous étonner. En effet, même considérant l'unicité du résultat que nous obtenons à partir du calcul des énoncés 2+2 et 2x2[8], je ne puis accepter que les opérateurs "+" et "x" aient même signification.
Au fond, j'ai défini mon point de vue dans Dutoit [1991] : *Quelle est la cognoscibilité de la signification ?* C'est-à-dire puis-je fabriquer un dictionnaire qui permette d'avancer dans la description des mots en tant que *signe*, c'est-à-dire en tant qu'objet causant quelque chose dans un certain espace défini par Saussure comme *psychique* et que nous nommerions aujourd'hui à l'aide du mot *cognitif*[9].
En définitive, ma **position linguistique** est celle d'un structuraliste[10] et d'un constructiviste. Considérant le dictionnaire, c'est-à-dire un ensemble de signes définis pour l'humain, et l'ordinateur, c'est-à-dire une machine traitant de l'information, je me demande quelle partie de la description des signes pour l'humain peut être transférée à l'ordinateur de telle manière que les signes y actionnent des programmes spécifiquement conçus pour réagir aux parties de description des signes que nous aurons identifiées lors d'inventaires systématiques desdites parties. Cet inventaire devrait être répété jusqu'à ce qu'aucune partie des signes ne soit oubliée. Nous ne nous demandons pas si la tâche décrite est possible puisque la détermination de cette faisabilité est une question ouverte, indépendamment du temps que nous pourrons lui consacrer. Par contre, il est intéressant de se demander ce que nous obtenons avec la démarche proposée au bout d'un certain temps. Ma thèse en 2000 propose un point sur la question, et les orientations proposées dans ce document dans le chapitre 6 Intégration structurale des points de vue componentiels et compositionnels, page 73, ouvrent sur des possibilités nouvelles qui résolvent l'essentiel des difficultés que nous avons rencontrées et que nous présentons dans ce document. Pour conclure sur cette présentation du domaine, tel que perçu à l'époque, nous détaillons les trois suivants :

- stratégie en matière linguistique

---

[7] définition de *synonymie*, Jean Dubois. Le texte vaut également si *plusieurs* remplace *un seul*.
[8] lesquels existent en nombre infini : 2=3-1 et 2=5-3 et 2=π - π +2 etc.
[9] Ce qui me semble finalement moins pertinent puisque la connaissance ni plus que le signe n'ont d'existence en-dehors d'une pensée c'est-à-dire de quelque chose entre autre susceptible de réagir aux signes linguistiques
[10] je suis favorable à la notion de structure : *ensemble, système formé de phénomènes solidaires, tels que « chacun dépend des autres et ne peut être ce qu'il est que dans et par sa relation avec eux* (Lalande), mais je préfère la phrase de Pascal : *toute chose étant aidée et aidante, causée et causante, et les plus éloignées étant liées de façon insensible, je tiens pour impossible de connaître la partie si je ne connais le tout et de connaître le tout si je ne connais pas la partie*. Enfin, la caractérisation de G. Deleuze [1973] me convient tout à fait s'il doit servir à qualifier mon travail exploratoire : *Ce qui est structural, c'est l'espace, mais un espace proprement structural, pré-extensif, pur spatium constitué de proche en proche comme ordre de voisinage (...) L'ambition scientifique du structuralisme n'est pas qualitative mais topologique et relationnelle*.

- situation par rapport aux réseaux sémantiques et aux logiques de description
- usages prévus des premiers résultats.

a) Stratégie en matière linguistique

Le premier "modèle" sémantique que nous avons retenu est voisin de celui des "traits sémantiques" du fait de leur grand pouvoir classificatoire appliqué à l'ensemble du lexique. Les traits sémantiques sont des sacs généralement considérés comme primitifs en cela qu'ils n'utilisent aucune relation, ni aucune hiérarchie entre eux. Voir par exemple la note 96 page 108 dans ce document un exemple chez Rastier. Le plus souvent aussi, la description des mots à l'aide de ces sacs repose sur l'idée qu'il faut définir et différencier par rapport à quelque chose au lieu de chercher à définir en soi. Pottier [1992] fournit un exemple célèbre de différenciation des *sièges* par rapport au "concept" de *siège.* Ainsi chaque lexème est représenté par des composantes sémantiques appelés *sèmes* chez Pottier[11]. Concernant ces deux propriétés fréquentes des approches componentielles que nous venons de souligner, à savoir l'existence de *primitive* et l'emploi d'un système du genre arbre de porphyre (*nouveau genre = genre proche + différence*), nous nous sommes simplement assuré que nous pouvions avoir des vues de cette sorte. Mais, chez nous les traits sont eux-mêmes décomposés, ce qui fait disparaître leur nature primitive et l'attrait de leur caractère booléen. En retour, cela apporte de la souplesse, et il devient possible de définir *fauteuil* par *meuble sur lequel une personne seule s'assied, ce meuble comportant des bras, pieds et dossier*[12], même si cela est moins efficace que de le définir comme Pottier le fait comme *siège pour une personne seule comportant des bras, pieds et dossier.* En fait, il devient possible en définissant *fauteuil* comme *Pottier* le fait à juste titre de le retrouver à partir de notre exemple de définition utilisant *meuble*[13]. . En pratique, les services apportés sont voisins de ceux du *thesaurus* et l'appellation du Thésaurus de Larousse[14] *des mots aux idées, des idées aux mots* aurait pu être retenu. Mais notre projet a été d'emblée plus large du fait de l'héritage des propriétés, ou plutôt dans notre cas, des localisations et de l'ambition de définir les mots au moyen des traits sémantiques.

b) Situation par rapport aux réseaux sémantiques et aux logiques de description

Nous examinons l'un et l'autre de ces objets en nous reposant sur les définitions proposées par Russel et Norvig [2006], pages 393 à 398.

Les réseaux sémantiques sont souvent associés aux notations graphiques par nœuds et arcs proposées par Charles Pierce [1909] et appelés *graphes existentiels.* Les réseaux sémantiques sont capables de représenter des objets individuels, des catégories d'objets et des relations entre les objets au moyen d'arcs portant des étiquettes. Sans aller plus loin, la définition fait apparaître l'absence de toute relation entre notre travail et ces réseaux :

- les réseaux sémantiques décrivent des objets tandis que nous décrivons des mots, des conceptualisations dénotées par un mot
- ils emploient les moyens de la logique (quantification, conjonction, arcs étiquetés) qui sont essentiellement hors de notre domaine.
- ils ont la nature booléenne du vrai et du faux, et nous sommes flous

[11] Ou *markers* chez Katz

[12] Il faut qu'une application comme le *dictionnaire à l'envers* (voir paragraphe 5.1.2.2, page 55) soit aussi capable de lire cette définition et de retrouver *fauteuil.*

[13] 5.1.2.2, Le dictionnaire à l'envers, page 55, comme exemple d'utilisation de cette élargissement du paradigme des traits sémantiques. Nous travaux sur le dictionnaire à l'envers ont commencé en 1992, et servaient à réfléchir sur le fonctionnement du réseau lexico-sémantique.

[14] Sous la direction de Daniel Péchoin, 1991.

De leur côté, les logiques de description sont des notations élaborées pour faciliter les énoncés sur les objets. Elles travaillent sur la description des définitions et des propriétés des catégories d'objets. De même, en première analyse, elles n'ont pas grand chose à voir avec nos travaux même si, et cela justifie l'importance de la présente note, bien des termes sont communs à nos travaux et à cette forme de logique. Par exemple, nous parlerons de classification ou d'inférence et nous pourrons avoir l'impression de retrouver la notion de subsomption à travers un "est impliqué par" trouvé dans le système de traits sémantiques, ou bien un "contient" si nous considérons le même système de traits sémantiques selon un mode ensembliste comme dans Dutoit [1992].
Enfin, ces systèmes se distinguent peut-être aussi du nôtre du fait de la taille relativement petite qu'ils ont atteinte. Aujourd'hui, le Dictionnaire intégral comporte environ plus de 1,5 millions de nœuds en 27 langues qui sont traités de manière égale.
En dernier lieu, ces systèmes satisfont à de nombreuses conditions de logique du premier ordre et nous ne pouvons aucunement simuler cette logique : nous ne pouvons nous contenter que de degré de vérité ou de vraisemblance. Par exemple, si dans le dictionnaire nous décrivons *samouraï* comme *guerrier* cela n'implique aucunement qu'une occurrence particulière de *samouraï* dans un texte renvoie à une occurrence du concept de *guerrier*. Nous suivons sur ce point Wittgenstein, cela dépend effectivement du contexte. En admettant maintenant que le problème de polysémie sous-jacent serait partiellement résolu, alors, peut-être, dans un second temps, un objet *guerrier* pourrait être conçu et utilisable. Ce genre de question nous intéressera à partir de 2000 (voir 4.1.2.2.3 *L'activation componentielle*, page 38) et nous conduira à réfléchir plus globalement en terme méréologique (il semble qu'une partie de telle occurrence de *samouraï* dans un texte recouvre le signe *samouraï* et cela devrait induire une certaine représentation nouvelle) qu'en termes de nature ou de logique.

c) Usages prévus des premiers résultats.
Nos objectifs étaient doubles :
- disposer d'un inventaire à plusieurs facettes de faits lexico-sémantiques
- examiner quelles utilisations nous pouvons effectuer, par exemple en termes de désambiguïsation, ou de constat sur le contenu du dictionnaire destiné à l'homme de cet inventaire.

Parmi ces usages, un usage particulier était prévu : distribution de l'inventaire en tant que tel, à destination du grand public, pour faciliter le passage de l'idée aux mots. Le dictionnaire sera effectivement commercialisé dès 1992, à la FNAC, sous le nom de Dicologique.
En conclusion, le travail d'analyse sémantique des définitions du dictionnaire de langue, peut-être du fait de l'outillage technique dont nous disposions, nous a rallié au courant componentiel. De fait, notre point de vue a été le suivant : *il est possible d'établir les propriétés sémantiques d'un terme hors contexte*, et nous avons considéré le sens comme une *donnée initiale indépendante du texte environnant* dans lequel *certaines propriétés sont activées ou désactivées* [Gayral, 1998]. Le but de la société MEMODATA a été la création d'un nouveau dictionnaire de la langue française exploitant les possibilités de l'informatique pour proposer un dictionnaire nouveau dans le sens de *un genre nouveau*.

2.2 Lexilog et les premières hypothèses de signification lexicale

Le logiciel résultat prit pour nom Lexilog, puis à partir de 1996 celui de Lexidiom pour rendre compte dans son nom des fonctionnalités de gestion mutilingue. Lexilog permettait de gérer un graphe orienté acyclique de mots et de concepts. Nous donnons les principales définitions en usage pour cette version du dictionnaire puis nous fournissons quelques figures les illustrant.

Premières définitions

Concept : tout objet artificiel structuré et structurant. Un concept est noté [concept].
En général, *concept* est défini par *représentation mentale*[15] *d'un objet* et s'oppose à *signifié* et à *référent.* Le plus souvent, un signifié renvoie à plusieurs concepts c'est-à-dire à diverses représentations mentales d'objets différents. Par exemple, *samouraï* renvoie à plusieurs concepts comme [Japon] et [guerrier], c'est-à-dire aussi [Asie] et [personne] et [guerre] etc. En tant que *représentation mentale*, le concept n'est pas observable et n'est aucunement un mot. Ainsi, en écrivant [Japon] nous souhaitons indiquer que nous ne renvoyons pas au mot Japon mais à une représentation mentale que le lecteur devrait se faire du référent *Japon.* Comme il ne nous est pas possible de démontrer que cette représentation existe bien et se trouverait dans la nature, nous insistons sur le caractère artificiel, c'est-à-dire produit par l'activité humaine finalisée de modélisation ou de service éditorial[16], du concept dans notre acception. Par structuré, nous entendons qu'un concept n'est pas un simple trait de sens ; en fait, il renvoie à d'autres concepts comme par exemple [Asie] ou [pays] pour [Japon]. Réciproquement, par structurant nous entendons qu'un concept n'a d'intérêt que s'il est impliqué par d'autres éléments, comme par exemple le concept [Tokyo] ou le mot *Tokyo n.p* pour [Japon].

Libellé d'un Concept : commentaire rédigé en texte libre destinée à renseigner l'humain sur l'usage dudit concept.

Mot-sens : mot de la langue dans une acception particulière[17].
Une acception i d'un mot, après analyse lexicologique est déclarée à l'aide de l'énumération des concepts [A], [B] et […] perçus par le lexicographe au moment de l'étude de i. En 1989, le mot-sens était défini par une simple énumération de concepts; cette énumération s'appelait quasi-définition.

Quasi-définition : Pour un mot-sens, ensemble de ces concepts immédiats, chacun de ces concepts jouant le rôle de trait définitoire.

Trait définitoire : concept componentiel doté d'un type.

Nous donnons ci-dessous quelques-uns des types de traits sémantiques que nous avons utilisés.

Classe : trait définitoire groupant des mots semblables en termes de nature. Par exemple, [*renard]*$_{classe}$ contient différentes énonciations du concept : *renard, goupil, isatis, renardeau, renard noir, renard bleu, renard polaire, renard commun, renard blond des sables, renard gris argenté, renard crabier, renard à petites oreilles, renard de la pampa, renarde, renard*

[15] dès qu'il y a *mental*, il faut inclure la notion de durée, dans le sens de la note 2 page 8. Notre concept exclut ici cette notion de durée. Il est bien un artefact conçu uniquement pour peupler un espace.
[16] La production d'un thésaurus suppose la dualité concept/mot. Par exemple, dans le thésaurus Péchoin, l'article 508 [courage] commence par le nom *courage.* La dualité y est naturelle, et nous trouvons toujours étonnant que les USA qui ont eu le thésaurus de Roget (1852) n'aient pas prolongé ce travail par exemple dans WordNet (Fellbaum. [1998]).
[17] Dans la théorie Sens↔Texte, les termes *unité lexicale* ou *lexie* sont utilisés en lieu et place de *mot-sens*. Il nous semble toutefois que *mot-sens* est moins sujet à diverses interprétations et c'est pourquoi nous le retenons.

argenté, renard blanc[18].
NB : Un même mot-sens peut appartenir à plusieurs classes (ex. *renardeau n.m.* appartient à $[renard]_{classe}$ et à $[petit\ d'une\ espèce]_{classe}$.

Thème : trait définitoire groupant tous les mots partageant un même élément de définition, indépendamment de leur partie du discours. *Renard[thème]* comprend $[renard]_{classe}$ et d'autres mots plus isolés : *glapir, renardière, hydrophobie, rage, piège à renard,* et les elliptiques de *fourrure de renard.*

Classe d'opposition : ex : $[monter/descendre]_{classe}$

Thème d'opposition : ex : $[monter/descendre]_{thème}$

Termes liés : trait non définitoire groupant tous les mots appartenant à un même thème mais non susceptibles d'appartenir à une des classes du thème du fait de leur isolement dans le thème (ex. *glapir, renardière, hydrophobie, rage, piège à renard* sont placés dans $[renard]_{termes\ liés}$ lui-même contenu dans *Renard[thème]*.

Caractéristique : trait définitoire groupant tous les mots dont la définition présente un modifieur relativement simple par rapport à leur espèce pour renvoyer à un thème existant (ex. projeter vt. --> $[jeter]_{classe}$ + $[fort]_{caractéristique}$).

Dans les figures suivantes, nous donnons accès, à titre d'exemple, aux premiers niveaux d'ancêtres pour les mots *landgrave n.m. samouraï n.m.* et *projeter v.tr.* Un graphique plus profond est donné pour *renard (en tant qu'animal).* Les graphiques proposés correspondent à l'état actuel du modèle du réseau lexico-sémantique et présentent des types de relation que nous n'avons pas encore décrits puisqu'ils n'existaient pas entre 1989 et 1992. Nous présenterons ces types de relation dans le paragraphe 3.1.2 *L'enrichissement du modèle : le Dictionnaire Intégral (LDI),* page 24. L'observation des relations pourra, à raison, donner l'impression au lecteur de redondances. En fait, sauf exception, le type de relation affiché correspond à une valeur par défaut du trait définitoire et est rempli automatiquement.

[18] La classe fournit souvent une indication de l'appartenance à un classème dans la terminologie de Pottier [1992].

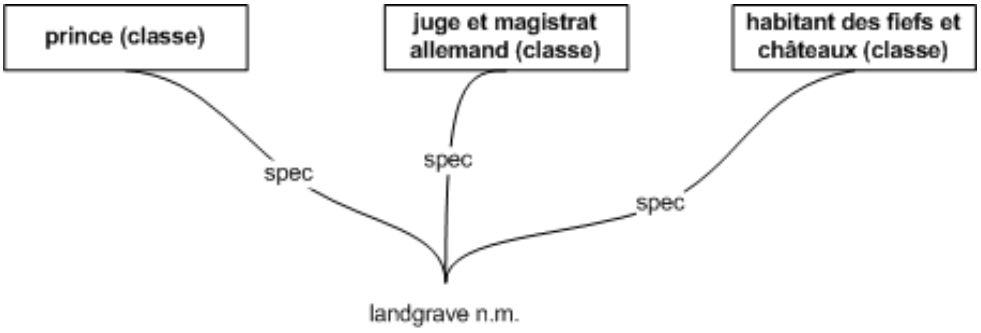


Figure 1 Description de *landgrave* à l'aide des classes.

La figure ci-dessus approxime - il s'agit d'une quasi-définition - la définition encyclopédique fournie par le Larousse encyclopédique. Voici cette définition:

> *landgrave* *n.m.*
> *Titre porté au Moyen-âge par plusieurs princes germaniques relevant immédiatement de l'Empereur, dont les comtes d'Alsace... etc.*
> *Magistrat qui rendait la justice au nom de l'empereur germanique… (Source : Larousse, encyclopédie en couleurs).*

On note dans notre graphe l'agrégation en un seul sens des deux sens proposés par le Larousse encyclopédique. Ici, étant donné le caractère général de la quasi-définition, la portée est mineure et le Robert en a jugé de même puisque, sans être contraint par un modèle de représentation, il a regroupé les deux sens dans son article :

> *landgrave n.m.*
> *Titre de princes souverains en Allemagne, dont l'État était un landgraviat*.

Dans la figure, les traits [prince], [allemand], [magistrat] sont représentés. L'exemple de *landgrave* fournit un exemple simple concernant la variété des découpages en sens du signifié d'une unité lexicale dans les différents dictionnaires.

Observation 1 ***Chacun [des dictionnaires] est une tentative de décrire un objet, ils ne peuvent être confondus avec cet objet.*** **Dubois et Dubois-Charlier [1990, p.10]**

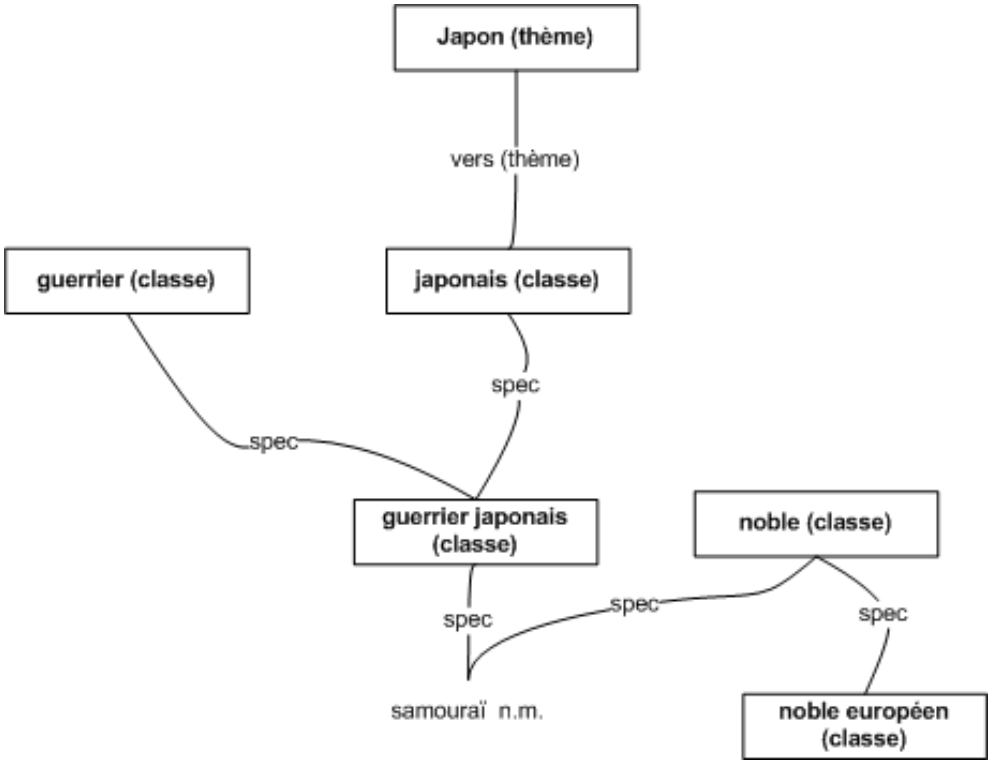


Figure 2 Le caractère fortuit des unions de classe : l'exemple de *samouraï.*

Considérant *samouraï* peu différent de *guerrier noble japonais*, nous observons que le réseau lexico-sémantique comporte des classes avec plusieurs traits regroupés (ici [*guerrier japonais]*) qui se voient immédiatement décomposées (ici en $[guerrier]_{classe}$ et $[noble]_{classe}$) tandis qu'un autre reste isolé. Les raisons de telles organisations sont empiriques et ont pour origines :

- l'importance quantitative d'une classe : la classe *guerrier japonais* est potentiellement intéressante pour un utilisateur du dictionnaire si nous pouvons y grouper plusieurs dénominations.
- une classe peut regrouper des dénominations partageant un ou plusieurs concepts que ce ou ces concepts aient ou non une dénomination générique. On dit d'un mot qu'il est *générique (ou qu'il prend un sens générique) quand il sert à dénommer une **classe** naturelle d'objets dont chacun, pris séparément, reçoit une dénomination naturelle* (Dubois [1973]). Ainsi, par définition, l'existence d'un générique pour une classe donnée entraîne la création de la classe concernée.

Le mot *samouraï* est intéressant entre autre du fait qu'en raison du matériel lexical disponible en français, il peut se décomposer de plusieurs façons, comme par exemple :

- *[japonais] + [guerrier] + [noble]*
- *[guerrier japonais] + [noble]*
- *[noble japonais] + [guerrier]*
- *[guerrier noble] + japonais*
- *[guerrier noble japonais]*

Il faut que toutes ces façons de décomposer donnent un même résultat et que les différentes voies de décomposition ne conduisent pas à des ambiguïtés.

Observation 2 : avec la componentialité de samouraï, nous avons observé que les multiples décompositions possibles d'un même sens doivent être également disponibles.

Cette observation présentée ici comme accidentelle rend compte en fait de deux besoins omniprésents:
- au plan de la conception, le travail sur un très grand chantier comme l'est celui du dictionnaire, nécessite une grande souplesse dans l'accès à une description particulière et dans l'interprétation de ladite description. En fait, tous les modes de décompositions imaginables doivent pouvoir être décrits pour éviter que soit recréé ailleurs un artefact similaire susceptible de créer une ambiguïté artificielle
- au plan de l'analyse componentielle d'un énoncé qui a besoin du fait du texte d'activer en contexte l'une ou l'autre de ces décompositions d'un même sens.
Ce genre de remarque nous a beaucoup inspiré et nous proposerons de l'adapter à partir de 1998 à la syntaxe. Voir l'exemple fourni, qui est le plus simple possible, à la *Figure 15 L'organisation de l'onomasiologie d'une instance d'une classe [date] n'est pas particulièrement triviale* (page 90). La question posée est naturelle dans les perspectives sémasiologique et onomasiologique que nous avons du fait que toute *instance* d'une date particulière dans un texte est susceptible d'être considérée dans un texte comme un spécifique du lexème *date.*

Ex. - Tu viens vendredi?
– Non, je suis pris à cette date.

Dans cet exemple, *date* asserte que *vendredi* qui est ordinairement une dénomination générique d'un certain *jour* peut-être pris comme une *date* à déterminer dans le reste du contexte.
Evidemment, l'exemple avec *samouraï* est componentiel et l'exemple avec *date* est compositionnel mais le résultat est le même : nécessité de multiplier les points de vue sur n'importe quel objet. Cela est bien évidemment compatible avec l'Observation 1 ci-dessus : par exemple, si un système souhaite utiliser ses connaissances pour valoriser un dictionnaire, ce système devra être doté de plusieurs points de vue sous peine de juger que la nouvelle ressource qu'il souhaite utiliser est inadaptée. Ainsi, c'est parce que notre ressource est dotée de plusieurs points de vue qu'elle a pu calculer automatiquement la traduction de la plupart des synsets requis pour le français dans le cadre du projet EuroWordnet (page 48).

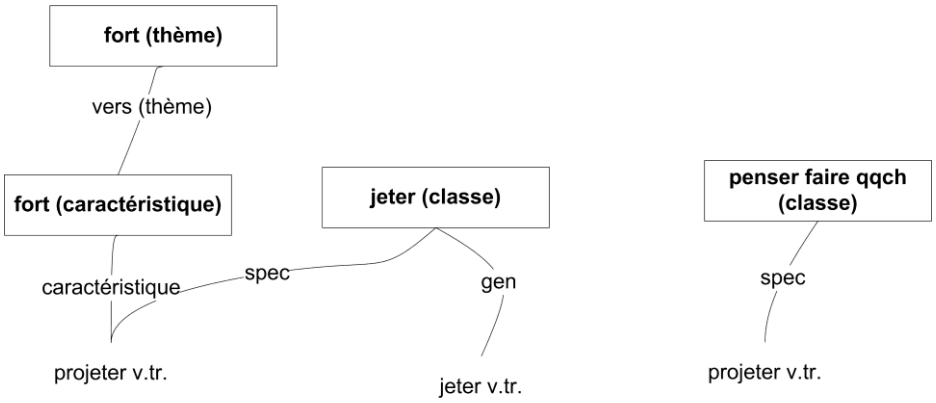


Figure 3 Deux sens de *projeter* dans le dictionnaire (extrait)

La figure présente deux sens de *projeter* en créant deux nœuds différents pour ce mot. Elle emploie de nouveaux éléments formels comme *caractéristique* que nous présenterons au paragraphe 3.1.2, page 24.

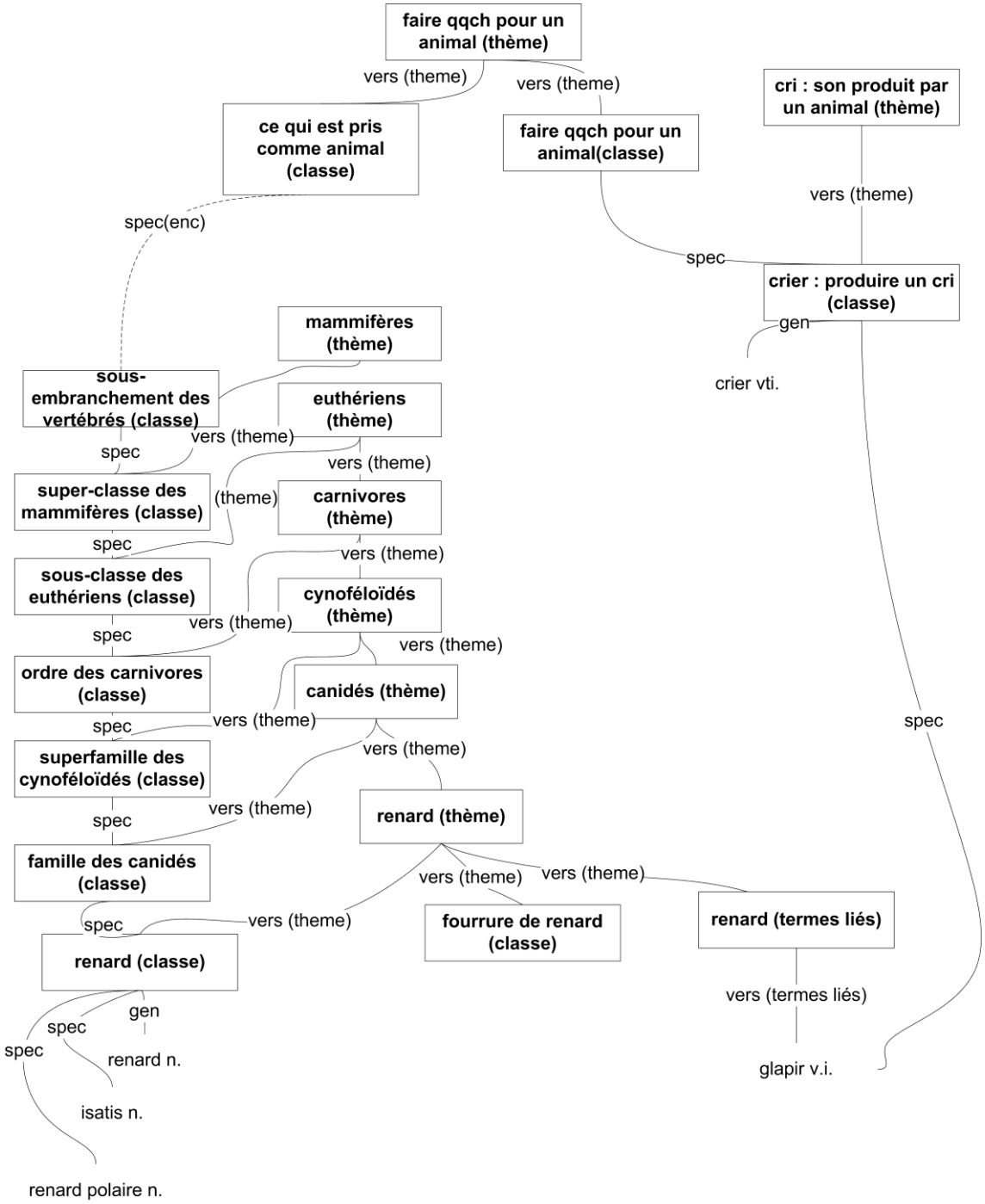


Figure 4 Une illusion d'ontologie concernant *renard*

La figure 4 présente une classification systématique pour illustrer les emboîtements *classe/thème* : une zoologie depuis *renard.* Le mot-clé est situé en bas de la figure. La figure montre immédiatement un double héritage : chaque classe est présentée dans une autre classe (héritage 1) et dans un thème (héritage 2). De leur côté, les thèmes, qui sont les concepts les plus larges, ne sont présents que dans d'autres thèmes. Il est possible de considérer que les thèmes autour de la zootaxie servent à indiquer les propriétés spécifiques de chaque classe. Pour autant, il ne s'agit pas de différence spécifique comme dans un arbre de Porphyre (234-305, de notre ère, *nouveau genre = genre proche + différence*) mais seulement de l'ajout progressif d'information créant du contexte. Dans le dictionnaire, l'intension de la zootaxie

utilisée n'est malheureusement pas représentée. En explorant le graphe de haut en bas, nous verrons toutefois apparaître les contextes de *cellule*, puis de *os*, puis de *colonne vertébrale*, puis de *reproduction sexuée*, puis de *mamelles etc.* sans que ces contextes soient décrits comme des conditions booléennes.
Avant de clore ce chapitre décrivant nos premiers travaux, nous souhaitons répondre à la question suivante de Pierre Zweigenbaum :

[renard]$_{thème}$ est-il autre chose que l'ensemble de tous les liens "vers(thème) qui partent d'un concept classe? (réification des relations). Si c'est le cas; il n'est pas nécessaire fonctionnellement de créer une "boîte de concept" thème pour le distinguer du concept classe?

Nous pouvons trouver une dizaine de réponses à cette question. La première réponse est en relation avec l'histoire technique : à l'époque, nous ne savions pas typer les relations, et seul l'usage des "boîtes" était possible pour nous. Il faut comprendre pourquoi seul cet usage était possible. La réponse est simple : notre modèle de référence était le dictionnaire analogique du type Roget, et nous voyons que le Roget est organisé autour de notions qui s'apparentent grandement aux thèmes. Le Péchoin [1991], qui est la transposition française du Roget, le montre bien : par exemple, nous y trouvons un article MAMMIFERES de quelques pages et cet article commence par l'unité lexicale *mammifère*. Dans le Péchoin, MAMMIFERES signifie *tout ce qui à trait onomasiologiquement à la* ***notion*** *de* [mammifère]. Cette expansion est par exemple différente de celle que nous pourrions faire en donnant les dérivés lexico-sémantiques de *mammifère n.m.* (par exemple *mammalogie n.f.).* D'autre part, le mot *mammifère n.m.* a dans le Péchoin plusieurs localisations pour le même sens. Par exemple, nous retrouvons *mammifère n.m.* dans l'article ZOOLOGIE. **Finalement, la distinction entre *unité lexicale, classe* et *thème* permet de différencier les *points de vue* tandis que la réification des relations ne permettrait pas aisément d'atteindre ce résultat.** Je souhaite conclure ce point en proposant :

Observation 3 : a) La première opération sémantique est une opération de localisation qui vise précisément à être capable de sélectionner parmi différents points de vue, un point de vue particulier portant sur l'emploi d'un signe dans un énoncé. b) En retour, la première tâche du lexicographe structuraliste utilisant les moyens de l'informatique est de préparer pour l'ordinateur les données et mécanismes qui lui permettront de définir l'*intérieur*, le *bord* et *l'extérieur* d'un *point de vue* selon un contexte.

Cette opération sémantique pourra rappeler au lecteur l'article *La frontière* d'Antoine Culioli [1990]. En effet, l'idée est très proche et le changement terminologique (*localisation* dans ce document et *centrage* chez Culioli) n'est là que pour rendre compte qu'avec notre spécification de *localisation* nous restons bien grossier par rapport au centrage qu'il *faudrait* pouvoir effectuer.
Mais il faut bien commencer par un bout. Et nous avons considéré que le bout que nous venons de décrire, tout imprécis qu'il soit, devrait avoir une certaine pertinence même à long terme et ainsi être durablement réutilisable.

Observation 4 : le travail de lexicographie est dirigé à long terme et les principes d'organisation de la modélisation doivent être capables de couvrir tout le lexique. C'est une fois que la couverture tend vers celle de tout un dictionnaire que des travaux d'évaluation des résultats peuvent être menés et des propositions d'élargissement du modèle effectuées.

2.3 Projets et documents

Projet
Le projet "Etudes des caractéristiques d'un nouveau dictionnaire de la Langue Française" fut accepté par le Ministère de la recherche et nous a fourni encouragements et premiers moyens. Grâce à cet apport, notre équipe intégra en CDI un linguiste : Philippe HERR, aujourd'hui agrégé de lettres.

Stage
Des ressources favorisèrent la collaboration avec la recherche académique :
- Soutenance de DEA de K. Dubaele (ISMRA/laboratoire d'informatique de Caen) : caractérisation informatique de Lexilog et de ses données.
- Critiques de Anke Eilers (alors doctorante de linguistique à ELSAP) : caractérisation linguistique du dictionnaire en cours de réalisation.

Publication, dissémination
Le travail accompli a permis de réaliser une première publication :
Dutoit [1991], *Dicologique : un nouveau type de dictionnaire*, revue La banque des mots, Conseil International de la Langue Française.

2.4 Conclusion et perspectives

Nous rappelons nos objectifs de la période
- disposer d'un inventaire à plusieurs facettes de faits lexico-sémantiques
- voir quelles utilisations nous pouvons faire, par exemple en termes de désambiguïsation, ou de constat sur le contenu du dictionnaire destiné à l'homme de cet inventaire.
Ces deux objectifs ont été satisfaits à l'issue de la période puisque la période suivante verra apparaître une première diffusion de l'inventaire et des hypothèses de calcul sémantique.
Cependant, nous notons des pertes d'information dans le graphe. Par exemple, l'observation de la *Figure 1*, page 14, fait apparaître certaines de ces pertes par rapport aux énoncés du dictionnaire Larousse. Il s'agit principalement de:
1. la perte du lien "relevant directement de l'Empereur".
2. la perte de la syntaxe des définitions. Ici, le lien existant par exemple entre $[\text{comte}]_{\text{classe}}$ et $[\text{Allemagne}]_{\text{termes liés}}$ est perdu.

Pour l'essentiel, ces deux pertes ont pour origine commune la perte (2) que nous venons de signaler. Soulignons cette anomalie importante : *le graphe orienté défini ne permet pas d'enregistrer les liens syntaxiques.* D'une façon plus générale, notons le problème suivant : *aucun expression propositionnelle ni aucun concept structuré ne peut être enregistré dans le graphe*. Gardons finalement en mémoire :

Question 1 Comment enregistrer des concepts structurés[19] dans le DAG ?

Malgré cette limite, la version du dictionnaire a été et reste la plus fine construction lexico-sémantique de grande taille fondée sur des rapports de similitude entre des objets (les signes). Cette ressource constitue un développement extensionnel de la démarche componentielle et permet dès 2000 d'imaginer des applications en traitement automatique du texte.
En 1991, les projets de recherche et développement pour la nouvelle période ont été :
- réaliser une version commercialisable de Dicologique
- réfléchir aux traitements automatiques que le dictionnaire pourrait permettre de mettre en œuvre. Faire une liste de ces traitements.
- présenter le travail à des chercheurs concernés par l'activité, en particulier des linguistes.
- évaluer le point d'entrée qui permettrait des extensions multilingues
- rechercher d'autres manières de structurer le lexique telles que ces manières permettraient de nouvelles utilisations (inférences) ou une économie dans nos coûts de maintenance.

En respect de notre **position linguistique** (page 9), nous évitons les applications proprement industrielles qui biaisent le travail tant que le modèle n'a pas suffisamment de capacité de localisation (Observation 3 page 18) et à l'époque c'est Cap Gemini qui a utilisé le dictionnaire pour réaliser à partir de Dicologique une maquette de réseau sémantique (page 10) à destination de l'interrogation en langage naturel des Pages Jaunes[20].

[19] Nous devons cette dénomination à Pierre Zweigenbaum ; elle n'est pas contradictoire avec la définition de *concept*, page 11. Les concepts du DAG sont déjà structurés entre eux du *point de vue* de la totalité du DAG. Cependant *structuré* appliqué à *concept* peut aussi signifier *concept* organisant d'autres concepts dans un jeu de relations qu'ils ont entre eux à travers lui-même. C'est cette valeur particulière que nous évoquons ici.
[20] Cette application pose évidemment le problème de la polysémie au sens lexical ; mais l'application pose aussi un problème de pertinence par rapport à l'utilisateur. A raison, le deuxième point l'emporte sur le premier. A l'époque, dans l'application Page Jaune, cela s'est traduit chez l'utilisateur par la suppression pure et simple du mot-sens *avocat-fruit* dans sa version de notre dictionnaire. Nous jugeons ce type de régression contraire à l'Observation 4 page 18 et nous ne pouvons pas la prendre en charge. Aujourd'hui un tel cas ne poserait plus vraiment problème; mais le principe de l'observation reste valable.

3 PREMIERS CALCULS COMPONENTIELS (1992-1996)

La période voit apparaître chez nous les premiers calculs componentiels de taille réelle, hors domaine. C'est la naissance du Sémiographe. Dans le même temps, nous sommes amené à enrichir le formalisme du dictionnaire de nouveaux attributs. Evidemment, s'agissant d'une recherche appliquée, il ne suffit pas de compléter intensionnellement le modèle. Au contraire, il convient de l'appliquer à la plus grande part du lexique d'une part pour vérifier l'efficacité du principe organisateur (Observation 3 page 14) et pour comprendre son impact sur la structure (voir **position linguistique** page 9). Cette période est aussi celle de notre premier projet européen (Projet 3 CRISTAL, page 29).

3.1 Les travaux de recherche

Quatre années après le début du dictionnaire, il devient possible de réaliser les premiers calculs automatiques. D'un autre côté, le gestionnaire de dictionnaire se voit enrichi de nouvelles fonctionnalités. Enfin, entre 1992 et 1994, en vue de se doter d'un corpus de textes français pour différentes travaux de *text-mining*, nous avons organisé la saisie d'environ 140 œuvres[21] du domaine public.

3.1.1 La naissance de l'idée du Sémiographe

Cette époque était marquée par les projets phare d'importants consortiums et des ambitions financées. Il s'agissait par exemple de GENELEX (470MF) suivi de GRAAL

[21] Voir la liste de ces ouvrages sur http://www.memodata.com/2004/fr/livres_en_ligne/index-svg.shtml. Les éditions qui ont été recopiées datent d'avant 1920. Nous avions acheté ces éditions chez différents bouquinistes, ventes aux enchères, vide-greniers etc. Les personnes qui ont réalisé ces saisies étaient en grande difficulté professionnelle : la saisie s'est faite dans le cadre de stages d'insertion ou de réinsertion préalables à des stages de professionnalisation comme le secrétariat. Ces personnes ont appris à utiliser un ordinateur, le logiciel OCR Omnipage, le traitement de textes Word; en outre ces personnes apprirent ou réapprirent à arriver à l'heure à un travail, à faire un travail soigné, à travailler en équipe et à discuter des *humanités.* (NB : nous n'avions pas pu avoir accès à Frantext).

(140MF). De même, avec le minitel et les pages jaunes, des projets industriels influents définissaient bien les principales directions.
Observant ces mouvements, et particulièrement les centaines d'années-hommes dépensées sur ces projets avec, il s'agit ici de notre jugement de l'époque, une orientation domaine marquée de telle manière que sa transposition à une nouvelle application n'aurait que peu de chance d'impliquer des économies d'échelle, nous avons continué dans une optique libre de toute contrainte applicative ou de domaine. La question que nous posions était alors la suivante : quels calculs pouvons-nous réaliser depuis la base de données constituée et comment effectuer ces calculs ?
C'est à cette époque que s'est installée une pratique que nous avons toujours :
- définition des principes d'organisation des entrées du dictionnaire
- estimation de leur intérêt pour différents calculs
- cohérence et compatibilité conceptuelles (en termes de modèle de données) et inférentielles (en termes d'inférences permises) avec le modèle courant.
En fait, pour nous, chaque représentation de dictionnaire est légitime conformément à l'Observation 1 page 14. La question principale qui se pose est l'intégration et la cohérence du tout. C'est à cette période que naît le terme *dictionnaire intégral* pour désigner notre projet.

Observation 5 Le point clé est l'intégration de différents points de vues linguistiques dans un tout dans lequel des inférences homogènes[22] peuvent s'effectuer naturellement.

Ainsi, à cette époque, c'est en évaluant le contenu du dictionnaire que nous avons déduit plusieurs **opérations** calculables à base d'inférences homogènes. En 1992, un article publié à Coling [Dutoit, 92] résume la situation :
(a) il devrait être possible de calculer le mot-sens associé à certaines occurrences. Il s'agit par exemple de calculer quelque chose voisin de *[document comptable]* pour *brouillard* dans : *le solde du brouillard[23] est incorrect*.
(b) appliquant ce même processus, il devrait pouvoir être possible de dégager des thématiques d'un texte.
(c) appliquant ce même processus, il devrait être possible de réaliser un dictionnaire à l'envers.
Prenons, le chapeau suivant paru dans un Ouest-France de cette époque :

> *L'accident a eu lieu par temps de brouillard. Les deux voitures qui se sont percutées sur les deux voies de la nationale ont provoqué un carambolage d'environ cinquante véhicules.*

Ce texte soumis aux calculs devrait retourner *carambolage* du fait que ce mot particulier rend compte, au plan componentiel de la plupart des autres mots de la phrase. Ici, intuitivement et par exemple, *carambolage* rend compte des mots *{accident {avoir lieu, provoquer, percuter}}* d'un côté et des mots *{voiture{véhicule},{voie, nationale}}* de l'autre.
Des observations marquent le caractère toujours actuel de ces questions. Il s'agit de considérer par exemple :
- la toute récente machine à traduire de Google[24] comme les différentes versions de Systran™

[22] Qui ne comprend que des éléments appartenant à un même moteur d'inférence.
[23] brouillard n.m. Livre de commerce, où l'on note les opérations à mesure qu'elles se font (cf. Main courante*) .(Le Petit Robert; ce sens existe d'après le Littré depuis au mois le XVIème siècle et est d'usage courant dans la gestion des organisations).
[24] http://www.google.co.uk/ language_tools

et d'autres restent pourtant d'accord sur un même résultat. Vers l'anglais, nous avons inévitablement : *the balance of the fog is incorrect*.
- les topiques : la recherche sur Google France de *samouraï* retourne un article sur trois présentant le topique de façon claire.
La première partie de ces années a été largement occupée par des tentatives de calcul qui n'étaient pas optimales. Je me rappelle des deux raisons suivantes :
- technique : la puissance de calcul à notre disposition (un 286) ne nous permettait pas d'imaginer des algorithmes "holistiques" exploitant l'ensemble des combinaisons du DAG.
- les contraintes pratiques : pour traiter du texte, il convenait au moins de traiter un peu la morphologie, avec en particulier la lemmatisation du français.

Au début de ces années, la technique nous a conduit à compiler (mettre dans une certaine forme pour permettre des temps de calcul raisonnables, de moins d'une journée pour un texte simple) une première version du Sémiographe. Cette version exploitait les données du DAG en perdant beaucoup d'informations puisqu'elle effectuait une projection sur une simple droite. Ses résultats ont toutefois été intéressants et débouchèrent sur le projet européen CRISTAL (page 29).

Durant cette période, nous nous posions des questions qui nous paraissent étonnantes aujourd'hui. Nous reproduisons ici trois de ces questions qui ont été discutées plus tard par la communauté :
(a) quel est le mot-sens impliqué dans un énoncé? Un jeu de catégories sémantiques est-il suffisant pour le caractériser et l'indexer [Wilks 1999]?
(b) pour résumer un texte, devons-nous rechercher des concepts généraux [Chauché 2003]?
(c) le parcours d'un réseau lexico-sémantique comme WordNet se fait-il de bas en haut ou de haut en bas [Agirre 1996]?

La question (a) est pour nous toujours absurde. Prenons par exemple : *l'avocat dont_je_parle aime les femmes*. Pour *avocat*, avons-nous un juriste, un défenseur, un plaideur, un être vivant, une personne, un homme etc. ou bien tout autre chose qui se ferait appeler *avocat*. Tout dépend en fait du contexte marqué *ce_dont_je_parle*. Voyons simplement qu'il pourrait s'agir de presque n'importe quoi d'autre, et en particulier d'un *fruit*, pourvu que ce n'importe quoi d'autre soit susceptible de supporter la prédication *aimer les femmes*. Du fait que par des figures courantes, nous trouvons nombre de marques, de produits et de services qui aiment les femmes (ou les enfants, ou les ados etc.), l'*avocat* co-défini par *ce_dont_je_parle* pourrait être celui-ci :

> *HUILE D'AVOCAT MELVITA (8€40)*
> *L'avocat aime les femmes! Excellente huile anti-ride, l'huile d'avocat principalement recommandée pour les peaux très sèches, elle peut être utilement préconisée pour le contour des yeux et les soins du cou. On la recommande également pour la prévention des vergetures en association avec le beurre de karité dont les insaponifiables sont remarquablement complémentaires. Elle présente parfois un dépôt tout à fait naturel.*

Pour bien comprendre notre propos, comparons l'énoncé amusant que nous venons de donner à quelque chose comme *l'avocat aime les sols argileux*, ou, pour se prêter à encore moins d'interprétation, *l'artichaut aime les sols sablonneux*. L'important devrait maintenant mieux apparaître. Le problème n'est pas un problème de catégorie ou de nature – du type, s'agit-il d'un homme de loi, d'un fruit ou d'un légume ? Mais d'un problème voisin de *selon quels points de vue* avocat *est-il un homme de loi* ou *un végétal étant donné le co-texte considéré*.
Ce sera seulement à partir de 1996 que nous commencerons à avoir pour certains cas une hypothèse directrice (voir 4.1.2.2.3 *L'activation componentielle,* page 38). Mais ce ne sera qu'en 2005 qu'une technique plus générale sera conçue ; nous présentons cette technique en 6.3. Ce dernier chapitre proposera une prise en charge minimale et endogène de la prédication.

De nombreux travaux postérieurs à 1996 ont pris pour hypothèse une réponse positive à la question (b). C'est ce que nous fîmes en 1992/1993 pour finalement rejeter l'approche dès 1994. Les raisons rétrospectivement peuvent être formulées très simplement. Il suffit de rapporter la question (b) à la solution proposée dans l'article Coling 1992. Nous voyons clairement dans l'exemple sur les accidents d'automobile qu'indexer les concepts *[accident]*, *[voiture]* et *[route]* (par exemple) serait bien moins précis qu'indexer un mot congruent à l'ensemble de ces concepts, quand ce dernier existe. Or, précisément le mot *carambolage* existe. Et c'est bien ce que nous avons proposé. Il reste à comprendre pourquoi nous avons voulu, malgré cela, indexer des généralités. La raison est finalement toute simple : nous n'avions pas alors de méthode de calcul de la solution proposée.
Nous présenterons cet algorithme comme un résultat d'étude de la période 1996-2000 : les limites imposées par les temps de calcul ne nous ont permis de travailler dans cette direction qu'à partir de 1996. Ce point est compatible avec notre réponse (a) donnée ci-dessus.

La question (c) trouve la réponse suivante : il ne faut pas parcourir le DAG. L'organisation des concepts définit une topologie relativement simple et il convient de réaliser des calculs de *repérage* rapide sur cette topologie indépendamment des situations haut-bas ou de la taille (en nombre de feuilles) d'un concept. En particulier, les calculs ne sont aucunement matriciels ou vectoriels (voir 4.1.2.2 *L'API de calcul de distance sémantique*, page 36).

3.1.2 L'enrichissement du modèle : le Dictionnaire Intégral (LDI)

Comme son nom l'indique, Dicologique était une ressource de sémantique lexicale plutôt simple. A propos des constructions fortement ensemblistes et organisées sous la forme de treilles Jean-Pierre Desclés [1981, p 134] a pu écrire : *seule une présentation formelle sous forme de treille permettrait de montrer comment se constitue un énoncé (plus généralement une famille structurée d'énoncés.* Cependant, nous commencions à faire attention à des phénomènes de circularité récurrente. C'était une constatation liée à une expérience et non à ce moment-là à une hypothèse théorique admise. Soit une série d'inclusions valant un jour *A inclus dans B inclus dans C*. Cette série pouvait prendre un peu après une forme *C inclus dans B inclus dans A*, et, au prix d'un nouvel effort, d'un nouveau point de vue ou d'une nouvelle hésitation, reprendre la forme *A inclus dans B inclus dans C*. Il est alors devenu plus favorable d'accepter ces circularités, et, à moins d'en faire une amie, de se contenter de programmer les parcours récursifs de listes de telle manière que ces dernières soient rompues[25] une fois donné leur contenu.

Mais à l'époque, la réponse technique fournie par Dicologique convenait parfaitement et nous laissait du temps pour nous préoccuper d'autres questions de sémantique lexicale. À partir de 1992, considérant les faits que nous allons énumérer, d'autres besoins sont apparus. Nous avons introduit : les fonctions lexicales, les génériques, les liens potentiels, le multilinguisme et les niveaux de langue, la morphologie et les constructions.
Chacun de ces ajouts a été motivé par des problématiques particulières. Pour chaque ajout,

[25] Aujourd'hui, nous savons que loin d'être à éviter, les circularités pourraient être, dans un schéma plus général que la théorie des ensembles, un moyen efficace pour engendrer une dynamique dialectique susceptible de conduire à une description économique de bien des phénomènes. Considérant une circularité apparente comme celle de *arbre / fruit*, il faut arriver à noter l'information implicite pour nous que l'arbre obtenu du fruit ne donne pas le fruit dont il est issu mais d'autres fruits. La théorie analytique des individus ($Pa \wedge \exists a \implies \exists x Px$) de Strawson [1959] fournit un cadre particulièrement bien défini de toutes ces questions et en souligne l'importance. C'est un élément qui nous amènera à nous intéresser plus tard aux hypergraphes.

nous donnons les principales motivations et le résultat.

A] Introduction des fonctions lexicales

Fondamentalement, les fonctions lexicales proposées par Mel'çuk dans la Théorie Sens↔Texte (TST) constituent le noyau technique de cette théorie dont nous rappelons la thèse centrale : *une des tâches primordiales de la linguistique théorique contemporaine est l'élaboration d'une théorie de la paraphrase langagière* [Mel'çuk, 1992, p10]. Rappelons-en le premier postulat : *La langue naturelle est (considérée comme) une correspondance multi-multivoque entre un ensemble dénombrable de sens et un ensemble dénombrable de textes* [page 14].
Cette théorie s'inscrit donc dans une perspective strictement compositionnelle à laquelle nous avons vu que nous sommes faiblement liés. Cependant, observant que le cadre componentiel permet mal, à lui seul, dans les moyens que nous avions, d'attribuer la description faite d'un mot-sens (*Ile-de-France* n.p ou *acheter* v.t) à celle d'un autre mot-sens (*francilien* adj, *achat* n.m), les fonctions lexico-sémantiques les plus courantes d'une langue donnée nous ont semblé très adaptées pour pallier cette difficulté.
Parmi les FL proposées par Mel'çuk pour le français, nous n'avons retenu que celles qui sont directement en relation avec la syntaxe de français. Ainsi, nous avons éliminé :
- les FL qui conduisent à la création d'une métalangue importante comme par exemple : CRIER(dindon) = glouglouter, considérant que celles-ci sont calculables automatiquement dans le graphe dans le graphe de concepts.
- les FL redondantes des nôtres.
Actuellement, les 30 FL gérées correspondent à environ 50.000 instances de fonction. Une part importante de ces relations a été instanciée automatiquement dans notre DAG, depuis une exploration de dictionnaire lui-même.

B] L'introduction des génériques

La *Figure 3 Deux sens de projeter dans le dictionnaire*, page 16, fait apparaître une telle relation. Un générique est un mot qui désigne une classe. Voir la définition de Dubois page 15. Dans l'exemple de Pottier (page 10), *siège* fait figure de générique de la classe [*siège*]. Il est possible avec les génériques de retrouver ses spécifiques : le terme générique d'une classe entretient avec les termes spécifiques de la classe une relation d'hyperonymie : cela dépend de ce que nous souhaitons faire. L'utilisation de l'implémentation de générique plutôt que celle d'hyperonyme présente quelques avantages. Nous ne pouvons citer ici que deux d'entre eux. Premièrement, il est possible de pointer en cas de besoin sur le terme générique lui-même plutôt que sur sa classe. Cela peut éviter des héritages peu idiomatiques ou tout-à-fait faux. C'est à notre sens ce genre de confusion, d'absence de frontière (voir Observation 3, page 18) qui fait échouer encore aujourd'hui les grands réseaux sémantiques. Pour la même raison, les travaux d'extraction automatique de clusters statistiques devraient continuer à donner, comme il le font depuis toujours, des résultats localement bons et devenant mauvais dans le passage à l'échelle. Deuxièmement, il est possible de considérer qu'un mot-sens est générique pour plusieurs classes ; cela ajoute de la souplesse sans créer d'homonymie artificielle.

Dans l'idéal, le modèle componentiel que nous développons aurait pu se passer de la relation de généricité : après tout si un mot-sens appartient à une classe et à rien d'autre, c'est qu'il est entièrement défini par la classe. En cela, il recouvre la classe et en devient un générique calculable. Mais la deuxième remarque précédente rend impossible l'application de cette heuristique. De plus, l'heuristique suppose que le dictionnaire soit juste et complet, ce qui évidemment ne sera jamais le cas (voir Observation 1 page 14). En définitive, la détermination du caractère générique d'un mot-sens pour un concept donné revient pour le moment à l'expertise humaine. Cela n'empêche pas qu'une partie des génériques du Dictionnaire Intégral aient été proposé à la validation humaine par l'ordinateur travaillant sur sa base de données.

C] L'introduction des liens potentiels

Certaines relations entre mots et concepts apparaissaient clairement comme définitoires et d'autres comme potentielles, presque encyclopédiques. C'est le cas de *bras* dans l'acception *pièce allongée plus ou moins mobile* qui s'applique à la description d'une série d'objets (*fauteuil, brouette, grue etc.*). Ce qui est en cause est cette série d'objets. Nous obtenons alors le graphe suivant :

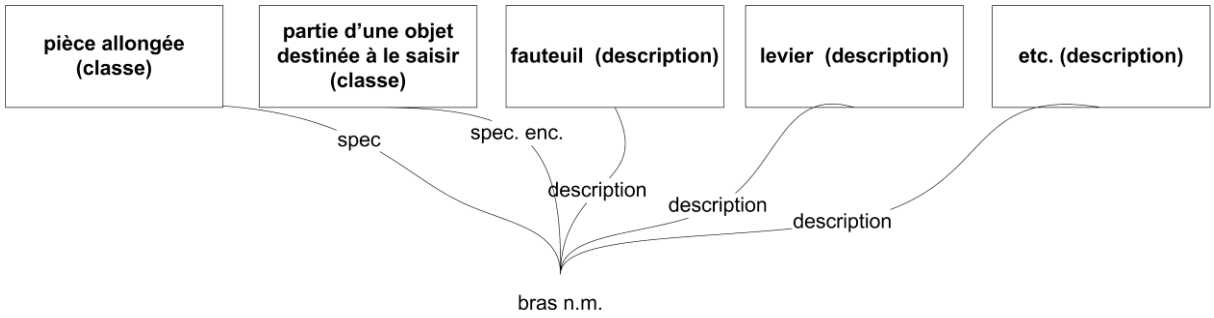


Figure 5 *bras* dans le dictionnaire

Dans la figure 5, les relations *spec.enc(yclopédique)* et *description* sont dites accessoires : la définition de *bras* dans ce sens n'implique pas que nous soyons en attente nécessaire du trait *fauteuil.*

D] L'introduction du multilinguisme

L'introduction des fonctions lexicales entre mots (comme *relatif à/qui appartient à etc.* pour *francilien* dans sa relation avec *Ile-de-France*) et l'introduction des types de relation (comme ici *générique* et *spécifique*) a entrainé une modification du modèle informatique qui permet également d'envisager le multilinguisme.
Le projet européen CRISTAL (page 29) nous a permis de proposer le modèle componentiel à une équipe de recherche italienne et à une équipe anglaise (un directeur de recherche de chacune de ces équipes vint participer à mon jury de thèse). Ce modèle est bien plus souple que celui de WordNet [Fellbaum 1998] et il est assez dommage qu'il ne soit pas implémenté par l'équipe de Princeton. En effet, en alignement des concepts multilingues, avec WordNet, nous sommes souvent amené à choisir entre mettre ou ne pas mettre une équivalence. Cette difficulté est directement liée à l'absence d'un système de traits conceptuels (non lexicaux) dans WordNet.

L'exemple du gâteau des Balkans

Par exemple, les turcs font un gâteau traditionnel, en suivant une recette défini, qu'ils mangent le matin. De leur côté, les grecs utilisent exactement la même recette, mais mangent le gâteau le soir. Nous voyons bien que la définition du gâteau n'est pas la même quand bien même le gâteau est matériellement semblable. Le modèle componentiel permet de noter ce genre de chose en procédant ainsi : création d'une classe très précise qui ne contient que les deux gâteaux, et distribution des différences : l'un est du matin, l'autre est du soir. Nous remarquons ici la mise en œuvre de traits conceptuels bien utiles. Des réseaux strictement lexicaux comme le DEC ou WordNet rencontrent des difficultés pour noter ce genre de chose. Dans WordNet, il est possible d'employer la relation SIMILAR et dans le DEC il faudrait créer deux entrées hyponymes de gâteaux. Dans les deux cas de figures, il n'est pas possible d'aller plus loin pour noter la différence componentielle.

E] L'introduction du niveau de langue, de la morphologie et des constructions

Pour de plus amples détails sur cette section, voir Dutoit [2000].

E1) Niveau de langue
Nous l'entendons de quatre façons :
- langue de spécialité : sens ou emploi particulier d'un mot dans un domaine donné tel que ce sens l'éloigne de son emploi courant et que l'emploi soit réservé à ce domaine (ex. *racine* en mathématique).
- niveau de langue proprement dit : introduction d'une marque concernant la manière de s'exprimer du locuteur (tournure de style, par ex. *littéraire : extraction* pour origine sociale, *vulgaire : caisse* pour voiture) ou son rapport à l'objet en référence (par ex. péj. *poubelle* pour voiture).
- lieu d'emploi : régionalisme (Québec : *dépanneur*)
- datation (ancien : *orthopnée*)

E2) Morphologie

Elle détaille le paradigme des mots-sens simples et des mots-sens composés. Le rattachement de la morphologie aux mots-sens plutôt qu'au mot a été expliqué dans ma thèse : il permet d'exprimer des restrictions éventuelles en ce domaine.
Pour le français, la forme prise par cette morphologie est un code de modèle flexionnel voisin de celui du DELAS [Courtois, 1990]. La forme a été reprise par la suite pour de nombreuses langues européennes (italien, anglais, espagnol, portugais, russe, allemand et néerlandais).
La morphologie des formes composées n'a pas été parfaitement décrite et est restée à part, du fait que nous ne pouvions pas faire de lien entre "l'intérieur" d'une expression et un point particulier du DAG pour reprendre un code de flexion simple existant (voir 6.1.1 *Intégration de la morphologie compositionnelle,* page 77, et 6.1.4 *Intégration du terme* , page 87 pour une mise en perspective).

E3) Construction
Un verbe (ou un nom, ou un adjectif) connaît pour chaque sens des constructions particulières qu'il est bon de donner pour différents usages (information au lecteur, analyseur syntaxique). Cette propriété que nous venons de donner n'est pas compatible avec celle utilisée par Caput [1969]: *chaque verbe peut avoir une ou plusieurs constructions indépendamment de son sens.* Environ 30.000 descriptions de construction ont été données en relation avec le sens mais, à l'instar de la morphologie des formes composées, et pour une raison voisine de gestion d'éléments à plusieurs éléments, ces résultats n'ont pas encore été utilisés.

Considérant l'ensemble de ces apports, et la variété des points de vue sur la langue qu'ils apportent, nous avons forgé le terme *Le Dictionnaire Intégral* (LDI) pour refléter l'idée de la confection d'un objet plutôt formel capable de rendre compte de ces différents points de vue

lexicologiques ou linguistiques.

Une autre raison nous fit retenir le terme LDI. En effet, *intégral* rappelle le fondement componentiel – intégrant - de l'ossature fonctionnant par décomposition/recomposition.

Définitions complémentaires

Fonction lexico-sémantique (FL) [Mel'cuk, p31] *Une fonction lexicale (FL) est une dépendance, ou correspondance, **f** qui associe à une unité lexicale* (pour nous, un mot-sens) L, appelée l'argument de ***f*,** un ensemble d'unités lexicales ***f***(L).

Relation générique Relation mot-sens à concept. Un mot-sens est générique pour une classe s'il exprime sans ajout ni retrait la classe considérée.
Si un même concept [C], comporte un générique G et un spécifique S, alors G est hyperonyme de S.
Il existe une relation *générique taxonomique* qui est réservée aux classifications systématiques type zoologie. Cette relation distingue un générique comme *animal* ou *renard* (terme courant) d'un terme comme *eumétazoaire* (taxème technique causant un emploi généralement pluriel, d'emploi rare et de dérivation adjectivale en français systématique).

Modifieur de saturation Relation mot-sens à concept. Exprime l'idée que le mot-sens considéré n'a pas nécessairement à être saturé[26] par le co-texte.
Par défaut, les traits de sens doivent être saturés. Un modifieur de saturation change toujours cet état des choses[27].

Relation interlingue Ces relations sont données par une fonction lexicale *Se traduit par* ou par une relation différentielle entre concepts (voir l'exemple *gâteau des Balkans* page 26).

[26] Il y a deux types d'éléments saturés dans le sens de Frege [1892] : des phrases complètes et des noms propres, parce qu'ils ne prennent pas d'arguments et ne sont pas des fonctions. Toutes les autres unités sont considérées comme non saturées. Nous remémorant notre réponse à notre question "*quel est le mot-sens impliqué dans un énoncé?"*, nous prévenons que sur cette question des éléments saturés nous allons obtenir un résultat presque opposé à celui de Frege : les noms propres impliquent un très grand nombre de relations qui leur sont précisément propres.

[27] Il est possible de concevoir que le modifieur de saturation est sur le plan des isotopies lexico-sémantiques un équivalent des attributs encyclopédiques qui ont fait glisser la sémantique du prototype du statut de standard à celui de étendu (dans la terminologie de [Kleiber 1990, p156]. Notons que si le motif est le même, notre point de vue est strictement inverse de celui de la sémantique du prototype. En effet, nous ne nous intéressons aucunement au statut des catégories, en-dehors de leur pouvoir de catégorisation, mais seulement au statut d'un mot dans une catégorie, étant entendus les cadres d'utilisation dudit mot. Or, pour reprendre l'exemple parfaitement classique du *pied* et de la *chaise*, le point de vue du *pied* qui est le nôtre, nous amène à considérer l'ensemble des utilisations de ce mot en tant que *support vertical* par lequel *chaise*, *falaise*, *mur*, *escalier*, *fauteuil* etc. touche idiomatiquement le sol. Ce qui compte pour nous n'est aucunement le prototype de *chaise* mais la description de toutes ces occurrences du mot pied. Nous ne reviendrons pas ici sur la sémantique du prototype sauf parfois pour rappeler cette différence essentielle de point de vue. Pour nous, la sémantique du prototype reste une école idéaliste et aristotélicienne en cela qu'elle s'intéresse aux abstractions dues au classes et cherche à les justifier au plan cognitif. De notre côté, nous ne nous intéressons qu'aux signes linguistiques et cherchons à en rendre compte depuis les mots et les textes. Dans notre document, l'utilisation faite des artefacts que sont les concepts ne relève que du procédé et de l'économie et non d'un quelconque statut psychologique.

Niveau de langue et domaine : spécification de contraintes ou d'effets particuliers liés à l'emploi d'un mot (langue de spécialité, niveau de langue proprement dit, lieu d'emploi, datation).

L'ensemble des adaptations du dictionnaire que nous venons de présenter ont permis de réaliser une première version intéressante, c'est-à-dire non limitée à un domaine (la langue est une structure – voir **position linguistique** page 12) et vérifiant l'ensemble des opérations qui avaient été annoncées en 1992 (voir chapitre 3.1.1 ci-dessus) comme nous le verrons entre 2002 et 2004 (voir 5.1.2, page 54). Durant les années 1992-1996, environ 70.000 mots-sens de LDI français ont été mis à jour.

3.2 Projets et documents

Projet

En début de période, nous avons obtenu deux soutiens du Ministère de la Recherche qui nous ont permis de débuter sérieusement le travail :

Projet 1 Dicologique

Définition des principes du dictionnaire.

Projet 2 Amélioration de Dicologique

Mise à l'épreuve d'une instanciation plus large ; les critères de qualités sont la progression (absence de régression) et la productivité du poste de travail lexicographique.

Les efforts précédents ont permis de déboucher sur un projet européen en indexation conceptuel trilingue : CRISTAL (Références 92K6451 et FRT9501).

Projet 3 CRISTAL

A cette époque où régnaient en France GENELEX et GRAAL, nous avons eu la chance de construire, de gagner et de gérer au plan scientifique le projet européen Conceptual Retrieval of Information using a Semantic dicTionary for Access in three Languages (LRE62059, 1994). Les partenaires de CRISTAL étaient l'Université de Manchester, le CNR de Pise, la société CAP GEMINI INNOVATION (intégrateur) et l'Européenne des Données (utilisateur).
CRISTAL se proposait deux tâches :
- d'une part, l'usage d'un module de synonymie interlinguale permettant d'interroger en français et d'obtenir des résultats en français, en anglais ou en italien. Ce module était limité parce qu'aucun partenaire du consortium ne possédait de données multilingues au démarrage et qu'il nous fallait de plus réécrire pratiquement tout Lexidiom pour prendre en charge les modifications que nous venons de décrire.
- d'autre part, une indexation reposant sur les concepts des documents. Nous avons déjà signalé les limites de ce type d'indexation. Elle nous était toutefois imposée par le rythme des livraisons du projet et par la technologie proprement dite.

Stage, équipe

Equipe : 1993 a vu le retour en son Alsace de notre ami Philippe Herr et l'accueil en bourse CIFRE de Yann Picand. M. Picand partage toujours nos travaux avec excellence mais n'a jamais voulu réellement rédiger de thèse.

Publication, dissémination
Trois articles nous ont cités d'une façon très importante :

Jean-Pierre BALPE [1992] Comment l'informatique donne du sens aux mots (La Recherche)

Cet article a présenté sur 3 pages le Dictionnaire Intégral en fournissant des exemples et en fournissant des hypothèses sur ses utilisations.

H. BEHAR [1994] Pour une problématique des odeurs : des essences pour des Esseintes (Etudes Françaises, N°31, 1995).
Cet article utilisa Dicologique pour constituer une partie de son corpus sur les odeurs et réaliser son étude.

Philippe HERR et Yann PERRAIS [1995], La représentation/construction du sens dans les dictionnaires électroniques, édition Scolia, sciences cognitives, linguistique & intelligence artificielle, sous la direction de G. Reb, publication de Proparlan, ERS du CNRS n°125 et Université des Sc. Humaines de Strasbourg.

Pour notre part, nous avons publié:
Dutoit [1992], A set theoritic approach to lexical semantics, International Conference on Computational linguistics (CoLing, Nantes)
L'article décrit comme son titre le montre un algorithme ensembliste de localisation de la trace laissée par une conjonction de mots d'un texte dans le DAG (que nous appelons aujourd'hui activation) et quelques utilisations potentielles : désambiguïsation, dictionnaire à l'envers, thématique etc.

3.3 Conclusion et perspectives

Entre 1992 et 1996, nous avons :
- réalisé une version commercialisable de Dicologique
- réfléchi aux traitements automatiques que le dictionnaire pourrait permettre de mettre en œuvre et nous avons publié à propos de ces hypothèses.
- présenté notre travail à quelques chercheurs en linguistique.
- évalué le point d'entrée permettant une extension multilingue
- recherché d'autres manières de structurer le lexique telles que ces manières permettent de nouvelles inférences et une économie dans nos coûts de maintenance.

Les objectifs que nous nous étions donné ont donc été satisfaits. Mais la période qui s'achève en 1996 laisse apparaître un grand chantier lexicographique qui comporte quatre directions :
- une dimension morphologique puisqu'il nous faut si l'on souhaite traiter de textes gérer convenablement le paradigme flexionnel
- une direction compositionnelle avec la TST qui nous invite à nous intéresser à la paraphrase
- une direction componentielle dont il faut affiner la définition du contenu par des expériences

de traitement automatique (d'inférences automatiques)
- une mise en œuvre de ces inférences automatiques à des fins de TAL
- une direction extensionnelle avec le projet de dictionnaire conceptuel multilingue.

Dans ces conditions, les travaux de la période suivante devront contribuer à stabiliser ces directions et à mieux définir certaines notions.

En ce qui concerne le traitement automatique des langues utilisant le réseau componentiel, la question qui nous intéressait le plus était en étroite relation avec l'Observation 3.a, donnée page 18. Cette question peut se formuler ainsi :

Question 2 Considérant plusieurs mots, comment extraire automatiquement ce qui les rapproche et ce qui les distingue sémantiquement ?

Nous observons toutefois que la *Question 1 Comment enregistrer des concepts structurés dans le DAG ?*, page 20, n'a toujours pas de réponse.

4 STABILISATION DES TRAVAUX, NORMALISATION ET MULTILINGUISME (1996-2001)

Cette période est marquée principalement par des travaux de consolidation ou d'évaluation. Le temps n'est pas aux idées nouvelles car la période précédente a laissé beaucoup de friches à valoriser. Il s'agit aussi de réunir les matériaux qui permettront la rédaction de ma thèse qui sera effectivement soutenue fin 2000. En matière de calcul automatique, il est temps de concevoir les programmes qui utiliseront convenablement les données disponibles depuis la morphologie jusqu'aux évaluations sémantiques. Par ailleurs, en matière de multilinguisme, nous considérons que notre visibilité et donc notre survie suppose qu'à l'horizon 2000 nous ayons un système gérant, en terme de synonymie interlinguale au moins six langues européennes.

4.1 Les travaux de recherche

La période aboutit à une version relativement stable du Sémiographe. Cette version utilise plusieurs des modifications fonctionnelles du Dictionnaire Intégral conçues précédemment. La période ne produit pas de modification du modèle de dictionnaire : il s'agit déjà de tester tout ce que nous pouvons calculer à l'aide de la ressource en fabrication.

4.1.1 Le développement du dictionnaire

Ce développement concerne surtout le multilinguisme. Il s'est agi d'introduire environ 40.000 mots-sens pour chacune des cinq langues suivantes : anglais, italien, espagnol, portugais et allemand.

Le développement de ce dictionnaire a débuté par la sélection des 40.000 mot-sens français. Nous avons procédé ainsi :

- processus de sélection des mots-sens [28]:

* tous les génériques non taxonomiques (voir *générique taxonomique* page 28)

[28] La méthode est assez comparable à celle employée par Vossein [1999]

présents dan le dictionnaire.

* complément par les lemmes les plus fréquents présents selon la base textuelle décrite à la note 21, page 20.

- recrutement sous forme de stages rémunérés (logement, frais) d'au minimum 6 mois d'étrangers désirant se perfectionner en français.

Les consignes générales étaient d'abandonner la tentative de traduction d'un mot-sens:

- si la tentative de traduction dudit mot-sens est mal reconnaissable à travers les concepts et extension du DAG et que les traductions viennent mal à l'esprit.
- si l'emploi semble technique et qu'il a peu de chances d'être attesté dans les dictionnaires bilingues papier généraux que nous utilisions régulièrement (pour ne perdre de temps en recherche infructueuse).

L'objectif de 40.000 mots-sens par langue n'a pas été tout-à-fait atteint pour chaque langue (en moyenne 37.000 mots sens avaient été traduits).

4.1.2 La fabrication du Sémiographe

Le Sémiographe (LS) considère LDI comme un ensemble de lieux sémantiques et évalue la cooccurrence de ces lieux dans les textes ou leur succession, à travers les fonctions lexicales dans une optique de génération. Dans cette section, nous ne détaillons pas l'ensemble des opérations et outils exploitant le dictionnaire et utiles à la réalisation d'applications. De nombreux outils sont seulement fonctionnellement définis en tant qu'ils participent à la réalisation des applications que nous décrivons tandis que, à contrario, quelques outils comme l'outil de "distance" sémantique, sont plus détaillés car ils sont centraux.

L'ensemble des outils du Sémiographe sont organisés comme une bibliothèque au service du développement d'applications de traitement automatique de l'écrit. Ces API (*application programming interface*) sont écrites en Java.

Nous présentons successivement ces API telles qu'elles étaient disponibles en 2000 :

- les APIs phonétiques, morphologiques, morpho-syntaxiques et d'expansion lexicale du Sémiographe
- les API de distances sémantiques
- quelques exemples d'« application » enchaînant ces API.

Certains des exemples donnés retracent des résultats de projets industriels ou de recherche.

4.1.2.1 APIs phonétiques, morphologiques, morpho-syntaxiques et d'expansion lexicale

A] Le reconnaisseur de langue

Cette API est capable d'identifier la langue d'un texte écrit dans l'une des cinquante langues pour lesquelles il a été conçu. Il fonctionne à l'aide de n-grammes [Grefenstette 1995].

B] L'API phonétique

Cette API est disponible pour deux langues : l'anglais et le français. Elle sert principalement à effectuer des opérations de correction orthographique[29].

[29] Dans un premier temps, le jeu d'environ 1000 règles prédisait 100% du corpus de Pérennou (environ 300.000

C] L'API morphologique

Elle lemmatise/fléchit les mots des cinq langues suivantes : français, anglais, italien, espagnol et portugais. Le site sensagent.com illustre son fonctionnement. La thèse que Suzanne Pereira soutiendra en décembre 2008 fait un usage important de cette API.

D] L'API séquence répétée

Elle extrait et compte toutes les séquences répétées trouvées dans un fichier. Cette API est pilotée par un fichier de configuration assez riche.
Le fichier de configuration permet de déterminer :
- sur quoi porte l'extraction (des lettres, des mots, des lemmes, des parties du discours, des concepts de LDI)
- de mêler des données en une même séquence, par exemple pour constituer des séquences mélangeant des concepts et des parties du discours.
- de définir, puis d'accepter ou de refuser automatiquement des séquences ambiguës.
- de définir les longueurs maximales des séquences enregistrées
- de définir des débuts et des fins de séquence

L'API mémorise des séquences de longueur maximale paramétrée (nous utilisons en général 9 comme valeur de ce paramètre) et nous sert à construire des modèles de langage depuis des corpus non étiquetés.

E] L'API morpho-syntaxique

L'API fonctionne à partir de séquences collectées par l'API précédente et résout les conflits à la manière d'un modèle de Markov en travaillant sur la perplexité (l'écart entre phénomènes certains et phénomènes incertains). Il s'agit d'apprentissage non supervisé, stochastique et sans corpus d'apprentissage dédié.

Pour le français, la base d'apprentissage est décrite en note 21 page 20 ; dans cette base nous avons tout de même éliminé des textes écrits en vieux français comme Pantagruel. Pour le reste, la base d'apprentissage est restée hétérogène (poésies, romans, essais, pièces de théâtre).

F] L'API d'expansion lexicale

Cette API permet de générer, pour l'expansion le développement de paraphrases simples, un lexique à partir d'un mot, suivant un script donné. Les fonctions disponibles dans le script sont données en note de bas de page[30].

---

formes) puis a été dégradé pour mieux satisfaire à des besoins de correction (phénomènes irréguliers mais fréquents de translittération etc.). On la trouve aujourd'hui sur certains services grands publics offerts (Universalis, Orange, Sensagent etc. dans des versions plus ou moins complètes). Aujourd'hui, Alexandria ou le site sensagent.com en implémentent une version rapide et simplifiée.

[30] **ALIAS**: La fonction permet de récupérer les synonymes stricts d'un mot. Par ex., les variantes graphiques : acuponcture ↔ acupuncture.
DERIVED: La fonction permet de récupérer les dérivés lexico-sémantiques d'un mot. De Gabon→gabonais à citronnier↔citron en passant par rêve→onirique ou alimenter→aliment. (la flèche → représente une fonction lexicale). Nous n'avons pas indiqué ici la nature de cette fonction).
TRANSLATION: La fonction permet de récupérer les traductions d'un mot dans une langue donnée. Cette langue doit être passée en paramètre.
INFLECTED: La fonction permet de récupérer les différentes formes fléchies d'un mot.

Le comportement de certaines fonctions d'expansion dépend des paramètres de configuration du système : il s'agit des fonctions DERIVED et BROTHER puisque toutes deux sont dépendantes du paramétrage de l'exploration de la structure.
Une variable globale agit également sur l'ensemble des fonctions. Il s'agit de la variable *sens uniques seulement*. Si *sens uniques seulement* est coché, la fonction SYNONYMS appliquée à *automobile* ne retournera pas *voiture* puisque le dictionnaire connaît un autre sens pour *voiture*. L'API d'expansion lexicale est par exemple employée pour l'expansion de requêtes. Elle intervient aussi dans le dictionnaire à l'envers.
La thèse de Thierry Poibeau utilisa largement cette API dont la fonction BROTHER pour évaluer son application d'extraction d'information avec dictionnaire, sans reconfiguration du dictionnaire ni apprentissage.

4.1.2.2 L'API de calcul de distance sémantique

L'emploi du terme *distance sémantique* est courant et nous le rencontrons dans des travaux d'appartenances diverses, en représentation des connaissances, en analyse lexicale et en analyse du texte. L'emploi de ce terme n'implique pas nécessairement chez les auteurs les trois restrictions euclidiennes portant sur la distance du même nom mais plutôt différents rapports de proximités qui sont quantifiés à l'aide de symboles ou de quantités. Nous donnons d'abord quelques-unes de ces définitions, puis nous fournissons notre définition et nous détaillons son implémentation en distance interlexicale.

4.1.2.2.1 Définitions de "distance sémantique"

En représentation des connaissances, nous pouvons rencontrer ce terme quand nous comparons deux ontologies. L'article d'Euzémat [2004] constitue un bon exemple de cet usage. Dans l'article, différentes définitions formelles de *distance sémantique* sont données puis appréciées qualitativement (distance au sens propre, proximité, similarité etc.). Finalement, le mot *distance* y apparaît comme un terme commode pour désigner un champs de travail où la notion de distance est redéfini en fonction de ce que nous voulons mesurer ou rapprocher, c'est-à-dire en fonction de points de vue particuliers.

En analyse lexicale, Mel'çuk [1995, page 87] utilise la définition suivante : *la distance sémantique entre les lexies L1 et L2 est mesurée par deux paramètre considérés ensemble : 1) la taille de la composante sémantique commune à L1 et L2 (plus cette taille est grande plus L1 et L2 sont proches) 2) la régularité de la distinction sémantique entre L1 et L2 (plus élevé est le nombre de paires de lexies où la même distinction apparaît dans la langue en question, plus L1 et L2 sont proches).* Soulignons que dans cette définition la *distance* intègre par (b) la notion non-symétrique de différence qui ouvre les champs à la pluralité des points de vue.

Il me semble que Brunet [2004] résume bien cette notion de points de vue à l'issue d'un

---

GENERICS: La fonction permet de récupérer les génériques d'un mot. Par défaut, la hauteur de remontée est fixée à 1. Ce comportement peut-être modifié en rajoutant le paramètre : H:n.
SPECIFICS: La fonction permet de récupérer les spécifiques d'un mot. Par défaut, la profondeur de descente est fixée à –1 (c'est-à-dire, pas de limites). Ce comportement peut-être modifié en rajoutant le paramètre: P:n.
SYNONYMS: La Fonction permet de récupérer les autres synonymes d'un mot.
GEOGRAPHY: La fonction permet de récupérer les toponymes associés à un toponyme.
BROTHER: La fonction permet de récupérer les mots situés dans la même classe et dont la distance sémantique (autrement dit, les différences) n'excède pas une certaine valeur.

parcours d'expérimentations informatiques où différentes approches (algorithmes) sont évaluées en terme qualitatif pour conclure par la définition suivante de *distance* : *La distance dans le discours est ce qu'elle est en peinture : une perspective, un point de vue*[3132].
Finalement, les trois définitions de distance que nous venons de donner nous conviennent assez puisque toutes incluent les notions de pluralité des *points de vue* dans un contexte où, de toute façon, tout critère que nous pourrions proposer entretiendrait, d'une manière ou d'une autre, qu'elles que soient les efforts que nous pourrions faire, des liens avec d'autres critères.
Voici notre définition :

distance sémantique *:* toute grandeur signalant par sa valeur l'existence d'un ou de plusieurs points de vues caractérisant des ressemblances et tel que depuis chacune d'elles il est possible de caractériser des dissemblances (considérant un grand nombre de points de vues qui tous ensemble caractérisent un tout d'ordre sémantique).

Nous remarquons que cette définition de distance sémantique est en rapport avec celle de structure (voir note 10, page 9) du fait que, par définition, tous points A et B présentant une ressemblance doivent être capables de caractériser leur dissemblance depuis cette ressemblance (1). Selon cette note, l'affirmation "il n'y a pas de primitive" prise sur le plan de la comparaison de deux signes est fausse car l'acte même de comparaison produit des primitives dans chaque instant où il aboutit (2). A contrario, dire qu'il n'y a pas de primitives dans la structure elle-même n'est pas faux : nous pouvons toujours croire en voir une, puis une autre et encore une autre (3).

Quelle est le statut de notre DAG par rapport à cette conception générale?
Premièrement le DAG décrit un certain nombre de lieux conçus sur les critères de dissemblance et de ressemblance; cela valide le critère (1) ci-dessus. Mais (2) et (3) ne sont pas validés : le système sémantique ne boucle pas sur lui-même. Nous verrons (chapitre 6 Intégration structurale des points de vue componentiels et compositionnels page 73) comment ce bouclage va devenir possible et nous donnerons des cas de l'utilisation de ce bouclage.

4.1.2.2.2 Les distances sémantiques chez nous

Du fait de l'Observation 5 page 22, il est important pour nous que les distances que nous établissons, modulo quelques ajustements justifiés particulièrement par le temps de calcul, puissent être utilisées dans différents contextes. Ainsi, nous avons utilisé les distances que nous présentons maintenant en alignements d'ontologies et en distance portant sur le discours. Nous verrons des exemples applicatifs en 5.1.2, *Le Sémiographe touche les applications*, page 54. Nous ne nous intéresserons ici qu'aux distances interlexicales pour en faire comprendre le fonctionnement.

Comme nous l'avons vu, le Dictionnaire Intégral surimpose deux graphes. Le premier dessine un graphe plutôt acyclique dans lequel les nœuds terminaux sont des mots, les autres nœuds des concepts et les arcs des relations. Le deuxième met en relation des mots à l'aide de

---

[31] Il ajoute, et nous soulignons les traits qui nous satisfont le plus: *Plus encore que le monde physique, l'univers du discours est soumis à la relativité. Faute d'un point d'appui unique, les mesures varient selon l'objet isolé, et la méthode choisie. Pourtant les paramètres qu'on croit isoler sont souvent liés entre eux, par l'effet d'une redondance ou surdétermination qui explique la convergence des résultats, comme si l'on photographiait une boule en variant les angles et les points de vue*.
[32] Voir notre citation de *Georges Braque* page 3.

fonctions lexicales. La figure 6 illustre dans une simplification extrême cette structure.

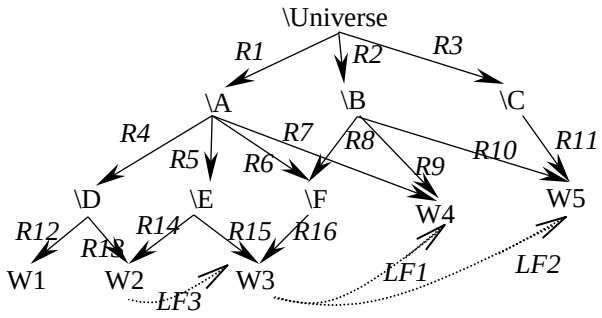


Figure 6 Un graphe de concepts, mots et fonctions lexicales pour illustrer le fonctionnement du Sémiographe

Dans la figure, les nœuds commençant par un « backslash » '\' sont des concepts tandis que W1, W2, W3, etc. figurent des mots. La racine du graphe est noté \universe : c'est l'ancêtre de tous les concepts. Il a trois fils, respectivement \A, \B, et \C, qui peuvent également être des classes ou des thèmes.
Les relations notées Rn relient les concepts entre eux et les relations LFn symbolisent les fonctions lexicales. Dans la figure, W3 a deux parents : R15(\E) = W3 et R16(\F) = W3. LF1 est une fonction lexicale reliant W3 à W4: LF1(W3) = W4. L'inverse des relations est accessible. Ainsi il est possible d'obtenir W4 de W3, par exemple : $LF1^{-1}(W3)=W4$.
Dans le dictionnaire, le nombre moyens de pères d'un concept est 2,1 ce qui, en soi, reflète la notion de différenciation au niveau local : les distances entre mots ou textes sont dérivées de la distribution locale des traits sémantiques. C'est la somme de deux mesures que nous avons pris l'habitude d'appeler *activation componentielle* et *différence componentielle*. Nous décrivons une version simplifiée de ces mesures qui ne prendra pas en compte les importantes altérations issues de la nature des relations[33] et de la nature des concepts.

4.1.2.2.3 L'activation componentielle

Cette mesure est la plus simple à effectuer. L'activation componentielle de deux mots M et N, est définie par la règle des plus petits ancêtres communs (Least Common Ancestors ou LCA) dans le graphe. Les plus petits ancêtres communs sont, parmi les ancêtres, ceux qui sont en position de fils et jamais en position de père. L'activation sémantique entre deux mots ou deux textes M et N est constituée par l'ensemble LCA de ces deux mots.
Dans la figure, nous avons

LCA(W2, W3) = {\E} et
LCA(W3, W4) = {\A, \B}

Voir *http://en.wikipedia.org/wiki/Tarjan's_off-line_least_common_ancestors_algorithm* pour une présentation détaillée de l'algorithme..
Le chemin d'activation entre W2 et W3 consiste en les nœuds W2 \E W3 avec les fonctions $R14^{-1}$ et R15. Les chemins entre W3 et W4 consistent en W3 \E \A W4 et W3 \F \B W4. Nous

[33] En réalité, la racine de LDI comporte une sous-racine des fonctions lexicales, une sous-racine des constructions, une sous-racine des niveaux de langue et domaine, et, pour l'essentiel, la sous-racine onomasiologique.

observons que nous obtenons toujours des sortes de « chapeau chinois ».
En considérant égales toutes les relations (ce qui ne vaut que pour simplifier le problème), nous définissons l'activation componentielle comme le nombre total d'arcs dans ces chemins divisés par le nombre de chemins :

d^(W2, W3) = (1 + 1) / 1 = 2
d^(W3, W4) = ((2 + 1) + (2 + 1)) / 2 = 3

Les LCA permettent d'extraire un espace de recherche componentiel et fournissent une sorte de mesure quantitative de quelque chose de qualitatif (voir la remarque de Deleuze note 10, page 9).

Observation 6 Nous faisons l'hypothèse que les LCA définissent tous les lieux relatifs à deux nœuds où quelque chose d'intéressant est susceptible de se produire. Ils sont les localisations (voir Observation 5, page 14) que nous recherchons et ils produisent des mesures en rapport avec la structure. Il est toutefois intéressant d'imaginer s'ils pourraient comporter d'autres types de productions[34].

Le principal problème que nous rencontrons pour parler de la *structure* et de ce que les LCA peuvent retourner et moins un problème d'expression mathématique, qui pour clair qu'elle soit n'illustre rien[35], qu'une difficulté à transmettre une expérience. Ainsi, nous attendons que le lecteur soit capable d'imaginer, prenant deux mots A et B, ce que le système des LCA devrait répondre étant donné le type d'information[36] représenté dans la structure sur lequel il repose. Ceci importe sinon pour une évaluation[37] du moins pour une prise en main.
Par exemple, que valent
1) LCA(renard_animal, glapir)
2) LCA(voleur, glapir)
et
3) LCA(samouraï, Tokyo_ville)?
Nous discutons de cela page suivante. Nous réutilisons la *Figure 4 Une illusion d'ontologie concernant renard,* page 17, pour présenter la solution de LCA(renard_animal, glapir).

[34] Voir chapitre 6 page 71 des exemples de ces productions.
[35] surtout pour un espace construit de telle manière qu'il représente toute l'hétérogénéité possible, et tous les cas possibles d'intrication et de critères liés tellement qu'il est impensable de les délier (voir note 31, page 37).
[36] des composantes sémiques.
[37] Nous utilisons d'ailleurs le système des LCA pour repérer des incohérences dans le graphe.

En étudiant la *Figure 4*, nous voyons que la comparaison entre *renard n.m.* et *glapir v.t.i.* retourne d'abord [renard]$_{(thème)}$ puisque :
- premièrement, [renard]$_{(thème)}$ est subsumé par tous les thèmes comportant [renard]$_{(classe)}$.
- deuxièmement, [renard]$_{(thème)}$ est subsumé par [faire qqch, s'agissant d'un animal]$_{(thème)}$ qui comporte les cris d'animaux.

Si le vrai graphe est conforme à cet extrait, il n'est pas possible que d'autres LCA apparaissent : en effet, en suivant le chemin de décomposition de *cri* (vers le son par exemple), il faudrait qu'une de ces décompositions comporte également *renard.* Il est plus vraisemblable de penser que cela ne se produit pas, et que seule la racine du graphe groupe à nouveau ces deux mots. Or, cette racine n'est pas LCA puisqu'elle contient [renard]$_{(thème)}$.
Cependant, il arrive souvent que le graphe soit plus riche et plus déséquilibré que notre exemple. Dans ce cas là, des LCA très hauts peuvent apparaître. Nous l'appelons le *bruit de fond* et nous acceptons volontiers de faire avec puisqu'il est aussi propre au champ de notre étude (voir note 31, page 37) qu'il l'est à celui de l'astronomie.
Prenant maintenant un autre mot que *renard,* par exemple *voleur* n.m., selon ce que nous trouvons dans [*ce qui est prit comme animal*]$_{(classe)}$, une comparaison d'un mot comme *voleur* et *glapir* retrouve [*faire qqch pour un animal*]$_{(thème)}$ qui est très haut mais éventuellement, aussi d'autres LCA (par exemple à travers l'*activité humaine*).
En définitive, si nous considérons tout LDI comme un système de règles du type (*if* SAMOURAÏ *and* SABRE *then* SAMOURAÏ-->[GUERRIER]), les LCA seraient pour le système la manière de retrouver les priorités dans l'application de ces (millions de) règles. Ils fournissent une localisation.
Que donnerait LCA(*samouraï*, *Tokyo*) ? Un LCA prévisible est [*Japon*]$_{(thème)}$. Mais nous pouvons toutefois imaginer d'autres LCA. Par exemple, considérant *samouraï* comme élément de l'organisation sociale (*noble*) et *Tokyo* de la même manière (*organisation citadine),* nous pouvons anticiper l'apparition de LCA éloignés mais toutefois plus ou moins prévisibles. Au fond, nous sommes capables de prédire avec une grande présomption les plus petits LCA tandis que les plus grands présentent un degré de présomption moindre[38].

Il nous faut maintenant considérer le "poids" spécifique aux relations.

a) Les modifieurs de saturation.

Nous avons introduit ces modifieurs page 28.
Ces relations sont traitées comme optionnelles et ne sont évaluées que si elles améliorent le score global du mot-sens étudié.

b) Spécifiques et génériques

Il est utile que l'activation componentielle trouve une différence de résultat entre *monnaie* comparé à *yen* et *yen* comparé à *dollar*. Pourtant, dans les deux cas, l'ensemble des LCA risque de se limiter à [monnaie]$_{classe}$. Toutefois les distances sont différentes : en effet, la longueur de la relation entre un terme générique et la classe dont il est générique est définie comme nulle (dans SEMREL).

[38] Dans la pratique, nous utilisons un jeu de meta-données appelée SEMREL qui, à l'aide des types de relation, décrit les profondeurs du graphe que l'on accepte d'explorer Les paramètres de SEMREL sont définis en fonction des contraintes de calcul. Une contrainte fréquente est le temps de calcul. SEMREL n'interfère pas sur le mode de calcul donné par 39. A contrario, en opposition avec la moyenne arithmétique que nous avons suggérée page 39, nous utilisons la moyenne harmonique (qui favorise les plus petites valeurs) et une moyenne de position : le premier quartile ou la médiane. C'est un moyen commode d'élimination du *bruit de fond* qui permet de se concentrer sur les résultats susceptibles d'être interprétés facilement : ceux qui viennent le plus directement.

4.1.2.2.4 La différence componentielle

La différence componentielle entre deux mots M et N utilise les plus petits ancêtres asymétriques LAA (Least Asymmetric Ancestors). LAA(M, N) est l'ensemble des nœuds communs aux deux mots qui ne sont pas membres des LCA, et pour lesquelles ces nœuds ont un fils qui est un ancêtre de M et n'est pas un ancêtre de N.
La plupart du temps, les ensembles LAA(M, N) et LAA(N, M) sont différents. Ce caractère essentiel de cette mesure reflète une différence componentielle qu'il est possible de dégager (voir 4.1.2.2.7 ci-dessous).
Dans la *Figure 6*, page 38, l'ensemble des ancêtres communs à W2 et W3 qui ne sont pas des LCA est {\A, \Universe}. \A a un fils \D qui est un ancêtre de W2 et qui n'est pas un ancêtre de W3, aussi LAA(W2, W3) = {\A}. L'ensemble LAA(W3, W2) = {\A, \Universe} parce que \F et \B sont fils respectifs de \A et \Universe et ancêtres de \W3 mais non de \W2.
Les LAA peuvent être très nombreux ; ils incluent des différences qui éventuellement s'additionnent et le bruit de fond dont nous avons déjà parlé.
La mesure de différence sémantique est fonction de la somme des distances de M à ses LAA et la somme des distances de N aux LAA trouvés pour M. Si nous posons E= LAA(M, N) nous avons pour une mesure de M vers N :

$$SD(M,N) = \frac{\sum_{E \in LAA(M,N) \cup LAA(N,M)} d(M,E) + d(N,E)}{Card(LAA)}$$

Dans la figure 3 :
SD(W2, W3) = (2+2) / 1 = 4 par \A
SD(W3, W2) = ((2 + 2) + (3 + 3)) / 2 = 5 par \A et \Universe.
Utilisation concrète de la mesure des LAA.

a) Des différences componentielles nulles d'un côté et importantes de l'autre
Ce résultat est obtenu en comparant *monnaie* et *yen*. La différence de *monnaie* à *yen* est vide, comme il se doit, et la différence de *yen* à *monnaie* peut être importante puisqu'elle emprunte des chemins partant *Japon(termes liés)* pour arriver dans le bruit de fond (voir 40). L'extraction de la source des différences et leur valeur (ici *Japon(termes liés*) depuis les nombreux LAA obtenus est difficile mais possible (voir 4.1.2.2.7 *Exemple commenté d'une extraction des ressemblances et différences spécifiques*, ci-dessus).
b) Des différences componentielles nulles de chaque côté
Les termes sont synonymes depuis le point de vue des LCA considérés ; souvent cela arrive dans des embryons de taxonomie que nous n'avons pas pris le soin de traiter. Ces différences vides sont le propre de ce que nous trouvons dans les ontologies fondées sur les seuls Is_a si bien que toute inférence y devient incontrôlable.
c) Des différences componentielles nombreuses de chaque côté
C'est la situation la plus courante.

4.1.2.2.5 La proximité componentielle

Au final, la proximité componentielle d est un agrégat construit en intégrant l'activation et la différence. Il s'agit par exemple de : d = d^ + SD:
d(W2, W3) = (2 + 4) / 2 = 3
d(W3, W2) = (2 + 5) / 2 = 3.5
Ainsi, nous disons que d'une part W2 est plus proche de W3 que ne l'est W3 de W2 et d'autre part que W2 et plus proche de W1 que de W3. Comme tout agrégat, la proximité componentielle présente une perte d'information importante mais certaines expériences,

comme le dictionnaire à l'envers (voir 5.1.2.2, page 55) se satisfont de son emploi ; en effet, il s'agit juste de dire que les distances de *yen* à 1) *monnaie du Japon*, 2) *monnaie de Hiro-Hito*, 3) *unité monétaire d'Asie*, 4) *monnaie des USA* vont croissantes.

4.1.2.2.6 Les mêmes mesures en incluant les fonctions lexicales

Les fonctions lexicales sont conçues pour faciliter la génération de textes. Il ne vaut mieux pas les utiliser pour calculer des différences sémiques. Nous ne pouvons discuter ici cette affirmation. Disons seulement que le simple fait qu'elles n'ont pas le même *point de vue* que ce dont nous parlons en ce moment laisse supposer le résultat. Au paragraphe 4.2.1.2 *Hiatus "observations sémantiques" et observations dans le syntagme* ci-dessous nous discutons un cas d'inférence très désirable et pourtant impossible à replacer dans un cadre sémique quelconque sans porter atteinte d'une façon définitive à ses qualités structurantes essentielles.

4.1.2.2.7 Exemple commenté d'une extraction des ressemblances et différences spécifiques

Dans ce paragraphe, nous étudions les mots *fleuriste (nom)* et *fleur (nom)* pour illustrer avec un exemple concret ce que fournissent les LCA et les LAA. Les résultats permettent de souligner la structure componentielle du dictionnaire et montrent des résultats intuitifs.
Nous obtenons :

LCA(fleuriste, fleur) = {\fleur (T), \Racine des noms (Grammaire)}

En effet, dans la structure réelle du dictionnaire, il est normal d'obtenir du fait que les thèmes n'appartiennent jamais aux classes au moins deux types de LCA : l'un pour les classes, l'autre pour les thèmes. Ici, nous obtenons le thème *fleur* et la racine des classes de noms [*Racine des noms*].

LAA(fleuriste, fleur) = {\monde du vivant [T],\homme et société [T],Xi [T], …}[39]

Xi [T] indique que nous ne fournissons pas la liste complète des LAA qui est trop importante. La plupart du temps, la racine générale du dictionnaire apparaît comme LAA. Dans notre cas, nous avons obtenu 107 LAA depuis *fleuriste*. Pour vérifier ce point, suivons les ancêtres communs LA (LCA ∪ LAA) jusqu'aux premières classes qui définissent *fleuriste*.
Depuis le LCA *Racine des noms (Grammaire),* le programme trouve immédiatement une classe : *les noms (classe grammaticale)* elle-même. Cela signifie que les deux mots *fleuriste* et *fleur* partagent le trait grammatical nom. Nous résumons cette information par la notation :
LCA(\les noms (classe grammaticale), fleuriste.n)→ \les noms (classe grammaticale) dans laquelle '→' dit que *fleuriste* est un nom (puisque le plus petit ancêtre commun entre *fleuriste* et *nom* est le concept *nom*). Comme *les noms (classe grammaticale)* est un concept commun à *fleuriste* et *fleur*, nous savons aussi que :
LCA(\les noms (classe grammaticale), fleur.n)→ \les noms (classe grammaticale)
Ces résultats peuvent évidemment s'obtenir plus directement (par la lecture directe des catégories grammaticales comme propriété des mots) mais cela n'importe pas ici. Donnons maintenant les autres résultats :
LCA(\fleur [T], fleuriste.n) → \Personne en relation avec les fleurs [\N classe]
LAA(\homme et société [T], fleuriste.n) → \Personne qui vend qqch[\N classe]

[39] Les LCA et les LAA ne sont pas ordonnés.

Les autres 105 LAA fournissent les mêmes premières classes définissant *fleuriste*. En conclusion, cela signifie que *fleuriste* et *fleur* sont tous deux des noms et qu'ils partagent le *monde des fleurs*. La différence entre *fleuriste* et *fleur* est que *fleuriste* est une *personne qui a pour activité de vendre qqch*. La différence entre *fleur* et *fleuriste* s'obtient au travers du concept *\Le monde du vivant [T]*. On trouve la classe *\partie portant les organes reproducteurs de certaines plantes (Cl).*

4.1.3 Exemple d'application des distances : le dictionnaire s'enrichit tout seul depuis le dictionnaire à l'envers

Les fonctionnalités que nous venons de décrire ont été implémentées à compter du moment où nous avons pu accéder à des machines capables d'effectuer suffisamment de calculs. À cette époque[40], le Sémiographe n'était pas industrialisable. Toutefois, les premiers résultats du Sémiographe ont pris une forme intéressante : en effet, certains résultats ont été jugés suffisants pour pratiquer certains processus automatiques d'alimentation du dictionnaire.

A cette époque, nous avions à côté du Dictionnaire Intégral – énorme réseau sémantique de 200.000 mots-sens -, un petit dictionnaire d'environ 80.000 définitions courtes[41]. Il est aisé d'imaginer que cette séparation était agaçante, surtout en ce qui concerne le Dictionnaire Intégral. Sachant que le temps de la réalisation manuelle de l'appariement des deux dictionnaires prendrait plusieurs milliers d'heures, la question était de savoir en quelle mesure le Sémiographe pourrait faire cet appariement tout seul sans erreur, ou du moins sans faire plus d'erreur qu'un humain travaillant rapidement mais attentif.

Aujourd'hui, 85% des appariements présents dans LDI ont été réalisés par le Sémiographe. Il s'agit d'une application du dictionnaire à l'envers (voir 5.1.2.1). Seules 12.000 entrées du petit dictionnaire restent à traiter. Ce reliquat est dû essentiellement à des différences importantes de discrétisation des sens dans les deux dictionnaires, des éléments manquants dans le réseau LDI, une mauvaise reconnaissance des locutions par le Sémiographe et quelques fautes dans le dictionnaire de définitions.

4.2 Réflexions critiques sur les résultats obtenus

Au plan de la réflexion, l'apport du Sémiographe fut considérable. En premier lieu, il a légitimé les efforts passés, et donc notre démarche empirique et structuraliste en montrant qu'il est capable de nous aider dans des développements extensionnels qui sont extrêmement coûteux (voir paragraphe 4.1.3, page 43, ci-dessus). Issus d'une recherche exploratoire et empirique, la mise en œuvre des Sémiographe et structure vérifient leur pertinence dans le développement applicatif. En outre, nous trouvons, avec le dictionnaire à l'envers (voir page 55), des erreurs fortuites de structure que nous pouvons corriger à l'occasion[42]. En bref, le

[40] A cette époque, l'appariement automatique des lexiques était pour nous non de la recherche mais plutôt un instrument au service du développement industriel. Même si nous observons qu'Alexandria (voir Le développement d'Alexandria page 63), 1,7 millions de mots alignés en 27 langues, est largement redevable de cette pratique, nous ne voyons pas quel résultat intéressant pour la communauté nous pourrions proposer en publiant à propos d'une technique ad-hoc conçue par exemple pour deux lexiques particuliers. Nous ferons donc dans ce document largement l'économie de ce genre de références.

[41] Ce dictionnaire avait été rédigé entre 1991 et 1994 dans le cadre d'un contrat commercial pour réaliser un dictionnaire électronique de poche

[42] Nous ne nous sentons pas obligé d'effectuer toutes ces corrections car en application de l'Observation 1 page

Sémiographe participe pleinement aux développements des travaux.
Sont-ce pourtant les seules contributions que cette mise en œuvre du Sémiographe a pu offrir ? Nous ne le pensons pas. En effet, nous défendons que la principale contribution du Sémiographe est d'ordre théorique. Elle concerne deux points que nous résumons ainsi :
- quelle est l'expressivité du DAG?
- quelles sont les inférences que l'on peut ranger en sémantique componentielle liée aux notions de définitions et d'extensions et lesquelles sont étrangères à ces notions.
Cette contribution repose sur l'observation d'erreurs prédictibles dont les différents efforts pour les corriger d'une manière ou d'une autre montreront qu'elles ne peuvent être résolues dans la structure que nous avons définie jusqu'ici. L'une des sources d'erreurs était prévisible du fait de la *Question 1*, page 20. Cependant, nous ne nous y étions pas encore attardés. L'autre source d'erreur est plus surprenante.

4.2.1 Les deux hiatus

Nous présentons ci-dessous ces deux hiatus entre modèle et réalité et nous généralisons.

4.2.1.1 Hiatus "dictionnaire à l'envers" en rapport avec l'absence d'organisation entre les concepts des quasi-définitions

Considérant une remarque de B. Victorri, nous avons pu vérifier que des requêtes dictionnaire à l'envers utilisant les mêmes ensembles pour référencer plusieurs objets ne fonctionnent pas.
Soient par exemple *négrier* comme marchand d'esclaves, c'est-à-dire de toutes les personnes ayant le statut d'esclave, et *fleuriste,* marchand de fleurs, c'est-à-dire de toutes les fleurs donnant lieu à ce commerce. La figure suivante montre clairement la situation courante dans laquelle nous perdons des éléments de définition.

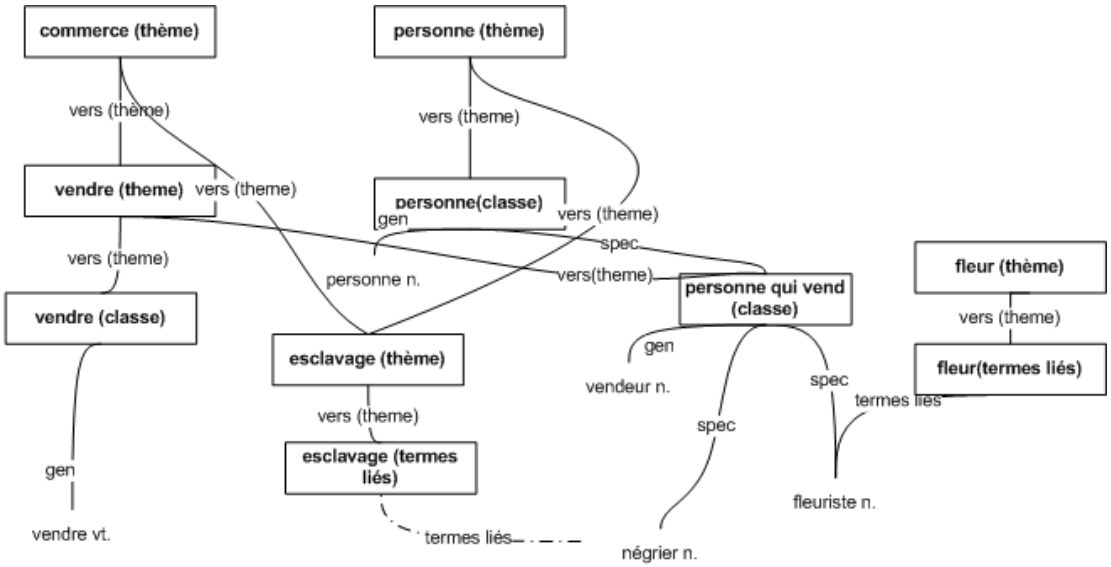


Figure 7 Personne qui vend des fleurs *versus* personne qui vend des personnes

---

14 nous considérons que le système comportant nécessairement des approximations doit toutefois être capable de fonctionner.

Dans les deux cas, *fleuriste* et *négrier* ont deux ancrages dans le graphe :
- *fleuriste : vendeur (personne + vendre) + fleur*
- *négrier : vendeur (personne + vendre) + esclave*

En remontant jusqu'aux ancêtres pour chacun de ces deux mots, ces ancrages devraient laisser une trace assez précise. C'est le cas de *fleuriste* qui permet *:*

- *vendeur + [fleur]* $_{(thème)}$*,*
- *personne + [vendre]* $_{(thème)}$ *+ [fleur]* $_{(thème)}$*,*
- *personne + [commerce]* $_{(thème)}$ *+ [fleur]* $_{(thème)}$
- *personne + [commerce]* $_{(thème)}$ *+ [plante]* $_{(thème)}$
- *etc.*

Mais ce n'est pas le cas de *négrier* dont la trace s'arrête à [esclave] $_{(thème)}$ puisque le reste est totalement subsumé par [commerce] $_{(thème)}$ et [personne]$_{(thème)}$ qui subsument déjà *vendre* et *personne*. Pour le DAG, il y a simplement redondance alors même que les instances diffèrent. Ainsi, *personne vendre personne* ne peut donner aucune solution spécifique. La question 1 peut se reformuler ainsi :

Question 3 Comment reformuler le Sémiographe pour qu'il puisse gérer des instances et comment reformuler Lexidiom pour qu'il prévoie la gestion de ces instances ?

Cette question est une reformulation de la *Question 1 Comment enregistrer des concepts structurés dans le DAG ?* (page 20). Nous avons maintenant un premier motif pour tenter d'y répondre. Il convient cependant d'élargir le domaine avant de tenter d'y répondre en raison de l'Observation 5 page 22.

4.2.1.2 Hiatus "observations sémantiques" et observations dans le syntagme

Le problème général qui nous avons rencontré est le suivant : nous n'avons trouvé pratiquement aucun signal componentiel entre deux mots situés dans un même syntagme. Cette absence de résultat est en opposition directe avec tous les modèles de propagation d'isotopies sémantiques comme celui de Rastier. Par exemple, chez Rastier un syntagme comme *samouraï mange* voit se propager le trait *\animé*. Pourquoi ne trouvons-nous pas ce trait dans notre résultat?

Nous allons essayer de faire sentir la raison de cette absence par trois voies différentes :
- le rappel d'une idée fondamentale concernant le signe
- une suggestion d'inférences
- une suggestion d'expérimentation.

Une fois que nous serons sensibilisés par le propos, nous montrerons l'impact potentiel d'une mauvaise résolution du problème sur le comportement du DAG.

4.2.2 Les changements de points de vue concernent les choses les plus simples

Nous essaierons ici de faire partager l'idée que *samouraï* n'est point Is_a *humain* mais seulement Is_a *personne.*

1) Fondamentalement

En tentant de mieux assimiler une idée fondamentale mais non originale concernant le signe.
En relisant l'Observation 3 page 18 et la note 31 page 37 nous supposons que l'absence de signal entre *samouraï* et *mange* dans le Sémiographe qui est fondé sur une structure

homogène manifeste le fait qu'il n'y a pas de lien componentiel, même ténu, entre *samouraï* et *mange.*

2) Inférentiellement
Les inférences suivantes sont immédiatement bizarres :
1) samouraï-->personne-->.....eumétazoaire....animal-->[tout truc qui mange] ^ manger vt.
2) samouraï -->personne-->mangeur-->[tout truc qui mange] ^ manger vt.
Pour (1), nous n'imaginons pas un *samouraï* dénommé *eumétazoaire* ni plus que nous ne pouvons imaginer descendant dans les *eumatozaires* trouver un *samouraï*.
Pour (2), nous n'imaginons pas un *samouraï* dénommé *mangeur* ni plus que nous ne pouvons imaginer descendant dans les *mangeurs* trouver un *samouraï*.
Les inférences suivantes ont le même défaut :
3) samouraï -->personne-->animal-->[tout truc qui mange] ^ manger vt.
4) samouraï -->personne-->humain-->animal-->[tout truc qui mange] ^ manger vt.
En effet *personne* signifie : *être humain considéré dans sa spécificité*. Et cela s'oppose précisément à la généricité de l'humain conçu en termes d'ensemble d'attributs génériques.
Il y a rupture de point de vue. Cette rupture est clairement marquée dans le dictionnaire : *être humain* **considéré dans** *tel point de vue.* Ainsi, *mangeur* est une *personne*, c'est-à-dire un *humain* **considéré dans** *un certain point de vue* ce qui ne peut se ramener seulement à *humain.* Nul locuteur français ne peut accepter la définition suivante de *mangeur* : *humain qui mange.*

3) Expérimentalement 1
En mettant samouraï à l'épreuve de la prototypicité, demander à son entourage 50 mots proches de *samouraï*. Compter le nombre d'occurrences de *manger* (en-dehors de la nourriture japonaise). Le signal résultant sera nul.

3) Expérimentalement 2
En comparant (text-mining) les contextes de *personne* et de *humain.* Mais je ne souhaiterais pas insister sur cette technique qui ne peut qu'aboutir à des indices et aucunement à une preuve du fait par exemple de la note 8 page 7 et de la note 31 page 37.

4.2.3 Effets sur une structure lexico-componentielle de la non-prise en compte du changement de point de vue

Nous faisons l'hypothèse que nous plaçons tout l'ensemble $[\text{personne}]_{\text{classe}}$ dans $[\text{mangeur}]_{\text{classe}}$ pour obtenir une signal dans une comparaison d(*samouraï*, *manger*). Cela entraîne une circularité mais, en principe, cette dernière n'est pas vraiment gênante. Les trois résultats sont les suivants :
- un *gréviste de la faim* devient *mangeur*
- un *gros mangeur* devient moins *mangeur* que le générique *personne* du fait qu'il entretient des différences spécifiques par rapport à *manger* alors que *personne* évidemment n'en a pas puisqu'il n'est pas définissable par rapport à *manger.*
- il faut de plus ajouter tous les *animaux* puisque ces derniers mangent aussi.
Evidemment, il serait possible d'empêcher ces résultats de survenir en inventant une relation ad-hoc. Mais la difficulté que nous soulevons dans cette section n'est qu'une partie d'une difficulté plus générale que nous voulons résoudre et l'emploi de cette relation ad-hoc ne résoudrait pas le problème général de localisation (voir Observation 3 page 18); cet emploi ne ferait que transformer un problème complexe que l'intelligence humaine embrasse

convenablement en un problème compliqué que nous aurions du mal à suivre.
D'une façon générale nous faisons l'observation suivante :

Observation 7 a) Nulle ontologie fondée en rapport à un domaine (c'est-à-dire à un point de vue) n'est susceptible de s'agrandir, en raison du fait même que la définition du domaine de l'ontologie est exogène à celle-ci. 7 b) Pour pouvoir s'étendre, une structure sémantique doit savoir gérer ses propres frontières.

Il est peut-être possible de défendre que l'IA a abondamment démontré (7.b). Remarquons toutefois le parallélisme de l'Observation 7.b ci-dessus avec l'Observation 3.a page 18. En conclusion, nous écrivons la question :

Question 4 Comment décrire dans notre DAG des données non componentielles qui appartiennent aux connaissances naïves[43] de telle manière que cette description soit clairement distincte des données componentielles que nous avons déjà représentées.

4.2.4 Synthèse critique

Le paragraphe 4.2.1.2 ci-dessus nous montre clairement l'occurrence d'un glissement au sein du triangle sémiotique[44] *signifiant/signifié* vers la direction de la *référence*. De son côté, le paragraphe 4.2.1.1, page 44, présente également un problème de référence. En application de l'Observation 5, c'est-à-dire de la nécessité de réaliser une intégration de différents points de vues linguistiques dans un tout dans lequel des inférences homogènes peuvent s'accomplir, nous avons décidé qu'il était urgent de ne pas chercher à résoudre le problème posé. Nous écrivons :

Question 5 Comment la Question 4 page 47 peut-elle trouver une solution homogène avec la Question 3 page 45?

4.3 Projets et documents

Projets
Quatre projets ont contribué au développement du Sémiographe sans qu'aucun ne cible ce développement. Le Sémiographe est un effort de réponse générique en rapport avec la **position linguistique** (page 40).

Projet 4 : MARLEN

MARLEN L'acronyme signifie *Multimedial ARchive and Learning ENvironment for creative writing*. C'est un projet européen type Leonardo.

[43] Ou autre appellation, par exemple celle de *Référence* ou de *connaissance sur le monde.*
[44] En pensant ici à Odgen et Richards qui écrivirent "*the meaning of meaning - A Study of the Influence of Language upon Thought and of the Science of Symbolism*. *"* en 1923.

En 1996, ce projet a besoin de contenus textuels et dictionnairiques. Nous sommes fournisseurs en matière de dictionnaires et nous en profitons pour constituer une petite bibliothèque électronique d'environ 250 œuvres classiques. Ces ouvrages devaient nous fournir un corpus à une époque où les corpus n'étaient pas très courants.
Bibliotexte est alors produit en DELPHI. Il permet de faire des recherches de type alignement ou extraction de séquences aujourd'hui courantes. Fin 1996, nous apprîmes anecdotiquement et statistiquement que Flaubert adore l'expression *de temps à autre*.
Les partenaires sont : CNR - ISRDS (Rome), DIPARTIMENTO DI LINGUISTICA E DI LETTERATURA UNIVERSITÁ "LA SAPIENZA" (Rome), DIPARTIMENTO DI SCIENZE DELL'EDUCAZIONE UNIVERSITÁ DI ROMA TRE (Rome), GOST (Rome), INSTITUTO MULTIMEDIA (Porto).

Projet 5 : AGIR

AGIR Ministère de l'industrie [1997]. Nous sommes au départ sous-traitant d'Alcatel puis nous reprenons la partie "texte" du projet. Il s'agit d'indexation de descripteurs de video et d'images. Nous essayons d'adapter le Sémiographe à ce contexte d'utilisation mais les temps de calcul sont assez mauvais. Toutefois, c'est grâce à ce projet que le Sémiographe a pu passer du stade de l'idée à celui de réalisation.

Projet 6 : EuroWordnet.

Nous avons été entre 1998 et 2000, responsables avec l'Université d' Avignon de la réalisation pour le français d'un projet de réseau lexical qui suivrait les orientations de WordNet [Fellbaum, 1998]. Dans EuroWordNet [Vossen, 1998], nous fûmes des fournisseurs de technologies pour l'appariement de réseaux sémantiques (entre les ressources multilingues incluses dans notre dictionnaire et la cible WordNet) .Le Sémiographe a été employé pour calculer automatiquement[45] une bonne partie des projections du référentiel LDI sur WordNet. La vérification manuelle a été réalisée par Avignon.

Projet 7 : IVOMOB

IVOMOB Ministère de la recherche [2000].
Le projet IVOMOB, financé par le RNRT, regroupe trois entreprises et un laboratoire de recherche (le LORIA) aux compétences complémentaires. Il a pour but final le développement d'un prototype d'accès vocal à un portail de services. Dans le projet, notre rôle est la génération de paraphrases pour produire des modèles de langage acoustique.
Nous réalisons les générations depuis Lexidiom, en gérant des relations ad-hoc que nous exportons dans le moteur d'inférence qu'est le Sémiographe pour enfin effectuer la génération. Les fonctions lexicales de LDI sont fortement utilisées, conformément aux prévisions de la THÉORIE SENS↔TEXTE. Nous gardons en mémoire de ce projet que pour exprimer *je voudrais (savoir/avoir) qqch*, les paraphrases courantes sont fort peu nombreuses (quelques dizaines) et rendent compte de 80% des besoins. En relation exponentielle de la loi de Pareto,

[45] Voir note 40. A la différence près des lexiques employés de référence, la technique globale fut assez proche de celle décrite dans Sagot [2008].

couvrir 100% des formules représente la production de quelques centaines de milliards de phrases. Les modèles de Markov ont pu choisir…

Stage
Deux masters II recherche.

Durant ces années, je m'occupe également de la rédaction de ma thèse. Ma thèse est soutenue le 30 Novembre 2000.

Publication, dissémination

D. Dutoit : « A text->meaning->text dictionary and process » [2000], acte de Language resource and evaluation, LREC.

D. Dutoit : Quelques operations sens→texte et texte→sens utilisant une sémantique universaliste apriorique, [30 Nov. 2000], thèse de doctorat.

Le dictionnaire intégral [1999], journée d'études du Centre National d'Etudes pédagogiques

Le sémiographe [1999], présentation à la journée Outils pour le Tal organisée par le groupe de recherche Information-interaction-intelligence en association avec l'Atala

Linguistique et apprentissage automatique, 10th european conference on Machine Learning, [April 1998]

En 2001, deux participations à l'organisation des journées de l'Atala

4.4 Conclusion et perspectives

La période qui s'achève avait débuté avec plusieurs orientations de travail qui aboutirent à des degrés divers :
- au plan morphologique, nous avons incorporé et enrichi la Base de Données LEXicales (BDLEX) de Guy Pérennou
- au plan compositionnel et de la TST nous nous sommes exercés à la génération de paraphrases (voir Projet 7 : IVOMOB page 48)
- au plan componentiel et de notre travail, nous avons commencé à caractériser le domaine des inférences componentielles
- cette caractérisation a été effectuée empiriquement, dans plusieurs mises en œuvre utiles du Sémiographe (Voir paragraphe 4.1.3, page 43 et *Projet 6 : EuroWordnet.*, page 48).

Enfin nous avons ouvert les travaux de lexicographie multilingue.
D'une façon plus particulière, nous avons répondu assez précisément à la *Question 2*, page 31, en caractérisant le mot *sémantiquement.* Mais nous observons que la *Question 1*, page 20, n'a pas trouvé de réponse technique. Cependant, sa reformulation dans la *Question 5*, page 47, élargi son champs d'application tout en y imposant des contraintes particulières justifiées par l'Observation 3.b, page 18.

Vers 1999, le temps n'est pas venu de casser, du fait de la *Question 5*, ce qui vient juste d'être accompli, à savoir un Sémiographe. Nous souhaitons d'abord le voir à l'œuvre en environnement monolingue ou multilingue dans les différentes applications que nous avions imaginées pour lui en 1992. De plus, nous espérons que cette mise en œuvre nous fournirons

d'autres perspectives concernant la *Question 5*.
Dans ce contexte, les objectifs de recherche pour la nouvelle période ont été les suivants :
- mise en œuvre du Sémiographe dans différentes applications cible afin d'évaluation
- choix et développement d'une vraie application qui sera diffusée
- développement concomitante d'une nouvelle version de Lexidiom qui devrait permettre de décrire des concepts structurés
- amélioration d'autres aspects de Lexidiom comme la prise en charge d'UNICODE. Il est prévu qu'à terme Lexidiom aurait entre autre pour rôle de permettre d'intriquer les relations et les nœuds de telle manière que l'on puisse transformer une relation en un nœud et réciproquement un nœud en une relation
- développement multilingue
- faire attention au développement d'Internet et suivre de près l'évolution des nouveaux standards techniques
Au plan de la recherche abstraite, à ce moment nous nous sommes posé la question suivante :
- quel rapport pourrait exister entre la *Question 5* et la résolution d'un énoncé aussi simple que *quelle est la question du cheval blanc d'Henry IV?*
- trouver d'autres problèmes qui présenteraient des ressemblances structurales avec celui que nous posons.
Concernant le *cheval blanc*, évidemment le problème soulevé pourra sembler bizarre même si sa qualité principale ne doit pas nous échapper : il semble exclure tout problème de référence, quelle que soit la manière dont nous pourrions concevoir cette dernière. Si l'argument ne convient pas, nous proposons au moins d'observer que l'énoncé est une question qui comporte sa réponse et que dans cette mesure le problème est de savoir quelles opérations strictement linguistiques et portant sur une structure (laquelle) sont capables de détecter cet état de fait. Cette question générale est pertinente en cela que n'importe quel discours peut être vu à différents niveaux comme un ensemble de questions (définitions de lieux) en relation avec un ensemble de réponses (lieux définis). Cela dit, ce qui nous intéresse le plus dans le *cheval blanc* est technique : nous ne pouvons pas trouver de discours ou de textes comme je viens de les caractériser qui soient plus élémentaires.

5 UNE PERIODE DE PROJETS INSTITUTIONNELS ET INDUSTRIELS (2002-2007)

Avec le retard que nous avons pris concernant le développement d'Internet, le temps s'accélère. Nous devons faire d'importants efforts dans le domaine du multilinguisme (LDI) et dans la maîtrise des technologies web. En TAL, l'ambiance est au tout statistique. L'expansion de la linguistique de corpus incline à ressentir que la situation n'est pas si différente en cette discipline. La phrase suivante extraite du résumé de la HDR de Juan Manuel Torres Moreno soutenue le mercredi 12 décembre 2007 résume assez bien le contexte : *Pendant ces années de recherche, plusieurs fois je me suis posé la question de savoir si la linguistique pouvait encore jouer un rôle dans le traitement de la langue naturelle.*
De notre côté, nous restons assez éloigné de tout ce remue-ménage de nombres et de lettres sauf quand nous utilisons le modèle de Markov pour réaliser un lemmatiseur[46] en français, anglais, italien ou espagnol ou quand nous demandons au Sémiographe de calculer des similarités dans des données sémasiologiques[47]. Finalement, comme tout le monde nous évitons d'affronter de front les questions théoriques et nous nous concentrons sur des validations à taille réelle, des acquisitions de savoir-faire technique, des améliorations techniques de notre logiciel lexicographique et des développements extensifs. Nous détaillons ces points dans les paragraphes suivants :
- du DAG à l'hypergraphe
- le Sémiographe touche les applications

5.1 Du DAG à l'hypergraphe

Tandis que LDI prend la forme et la terminologie d'un hypergraphe, le Sémiographe reste un graphe dont le nombre maximum de sommets des arêtes vaut 2.

[46] Voir API morpho-syntaxique page 35.
[47] Voir paragraphe 4.1.3, page 43 et *Projet 6 : EuroWordnet.*, page 48.

5.1.1 Le moteur de LDI devient un hypergraphe

Un des projets de recherche co-financé de cette période suppose que nous réalisions une association entre LDI et WordNet également au plan de la structure des données. Ce projet s'appelle Balkanet (voir Projet 9, page 66 ci-dessous). Dans Balkanet, six nouveaux WordNet (turc, bulgare, roumain, tchèque, grec et serbe) sont développés. Notre tâche est une tâche de conseil et de contrôle. Nous avons considéré que le plus simple pour avancer dans cette tâche en faisant un travail utile consistait d'une part à rendre compatible WordNet et LDI, d'autre part à mettre tous les WordNet créés ou à venir dans le conteneur Lexidiom, en fusion de réseau avec LDI. Il s'est agi d'une grosse modification de Lexidiom et la moindre de ces modifications était la gestion de Unicode et le passage à un nouvel SGBDr[48]. La pire de ces modifications a concerné les trois points suivants :

1) Fusionner les graphes mais être capable de les dissocier.
   Ainsi, chaque point et chaque relation du graphe est connue comme membre d'un ou de plusieurs réseaux (il y a plusieurs millions de ces points).
2) Etre évidemment capable de supprimer un graphe et de le ré-ajouter (maintenance)
3) Enfin, il s'agissait d'introduire un nouvel élément dans LDI. Cet élément présent dans WordNet et absent dans LDI est le Synset.

LDI voit alors apparaître de nouvelles définitions. Ces ajouts sont en rapport avec l'introduction de la notion d'hypergraphe dont la figure suivante suffit à illustrer les concepts utiles à notre présentation.

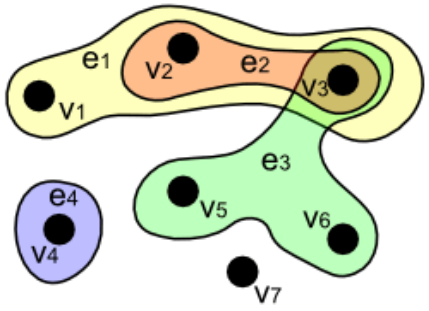


Figure 8 Exemple d'hypergraphe[49]

Les hypergraphes ont été ainsi nommés par Claude Berge en 1960. Les hypergraphes généralisent la notion de graphe dans le sens où les arêtes ne relient plus un ou deux sommets, mais un nombre quelconque de sommets (compris entre un et le nombre total de sommets de l'hypergraphe).
Dans la figure précédente, l'arête e3 réunit les sommets v3, v5 et v6 en une seule *partie* tandis v7 n'est relié par aucune arête et reste isolé : v7 n'est pas une partie de l'hypergraphe puisqu'il n'a pas d'arête.
Le nombre de sommets n d'un hypergraphe est son ordre. Ici, ce nombre vaut 6.
Le rang d'un hypergraphe est le nombre maximum de sommets réunis par une même arête (que nous appellerons hyper-arête par la suite pour les distinguer des arêtes ordinaires de rang 2). Dans la figure, ce rang vaut 3.

---

[48] Système de Gestion de Base de données relationnelles.

[49] Figure empruntée à *Claudio Rocchini* depuis http://fr.wikipedia.org/wiki/Image:Hypergraph.gif

Un hypergraphe est dit simple s'il n'a pas d'arêtes multiples c'est-à-dire plusieurs arêtes pour une même partie. La figure présente un hypergraphe simple.
On appelle *famille de Sterner* tout hypergraphe dans lequel aucune arête n'est contenue dans une autre. L'hypergraphe de la figure n'est pas une telle famille du fait que e2 est inclus dans e3.

Aujourd'hui LDI a la définition suivante :
- ordre : 1.500.000
- rang : 180.000 en considérant l'ensemble des synsets de WordNet comme une structure particulière (chaque version de WordNet réunit un nombre particulier de synsets).
- multiple : une arête élémentaire de WordNet entre deux synsets peut exister en même temps qu'une de nos fonctions lexicales (rem: les concepts sont une particularité de LDI par rapport à WordNet).
- famille de Sterner : non. Premièrement, tout concept de LDI définit une arête groupant un nombre particulier de sommets et deuxièmement tout concept de LDI, sauf par convention la racine, est inclus dans au moins un autre concept.

Pour réaliser ce passage de gestionnaire de graphe à celui de gestionnaire d'hypergraphe, Lexidiom a été enrichi par un champ que nous appelons communément "Lieu" mais qui est en fait une hyper-arête
Si par exemple nous avions dans la version précédente de LDI l'entrée suivante :

samouraï *[guerrier]*$_{classe}$ *spec*

nous trouvons maintenant :

samouraï *[guerrier]*$_{classe}$ *spec* *LDI.*

Ainsi LDI est maintenant défini comme une hyper-arête de rang environ 120.000 pour ce qui concerne son DAG de concepts (ni les mots ni les fonctions lexicales.

Il est possible de réifier cette hyper-arête en écrivant par exemple:

LDI *[dictionnaire]*$_{classe}$ *spec* *LDI*
LDI *[Memodata]*$_{thème}$ *vers(thème)* *LDI*

ce qui signifie :

dans LDI, .LDI est une sorte de dictionnaire
dans LDI, .LDI est spécifié par MEMODATA.

Puisque le Sémiographe manipule des graphes étiquetés et que nous sommes maintenant amené à utiliser diverses sortes de graphes, nous devons préciser la notion de *graphe* dans notre cas.

Définitions complémentaires

Graphe : Ensemble de nœuds et de relations, avec mode de fonctionnement. Un graphe est déclaré par son nom. Le nom d'un graphe est un concept particulier.
Dans la version actuelle de Lexidiom, il existe plusieurs graphes:
- les graphes à base de concepts et de FL de LDI
- les graphes à base de Synsets et de FL des WordNet
- les synsets

Synset : Ensemble type WordNet, hyper-arête comprenant une glose (une définition naturelle) dans une ou plusieurs langues et des littéraux (des mots-sens) d'une ou de plusieurs langues.

Si une même langue présente dans un même synset plusieurs mots-sens, ces mots-sens sont synonymes entre eux.

FL LDI (redéfinition) : Les anciennes FL de LDI sont susceptibles dorénavant de mettre en relation deux mots-sens (ex: H2O et eau), deux synsets ({achat} et {acheter} ou un mot et un Synset selon les besoins.

FL Wordnet : Toute relation de WordNet entre Synsets (méronyme, hyponyme, cause, antonymie de Wordnet etc.)

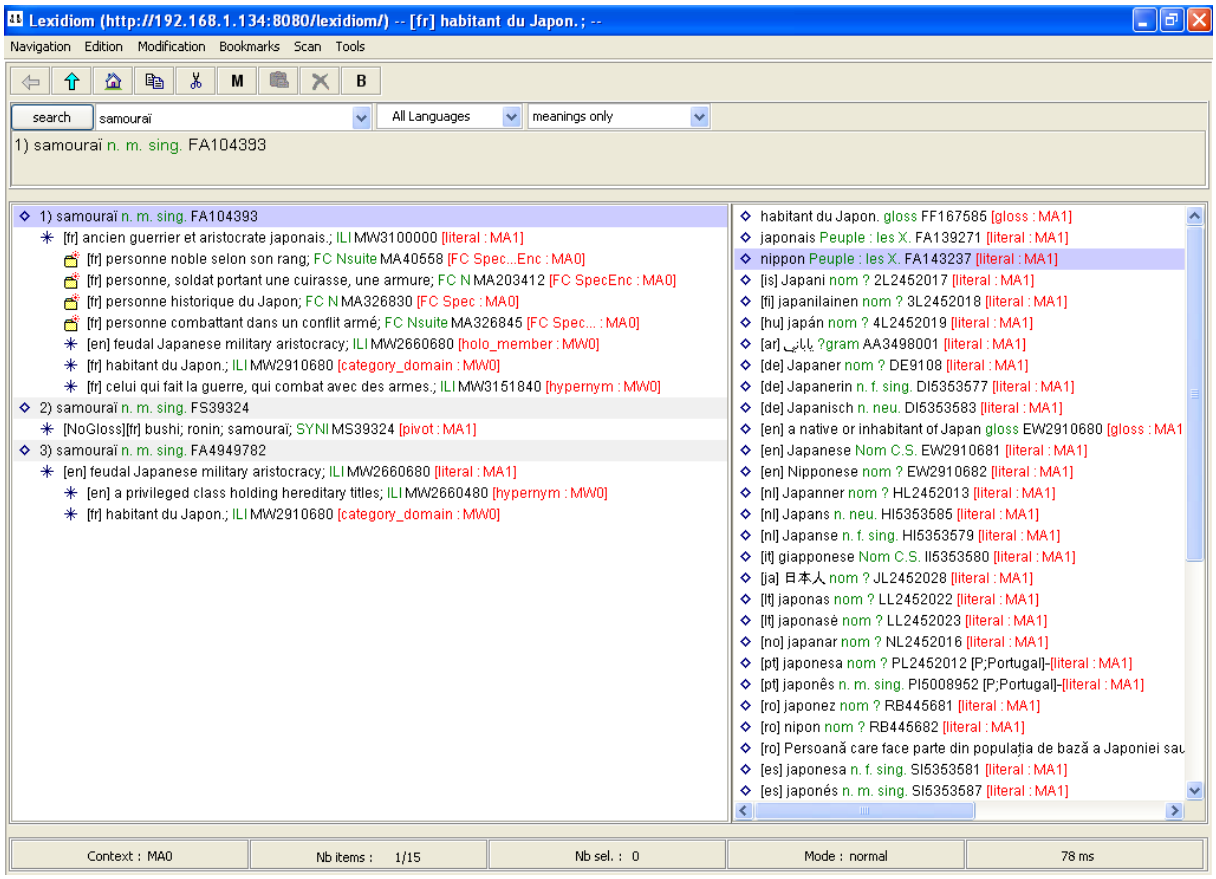


Figure 9 Une illustration Lexidiom

Le champ *Context* en bas à gauche a pour propriété : MA0. MA0 est l'identifiant de LDI. Le champ rappelle que l'écran présente une vue de l'hyper-arête LDI.
L'outil présente à gauche des littéraux pour le synset samouraï en plusieurs langues.

5.1.2 Le Sémiographe touche les applications

Tandis que nous modifions Lexidiom pour en faire un gestionnaire d'hypergraphe, le Sémiographe continue à être un utilisateur des graphes ordinaires de rang 2.
Nous présentons[50] trois applications concrètes du système et une application 'exploratoire'. Il

[50] A cette époque, nous avons décidé de faire l'impasse sur la désambiguïsation lexico-sémantique pour

s’agit du "dictionnaire à l’envers", "la gestion d’un fond documentaire", d'une "gestion documentaire multilingue" et d'une "aide à la navigation multimédia"
En-dehors du dictionnaire à l’envers, nous décrivons dans cette section trois services utilisant le Sémiographe. Les deux premières applications sont le fait de clients. L'un est une SSII française importante dans le secteur de la GED. L'autre est une multinationale intervenant comme éditeur de solutions GED au niveau mondial. La troisième application est davantage une maquette importante, effectuée par nous pour le compte d’un projet européen, pour avoir une idée des performances du Sémiographe dans sa partie proximité componentielle appliquée à une autre langue que le français : l'anglais.
La présentation des applications s’effectue en enchaînant les composants que nous avons présentés en 4.1.2.1, page 34.

5.1.2.1 Les applications non lexico-sémantiques

Le Sémiographe a fourni des outils pour différentes applications (veille, documentation, terminologie) que nous ne présentons pas ici en raison qu'elles emploient peu le réseau lexico-sémantique et donc ne sont pas susceptibles de nous éclairer sur le fonctionnement du système lexical.

5.1.2.2 Le dictionnaire à l’envers

Le but de cette application est l’extraction depuis LDI des mots répondant à l’énonciation d’un définissant fourni par l’utilisateur.
Introduction
L'application *dictionnaire à l'envers* est une application assez classique du TAL. Nos premières expérimentations remontent à 1992. Comme tout dictionnaire, les *dictionnaires à l'envers* souffrent de l'Observation 1 page 14 : ils sont incomparables deux à deux. Alors, si l'on ne considère que son titre, cela pourrait être particulièrement vrai avec le "dictionnaire *mental*" de Michael Zoch. Pourtant la lecture de Zoch [2006] montre plutôt des points de rapprochements. Enfin, comment ne pourrions-nous pas souscrire à : *Contrairement à une hiérarchie avec une seule voie d’accès, dans ce réseau hautement interconnecté il y a presque toujours un moyen d’accéder à l’information recherchée.* Car c'est bien ce que nous recherchons en nous interrogeant sur l'accessibilité de *samouraï* depuis *manger* ou de l'accessibilité inverse de *manger* depuis *samouraï.* Simplement, nous cherchons à définir cette accessibilité selon ce que l'on a déjà. Et l'on s'aperçoit que nous disposons de plusieurs types de dictionnaires (voir Hiatus "observations sémantiques" et observations dans le syntagmeb 4.2.1.2, page 45 et presque tout le restant de ce document). Au fond, l'affirmation de l'unicité du dictionnaire nous semblerait vraiment étrange. En attendant, pour le moment, nous

---

différents motifs. La principale raison est que, contrairement à Wilks [1999] nous doutions de la pertinence de la tâche elle-même. Aujourd'hui la désambiguïsation lexico-sémantique se pratique comme nous la pratiquions entre 1991 et 1993… A cette époque, nous opérions d’abord une discrétisation dramatique des sens du dictionnaire, puis nous cherchions à retrouver nos sens dans les textes. Quiconque pratique sérieusement cette expérience s’apercevra que 1) la discrétisation est généralement impossible par certains côtés (voir par exemple Dutoit [2004]), que 2) les résultats que nous obtenons en discrétisant par mot-sens sont moins des valeurs sémantiques utiles comme face à *samouraï*, suis-je en face de *noble accueillant* ou de *guerrier menaçant* que des génériques grossiers comme face à *samouraï*, suis-je en face de *personne*, de *animé* quand bien même il ne s’agirait que de *statue pétrifiée de samouraï.*

n'espérons aucunement que notre dictionnaire à l'envers retourne *samouraï* depuis *personne qui mange.*
Nous verrons dans ce paragraphe comment le Sémiographe réalise une opération de dictionnaire à l'envers sans rappeler que cette opération à des applications concrètes (voir 4.1.3, page 43) pour la maintenance du dictionnaire.
La figure suivante décrit l'implémentation du dictionnaire à l'envers.

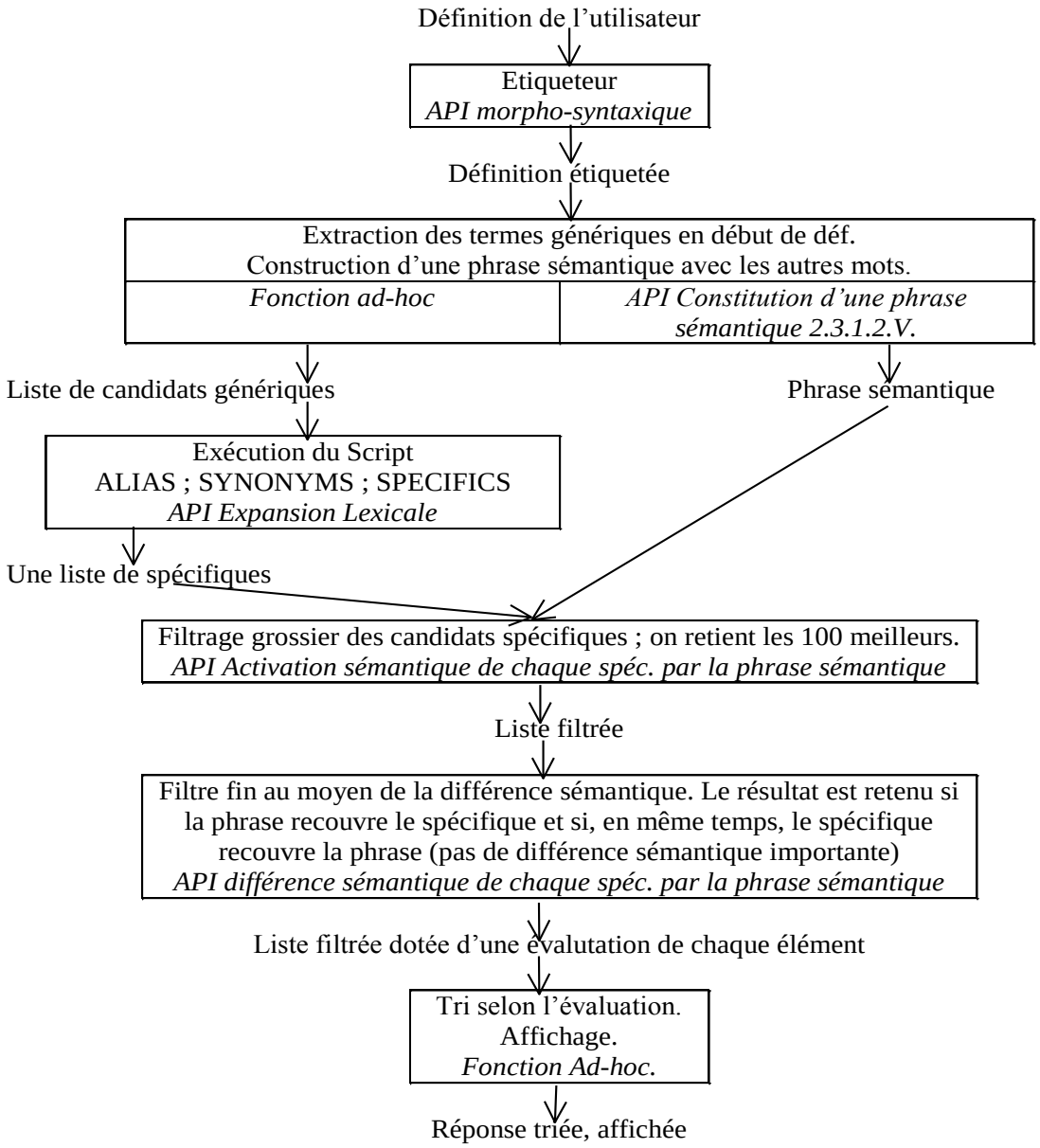


Figure 10 Les traitements[51] du dictionnaire à l'envers

[51] Le traitement "constitution d'une phrase sémantique" n'est pas décrit dans ce document dans l'état où il était à cette époque. Une phrase sémantique est une sorte *tableau noir* sur lequel nous notons pour chaque *token* de la phrase les relations qu'il peut avoir d'autres mots de la phrase à travers tel ou tel LCA (voir 4.1.2.2.3 page 38).

Soit *personne qui vend des hortensias* ? Le tableau suivant détaille les principaux résultats.

Libellé	Crit.1	Crit.2	Crit.3	Origine
Vendeur	817	1.10	1.59	(2-vend:333; 3-fleurs:2000)
Fleuriste	784	1.80	2.76	(2-vend:818;3-fleurs:750;)
Bouquetier	784	2.01	2.76	(2-vend:818;3-fleurs:750;)
Floriculteur	770	2.16	1.11	(2-vend:818;3-fleurs:724;)
Horticulteur	784	2.23	3.2	(2-vend:818;3-fleurs:750;)

Tableau 1 Les quatre meilleurs résultats du dictionnaire à l'envers pour *hortensia*

Le dictionnaire connaît environ 10.000 dénominations différentes de "personne" pour le français. Evidemment, aucune de ces personnes n'est connue comme vendant spécifiquement des hortensias. Pourtant, le dictionnaire arrive aisément à produire cinq solutions augmentées du mot *vendeur* pour lequel le Sémiographe prend un risque… nul.
Dans le tableau 1 ci-dessus, quels que soient les critères, les scores bas sont les meilleurs. Les valeurs dites d'activation utilisent seulement les LCA. Elles sont comprises entre 700 (minimum) et 2000 ; elles correspondent aux première et dernière colonnes.
Les critères 2 et 3 utilisent les chemins et établissent avec quelques variantes la proximité componentielle (minimum : 0 ; maximum : 140). Les échelles sont semi-logarithmiques.
Nous fournissons ci-dessous d'autres exemples :
- *fournir un aliment*, comme *fournir un croissant* produisent évidemment *alimenter* ou *nourrir*.
- *Monnaie du Japon* comme *unité monétaire de Tokyo* produisent *yen*
- *Guerrier noble japonais* ou *japonais guerrier et noble* ou *noble japonais faisant la guerre* ou *chevalier japonais* produisent également *samouraï.*

Il ne faut cependant pas s'étonner de certaines (mauvaises) réponses. Par exemple, *guerrier de sushi* produit *samouraï* avec toutefois un score moins bon du fait des différences visibles depuis *sushi* dans *samouraï* sachant *guerrier* : le côté *nourriture* de *sushi* n'arrive pas à être saturé. Cela confirme combien il est important que le niveau componentiel reste assez pur.

5.1.2.3 Une gestion documentaire multilingue

Le progiciel concerné est une plate-forme de gestion documentaire multilingue. Cette plate-forme comportait déjà des outils morpho-syntaxiques pour les différentes langues installées. Les buts de la plate-forme sont multiples :
- aide à la traduction (par extraction de documents voisins)
- aide à la lecture des documents (documents réactifs)
- extraction d'information, extraction terminologique multilingue etc.

La plate-forme d'origine comporte des outils syntaxiques en six langues (français, italien, anglais, espagnol, hollandais et allemand) et d'importants outils de text-mining.
La contribution de nos contenus était la suivante :
- fournir des relations de synonymie assez précises (automobile, voiture, auto)
- fournir des relations de traduction en mettant en rapport des "synsets" de chaque langue
- fournir des moyens d'accès à des voisins (par exemple *oncle*, *tante*, *neveu*, *nièce*).

La satisfaction du besoin est passée par la production d'une ressource multilingue obtenue par projection du Dictionnaire Intégral sur les vues choisies :

- les fonctions lexicales monolingues
- les fonctions multilingues
- les fonctions Interdep (A est défini par B et B est défini par A comme dans *bananier / banane*).

Le résultat est un dictionnaire multilingue organisé selon les listes analogiques obtenues dans chaque langue avec une projection pour chaque lexie vers un ou plusieurs mots dans les autres langues.

Ce genre d'application est suffisamment demandé pour :
- envisager un développement parallèle en d'autres langues
- augmenter la taille de la ressource
- entretenir des liens étroits avec les autres lexiques généraux comme WordNet [Fellbaum 1998].

5.1.2.4 Aide à la navigation multimedia.

Le plus souvent, les techniques de traitement de la langue sont utilisées pour comparer une requête avec un index texte intégral. Dans cette application, nous avons étudié une hypothèse où un utilisateur ne peut pas saisir de requête.

Ce contexte un peu dérangeant au départ nous a été fourni par le projet européen ITEA-EUREKA AMBIENCE [2000-2002]. Ce projet général étudie différents aspects de ce que pourrait être une "intelligence ambiante" à moyen terme. L'idée d'Ambient Intelligence a été forgée par Philips Eindhoven.

Dans sa phase de réalisation de démonstrateur, Ambience a été divisé en quatre sous-projets dont le démonstrateur Intelligent Multimedia Browsing at Home (MB) conçu par Thomson Multimedia (Rennes). C'est ce sous-projet qui nous intéresse. Dans le scénario d'utilisation, l'utilisateur peut naviguer avec sa voix ou sa télécommande parmi des choix de programme qui s'affichent à l'écran. Il ne peut cependant pas effectuer une recherche texte intégral du fait de l'absence de clavier. Si le nombre total de programmes tv est inférieur à 50, nous pouvons penser qu'une technique de parcours de quelques écrans est acceptable. Mais considérant que l'utilisateur peut accéder à des milliers de programmes téléchargeables il n'est pas possible d'effectuer un parcours exhaustif.

Plusieurs partenaires d'Ambience étaient impliqués dans ce sous-projet :

Partenaire	Pays	Tâche
Thomson Multimedia	France	- Spécification - Interface graphique - Base de donnée - Profil utilisateur - tests
Telisma	France	- reconnaissance de la parole
Epictoid	The Netherlands	- avatar
Vitec	France	- identification par reconnaissance de visage
Memodata	France	- Analyse de textes
VTT	Finland	- Classification

Tableau 2 Organisation du projet Ambience pour la France

Comme nous le voyons, l'analyse textuelle n'est pas au centre de ce démonstrateur très multimédia. Malgré cette position périphérique, nous avons montré comment cette activité de

contenu peut enrichir la qualité globale du service MB, quand bien même l'utilisateur ne peut s'exprimer par des mots.
Le corpus utilisé pour le démonstrateur s'appelle “Internet Movie Database (IMDb)[52]”. IMDb se présente lui-même : *the IMDb is the ultimate online movie database covering over 325,000 titles and over 1,000,000 people with facts, trivia, reviews plus multimedia links from the earliest films to the latest releases.*
En définitive, notre corpus contenait 301.908 émissions documentées par 48.871 résumés rédigés par des utilisateurs du site. Dans IMDb, d'autres types de données étaient disponibles. Ces données non textuelles ont été gérées par VTT qui s'en est servi pour calculer une carte de Kohonen. Il nous restait donc les titres de films et des résumés d'une taille moyenne de 71 mots (corpus total : 21 Mo).
Nous avons tout d'abord pris connaissance du corpus par sondage pour essayer de déterminer nos possibles contributions. Il est apparu que :
- de nombreux résumés n'étaient pas rédigés en anglais (suédois, français, etc.) ; cela nous a amené à utiliser le reconnaisseur de langue intégré à l’API du Sémiographe.
- les résumés de films comportent de très nombreux noms propres. Ces noms propres ont souvent un intérêt documentaire important dans ce genre de base (WW2, Charlot, James Bond, etc.). Une tâche d'extraction et de normalisation des noms propres s'imposait d'autant qu'elle était particulièrement attendue par l'équipe chargée du profil utilisateur (lieux, personnes, organisations sociales, événement etc.).
- l'existence de résumés rédigés en une langue différente de l'anglais montre que les résumés sont susceptibles d'être lus par des locuteurs non anglophones de langue maternelle. Un coup de pouce en aide à la traduction semble utile.
Il reste le besoin d'extraire des thèmes des films (attaque de fourgons, voyage d'exploration, etc.), afin de renforcer le calcul des profils.

En résumé :
- Identification de la langue
- Aide à la lecture des résumés d'IMDb pour un locuteur non natif
- Extraction des entités nommées
- Indexation des thèmes.

[52] www.imdb.com

5.1.2.4.1 Aide à la lecture

Le dictionnaire présenté dans la section précédente, enrichi de la morphologie, a fourni les éléments nécessaires au démonstrateur :

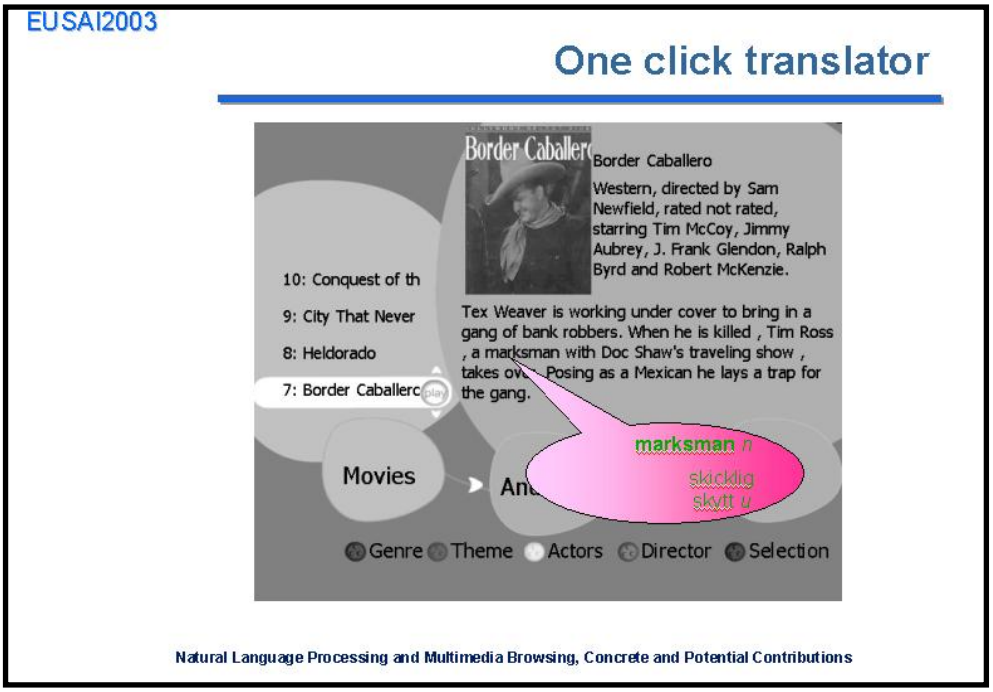


Figure 11 Le démarrage d'Alexandria à l'issue du projet AMBIENCE.

Ce démonstrateur a motivé nos orientations de développement industriel à partir de 2004. Il s'agit du projet Alexandria (il est aujourd'hui dans les 10.000 premiers sites au monde en termes de fréquentation).

5.1.2.4.2 Extraction et normalisation des entités nommées

La démarche est décrite dans Poibeau [2000]. Démarche hybride, cofondée sur des connaissances lexicales et morphologiques et des faits statistiques, elle s'imposait en définitive dans notre travail où était attendue, pour l'anglais, une langue que nous pratiquons moins, une distinction entre noms propres un peu particulière : en effet, il convenait entre autre d'effectuer une distinction depuis les seuls résumés entre nom d'acteur et nom de personnage.
Nous fournissons ci-dessous la liste des métadonnées à déterminer :

TitlesList: *title (title, subtitle, movie title, song title, etc.)*
EventsLis:t *events list*
PlacesLis:t *locations list (country, town, river, montain, building, planet, street, etc.)*

PersonsLis:t *persons list (actor name, band name, god name, wrestler name, etc.)*
CorpsList: *companies list (company, association, group etc.)*
TimesList: *times list (feast, century, date etc.).*

Nous fournissons ci-après deux textes et deux exemples de sortie.

Movie n°553 (english)
The film consists of four stories plus epilogue , set in 19th-century Sicily. THE OTHER SON – A mother spends her life waiting for news from her two sons (emigrated to America) while ignoring her third , because he is the reincarnation of the bandit who raped her. MOON SICKNESS - a newly-wed peasant girl discovers that her husband goes mad every full moon. She arranges for a male friend to protect her , but they end up in bed together just as the moon emerges from behind a cloud. THE JAR - a rich landowner hires a master craftsman to repair a giant olive jar , but the craftsman gets trapped inside. REQUIEM - villagers band together in an attempt to force their landlord to let them bury their dead. CONVERSATIONS WITH MOTHER - the writer Luigi Pirandello talks with his aged mother about a story he always wanted to write , but which he never managed to capture in words.

PersonList: *Luigi Pirandello*	*TimeList* *19th-century* *Time (normalized) : century : 19*
PlaceList: *Sicily* *America*	*TitleList:* *THE OTHER SON* *MOON SICKNESS* *THE JAR* *REQUIEM* *CONVERSATIONS WITH MOTHER*

Tableau 3 Exemple d'extraction d'entités finies

Movie 5503 (english)
Part fact and part opinion , mainly of Jim Garrison and director Oliver Stone , as to the events surrounding the proposed conspiracy of the assassination of President John Fitzgerald Kennedy on November 22 , 1963 in Dallas , Texas. New Orleans District Attorney Jim Garrison began a probe into the actions of The F.B.I. and other officials of whom he suspected where covering up information that could lead to evidence of multiple shooters. The motive is believed to be to escalate the United States involvement in the Vietnam War. President Kennedy was attempting to prevent any further involvement in this situation, but which Vice President Lyndon B. Johnson supposedly promised the United States government that he would "give them the war". Thus, the motive for eliminating President Kennedy. The movie also details the events of many people involved in the assassination, from Lee Harvey Oswald to Clay Shaw, a prominent figure in New Orleans.

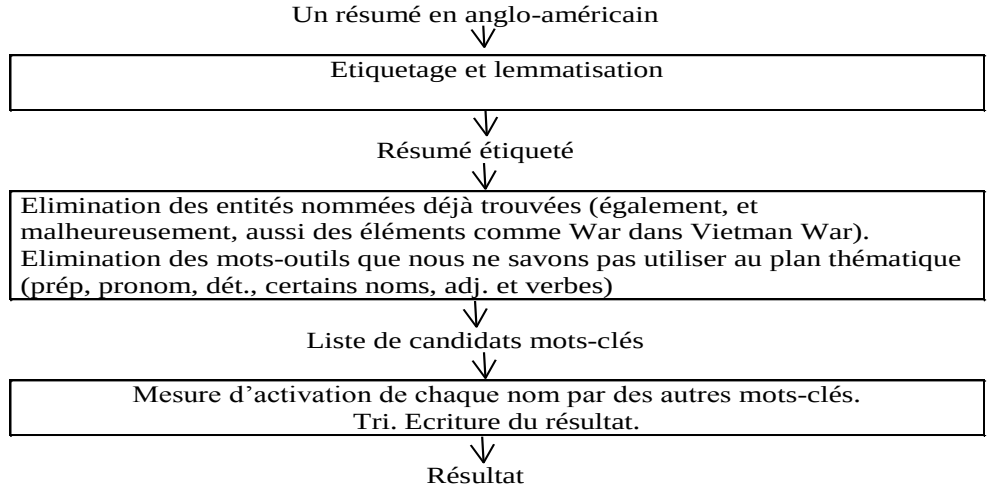
PersonList: *Jim Garrison* *Oliver Stone* *President John Fitzgerald Kennedy* *Attorney Jim Garrison* *President Kennedy* *Lyndon B. Johnson* *Lee Harvey Oswald* *Clay Shaw.*	*EventList:* *Vietnam War*
CorpsList: *F.B.I.*	*PlaceList:* *Dallas* *Texas* *New Orleans* *United States*
TimeList: *November 22 , 1963* *Time (normalized) :century : 20*	

Tableau 4 Autre exemple d'extraction d'entités finies

Les règles obtenues par ce projet, responsables de ces productions, doivent, quand, l'occasion nous sera donnée, être intégrées à l'étiqueteur déjà présenté : s'agissant de la gestion de phénomènes contigus, son utilisation semble adaptée. Aujourd'hui, ces règles ne sont pas intégrées aux API du Sémiographe disponibles pour l'anglais.

5.1.2.4.3 Extraction des thèmes

L'extraction des 'thèmes' des films depuis les résumés est effectuée à l'aide des distances sémantiques du Sémiographe, déjà présentées en 4.1.2.2. Nous avons présenté le principe général : parler de quelque chose suppose au moins l'utilisation de mots définis par ce quelque chose ou de mots définissant ce quelque chose. Nous avons utilisé l'activation pour effectuer cette mesure en raison de sa rapidité. Avant l'activation, nous avons enchaîné les traitements décrits dans la figure suivante.



Cinématique de l'extraction des thèmes

Au plan informatique, le résultat de notre intervention a évidemment pris la forme de fichiers XML enrichis de nombreuses métadonnées. La figure suivante montre un exemple de résultat obtenu :

A famous French filmmaker (Jean-Luc Godard) is hired by a major Hollywood producer (László Szabó) to make a documentary on the state of post-Cold War Russia. The filmmaker , though , subverts the project by stubbornly remaining in France and casting himself as the title character of Dostoyevsky's "The Idiot , " offering up a series of typically Godardian musings on art , politics , the nature of images and the future of cinema.

Personnes :

- Jean-Luc Godard
- László Szabó
- Dostoyevsky

Lieux :

- Hollywood
- Russia
- France

Titres :

- The Idiot

Approche activation :

filmmaker	985	hired:759; producer:593; documentary:600; state:674; filmmaker:250; casting:571; offering:646; series:788; cinema:333;
producer	1299	filmmaker:593; major:771; documentary:571; state:715; offering:683;
documentary	1047	filmmaker:600; hired:597; producer:571; state:643; casting:624; offering:621; series:788; cinema:573;
state	1384	filmmaker:674; producer:715; documentary:643; offering:624;
project	1633	state:795; character:795; offering:865; nature:754;
casting	1504	filmmaker:624; documentary:624; series:830;
character	1546	state:595; project:795; nature:597; images:550;
offering	1499	filmmaker:855; producer:922; documentary:784; state:817;
series	1541	filmmaker:788; documentary:788; casting:830; nature:721;
musings	1689	state:650; series:875; art:671;
nature	1463	state:571; project:754; character:571; series:624;
cinema	1464	filmmaker:771; hired:571; documentary:573; casting:595;

Figure 12 Extraction en anglais, dans *IMDB*, des thèmes et entités nommées

Afin de montrer un exemple plus lisible, nous présentons le résultat suivant dans un style XML assez relâché :

<movie>

<Summary Idfilm=”2564” language=”English”> Tex Weaver is working under cover to bring in a gang of bank robbers. When he is killed, Tim Ross, a marksman with Doc Shaw's traveling show , takes over. Posing as a Mexican he lays a trap for the gang. </Summary>

<PersonsList>

<Pe>Tim Ross</Pe>

<Pe>Tex Weaver</Pe>

</PersonsList>

<ClustersList>

<CL word="bank" value=2172> working:787 gang:1640</CL>

<CL word="gang" value=1342>bank:1754 robbers:1426 killed:1437 marksman:1634 </CL>

<CL word="robber" value=1243> gang:1426 bank:1754 killed:855 marksman:1034

trap:711 ...

<CL word=" trap" value="1330"> robbers:711 marksman:1582 robber:1034 gang:1342</CL>...

</movie>

Comme nous l'avons vu pour l'activation, plus le score est faible, plus l'élément est proche d'un autre élément ou est au centre d'un ensemble d'autres éléments. C'est une histoire de *robber* (CL value = 1243). Le travail d'extraction a été effectué automatiquement pour l'ensemble des résumés en anglais d'IMDB. Sur les 3,5 millions de tokens que comptait la base au départ, nous avons souligné 170.000 tokens comme faisant partie d'entités nommées et déterminées (*personnage* différent de *acteur*, etc.), et 370.000 candidats mot-clé. L'ensemble de ces données ont été ensuite incluses à un algorithme de 'profiling' organisant les films les uns par rapport aux autres, d'une part, et par rapport à l'utilisateur d'autre part.

5.1.2.5 Le développement d'Alexandria

Avec le développement du Sémiographe, plusieurs applications sont réalisables. En laissant aller notre penchant de collectionneur de mots, et puis aussi parce que les réponses à la *Question 5*, page 47, ne sont pas encore données et parce que nous sommes toujours en train d'accroître une liste de cas d'utilisations (voir section 6.1, page 5), nous décidons le développement extensionnel et multilingue du DAG. Cette décision a l'avantage de ne pas créer trop de tension en rapport à l'Observation 5 : le développement sous-jacent reste componentiel. Il s'agit de créer un dictionnaire analogique en autant de langues[53] que les occasions nous le permettront. Parmi les trois applications que nous venons de voir, nous avons donc mis l'accent sur le dictionnaire à l'envers. En opérant ce choix, nous avons choisi de destiner notre travail au grand public comme nous l'avions fait en 1992 avec Dicologique. Le produit retenu a pour nom Alexandria[54] et est servi par le serveur SensAgent[55].

Alexandria est donc un projet de valorisation de certaines technologies linguistiques dont nous disposions et de leur élargissement extensionnel. Cela n'empêche pas la tenue de certaines réflexions sur le fond. Au plan technique, Alexandria est ce que la communauté des développeurs web a appelé plus tard un pop-into : un composant de page web s'ouvrant à l'intérieur d'une page suite à une sollicitation par double-clic sur un mot d'une page web intégrant le composant. Au moment où nous avons proposé ce composant, Alexandria était, en 2005, une vraie innovation, et elle a étonné plus d'un spécialiste. La petite fenêtre web était censée apportée des services :

- correction et recherche phonétique
- définitions, synonymes, expressions, morceaux de LDI et de Wordnet en plusieurs langues
- traductions vers 22 langues.

[53] En relation avec l'ACALAN, voir www.acalan.org, il est par exemple question en ce moment (2008) de développement de l'haoussa.

[54] Il y a un grand nombre de lectures possibles pour ce nom. Nous laissons au lecteur le soin d'évaluer tous les arrangements compositionnels possible, dont le nom complet, pour deviner les différents sens du nom retenu.

[55] il faut lire *sense*A*gent* : voir http://www.sensagent.com

Figure 13 Alexandria sur TV5 monde : la traduction de *retransmise* vers l'arabe

Environ 1000 sites proposent aujourd'hui Alexandria sur leurs pages et trouvent des usagers quotidiens. Alexandria présente également des fonctionnalités "web 2.0" comme l'envoi et l'échange de documents : en effet, il est possible avec cette petite bibliothèque de redéfinir le contenu associé aux mots, par exemple pour afficher des éléments d'un cours, des notes, des annonces, des menus etc. Alexandria et Sensagent.com se placent tous les deux dans une compétition internationale dans laquelle les Nords-Américains ont pris l'habitude d'emporter par KO les marchés à long terme. À ce titre, ainsi qu'au titre de recherches propres à l'informatique (protocoles, services distribués avec répartition de charge etc.), notre équipe a bénéficié d'un premier petit soutien en 2005 de l'ANVAR Basse-Normandie et d'un soutien plus important dans le cadre des appels d'offre de l'Agence Nationale de la Recherche. En effet, nous avons déposé et emporté RNTL cette année-là avec le projet VODEL.

5.2 Projets et documents

Projets

Projet 8 VIVIAN

En 2000, nous avons participé sur l'invitation de Nokia à un projet, Vivian, concernant des services à distance pour les téléphones mobiles. Il s'agissait pour Nokia de concevoir et d'évaluer un middleware permettant d'échanger des données entre serveurs et mobiles, en vue, par exemple et en ce qui nous concerne, d'utilisation de dictionnaires à distance. [Tuominen, 2000] Il s'agissait pour nous de découvrir de nouvelles technologies et protocoles (SOAP, XML etc.). Le projet s'est achevé fin 2002.

Projet 9 Balkanet

En 2001, nous avons été demandés comme conseiller et évaluateur du projet Balkanet. Ce projet visait à l'établissement de réseaux sémantiques type WordNet pour les langues balkaniques (grec, serbe, tchèque, turc, bulgare, roumain). Nous y transmettons les connaissances et savoirs faire technologiques acquis. [Christodoulakis, 2000]

Projet 10 Ambience

En 2002, suite à un cours de 5 jours donné au Centre de Recherche de Thomson, nous avons été invités à participer comme fournisseur de technologie au projet Ambience.
Ambience est le mot retenu par Philips pour désigner tous les systèmes communicants (de la maison, du bureau etc.) comportant une intelligence embarquée. Nous y avons implémenté des solutions linguistiques avancées, dérivées de notre Sémiographe. Nous les décrivons dans ce rapport [Van Loenen, 2002]. L'application a été montrée et saluée par le Roi de Belgique et l'ITEA Office Board.

En 2004, nous avons collaboré avec la Pusan National University de Corée pour faire bénéficier de notre expérience en termes de réseaux sémantiques multilingues.

Projet 11 Le DES

Le DES : le Dictionnaire Electronique des Synonymes en collaboration avec le laboratoire CRISCO (Université de Caen)

Projet 12 OSEO ANVAR

Développement du dictionnaire Multilingue et déploiement d'Alexandria

Projet 13 VODEL

VODEL[56]*"Valorisation Ontologique des Dictionnaires ELectroniques"*. Nous sommes responsables du consortium qui comporte 6 participants.

VODEL se situe dans le cadre de la recherche d'information sur Internet. Son objectif principal est l'exploitation conjointe des ontologies et des dictionnaires électroniques afin de profiter pleinement de leurs atouts respectifs. Plus précisément, il s'agit de permettre à un utilisateur non expert d'accéder à des documents indexés par une terminologie (voire une ontologie) métier par le biais de définitions d'un dictionnaire métier adossé par des liens sémantiques à un dictionnaire généraliste. Une idée importante du projet est de rechercher les termes de la requête experte en utilisant un vocabulaire non spécialisé. Les résultats attendus sont des algorithmes d'analyse de dictionnaires informatisés, une architecture logicielle pour la coopération de tels dictionnaires avec des ontologies ou terminologies métiers pour la recherche d'information sur le Web ou dans des bases métiers et une interface de navigation contextuelle. Le projet VODEL a été évalué à travers 3 applications pilotes :
- dans le milieu hospitalier afin d'ouvrir aux patients l'accès à des informations médicales en ligne (site médical CISMeF du CHU de Rouen) ;
- dans le service qualité d'EADS où il s'agit de montrer comment une information contextuelle dotée de liens peut faciliter l'appropriation et la mémorisation des principes et normes concernés ;
- dans le cadre du progiciel documentaire Sinequa CS, pour créer une ergonomie nouvelle depuis les résultats du moteur d'indexation ; La proposition se situe dans l'orientation thématique « réseaux d'information et de connaissance » dans l'axe prioritaire 2.3.
Les objectifs de recherche consistent à renforcer globalement les trois caractères suivants :
- Polyvalence du document
- Ubiquité de la recherche
- Perméabilité domaine métier / langue générale
VODEL a réuni 6 partenaires (3 universitaires et 3 industriels) qui disposent de compétences complémentaires :
- Coordinateur : la société **Memodata** qui est reconnue pour trois outils : le Dictionnaire Intégral, le Sémiographe et Alexandria. Dans le projet, elle est spécialiste des dictionnaires électroniques et porteur de l'application Alexandria.
- Partenaire 2 : le laboratoire **LASELDI** est reconnu pour ses outils NooJ et INTEX. Dans le projet, il est spécialiste du traitement de la langue naturelle et intégrateur des applications INTEX et NooJ.
- Partenaire 3 : le laboratoire **LITIS (ex PSI)** a pour thème de recherche privilégié de recentrer les systèmes de recherche de document ou d'information sur l'utilisateur. Dans le projet, il apporte ses compétences dans ce domaine ainsi que ses connaissances en fouille de document.
- Partenaire 4 : la société **EADS** a conçu et développé une plate-forme ouverte basée sur des standards (XML, RDF). Cette plate-forme est constituée de composants indépendants garantissant la complétude de la chaîne de veille (recherche, acquisition, filtrage, extraction, distribution, visualisation et aide à la décision). Acteur reconnu dans le domaine du text-mining et de la veille économique, EADS apporte un cadre applicatif pour VODEL dans le domaine de la gestion de la qualité.
- Partenaire 5 : Les travaux de Stéfan Darmoni et de son équipe du CHU de Rouen portent sur le projet **CISMeF** (Catalogue et Index des Sites Médicaux Francophones). Ses objectifs sont de décrire et d'indexer les principales ressources en santé disponibles en français et d'assister

[56] Voir http://vodel.insa-rouen.fr

les professionnels de santé et les patients dans leurs recherches de documents sur l'Internet. En plus d'un rôle d'évaluateur, CISMeF apporte à VODEL leurs compétences en indexation et recherche d'information.
- Partenaire 6 : Depuis l'année 2000, une entité de Recherche a été créée sous le nom de **Sinequa Labs**. Le laboratoire apporte donc au projet VODEL son expertise, sa technicité et les besoins de ses clients en relation avec son logiciel documentaire Sinequa CS.

Projet 14 INTERSTIS

Le but du projet InterSTIS est de rendre interopérables au sein d'un « serveur terminologique multi-sources » les terminologies médicales francophones usuelles comme la SNOMED pour le codage d'informations cliniques, la CIM-10 et la CCAM pour le codage médico-économique, la CISP utilisée par les médecins libéraux, le MeSH pour la bibliographie, et d'autres terminologies propriétaires.
Partenariat :

VIDAL SA, coordination, Issy-les-Moulineaux, LERTIM, direction scientifique, Faculté de Médecine, Université de la Méditerranée, Marseille, Mondeca, industriel, Paris, Memodata, industriel, Caen, Equipe CISMeF, CHU de Rouen, LIMSI, équipe CNRS, Orsay, DSPIM, Faculté de Médecine, Saint Etienne, HON, Fondation Health On the Net, Genève, LabSTIC, Faculté de Médecine, Université de Nice-Sophia Antipolis.

Au plan des résultats, les financements obtenus nous ont permis d'industrialiser jusqu'à un certain point la solution tandis que nous étions prêts à attendre pour reprendre les travaux sur le Sémiographe et LDI des machines plus puissantes. Avec Vodel, Alexandria est passé de 250.000 à 30.000.000 de fiches.
Observons que nous finissons cette présentation concernant quinze années de travail un peu comme nous avons débuté :
- avec un produit public, visant cette fois un marché au moins européen, sinon mondial
- une mise en attente des travaux de recherche exploratoires en fonction de financements dédiés et d'une disponibilité de machines suffisamment rapides pour pouvoir travailler
- des perspectives de recherche nouvelles que nous allons évoquer dans notre chapitre 3 : du développement de nouveaux parcours interprétatifs universalistes et aprioriques. Cette présentation aura finalement exactement le même statut que notre article Coling 92 et parfait le parallélisme des situations entre les deux périodes tout en modifiant l'intensité, ampleur et échelle.

Stage, encadrement, collaboration

Participation à un jury de thèse :
14 Décembre 2007, Marianne Dabbadie "Recherche d'un méta-modèle d'évaluation basé sur le sens pour l'évaluation des systèmes d'accès à l'information". Université de Lille.
Ce travail présente EVALIR qui évalue le SEMIOGRAPHE comme métrique d'évaluation des moteurs de recherche.

Durant cette période, nous avons beaucoup collaboré. Sans donner une liste complète, nous soulignons notre collaboration avec Thierry Poibeau, alors en préparation de thèse, et aujourd'hui chargé de recherches au CNRS. La thèse, publication et ouvrage de Thierry font utilisation ou référence parfois importantes à nos productions.

Samuel Parfouru a effectué son stage de Master 2 chez nous. Aux dernières nouvelles, Samuel terminait sa thèse en CIFRE chez EDF (traitement automatique des langues, multimodalité, dialogue).

Michael Riotte. Université de Provence (Master 2).
Michael (mention Bien) a eu pour mission de spécifier les opérations de transfert à automatiser pour passer d'un énoncé de langue comme *quelle est la couleur du cheval blanc d'Henri IV?*, à sa résolution à l'aide d'un moteur de premier ordre tel que SUMO[57] (Niles [2001])

Formation entreprise : 2 sessions de formation en TAL et sémantique lexicale (2*5 jours)

Plusieurs participations à des jurys de conférence

Une organisation et deux co-organisations de journée de l'Atala.

Communications et dissémination
Les publications ont bénéficié de l'impact positif de mon poste de Directeur de recherche Associé au CNRS (contrat de trois ans, représentant une journée par semaine).

Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, Maria Grigoriadou [January 2002], Balkanet: A multilingual Semantic Network for Balkan Languages, In Proceedings of the First International WordNet Conference, Mysore India.

D. Dutoit , T. Poibeau : « Inferring knowledge from a large semantic network » [Août 2002], full paper, acte de Conference on Computational linguistics, COLING TAIWAN

D. Dutoit , T. Poibeau : « Generating extraction patterns from a large semantic network and an untagged corpora » [Août 2002], acte de Workshop, COLING, TAIWAN.

Dutoit D, P. Nugues : « A lexical network and an algorithm to find words from definitions », acte de European Conference on Artificial Intelligence [2002] ECAI, LYON.

Dutoit D, T. Poibeau : « Évaluer l'acquisition semi-automatique de classes sémantiques », [2002] acte de TALN.

D. Dutoit , T. Poibeau : « Evaluating resource acquisition tools for information extraction », [May 2002], full paper, acte de Language resource and evaluation, LREC, Las Palmas

[57] http://www.ontologyportal.org/

Dutoit D, P. Nugues : « The right word », [May 2002], full paper, acte de Language resource and evaluation, LREC, Las Palmas

Dutoit D, P. Nugues , P. de Torcy: « The Integral Dictionary : a lexical network based on computational semantics », [May 2003], Springer Ed., ICCSA International Conference on Computational Science and its Applications, Calgary, Canada

Dutoit D, Y. Picand , P. de Torcy, Roger G. [2003]: *Natural Language Processing and Multimedia Browsing, Concrete and Potential Contributions*, European Symposium on Ambient Intelligence - Eindhoven, The Netherlands.

F Soufflet, S Le Huitouze, Korpipaa P, D Dutoit, P Ten Hagen, F Kuijik, O Guye, JR Vigouroux, L Chevallier [2003] *Multimedia browsing*, European Symposium on Ambient Intelligence - Eindhoven, The Netherlands.

Dutoit D, P. Nugues , P. de Torcy: « The Integral Dictionary: An Ontological Resource for the Semantic Web » [May 2004], full paper, acte de Language resource and evaluation, LREC, Barcelona

D. Dutoit, P. de Torcy, Y. Picand, « Quelques contenus généraux au service des documents », [June 22 – 25, 2004], 17 pages, CIDE 7 Conférence Internationale sur le Document Electronique, La Rochelle, France.

J. François, D. Dutoit, [2006], *Compte-rendu de "Sémantique et traitement automatique du langage naturel", de Patrice ENJALBERT (dir.), publié chez Lavoisier / Hermès Science Publications [2005]*. Bulletin de la Société de Linguistique de Paris.

D. Dutoit, J. François [2008], *Changer et ses synonymes majeurs entre syntaxe et sémantique : le classement des verbes français en perspective* Revue Langue Française, édition Larousse, France.

5.3 Conclusion

La période qui s'est maintenant achevée a débuté avec plusieurs orientations de travail qui aboutirent de la manière suivante :
- mise en œuvre du Sémiographe dans différentes applications cible afin d'évaluation
- développement d'une vraie application qui est diffusée (Alexandria et plusieurs services associés)
- réalisation d'un éditeur d' hypergraphe et support du multilinguisme aux plans techniques et des interfaces
- extension multilingue du modèle

Concernant la recherche au plan fonctionnel (la définition du quoi faire), la période a abouti à des progrès concernant les deux questions que nous avons posées :
- d'une part, nous avons élaboré un corpus à mettre en rapport avec la question posée (voir 6.3.1.3 *Un corpus plus étendu de* cheval blanc, page 114)
- et en particulier, nous avons au moins une solution au problème du *cheval blanc* (voir chapitre 6.3.2 *La résolution du cheval blanc*, page 115) compatible avec tout ce corpus
- d'une façon générale, nous avons élaboré conçu un nouveau champ d'expansion du dictionnaire onomasiologique qui pourrait devenir progressivement intensionnel (voir chapitre 7 Conclusion, page 133).

6 INTEGRATION STRUCTURALE DES POINTS DE VUE COMPONENTIELS ET COMPOSITIONNELS : POURQUOI ET COMMENT

Il n'est pas plus possible de fabriquer un dictionnaire sans s'occuper de l'*usage* que nous trouvons en particulier dans les énoncées et les textes qu'il n'est possible de s'intéresser aux énoncés ou aux textes sans s'intéresser au repérage de l'*usage*. Ainsi, rien dans ce chapitre ne sera absolument nouveau par rapport à tout ce que nous avons déjà présenté dans les chapitres précédents. Dans les chapitres précédents, nous nous sommes intéressé à des sélections en contexte de traits dans un axe componentiel. Nous avons déjà observé dans le chapitre 4.2.1 *Les deux hiatus*, page 44, et dans de nombreux autres endroits des difficultés particulières à effectuer de manière acceptable ou systémiquement cohérente cette sélection. Ce chapitre propose une voie proprement structurale de résolution de la plupart de ces difficultés.

Au plan didactique, jusqu'ici, notre propos n'a été qu'une illustration que l'on voudrait de plus en plus fine de l'affirmation suivante : les mots actent dans des espaces ***psychiques*** (page 7) de nature langagière que l'on doit aussi soigneusement que possible (voir note 31, page 37) séparer de la référence, plus particulièrement, de la référence dans quelque chose d'extérieur qui serait le Monde. Jusqu'ici, nous avons toujours parlé de points de vue (voir paragraphe 4.2.2 *Les changements de points de vue concernent les choses les plus simples*, page 45). Par exemple, nous avons argué qu'un signe (par ex. *samouraï*) n'est jamais tout à fait lui-même quand il devient une occurrence ; cela signifie qu'en tant que tel il comporte à la fois un équilibre à travers la définition et un déséquilibre à travers l'occurrence (Dutoit [2004]), c'est-à-dire un déséquilibre entre sa définition et son développement extensionnel. [58]

---

[58] Par exemple, supposons que nous voulions bien considérer une règle de jeu, de jeu d'échecs par exemple, comme une définition, en fait, comme un ensemble de définitions constituant un tout qui est le jeu. Un état particulier de déroulement d'une partie peut aussi être considéré comme une occurrence pourvu qu'il soit historiquement atteignable par des étapes telles que chacune de ces étapes respecte les points de règle concernés par l'étape. Mais en soi un développement particulier, une extension de la règle du jeu dans une partie introduit deux déséquilibres :

1°) un état d'une partie peut correspondre à un grand nombre d'histoires du déroulement du jeu; de plus, dire que cet état peut être atteint par la règle du jeu n'implique pas que cette règle de jeu précise ait été impliquée dans le déroulement d'une partie utilisant les pièces que l'on perçoit. Nous pourrons toujours douter de la règle-tout

Brisons-là tous les développements dialectiques qui pourraient prolonger cette introduction. Nous reprendrons ces raisonnements ponctuellement ou synthétiquement seulement dans la mesure où nous en aurons besoin. Nous nous intéressons aux interactions entre signes. Nous nous intéresserons juste à ces interactions qui pourraient découler de la langue vue comme un système.
Le titre de ce chapitre est *intégration structurale des points de vue componentiels et compositionnels : pourquoi et comment.* Avant de présenter comment ce chapitre va se développer, intéressons-nous dans un premier temps à interpréter correctement ce titre un peu long.
Le mot *intégration* doit être pris pour tous les sens qu'il connaît, à savoir :

1°) opération inverse de la différentiation
2°) établissement d'une interdépendance plus étroite entre des parties
3°) Incorporation
4°) Coordination des activités de plusieurs organes, nécessaires à un fonctionnement harmonieux (d'après Robert).

Le sens 1, étant donné le complément dans notre titre de *intégration*, signifie que nous allons faire en sorte de déterminer une manière particulière de considérer la question de la distinction point de vue compositionnel/componentiel telle que dans cette manière la distinction n'opère plus, ou plus exactement n'a plus d'effet parasite gênant.
Le sens 2 renvoie au procédé de cette indifférenciation : nous ferons en sorte d'établir davantage de dépendances entre des parties d'un tout qu'il nous faudra préciser.
Le sens 3 insiste sur l'existence du tout : action de faire entrer (un élément) dans un tout.
Le sens 4 précise le but : il s'agit de coordonner plusieurs organes afin de permettre un fonctionnement harmonieux d'un organisme.
Dans notre cas, l'organisme est un système. Ce système comporte des organes dont les plus nombreux sont des agents réflexe simples dans la terminologie de Stuart Russel [2006, p. 53] : *agents qui sélectionnent une action en fonction du percept courant et ignorent le reste de l'historique des percepts*. Il vient qu'un agent qui a vérifié une perception[59], émet une action. Il nous faut donner cette action. Cette action est une modification du système. La plupart de ces modifications seront des ajouts dans le système. A chaque instant, le système a un certain état. Nous donnons à cet état le nom de structure. Cette appellation nous convient tout-à-fait puisqu'au plan linguistique chacun pourra vérifier la forme structuraliste de n'importe quel tout que nous allons découvrir.

---

depuis l'occurrence-partie.
2°) D'autre part, si à un état observé du déroulement du jeu, tous les fous ont été sortis du jeu, les règles concernant les fous ne sont plus accessibles. Cela ne veut pas dire que ces règles n'existent pas dans cet idéal qui est la règle du jeu prise globalement. Notre exemple de 2004 avait fait ce genre d'analyse sur un signe (une règle de jeu si l'on veut) très élémentaire : le signe graphie-son *i* du point de vue du dictionnaire.

[59] *Nos sensations sont purement passives, au lieu que toutes nos perceptions ou idées naissent d'un principe actif qui juge.* (Rousseau). Il est intéressant de compléter l'article du Robert : *Fonction par laquelle l'esprit se représente les objets; acte par lequel s'exerce cette fonction; son résultat*. Et les exemples : *Perception et imagination*. « *Quand je dis : "l'objet que je perçois est un cube", je fais une hypothèse que le cours ultérieur de mes perceptions peut m'obliger d'abandonner. ... Dans la perception, un savoir se forme lentement* » (Sartre).
Nous trouvons l'exemple du cube chez Bergson [1907], avec une tonalité complémentaire : *J'ai donné* en esprit *une forme à ce que j'entends; cette forme n'est pas dans les sons mêmes. Dans un cube dessiné en transparence ... chacune des deux faces peut indifféremment me paraître à l'avant ou à l'arrière du cube. Le dessin ne change pas ; c'est* en esprit *que je l'organise différemment. L'esprit détient le pouvoir d'organiser pour lui-même les choses sans y changer quoi que ce soit en réalité.* Voir note 112 page 89, la conclusion où nous reprenons ces notions après les avoir postulées pour notre analyse linguistique qui dégage un lieu de *pragmatique abstraite*.

Le titre du chapitre est finalement clair. Il reste à discuter du pourquoi et du comment. Nous en venons au plan du développement.
Pour procéder à l'élucidation des deux questions *pourquoi* et *comment*, nous traitons des questions analytiques variées couvrant un très grand nombre de phénomènes, qui appartiennent ordinairement à différents niveaux de l'analyse linguistique, et ont leur propre bagage théorique, leur propre input et output et leur propre formalisme. Selon les cas, nous traitons de ce que nous distinguons couramment sous les mots de morphologie, de syntaxe, de sémantique et de représentation des connaissances. Pour chaque élément de l'inventaire, pour chaque question analytique, nous montrons que la spécificité des inputs/outputs et formalismes introduit en tant que tel des effets parasites très gênants. C'est le niveau 1 de chacun de nos paragraphes. Il justifie localement, pour une question donnée, le *pourquoi*. Le niveau 2 est l'élimination du caractère spécifique des inputs/outputs et formalismes. Il répond pour chaque question analytique abordée au *comment*. Le niveau 3 est un résultat unique dans lequel nous disons la même chose concernant le caractère homogène de ce que nous avons fait selon les quatre critères suivants :

C1 : unicité de la Structure
Il y a ou non hétérogénéité des représentations (structure input, c'est-à-dire structure).
C2 : unicité de l'analyse
Il y a ou non hétérogénéité des analyses (analyses).
C3 : unicité du résultat
Il y a ou non hétérogénéité des résultats (structure output, c'est-à-dire structure).
C4 : unicité du Signe
Il y a ou non hétérogénéité de la perception du signe (le signe).

Il est important de rappeler ici que du fait que depuis 1992, nous travaillons sur un Dictionnaire Intégral[60] (c'est-à-dire plusieurs dictionnaires[61] qui selon Dubois sont tous des tentatives particulières de décrire un objet, et qui ne peuvent être confondus avec cet objet), la conservation de l'homogénéité dudit objet supposé (C4 : le signe selon un point de vue et sa représentation formelle) est le sujet le plus immédiatement sensible[62] pour nous.
En résumé, dans ce chapitre nous prenons individuellement différentes questions analytiques, nous montrons une ou deux difficultés inhérentes aux traitements directs de ces questions, nous levons ces difficultés par une intégration et nous concluons par la forme prise par l'intégration.
Le chapitre a aussi pour objet de faire apparaître progressivement le fonctionnement assez complexe de l'ensemble. C'est pourquoi, nous commençons par nous intéresser à des choses très simples qui posent déjà des problèmes d'intégration. Six choses simples sont traitées dans le chapitre 6.1 *Intégration d'énoncés compositionnels*, page 76. Le problème que pose ce chapitre est immédiatement perceptible dans le titre : si ces énoncés sont compositionnels, nous ne voyons pas bien pourquoi il faudrait les intégrer au niveau componentiel. Mais c'est entre autre ce qu'expose 6.1 : leur non-intégration au niveau componentiel crée toujours une catastrophe. Le chapitre 6.2, *L'intégration de faits sémantiques qui ciblent à la fois des points*

[60] voir paragraphe 3.1.1, page 21, et son pendant paragraphe 3.1.2, page 24.
[61] Dictionnaire morphologique qui donne des natures et des paradigmes flexionnels, dictionnaire de dérivation lexico-sémantique qui donne des emplois en rapport avec l'organisation syntaxique de la phrase, dictionnaire grammatical sans lequel le dictionnaire précédent n'aurait pas de socle, dictionnaire statistique des occurrences, dictionnaire onomasiologique, dictionnaire sémasiologique, dictionnaire des synonymes, dictionnaire de constructions, dictionnaire de dépendances etc.
[62] Le mot est ce par quoi débutent et aboutissent toutes nos analyses. Voir 2.1 *Positionnement de nos travaux* page 7.

de vue compositionnels et componentiels, page 108, aborde quatre exemples de problèmes complexes. Nous nous contentons de donner ces problèmes complexes et une direction structurale de leur résolution, sans leur donner une résolution complète, parce que par définition le traitement d'un objet complexe suppose la perception holistique de plusieurs phénomènes dont, dans ce chapitre, nous n'avons pas encore la liste. Ce chapitre traite donc davantage de *pourquoi* que de *comment*. A l'inverse, le chapitre 6.3 *La définition d'une microsyntaxe pour élargir un peu le champ perceptif de la Structure*, page 109, prend en entrée l'énoncé à résoudre le plus élémentaire que nous ayons imaginé (*quelle est la couleur du cheval blanc d'Henry IV*) pour répondre d'une façon assez générale à la question *comment*. La forme prise par cette réponse à cette question *comment* est déjà assez complexe et nous comprenons pourquoi, sauf à l'aide d'un ordinateur, nous serions très en difficulté pour répondre *convenablement*, sans théorie ad-hoc aux problèmes posés en 6.2. A un moment donné, il convient de réaliser des instruments. La réalisation de ces instruments est la partie terminale de la conclusion de ce mémoire.

6.1 Intégration d'énoncés compositionnels

Ce chapitre traite d'un premier niveau d'intégration en n'abordant que de choses simples. Au moyen de ces choses simples, il introduit certains mécanismes réflexes dont nous aurons besoin par la suite. En même temps, le chapitre raisonne sur un point limite de l'intégration : après tout, il est presque contre-intuitif qu'il faille intégrer des choses proprement compositionnelles dans un espace proprement componentiel. Le paragraphe 6.1 traite donc de deux questions : le mécanisme et l'opportunité même s'agissant de cas limite.
Pour traiter de ces questions d'une manière suffisamment exhaustive pour que cette manière puisse arrêter notre attention, nous avons retenu six classes de phénomènes. Ces classes sont toutefois considérées depuis un exemple. Le point commun de chacun des cas et des exemples est leur caractère de simplicité. Tous les cas et tous les exemples sont faiblement décomposables. En effet, tous semblent aboutir à une sélection contextuelle d'une partie assez autonome du signe. Selon les cas, cette partie est le signifiant, tout le signifié, une partie du discours etc. Ces cas sont titrés :

- Intégration de la morphologie compositionnelle
- Intégration d'énoncés compositionnels métalinguistiques
- Intégration d'une grammaire syntagmatique
- Intégration du terme
- Intégration d'une date
- Intégration d'une formule

Comme nous l'avons dit dans l'introduction du chapitre 6, chaque intégration sera évaluée selon quatre critères que nous rappelons ci-dessous :

C1 : unicité de la Structure
Il y a ou non hétérogénéité des représentations (structure input, c'est-à-dire structure).
C2 : unicité de l'analyse
Il y a ou non hétérogénéité des analyses (analyses).
C3 : unicité du résultat
Il y a ou non hétérogénéité des résultats (structure output, c'est-à-dire structure).
C4 : unicité du Signe
Il y a ou non hétérogénéité de la perception du signe (le signe).

Nous avons essayé de choisir nos exemples de telle manière qu'ils illustrent un point de vue unique. Malheureusement, comme indiqué par la note 31 page 37, cela n'est en général pas

possible. Ainsi, dans le chapitre, nous serons obligé d'attirer l'attention sur le point de vue qui nous intéresse.

6.1.1 Intégration de la morphologie compositionnelle

Nous traitons ici uniquement de morphologie des termes composés réputés strictement non componentiels.
Prenons l'exemple prototypique :

pomme de terre.

Nous n'abordons pas le fait que pour nous il puisse y avoir quelque chose de componentiel dans *pomme de terre.* Cela n'importe aucunement ici. Le problème posé est le suivant. Le traitement strictement compositionnel de la locution porte atteinte à :

C1 : l'unicité de la Structure

Pour décrire un groupe de signes reliés entre eux (ici *pomme+de+terre*) il faut remettre en cause l'homogénéité de la représentation, c'est-à-dire créer des agents réflexes de perception non atomistiques.

C2 : l'unicité de l'analyse

Pour repérer un groupe spécifique de mots reliés entre eux il faut créer un module spécifique de gestion de ces agents non atomistiques

C3 : l'unicité du résultat

Le traitement du résultat du module spécifique amène à devoir choisir entre trois signes et un signe.

C4 : l'unicité du Signe

Il est affirmé ici que certains signes n'ont pas de sens; dans notre exemple, il est affirmé que *pomme* dans *pomme de terre* n'a pas de sens. Nous ne parlons pas évidemment d'un sens lié à une interprétation componentielle possible. Nous ne parlons que de l'affirmation proprement dite dans la limite du champ perceptif d'un capteur syntagmatique.

Mise en contexte et effets

Débutons par C3]. Benoît Sagot[63] décrivant SxPipe souligne un important principe de son système : *Un des principes sur lesquels repose SxPipe est la préservation des* ***ambiguïtés****. En effet, une succession linéaire de traitements accumule progressivement des informations sur le texte. Mais certains traitements peuvent ne pas disposer de toutes les informations nécessaires pour effectuer certains choix. Dans ce cas, SxPipe fait le choix, autant que possible, de préserver les ambiguïtés, retardant ainsi la prise de décision à une phase ultérieure qui disposera de plus d'éléments*[64].
Le sémiographe connaît la difficulté soulevée par B. Sagot depuis de nombreuses années. L'observation même de l'architecture découpée en modules (voir 4.1.2.1 *APIs phonétiques, morphologiques, morpho-syntaxiques et d'expansion lexicale*, page 34) implique une cinématique comme celle donnée dans la *Figure 10 Les traitements du dictionnaire à l'envers,*

[63] Voir http://alpage.inria.fr/~sagot/sxpipe.html
[64] Il ajoute juste après *Ceci nécessite que les modules concernés sachent produire en sortie, mais aussi prendre en entrée des entrées ambiguës (des DAG, ou graphes orientés acycliques).* Nous sommes d'accord sur le fond avec cet ajout mais nous ne pensons que des DAG (lesquels) soient suffisants. D'autre part, les DAG de Sagot sont utilisés pour conserver des ambiguïtés alors que les nôtres sont conçus pour lever ces ambiguïtés.

page 56, et un problème insurmontable de choix : nous n'imaginons pas une boucle revenant en arrière depuis le module sémantique jusqu'au module morphologique. Nous trouverions cette boucle très insupportable à gérer. A contrario, nous pourrions envisager comme Sagot le propose de conserver l'ambiguïté. Mais, sauf à ajouter quelque chose de neuf dans la Structure (C1), cela n'est pas possible dans la structure courante. Il serait bien sûr possible de créer une structure supplémentaire mais cela toucherait à l'unicité de la Structure.

Nous sommes en définitive, comme le dit, Sagot amenés à décider entre deux signes recouvrant un même ensemble de places sans disposer de critères pour prendre cette décision. D'une façon générale, le calcul des locutions dans le Sémiographe de 1996 est rarement effectué parce qu'il ne concerne pas le cœur de ce que nous avons souhaité observer. Cependant, l'expérience montre que plus de 20% des erreurs (bruit ou silence) de l'outil vient de la non-gestion d'un certain degré de figement. Malheureusement, nous n'avons pas de mesure très exacte. Nous pensons que cette mesure dépendrait de toute façon du corpus et de notre capacité prétendue de juger d'une chose selon un unique point de vue. De toute façon, une mesure plus exacte est inutile en pratique : 20% d'erreurs c'est considérable. Cela mérite d'être considéré. C'est ce que nous faisons maintenant dans ce chapitre traitant d'une tentative d'intégration.

Solution

Dans notre discussion, nous avons mis de côté C4. La rupture C4 concerne ce que nous allons maintenant appeler Atteinte à l'Unicité du Signe : A4.
Nous avons indiqué que cette rupture A4 tient en l'affirmation suivante : dans la vue syntagmatique particulière de *pomme de terre, pomme* n'a pas de sens. Comment un signe tel que *pomme* peut-il n'avoir aucun sens dans une vue donnée qui le définit en terme d'emploi ?

Dans *pomme de terre*, nous proposons de considérer que *pomme* a pour sens[65] :

élément de *pomme de terre*.

Nous écrivons :

pomme n.f. **--e1-->** pomme de terre loc.f. **dans** pomme de terre loc.f

Détails sur le formalisme

A] Observations générales sur **dans** ou la notion de *lieu*

1) Nous avons déjà vu une notation très semblable (paragraphe 5.1.1 *Le moteur de LDI devient un hypergraphe*, page 52). Elle permettait de distinguer une relation de LDI et une relation de WordNet. Plus précisément, elle permettait de concevoir tout LDI et tout WordNet d'autre part comme des graphes différents.
2) Ici l'usage de la notation est exactement le même. En effet, la notation asserte que *pomme de terre* est un graphe (un lieu particulier de réalisation d'une certaine manière de concevoir

[65] Il n'y a pas ici de confusion entre *signe* et *graphie* ou entre *signe* et *représentation phonétique*. Dans tous les cas, il y a d'abord perception de quelque chose pris comme un tout : ici, ce tout est *pomme*. Nous ne pouvons imaginer que ce tout n'est pas de justification, ne puisse être pris dans un certain sens, c'est-à-dire selon *une certaine idée intelligible à laquelle un objet de pensée peut être rapporté et qui sert à expliquer, à justifier son existence* (Robert) puisqu'alors la perception même dudit tout perdrait son fondement. Voir note 59 page 72.

un sens) et que dans cette manière particulière, le mot *pomme* a, d'une façon intrinsèque à cette manière, un certain sens.
3) intrinsèque signifie que les deux éléments (ici *pomme* et *pomme de terre*) existent en même temps en un lieu *pomme de terre,* établissent l'existence de ce lieu, et sont tels que si l'un quelconque des termes manquait, le lieu lui-même sinon n'existerait plus du moins trouverait un autre degré de vraisemblance.

B] Notes sur la notion d'*élément*

1) Nous sommes en face d'un objet trouvé dans un lieu particulier. Nous avons donné à cet objet le nom d'élément pour éviter la confusion avec la notion informatique d'objet.
2) Nous avons indicé cet élément en l'appelant **e1**. Cet indiçage est arbitraire et n'a rien à voir avec une notion de priorité dans une série. Nous l'avons appelé e1 simplement pour le distinguer d'e2, e3 et de tout autre élément occupant et constituant le lieu que nous décrivons.

Portées du formalisme

Figure 14 a. Pomme dans pomme de terre **Figure 14.b pomme de terre**[66]

13a. *pomme n.* apparaît dans le DAG comme un simple élément : il n'a pas d'existence propre.
13b. *e1 ET e2 et E3* : la locution est potentielle ; la satisfaction de la contrainte "e1 *préc* e2 *préc* e3" entraîne l'émission d'un complément de graphe comme suit.

[66] Pour *pomme de reinette* et *pomme de pin*, la situation serait différente. *pomme de reinette* serait proche de *pomme de terre* en ce qui concerne *de* et *reinette* tandis que *pomme de pin* s'apparente plutôt à *cancer du poumon* (voir 6.1.4 Intégration du terme page 85).

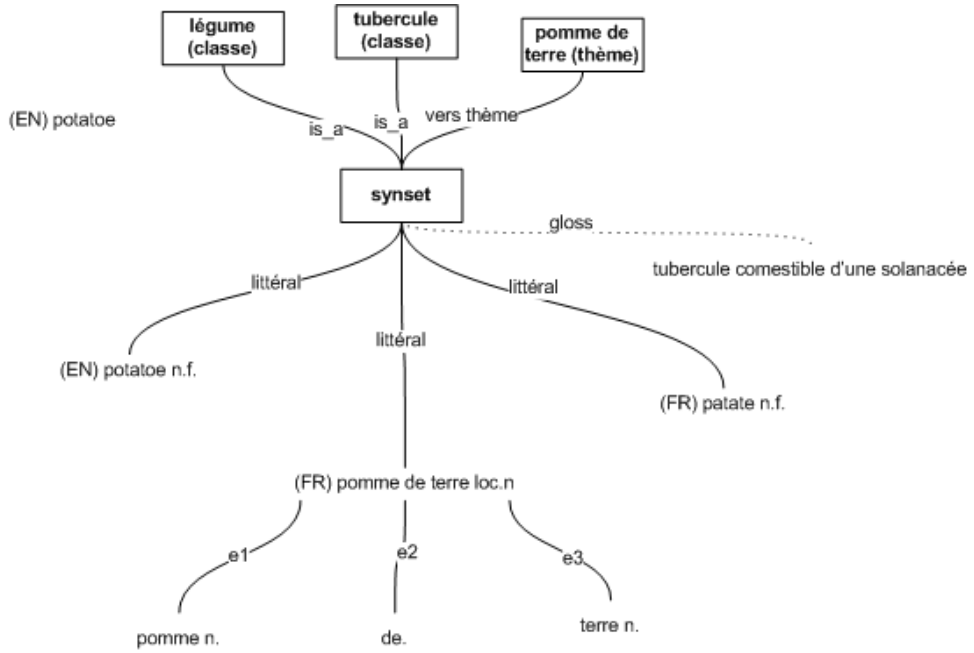


Figure 14 c. Structure présentant *pomme de terre* créée dans le graphe componentiel.

Portées du formalisme sur les critères C1 à C4.

C1) la structure obtenue reste atomistique. Soit nous n'avons pas encore *pomme de terre* et nous avons seulement des éléments, soit nous avons *pomme de terre* et nous trouvons alors un seul élément.
C2) l'algorithme des LCA continue à fonctionner dans la limite du tout petit graphe défini/définissant *pomme de terre*. La démarche analytique est la même que celle définie depuis 1996 voire depuis 1992.
C3) cela n'empêche pas que dans un autre graphe *pomme* ait une existence propre. Le fait qu'il y ait plusieurs perceptions possibles est une évidence. Cela n'implique pas l'existence de plusieurs Structures résultantes qui de plus suivraient des formalismes différents.
C4) *pomme* a aussi un sens dans *pomme de terre*. L'unité du Signe est maintenue. Par exemple, cette unité permet de partager un paradigme flexionnel entre *pomme* autonome et *pomme* de *pomme de terre.*

Pour la suite, ce que nous appelons :

- Structure comporte dorénavant la morphologie compositionnelle des mots composés.
- Analyse sait traiter de façon monotone des faits de morphologie compositionnelle. Il prend en entrée des éléments et produit des éléments.
- Résultat comporte aussi une représentation atomique des réalisations compositionnelles. Cette représentation ne supprime pas la représentation des composantes. Pour être tout-à-fait conséquent, il nous faut insister sur le fait que cette représentation influe sur la représentation des composantes en instanciant une composante. Par exemple, du fait que *pomme de terre* existe effectivement, *pomme_de_pomme_de_terre* existe dans les instances créées de la Structure résultat et pourra servir pour régler par exemple des questions d'accord comme dans *bonne pomme de terre*[67].

[67] Nous n'avons pas représenté dans la Figure 14-c l'émission de la composante *pomme* (par exemple, dans un

- Signe conserve une forme intangible[68].

6.1.2 Intégration d'énoncés compositionnels métalinguistiques

Un signe a toujours un sens. En admettant que ce sens soit strictement fondé sur une localisation (voir note 10, page 9, la remarque de G. Deleuze), nous ne voyons pas bien encore comment un sens ainsi défini agit concrètement, c'est-à-dire cause une action cognitive.
Pour répondre à cette question, le plus simple est de trouver un exemple. Le lieu où le discours agit le plus, quoiqu'en disent les exemples des conférences d'Austin, est le discours lui-même. En effet, la pragmatique est d'abord abstraite[69]. Par conséquent, c'est dans le domaine du vocabulaire traitant du langage que nous trouverons les plus fortes intrications. Prenons l'exemple suivant :

Le nom samouraï comporte 8 lettres.

Le problème posé est le suivant : s'agissant de métalangage, l'exemple traite de la Structure. Traitant de la structure, il est donc en-dehors de cette dernière : la grammaire est une activité métalinguistique.
Pour nous, le traitement métalinguistique du métadiscours porte atteinte à :

C1 : l'unité de la Structure
il fait apparaître une nouvelle structure : la structure métalinguistique.
C2 : l'unité de l'Analyse
si la structure métalinguistique diffère de la Structure, alors l'inférence dans la Structure métalinguistique diffère de l'inférence dans la Structure.
C3 : l'unité du Résultat
le résultat métalinguistique est incomparable, irréductible au résultat dans la Structure
C4 : l'unité du signe
le mot *samouraï* n'aurait pas de sens dans cet exemple.

Mise en contexte et effets

Il est évident que notre exemple n'est pas isolé. Comme nous l'avons dit, le métadiscours constitue une très large part du discours. *Comme nous l'avons dit*, que nous venons d'écrire,

---

lieu grammatical de gestion des accords) qui fait suite à l'événement *pomme de terre*. Nous nous sommes contentés de représenter l'émission du tout *pomme de terre* dans l'axe componentiel qui était ici le point de vue que nous souhaitions traiter.
[68] *À quoi on ne doit pas toucher, porter atteinte; que l'on doit maintenir intact* (Robert). Voir note 62 page 73.
[69] Loin de toute recherche d'oxymore, nous pouvons prendre l'assertion *pragmatique abstraite* pour tous les sens de *pragmatique* de la façon la plus littérale qui soit :
1°) *Étude des signes en situation* (Robert). Voir sur la devanture d'une boulangerie le mot *boulangerie*. Ce signe *boulangerie* va interagir d'une certaine manière avec le stimuli visuel de la boulangerie. Peu importe la manière. Ce qui compte est la chose suivante : dans quel espace se produit cette interaction ? Dans l'espace concret si nous généralisions Austin ? Aucunement! Tout juste dans le monde abstrait de la représentation, dans ce qui n'existe que sous forme d'idées.
2°) *Qui est adapté à l'action sur le réel* (Robert). Quel réel existe en-dehors des représentations que nous en avons ?
3°) *Qui concerne la vie courante* (Robert). Qu'est ce qui appartient davantage à la vie courante que la transformation des stimuli en signes ?
4°) *Relatif au pragmatisme, doctrine qui donne la valeur pratique comme critère de la vérité (d'une idée).* (Robert). Qu'est ce qui nous est plus important que de faire de l'utile le vrai ?

comme *que nous venons d'écrire*, que nous venons de lire comme *que nous venons de lire* qui vient juste d'être dit est un élément du métadiscours.
Du fait que l'ordinateur ne connaîtra jamais le monde comme nous le *vivons* et donc ne pourra jamais constater la ressemblance qu'il y a dans les *choses* entre une personne et un humain (voir paragraphe 4.2.2, page 45), que nous battons-nous pour la lui inculquer? N'est-il pas plus important de lui donner à étudier le métalangage (en particulier le dictionnaire) et le métadiscours qu'il serait peut-être en mesure de *percevoir* si nous lui en donnions les moyens.

Solution

Commençons donc comme nous l'avons déjà fait en 6.1.1 par C4 :

> le mot *samouraï* n'aurait pas de sens dans *le nom samouraï comporte huit lettres.*

La rupture C4 concerne ce que nous avons appelé l'unité du Signe. Ici la rupture tient en le fait que *samouraï* n'est plus un signifiant dans la Structure puisque nous avons retiré le métalangage de la Structure.

Nous disons donc que cette rupture tient en l'affirmation suivante : dans la vue métalinguistique *le nom samouraï*[70]*, samouraï* n'aurait pas de sens. Comment un signe tel que *samouraï* peut-il n'avoir aucun sens dans une vue donnée qui le définit en les termes particuliers de cette vue?

Dans *le nom samouraï*, nous proposons de considérer que *samouraï* a pour sens :

> Is_a *nom masculin* dans le *nom*.

Nous écrivons:

> samouraï n.m. **-->Is_a-->** [nom.masculin]$_{classe}$ **dans** [nom]$_{classe}$

Du fait de l'inexistence d'autres conditions e_i, [nom.masculin]$_{classe}$ est un succès et peut émettre son information:

> [nom.masculin]$_{classe}$ **-->Is_a-->** [nom]$_{classe}$ **dans** [Ontologie des POS] $_{classe}$
> [nom]$_{classe}$ --> **Is_a-->** [Part of speech]$_{classe}$ --> **dans** [Ontologie des POS] $_{classe}$

Par ailleurs, nous avons :

> selon la vue des génériques :
>
> > n.m. **-->générique** --> [nom.masculin]$_{classe}$ **dans** [nom]$_{classe}$
> > n. **-->** **générique** --> [nom]$_{classe}$ **dans** [nom]$_{classe}$
>
> et selon le point de vue du lexique
>
> > [nom]$_{classe}$ --> **Is_a-->** [mot]$_{classe}$ --> **dans** [lexique] $_{thème}$
> > mot **-->** **générique** --> [mot]$_{classe}$ **dans** [lexique]$_{thème}$
>
> et selon le point de vue qui nous a alerté de l'intégrité du signe (C4), nous avons :
>
> > signifiant **-->** **générique** --> [mot]$_{classe}$ **dans** [signe]$_{thème}$

[70] ou dans la vue métalinguistique du dictionnaire : *samouraï n.m.*

Détails sur le formalisme

A] Nous ne ferons pas de figure représentant les inférences précédentes : en effet, une telle figure serait presque une hérésie puisqu'elle montrerait sur le plan (c'est possible dans ce cas précis) des éléments d'un programme (un signifié) qui normalement ne se peut réaliser qu'en partie. Par ailleurs, nous n'avons pas introduit de concepts nouveaux.

B] Etant données les inférences fournies, *samouraï* est-il ou non un *n.m,* un *n*, un *nom*, un *mot***,** ou un *signifiant*? Dans les chapitres 1 à 5, avons-nous jamais dit que *samouraï,* pris tout seul, serait un *guerrier*, un *noble* ou un *japonais?* Nous n'avons jamais dit cela. Nous avons dit : pris tout seul, *samouraï* est défini par *guerrier + noble + japonais* et les inférences de ces éléments. Nous avons en outre précisé que cette définition est vraie dans le *dictionnaire componentiel.* La situation est rigoureusement la même ici : *samouraï* est tout ce nous venons de dire (y compris un groupe de 8 lettres) dans les lieux que nous avons régulièrement indiqués.

C] Alors quel sens précis a *samouraï* dans *le nom samouraï?*

Dans le syntagme *le nom samouraï, samouraï* est un nom et n'est rien d'autre : le co-texte métalinguistique a sélectionné les traits sémantiques pertinents comme c'est son rôle de le faire.

Portées du formalisme sur les critères C1 à C4.

C1 : la structure obtenue reste atomistique
C2 : l'algorithme des LCA continue à fonctionner dans la limite du tout petit graphe défini/définissant *le métalangage*
C3 : cela n'empêche pas que dans un autre graphe *samouraï* ait un autre sens
C4 : *samouraï* a aussi un sens dans *les parties du discours.*

Il faut en outre noter quelque chose d'important. En réalisant cette intégration, nous avons bien progressé dans la question 2) que nous rappelons ci-dessous :
Question 2 Considérant plusieurs mots, comment extraire automatiquement ce qui les rapproche et ce qui les distingue sémantiquement ? (page 31)
L'extraction automatique des éléments *rapprochant et distinguant sémantiquement* deux mots est une fonction unique du *lieu* d'exécution de ce rapprochement.
Nous notons de plus que pour un lieu, les mots ne sont pas nécessairement actifs de la même façon. Nous voulons dire qu'ils *n'agissent pas* semblablement dans ce lieu. Dans notre exemple, *nom* détermine, asserte, prédique un lieu et se comporte comme un agent cherchant à satisfaire ses objectifs tandis que *samouraï* se comporte comme un patient, un serveur réalisant un objectif. En 6.3.1.1 *Définition de la microsyntaxe*, page 110, nous parlerons plus généralement de **marqueur**[71]**.** Ici *nom* s'est comporté comme un marqueur métalinguistique et s'est réalisé comme tel. D'un autre côté, le programme componentiel de *samouraï* ne s'est aucunement réalisé.

[71] Aristote parlerait ici de *puissance active* pour *nom* et de *puissance passive* pour *samouraï.* Une puissance active est, du côté d'un agent, sa disposition à agir, à produire tel effet particulier. Une puissance passive est une disposition, pour un patient potentiel, d'avoir une réceptivité appropriée, une aptitude à accueillir l'effet. *Ces deux dispositions sont des puissances, l'une active et l'autre passive, dont le concours permet la réalisation effective d'une détermination, l'acte* [Bernhardt 1972, page 106].

Notons pour finir que les mots agissent, et que notre notion de localisation commence justement à les faire agir. Les mots agissent en marquant des références dans *un espace proprement structural et pré-extensif* (voir Deleuze note 10, page 9.

Pour la suite, ce que nous appelons :

- Structure comporte dorénavant la déclaration du métalangage des natures grammaticales.
- Analyse sait traiter de façon monotone certains faits métalinguistiques qui concernent les natures grammaticales.
- Résultat comporte aussi une représentation atomique des réalisations de ces faits métalinguistiques. Des instances d'emplois métalinguistiques y sont présentes.
- Signe conserve une forme intangible.

6.1.3 Intégration d'une grammaire syntagmatique

Il existe plusieurs types de grammaire ou de modèles réalisant le programme objectif de la grammaire : détailler les natures grammaticales et réaliser des relations entre ces natures. Il peut s'agir par exemple de chaînes de Markov s'appuyant sur des étiquettes, de grammaire syntagmatique, de grammaire de dépendance ou d'autres types de grammaires plus hybrides ou lexicalisées. En effet, le raisonnement que nous allons suivre s'appliquerait à l'identique pour chaque type de grammaire. Nous sélectionnons la grammaire la plus connue : la grammaire de réécriture (avec contexte libre ou non) qui s'intéresse aux descriptions syntagmatiques et à la phrase. Prenons donc l'exemple suivant :

Adjectif + Nom --> Nom
Déterminant + Nom --> Groupe Nominal

Le problème posé est le suivant : la *règle* de réécriture n'appartient pas à la Structure.
Le traitement hors la Structure de la règle de grammaire porte atteinte à :

C1 : l'unité de la Structure ; il fait apparaître une nouvelle structure : la règle syntagmatique.
C2 : l'unité de l'Analyse : si la structure syntagmatique diffère de la Structure, alors l'inférence dans le lieu syntagmatique diffère de l'inférence dans la Structure.
C3 : l'unité du Résultat : le résultat syntagmatique est incomparable, irréductible au résultat dans la Structure
C4 : la définition du signe : le mot *samouraï* n'aurait pas de sens au plan syntagmatique dans *le samouraï.*

Mise en contexte et effets

Nous savons que les grammaires fournissent des éléments importants dans la détermination d'une localisation. C'est le cas de la grammaire de réécriture appliquée à la bande syntagmatique et prise comme un récepteur particulier de cette bande. L'exemple considère l'intégration des points de vue de la Structure et de la grammaire syntagmatique. Il est évident que notre exemple n'est pas isolé. Il s'applique également à l'intégration des points de vue des grammaires de dépendance ou des modèles de langage appliqués à la Structure. Il s'applique évidemment aux relations qui existent entre eux.

Solution

Commençons donc comme nous l'avons déjà fait par C4 : le mot *samouraï* n'aurait pas de sens dans *le samouraï.*
La rupture C4 concerne ce que nous avons appelé *atteinte à l'unicité du Signe*. Ici la rupture tient en le fait que *samouraï* n'est plus un élément de l'organisation du discours (représenté ici

par la grammaire syntagmatique) puisque nous n'avons pas intégré la grammaire syntagmatique dans la Structure.
Nous disons donc que cette rupture tient en l'affirmation suivante : dans la vue syntagmatique *le samouraï*[72], *samouraï* n'aurait aucun sens. Comment un signe tel que *samouraï* peut-il n'avoir aucun sens dans une vue donnée qui le définit en les termes particuliers d'un emploi dans le discours?
Dans *le samouraï*, nous proposons de considérer que *le samouraï* a pour sens : Is_a *nom masculin* dans le *groupe nominal*.
Nous avons déjà (voir 6.1.2 *Intégration d'énoncés compositionnels métalinguistique*, ci-dessus) :

samouraï n.m. **-->e1-->** [nom masculin]$_{classe}$ **dans** [nom]$_{classe}$

Nous ajoutons d'abord :

le **-->e1-->** [dét. masc]$_{classe}$ **dans** [déterminant]$_{classe}$
[dét.masc]$_{classe}$ --> Is_a --> [déterminant]$_{classe}$ **dans** [Ontologie des POS] $_{classe}$

Puis, nous ajoutons :

[déterminant]$_{classe}$ --> **e1-->**[dét+nom] $_{graphe}$ **dans** [gram. syntagmatique] $_{classe}$
[nom.masculin]$_{classe}$ **--> e2-->**[dét+nom] $_{graphe}$ **dans** [gram syntagmatique] $_{classe}$

En vertu de l'état atteint par la Structure au paragraphe précédent et en respect d'un principe d'ordre que nous suivons ordinairement, nous ajoutons enfin :

[dét+nom] $_{concept\ structuré}$--> Is_a --> [groupe nominal] $_{classe}$ **dans** [Ontologie Syntag.] $_{classe}$
groupe nominal--> **Générique**--> [groupe nominal] $_{classe}$ **dans** [**Le dictionnaire**] $_{graphe}$

Détails sur le formalisme

A] Nous notons l'apparition d'une nature *graphe*. Vu par le Sémiographe, les *Classes*, les *Thèmes*, les *dictionnaires* sont tous des graphes et ce graphe qui s'appelle *graphe* ne nous dérange pas : c'est le nom par défaut de n'importe quelle description d'un tout.

B] Un habitué des grammaires syntagmatiques pourra être étonné que nous n'ayons pas représenté l'accord entre le déterminant et le nom. Cela est-il un choix, un oubli ou une impossibilité? En fait, aucune de ces raisons n'est la bonne. Ce qui nous a fait ne pas le représenter est seulement le besoin de tenir un discours pas trop "multifacé", pas trop complexe. Il est évidemment possible de représenter l'accord dans le graphe. Trois moyens simples sont à notre disposition :

a) soit nous utilisons l'hyper-arête [dét+nom] $_{graphe}$ et nous enrichissons les contraintes qui pèsent conjointement sur certaines *parties* de l'élément e1 et de l'élément e2.
b) soit nous créons (*) une hyper-arête [dét n.m+nom n.m] $_{graphe}$ plus précise puisqu'elle inclut l'accord. Dans ce cas, ce lieu (**) devra en outre être présent à l'intérieur du lieu [dét+nom] $_{graphe}$ pour garantir la qualité des LCA produits par la Structure.
c) soit nous créons un "lieu" grammatical de gestion des accords.
En fait toutes ces approches sont bonnes en même temps mais ne donnent pas exactement la même chose :

[72] ou dans la vue du métalinguistique du dictionnaire : *samouraï n.m.*

L'approche (a) présente l'avantage d'être familière et rapide mais rend la perception de la faute (la correction) plus difficile.
L'approche (b) (*) présente l'avantage d'être réutilisable pour donner des génériques de nature comme *nom masculin*, ou l'avantage inverse de réutiliser les génériques de nature grammaticale déjà connus. (b) (**) permet d'obtenir un LCA [dét+nom] même dans le cas erroné de *le samouraïs.* Ce LCA est relativement haut et est donc, au plan d'une métrique, moins bon que ce qu'il aurait fallu obtenir.
Enfin, l'approche (c) (*) permet d'exprimer intentionnellement la règle de grammaire du français : *en français, le déterminant et le nom sont accordés en genre et en nombre*.

C] Comme la boucle sur le métalangage est bien effectuée, nous apprécions que le système soit maintenant capable de percevoir un énoncé métalinguistique comme *le syntagme nominal "le samouraï"*.

D] Finalement quel sens précis a *samouraï* dans *le samouraï?*

Dans le syntagme *le samouraï, samouraï* est un nom masculin dans la grammaire syntagmatique. Cela ne veut pas dire qu'il ne signifierait rien dans le point de vue componentiel. Cela signifie juste que rien n'est actif/activé/émergent/ dans ce point de vue. Cela posé, un lieu particulier, en l'occurrence le lieu syntagmatique a sélectionné des traits [de sens] d'un Signe pertinents comme c'est son rôle de le faire.
En outre, au plan d'une espèce de logique des déterminants, il n'est pas grand chose d'autre du fait que nous n'avons pas encore spécifié le sens *agissant* de *le*[73].

Portées du formalisme sur les critères C1 à C4.

C1 : la structure obtenue reste atomistique
C2 : l'algorithme des LCA continue à fonctionner dans la limite du graphe défini/définissant *une grammaire syntagmatique*
C3 : cela n'empêche que dans un autre graphe *samouraï* ait un autre sens
C4 : *samouraï* a aussi un sens dans *la grammaire syntagmatique.*

Il faut en outre noter quelque chose d'important. En réalisant cette intégration, nous avons bien progressé dans la question 2) que nous rappelons encore :
Question 2 Considérant plusieurs mots, comment extraire automatiquement ce qui les rapproche et ce qui les distingue sémantiquement ? (page 31)
L'extraction automatique des éléments *rapprochant et distinguant sémantiquement* deux mots est une fonction unique du *lieu* d'exécution de ce rapprochement.
Nous notons de plus que pour un lieu, les mots ne sont pas nécessairement actifs de la même façon. Nous voulons dire qu'ils *n'agissent pas* semblablement dans ce lieu. Dans notre exemple, *le* détermine, asserte, prédique un *nom* et se comporte comme un agent cherchant à satisfaire ses objectifs tandis que *samouraï* se comporte comme un patient, un serveur réalisant l'objectif d'un tiers. Au chapitre 6.3.1 page 109, nous parlerons plus généralement de

[73] (a) Le + *nom* cause la création d'une classe *d'instance* [nom] et un emploi générique *d'instance* dans cette classe [*nom*]. Comme une classe n'est justifiée que si son contenu comporte plus d'un élément, Le + *nom*, pris isolément crée une instabilité dans le graphe : le graphe de l'instance générique de la classe est en construction et attend ses spécifiques. (b) Nous venons simplement d'exprimer que, formellement, un énoncé comme *quelle est la couleur,* à contexte nul, est incorrect. A contrario, de ce point de vue, *quelle est couleur du cheval blanc d'Henry IV?* est tout-à-fait correct.

marqueur[74]**.** Ici *le* s'est comporté comme un marqueur syntagmatique et s'est réalisé comme tel. D'un autre côté, le programme componentiel de *samouraï* ne s'est aucunement réalisé.
Notons pour finir que les mots agissent, et que notre notion de localisation commence justement à les faire agir. Les mots agissent en marquant des références dans cet *espace proprement structural et pré-extensif* appelé par les structuralistes (voir note 10, page 9).

Pour la suite, ce que nous appelons

- Structure comporte dorénavant la déclaration d'une ou plusieurs grammaires[75].
- Analyse sait traiter de façon monotone différentes sortes de grammaire.
- Résultat comporte aussi une représentation atomique des réalisations de ces faits grammaticaux et de leur impact dans d'autres analyses. Des instances particulières comme *samouraï* sachant *le samouraï* sont présentes.
- Signe conserve une forme intangible.

6.1.4 Intégration du terme

Nous traitons maintenant de morphologie des termes composés componentiels.
Prenons pour exemple : *cancer du poumon.*
Nous montrons ici que l'économie de la représentation componentielle de chacun des mots compris dans le terme *cancer du poumon* porte atteinte à :
C1 : l'unité de la Structure. Le terme n'est pas analysable componentiellement alors qu'il devrait l'être.
C2 : l'unité de l'Analyse. Si la structure du terme diffère de celle utile à ses composantes dans ce terme, alors l'inférence dans la structure diffère selon que nous partons du composé ou des composants.
C3 : l'unité du Résultat. Il s'ensuit deux analyses automatiques d'un texte comportant *cancer du poumon.* Ces analyses sont objectivement en concurrence alors que nous aimerions qu'elles ne le fussent pas.
C4 : l'unité des signes *cancer, de, le* et *poumon* est rompue.

Mise en contexte et effets

Il y a dans notre dictionnaire pour le français environ 75.000 locutions plus ou moins figées, nominales ou verbales. Les effets des ruptures précédentes sont tellement variés que nous ne tentons pas de les représenter ici. Considérons toutefois un thésaurus médical qui comporterait *cancer du poumon* relié à *poumon.* Supposons un document traitant de *cancer de poumon* mais ne comportant pas la citation exacte du terme. Dans ce cas, nous observons que tout algorithme de classification automatique des documents médicaux devrait alors casser

[74] Voir note 71, page 83.
[75] Pour le français, par exemple :
- un modèle de langage qui est particulièrement intéressant pour capter et fournir une évaluation des contiguïtés ; par exemple la contiguïté en français Det-Adj qui n'a pas tellement de sens dans une grammaire syntagmatique ou dans une grammaire de dépendance.
- une grammaire syntagmatique, qui est particulièrement intéressante pour décrire la bande verbale du français.
- une grammaire de dépendance qui pourra s'intéresser à d'autres phénomènes, et s'occuper de nombreuses interactions entre points de vue.
Toutes ces grammaires sont plus ou moins lexicalisées. Dans tous les cas, le bon sens méréologique impose qu'à toute forme très contrainte (très lexicalisée) corresponde un conteneur prenant une *forme* moins contrainte (moins lexicalisée) : c'est l'ordre des LCA qui est en jeu ici.

soigneusement le terme composé pour espérer obtenir une classification correcte.

Solution

Comme dans chacun des cas précédents, la solution tient en le rétablissement de l'unité du signe (C4) qui par conséquence rétablit l'unité de la structure. Il "suffit" de dire que, par exemple :

cancer -->*Is_a*--> $[cancer]_{classe}$ **dans** cancer du poumon n.m.

Nous observons que, du point de vue qui nous intéresse, peu importe que le *cancer* à gauche de la proposition ait ou non le même sens que l'élément *cancer* générique de la classe *[cancer]*. Ce qui compte est seulement qu'il existe et appartienne à la même classe que ce dernier.

Détails sur le formalisme

Le formalisme ne pose pas de problème particulier.

Portées du formalisme sur les critères C1 à C4.

C1 : la structure obtenue reste atomistique
C2 : l'algorithme des LCA continue à fonctionner dans la limite du tout petit graphe syntagmatique défini/définissant les parties *cancer du poumon* et dans le grand graphe componentiel défini/définissant les parties du tout *cancer du poumon* et le tout *cancer du poumon.*
C3 : cela n'empêche pas que dans un autre graphe *cancer* ait une existence propre
C4 : *cancer* de *cancer du poumon* a à la fois un sens dans *cancer du poumon* et un sens dans $[cancer]_{classe}$.

Pour la suite, ce que nous appelons

- Structure comporte dorénavant la morphologie compositionnelle des mots composés et l'information componentielle propre à leurs composants.
- Analyse sait traiter de façon monotone ces informations et les garde unies.
- Résultat comporte une double représentation compatible. Il y a double instanciation du terme et de ses composantes.
- Signe conserve une forme intangible.

6.1.5 Intégration d'une date

Tout ce dont nous avons traité dans ce chapitre peut revendiquer le nom de référence[76]. Quelle est la référence de *pomme* dans celle de *pomme de terre* ? Quelle est-elle pour *samouraï* dans le *nom samouraï* ? Que devient-elle dans *le samouraï* ? Que valent-elles pour

[76] Il est possible de prendre le sens courant ou le sens linguistique. Courant : *Action ou moyen de se référer, de situer par rapport à ; système de référence.* (Robert) Philo, Ling. : *Fonction par laquelle un signe renvoie à ce dont il parle, à ce qu'il désigne.* Comme nous avons la chance de ne pas avoir à traiter du Monde, référence pour nous renvoie toujours à ce dont un signe parle : lui-même, ou un autre (morceau de) signe. Nous n'avons pas de vraie boulangerie (voir note 69-1, page 78) à gérer. La citation de Ricœur est intéressante *: Alors que les signes n'ont de rapport qu'entre eux, le discours se rapporte aux choses d'une manière spécifique, qu'on peut appeler dénotation ou référence.* C'est exactement ce que nous sommes en train de mettre en place en développant un modèle portant sur les individus méréologiques.

cancer dans *cancer du poumon?*

Nous étudions maintenant les relations qui existent entre élément d'une date et la date proprement dite. Comme à chaque fois, cette prise en compte va s'effectuer non pas en considération de phénomènes extralinguistiques, mais seulement en considération d'un tout qui a plusieurs effets sur la Structure. Nous voyons alors qu'une date n'est pas quelque chose de *si* simple qu'elle pourrait se ramener à une quelconque métadonnée. Une simple date n'est pas une métadonnée. Nous montrons qu'il serait illusoire de la ramener à une métadonnée si nous voulons nous intéresser à l'interprétation du texte.

Mise en contexte et effets

D'une part, évidemment, tous les effets délétères que nous avons déjà vus des ruptures s'appliqueraient à cette rupture-ci si nous transformions une occurrence d'une perception de date en seulement un tout extralinguistique prenant la forme d'une métadonnée. Nous allons montrer une conséquence dommageable. Cette conséquence concerne à la fois un traitement de la coréférence dans le texte et la désambiguïsation d'un énoncé. Ce que nous montrons ici est très simple à mémoriser : si la date est représentée par une métadonnée, alors il faudra développer un module spécifique de gestion de la référence temporelle.

Soit l'élément de dialogue suivant :

Le médecin : Je vous propose le mardi 25 avril 2008.
Le patient : non, je suis pris le 25 etc.

Comment interpréter la séquence *le 25* de la réponse du patient sans prendre en compte l'instance de date *mardi 25 avril 2008* ? Il faut s'interroger sur les significations compositionnelle et componentielle de *le 25* dans :

Le 25 est sorti.
Le 25 est occupé.
Le 25 m'intéresse.
J'ai réservé *le 25*.
Je serai sur *le 25* lundi.
J'enchéris sur *le 25*.
Il faut revoir *le 25*.

Nous ne pouvons évidemment pas élucider ces significations sans leur contexte. Mais si ce contexte est enfermé dans une métadonnée, alors il nous faudra pour chacun d'eux développer un module spécifique. Nous comprenons bien que cela est impossible puisqu'il existe une infinité de contextes.

Solution

Il nous faut considérer la date construite comme un élément de la Structure qui a de plus causé l'existence dans la Structure d'autres éléments. Ces autres éléments sont tous les éléments qui peuvent faire référence au tout construit, à savoir la date complète observée. Cela se fait en respect des considérations d'usage et de signification qui sont l'objet même du dictionnaire. Dans ce cas, toute perception d'une possibilité de coréférence pour une date (et pour les exemples non élucidés ci-dessus) s'effectuera de la façon la plus monotone qui soit. Voyons comment procéder pour notre exemple.

Dans tous nos échecs, nous avons mis dans une métadonnée non accessible depuis l'instance *le 25* tout ce qui permettrait de désambiguïser cette instance. C'est ce qu'il nous faut corriger.

La figure ci-dessous représente différents états méréologiques de la date *mardi 25 avril 2008*. Nous adaptons ici la représentation à l'aide d'un hypergraphe (voir *Figure 8 Exemple d'hypergraphe*, page 52) parce qu'une représentation sous forme d'un DAG, avec cycles et événements, serait totalement illisible.

Cet hypergraphe est d'ordre 6 et de rang 6 et présente des arêtes multiples.

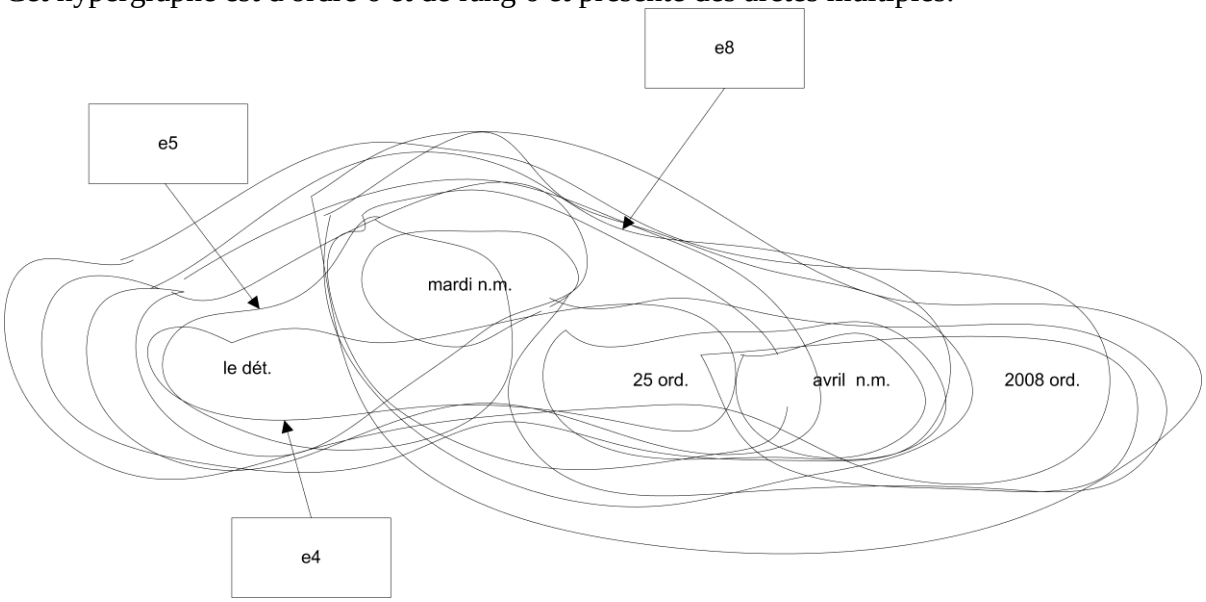


Figure 15 L'organisation de l'onomasiologie d'une instance d'une classe [*date*] n'est pas particulièrement triviale

Dans le graphe, nous voyons que l'hyper-arête e4 qui correspond à *le 25* peut être perçue comme une instance de [date]. Cela signifie-t-il pour autant que toute occurrence de *le 25* est une instance de date? Il serait dommage que cela soit ainsi. Alors que signifie ce graphe méréologique?

Pour essayer de répondre à cette question, dessinons un nouveau graphe. Ce graphe présente méréologiquement ce que nous admis savoir de *samouraï* dans le *nom samouraï.*

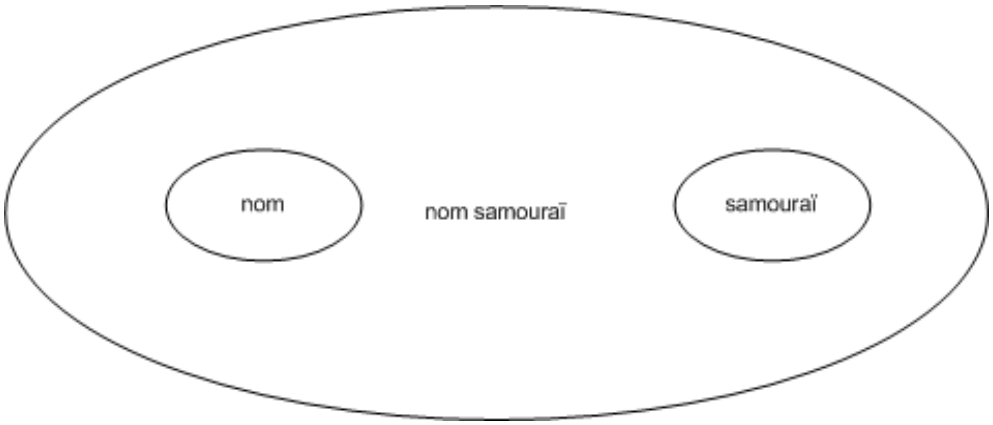


Figure 16 Le nom samouraï

En lisant la figure précédente, disons-nous que, par exemple, tout l'ensemble des noms se limite à *samouraï?* Aucunement. A contrario, disons-nous davantage que tout ce qui concerne le signe *samouraï* se limite à *nom?* Nous ne le disons pas davantage.

En fait, dans ce genre de graphe, rien de ce que nous pouvons dire comprend une quantification universelle ni plus que comprend une logique du premier ordre. Au contraire, nous lisons seulement : **dans** *nom samouraï, samouraï* est un élément et *nom* en est un autre. Il n'est pas possible de dire davantage. La *Figure 15* se lit exactement de la même façon. Dans cette figure *le 25* se lit :

dans *date, le 25* est un élément et *mardi 25 avril 2008* en est un autre.

C'est pourquoi nous pouvons parler, d'une certaine manière, d'onomasiologie de la date : *le 25* n'est rien d'autre qu'un moyen commode, rapide et courant de faire *référence* en français à la sémasiologie d'une instance de date.
Par ailleurs, la *Figure 15*, telle que nous l'avons représentée, ne dit aucunement, si *le 25* ou *mardi 25 avril 2008* présente ou non des différences fonctionnelles au plan de l'analyse du discours. La question qui se pose à nous alors est la suivante :

le 25 ou *mardi 25 avril 2008* présentent-ils ou non des différences fonctionnelles au plan de l'analyse du discours?

Il est évident que nous pouvons répondre positivement à cette question :

a) *le 25* ne présente pas les éléments componentiels et compositionnels minimaux pour faire causer avec vraisemblance une hypothèse une date. En fait, en tant que partie d'un tout, il ne cause une date, que si le tout *date* préexiste comme instance.

b) à rebours, le *mardi 25 avril 2008* présente tous les éléments componentiels et compositionnels susceptibles d'instancier le concept de date.

Ainsi donc, sur un plan événementiel, nous trouvons dans notre structure, avec (b) l'activation d'une *date* tandis qu'avec (a) nous n'avons rien.
Mais précisément, comme nous avons dit que dans une localisation *date*, *le 25* est une date, alors bien évidemment, maintenant *le 25* ainsi que toutes les autres composantes méréologiques de date indiquées dans la figure en deviennent potentiellement. Avec la figure, c'est par exemple le cas de *mardi*, de *en novembre* etc.

Mais cette figure est extrêmement rudimentaire et ne comporte pas toutes les onomasiologies **référentielles** de *mardi 25 avril 2008*. Ces autres signes qui pour le moment ne sont pas là sont :

- jour, jour-ci, jour-là
- moment, moment-ci, moment-là
- journée, journée-ci, journée-là
- etc.

Ainsi, nous venons de monter pour la cinquième fois, comment, en maintenant les consistances compositionnelles et componentielles du signe (ici *mardi 25 avril 2008)* et en maintenant en même temps la consistance componentielle de la *classe*, dans une même structure, nous résolvons sans processus particulier un cas typique de gestion de la coréférence qui fait aujourd'hui l'objet de recherche spécifique domaine par domaine (par exemple, la coréférence. et le temps comme ici). Ce peut être également, la coréférence et l'espace, la coréférence et les entités nommées etc.
C'est ainsi que dans le dialogue suivant

Le médecin : Je vous propose le mardi 25 avril 2008.
Le patient : non, je suis pris (ce jour) (le 25) (mardi) etc.

nous **percevons** très naturellement la coréférence.

Détails et conséquence sur le formalisme

A] Nous avons utilisé une sorte de *tableau noir :* une *date_localisation* vient dans le tableau noir qu'est la Structure et du fait que cette date comporte ses inférences, ses composés et ses composantes, elle est à la fois localisée et localisation. Réceptrice d'une référence, elle intervient comme une localisation pour d'autres instances de signes.

B] La modélisation que nous venons de faire ne s'occupe pas du tout de la pertinence ou de l'introduction de critères pragmatiques particuliers. Nous imaginons bien que ce genre de choses puisse donner lieu à des efforts spécifiques mais nous attendons la démonstration.

C] La modélisation que nous venons de faire ne sert qu'à rendre endogène une certaine perception dans un modèle ; cette endogénéisation est rendue nécessaire pour respecter les critères C1 à C4 :

C1 : la Structure obtenue reste atomistique
une date n'est rien d'autre qu'un point, un élément vu de l'extérieur et n'est rien d'autre qu'un DAG vu de l'intérieur.
C2 : l'Analyse reste homogène
elle prend en entrée une Structure atomistique dans lequel l'algorithme des LCA continue à fonctionner dans la limite du graphe défini/définissant une *date*
C3 : la Structure résultat reste atomistique
cela n'empêche que dans un autre graphe *le 25* ait un autre sens
C4 : Le Signe est conservé
le 25 a un sens dans *la date.*

Remarque : ce n'est pas parce que tous les liens compositionnels et componentiels de *date* sont vraisemblablement chez un humain actifs dès la prise de rendez-vous, que nous les rendons actifs. Nous les rendons actifs uniquement pour ne pas atteindre aux intégrités des Structure, Analyse, Résultat et Signe de notre modèle.

D] le formalisme courant est suffisant pour traiter complètement le cas que nous venons d'exposer.

Observations complémentaires sur le formalisme

A] Dans la *Figure 15*, nous supposons que nous avons l'arête e8 (*mardi 25 avril*) construite compositionnellement à partir de *mardi 25 + avril* [méthode 1] d'un côté et de *mardi + 25 avril* [méthode 2] de l'autre. Nous sommes très éloignés d'une famille de Sterner puisque la plupart des arêtes sont contenues dans d'autres arêtes. Notons aussi l'importance qu'il y aurait dans une grammaire à établir automatiquement que le résultat obtenu par la méthode 1 et le même **à l'identique** que celui qui serait obtenu par la méthode 2. Ce point a fait l'objet du DEA de François Duchène (mention très bien, encadré par nous) soutenu à Paris VII. Nous ne détaillerons pas les résultats de cette étude ici.

B] Un autre exemple simple de coréférence.

Toujours concernant des problèmes de référence, nous pouvons étudier des textes courts comme cet extrait de dialogue où *réserve* peut être considéré comme anaphorique de *gisement* et, en tout cas, désigne un même objet.

Soit le dialogue suivant :

- *La Shell possède les plus riches gisements de la région*
- *Non, c'est BP qui possède ici les réserves les plus importantes*

C'est évidemment plus complexe qu'une simple date et nous ne chercherons pas à l'explorer ici. Des pistes ont été données avec les LCA componentiels (voir paragraphe 4.1.2.2.4 *La différence componentielle*, page 41) et une analyse plus fine sur *riche* sera proposée au paragraphe 6.2.1.1 *Intégration de la définition prenant une forme schématique*, page 102.

Pour la suite, ce que nous appelons

- Structure comporte dorénavant la date et l'information componentielle propre à ses composantes.
- Analyse sait considérer que certaines composantes ne sont activables que si l'existence du tout est établie.
- Résultat comporte une représentation dans laquelle des composantes libres (comme *le 25*) sont perceptibles comme en rapport avec des composantes liées (composantes formant le tout établi).
- Signe conserve une forme intangible.

6.1.6 Intégration d'une formule

En traitant d"une date, nous avons conservé les composantes possibles d'une date dès qu'une date bien formée a été construite afin de conserver l'onomasiologie du concept et de donner consistance au Signe pris comme référence. Mais que se passe-t-il avec un énoncé algébrique? Que pourrait faire la Structure d'un énoncé comme

*Calculer 4+5*6*91+23*

En ce cas :

- la système pourrait-il extraire exactement la taille de la formule – ce qui importe plus que n'importe quoi d'autre pour identifier le sens de la formule
- et calculer le résultat de cette formule puisque
  a) dans l'énoncé, le verbe *calculer* demande explicitement au système de le faire
  b) ce résultat est dans une certaine mesure le sens de la formule
  c) ce résultat est peut-être nécessaire pour repérer un énoncé coréférentiel.

Ces spécifications sont-elles susceptibles de casser la structure? Avons-nous besoin de définir à l'intérieur de la Structure ou de l'Analyse de nouveaux éléments que ceux que nous manipulons déjà (en dehors de l'appel au processeur mathématique proprement dit)?

6.1.6.1 Le bornage strict d'une séquence et l'insertion de lieux nommés[77].

En effectuant le calcul, nous souhaitons éviter l'apparition de certains résultats intermédiaires incorrects. Dans notre cas, cela peut être *4+5* qui ne participe pas au résultat final. La question se pose pour nous, puisque contrairement aux situations ordinaires, nous sommes en face d'un vrai texte : personne ne pressera <Retour> comme sur une calculette

[77] Cette étude de faisabilité a été réalisée avec Yann Picand qui s'intéresse particulièrement aux grammaires syntagmatiques et voulait vérifier le formalisme sur les besoins propres de ses modules.

pour indiquer que la *composition* de la formule est réalisée.
Pour dessiner notre algorithme dans la Structure, nous allons utiliser un graphique *élément-processus-résultat* comme celui de la *Figure 14*, page 79. Nous ne présentons dans la figure que le vocabulaire suivant : *calculer*, $_{+,}$ -, * et / et quelques nombres. Nous ne nous intéressons pas pour le moment à la priorité des opérateurs puisque nous savons que nous devons déjà répondre à la question concernant l'étendue de la formule.

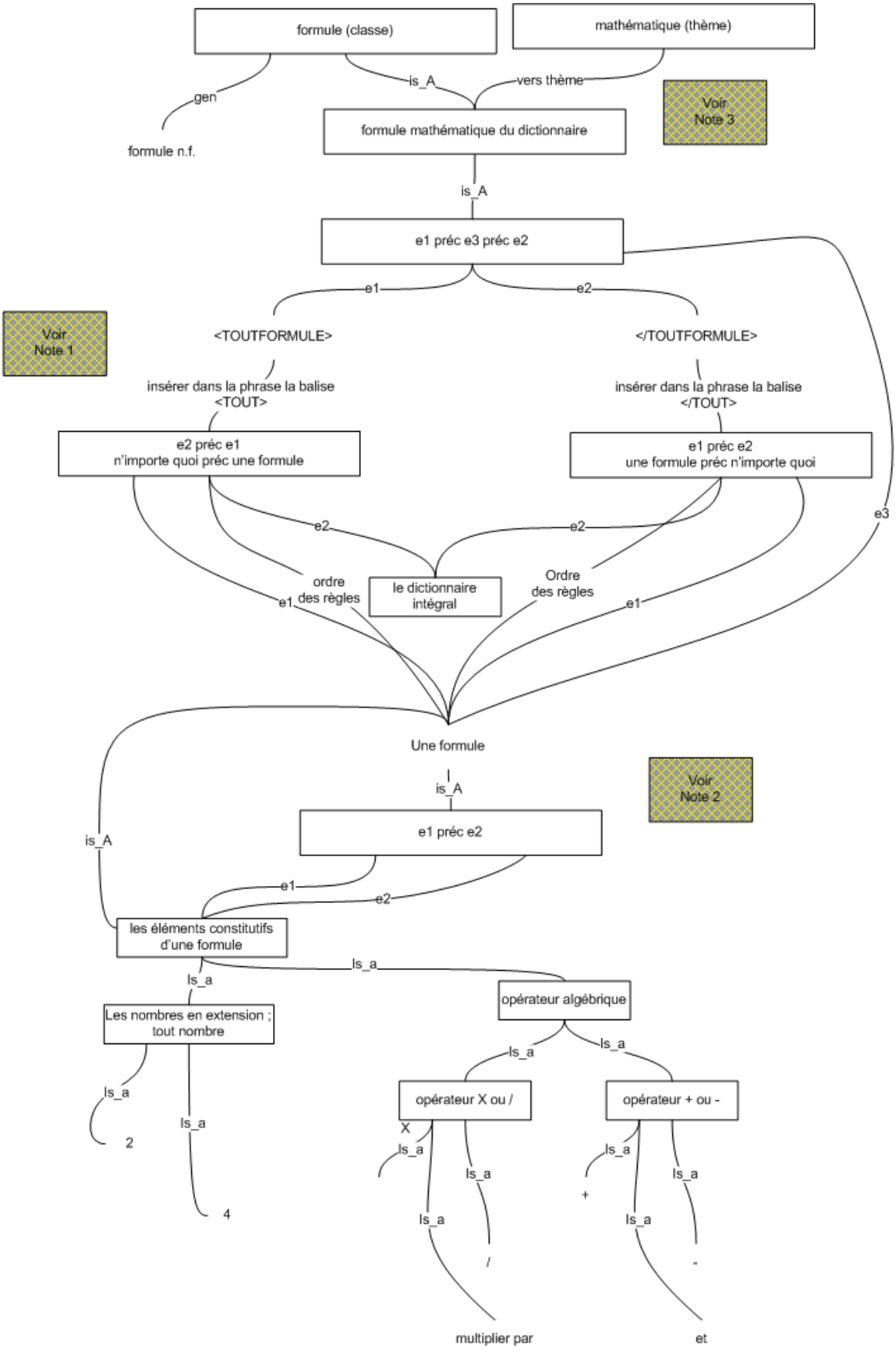


Figure 17 Définition d'un TOUT précis et balisage.

Nous détaillons un peu ce graphe. Nous pourrions ratiociner l'analyse de *une formule*[78] mais cela n'importe pas ici.
L'interpréteur de la Structure travaille en prenant en entrée cette Structure et l'énoncé exemple *Calculer 4+5*6*91+23*. Il construit au fil de l'eau un hypergraphe mêlant la structure en extraction et la phrase en lecture. Cet hypergraphe est la Structure résultat.

En particulier :
1) il instancie tout ce que la structure reconnaît dans les tokens qu'il lit
2) du fait d'une certaine disposition de la phrase, il constate la création d'objets de différents niveaux d'analyse et d'abstraction
3) ces objets créés sont eux-mêmes susceptibles de créer d'autres objets plus **ou** moins complexes ou plus **ou** moins compliqués : plus complexes s'ils se résument les dans les autres (c'est le but recherché ici), plus compliqués s'ils restent les uns aux bouts des autres sans permettre de création "sémantiques".

NOTE 1 : le processus débute par le haut à gauche. Il commence par un mot du dictionnaire intégral et un élément de formule. Il insère la balise <TOUTFORMULE> dans les graphes[79] qu'il gère. En insérant cette balise, il prend bien soin de dire que dans le sens qui nous occupe, le mot <TOUTFORMULE> s'applique à un élément de formule particulier situé en un lieu donné. En particulier, un nœud fictif père commun entre TOUTFORMULE et ce mot est créé.

NOTE 2 : Considérant un élément d'une formule et un autre élément, il crée une formule qu'il considère alors comme une instance d'une formule (en bas à gauche). Cette nouvelle formule reste connexe à l'élément de formule créé précédemment et donc à la balise. Ensuite, l'instance de formule est complétée par un nouvel élément de formule tant que cela est possible.

NOTE 3 : Quand l'ensemble des termes de formule ont été examinés, l'interpréteur rencontre un autre mot du dictionnaire intégral. Cela peut-être <findetexte> puisqu'il s'agit d'un mot du dictionnaire intégral que l'interpréteur peut employer pour signaler que des actions éventuellement en attente (comme ici) peuvent être réalisées. Le résultat de la rencontre de ce mot terminal pour un TOUTFORMULE est l'insertion d'un /TOUTFORMULE.
Maintenant il vient de créer une formule complète. C'est le but recherché de notre présentation. Il insère ce tout comme instance du dictionnaire componentiel comme spécifique (Is_a sur la figure) de la classe [formule mathématique du dictionnaire]. Ce qui est intéressant est que cette classe comporte des éléments calculables et qu'elle peut prendre en charge ces calculs si le texte lui demande de le faire comme c'est par exemple le cas ici.

OBSERVATIONS

Devons-nous trouver surprenant que le système n'effectue aucune opération quand il rencontre des éléments du Dictionnaire Intégral qui seraient aussi des éléments de formule ?

---

[78] Le processus décrit ici est susceptible d'admettre une formule incorrecte comme 3+*.
[79] Dans l'interpréteur tous les graphes sont liés, depuis la gestion des tokens, en passant par les groupes, les syntagmes, les balises, la profondeur componentielle et toute sorte d'événements qui seraient susceptibles de surgir. L'interpréteur classe ces événements d'une façon componentielle et compositionnelle, en gérant d'un côté la composition de la phrase (les tokens, les balises créées), l'axe componentiel partant de chaque point, et tous les relations axe compositionnel / componentielle /partie tout qu'il trouve de telle manière que tout ce qu'il trouve puisse servir de référence, d'élucidation à un autre objet, à l'exception des composantes complètes des TOUT que nous voyons ici.

A priori non. Voici pourquoi :
- il est vrai que tout nombre et tout opérande font partie du dictionnaire et que donc les opérations présentes au niveau de la note 1 devraient se réalisées. En effet, ces opérations ne demandent comme paramètres que :
    - un élément formule
    - un élément dictionnaire intégral.

Or, du fait que tout élément de formule est élément du dictionnaire intégral, ce contrôle devrait se réaliser.
En fait, il n'en est pas ainsi du fait de la régularité de la Structure. Nous observons que tout ce qui règle la perception de l'intérieur des formules est rangé dans ce qui conditionne les bornes des formules : cela est naturel puisqu'un intérieur est défini par une limite. Ainsi, la règle des LCA s'applique : si le système n'effectue aucune opération quand il rencontre des éléments du Dictionnaire Intégral qui seraient aussi des éléments de formule c'est qu'il ne voit pas ces deux conjectures à la fois. Du fait des LCA, quand il est à l'intérieur, il ne perçoit rien de l'extérieur.

6.1.6.2 Réduction algébrique : calculs utiles à l'analyse de texte

Dans une application de gestion de biens immobiliers, il peut se produire que l'on trouve une expression comme *salon de 4*3 m^2*. Si un client nous demandait un salon d'un minimum de 10m^2, comment pourrions-nous, sans briser la structure, proposer ce salon? Il s'agit d'un cas de paraphrase.

Nous présentons les calculs de base en s'appuyant sur notre exemple :

*Calculer 4+5*6*91+23*

Nous ne présentons pas le graphe lié qui ne présente aucune difficulté maintenant que nous avons borné notre formule.

1. la multiplication

créer un concept structuré e1 e3 e2, avec e1 instance de nombre, e3 instance de multiplication et autre e2 instance de nombre. Emettre le résultat comme un nombre occupant l'ensemble des positions des tokens dont il est issu.

2. l'addition

a) créer un concept structuré e1 e2 e3 e4 e5 avec :

- e1 <TOUTFORMULE>
- e2 instance de nombre
- e3 instance de + ou –
- e4 instance de nombre
- e5 instance de + ou –

Puis :

- calculer e2 e3 e4 pour de vrai
- émettre le résultat comme un nombre occupant l'ensemble des positions des tokens dont il est issu

b) créer un concept structuré e1 e2 e3 e4 e5 avec :

- e1 <TOUTFORMULE>
- e2 instance de nombre
- e3 instance de + ou –
- e4 instance de nombre

- e5 </TOUTFORMULE>

puis

- calculer e2 e3 e4 pour de vrai
- émettre le résultat comme un nombre occupant l'ensemble des positions des tokens dont il est issu.
- mettre à toutes fins utiles ce résultat comme instance dans la classe [résultat] du dictionnaire.

En observant que CALCULER FORMULE signifie CALCULER RESULTAT FORMULE, laisser l'ordinateur émettre où il veut :

Générique[Résultat]$_{classe}$: <RESULTAT>, c'est-à-dire :
résultat : 2757.

6.1.6.3 Quelques remarques sur l'exemple

Nous avons montré cet exemple pour :
- rappeler le fonctionnement des LCA.

Il est impossible d'aller plus loin, sans imaginer le comportement des LCA.
- montrer que, dans la mesure d'une certitude, il est possible de baliser un <TOUT> et de le manipuler de l'extérieur sans pouvoir par la suite toucher à ses parties.

Il nous semble que l'algèbre est le seul lieu où ce besoin est parfaitement justifié. C'est pourquoi, après hésitation, nous avons choisi de montrer cet exemple et non un autre, comme une date, dont nous avons vu qu'il serait le contraire de l'idéal. En effet, dans tous les autres cas, il est nécessaire de conserver la structure partie-tout du texte et même, comme on le verra par exemple avec le traitement du *cheval blanc*, de l'enrichir.

En outre, dans cet exemple nous avons montré que :

C1 : l'unité de la Structure
un terme non analysable avant un certain événement peut être conservé et réutilisé par la Structure.

C2 : l'unité de l'Analyse
à TOUT donné, nous n'avons qu'une seule analyse.

C3 : l'unité du Résultat
à TOUT donné, nous n'obtenons qu'un seul résultat.

C4 : l'unité du Signe
l'unité du signe est maintenue, que celui-ci apparaisse dans le texte ou à la suite d'une inférence ou d'un calcul, il reste ancré dans le discours et dans les choses dénommables.

Pour la suite, ce que nous appelons

- Structure comporte des cycles et des ordres d'application des règles. Mais nous savons déjà cela depuis le chapitre 4.1.2.2 page 36 même si nous n'en avons pas encore profité. En effet, ces éléments sont au cœur des LCA.
- Analyse sait concevoir un tout comme un tout et ne pas se perdre dans des sous-touts isomorphes.
- Résultat comporte une représentation dans laquelle des opérations non-linguistiques peuvent être effectuées et récupérées[80].
- Signe conserve une forme intangible.

6.1.7 Conclusion

Les six classes de phénomènes que nous avons examinées dans cette section examinent principalement l'intégration d'informations plutôt compositionnelles.
Cette intégration s'est réalisée

- en acceptant l'indifférenciation, ce qui a permis de maintenir l'unité de la Structure
- en acceptant les interdépendances utiles entre les parties, ce qui a permis de maintenir l'unité de l'Analyse
- en acceptant de voir se former un corps particulier qui serait semblable à la Structure d'entrée, ce qui nous a permis de maintenir l'unité du Résultat

L'ensemble a été rendu possible en acceptant de considérer le Signe comme un objet intangible. C'est en tant que tel qu'il figure dans un système de Signes susceptible de se mettre en œuvre quand le système en question est stimulé par un énoncé.
Cette intégration aboutit à quelque chose de plus complexe que ce que nous avons manipulé jusqu'au chapitre 5. En effet, nous n'avions pas alors besoin d'étudier un système dynamique : nous nous contentions de relever les concepts activés pas le Sémiographe sans considérer les conséquences possibles de ces activations. Ce relevé n'entraînait pas de nouveaux calculs inhérents, directement déclenchés pas le relevé lui-même.
Ici, nous avons fait apparaître des instances qui entraînent de nouveaux calculs. Il s'agit d'un système dynamique dans lequel des instances sont disposées selon le double point de vue méréologique des informations connues du système et la perception[81] de l'énoncé.
Maintenant que nous avons intégré des énoncés compositionnels relativement purs à la Structure et à l'Analyse, nous nous intéressons à des cas hybrides où cette intégration impacte ni tout le signifiant ni tout le signifié. Il s'agit de l'intégration hybride des points de vue compositionnels et componentiels.

---

[80] a) Il nous faut bien gérer a minima l'algèbre de base pour gérer la coréférence dans des énoncés comme : *Pierre a mangé un carambar et une sucette. Les deux bonbons mangés...etc*
b) Nous ne discuterons pas de savoir si cet algèbre de base appartient ou non au langage. Ce qui est toutefois certain est que des opérations énonciatives de gestion de certaines coréférences passent indubitablement par la connaissance de certains calculs algébriques

[81] Nous sommes resté avec des agents réflexe simples. Voir l'introduction du chapitre 6.

6.2 L'intégration de faits sémantiques qui ciblent à la fois des points de vue compositionnels et componentiels

Toutes les discussions que nous avons développées pour l'instant ont ceci de commun qu'un seul élément du signe est modifié et que cet élément est libre, c'est-à-dire entretient une relation Is_a avec une partie d'un Tout particulier. Il s'agit par exemple de :

Samouraï et *défendre* qui sélectionne le côté *guerrier* (désintégration componentielle)
nom samouraï qui sélectionne le côté *partie du discours* (désintégration métalinguistique)
3+2 qui donne *5* (désintégration compositionnelle)

ou a contrario

cancer dans *cancer du poumon* qui est un *cancer* (réintégration componentielle)
le 17 sachant *17 janvier 2004* qui est une *date* (réintégration componentielle)
<formule>3+2</formule> qui est une *formule* (réintégration componentielle et compositionnelle)

Comme les exemples "purs" ci-dessus ont déjà conduit à des exposés parfois peut-être difficiles à suivre, puisqu'il faut toujours avoir une structure[82] intégrée en tête pour percevoir la désintégration ou la réintégration dans l'intégration, nous imaginons bien que des exemples hybrides, pluriels, susceptibles d'agir de plus dans plusieurs lieux liés et en même temps vont être difficiles à présenter et de surcroît très pénibles à lire. Pour ces raisons, nous décidons d'abandonner le plan détaillé que nous avons retenu dans le paragraphe précédent. Ce plan nous a permis toutefois d'attirer l'attention sur les conséquences néfastes insurmontables des pertes d'unité de la Structure, de l'Analyse, du Résultat et du Signe. Nous mémorisons l'importance du maintien de ces unités pour aborder des cas plus complexes.
Dans un premier temps, nous regroupons quelques exemples susceptibles de faire sentir le terrain. Nous nous contenterons de cela. Dans le paragraphe suivant (voir chapitre 6.3 *La définition d'une microsyntaxe pour élargir un peu le champ perceptif de la Structure*, page 109) nous développons une "théorie" très simplificatrice mais qui nous permettra de progresser encore un peu. Enfin, nous appliquons notre "théorie" au problème du *cheval blanc* dont nous apercevons maintenant la principale qualité : celle d'être l'exemple le plus élémentaire que l'on puisse imaginer traiter dans une structure intacte et en en montrant toute l'organisation sur une même figure, ceci sans utiliser le moindre procédé extraordinaire qui échapperait à la signification lexicale.

[82] C'est à dire un certain état photographiant le développement d'un système à un instant t.

6.2.1 Quatre cas complexes mais solubles

Soluble : 1) Qui peut se dissoudre (dans un liquide).
2) Qui peut être résolu.

Dans le titre, les deux sens de *soluble* sont assez pertinents pour notre propos. Après tout, c'est le mot *soluble* qui nous est venu et non pas, celui bien plus rassurant, de *résoluble*[83]. Dans ce paragraphe, nous choisissons quatre cas[84] que nous expliquons avec quelques indices sur la structure menant à une solution. Chacun pourra prolonger la réflexion structurale ou bien dissoudre[85] le matériau fourni dans l'exogénéité[86] de son choix.
Pour l'exposé de chaque classe de phénomènes, nous appliquons le plan suivant :
- description du cas
- direction pour une solution structurale
- localisation de l'impact et formes résultantes

Nos classes de phénomènes sont intitulées :

- Intégration de la définition prenant une forme schématique
En nous plaçant dans le cadre de la théorie des opérations énonciatives d'Antoine Culioli, nous étudierons dans ce paragraphe le cas très général de la prise en compte de la signification d'une définition d'un Signe qui interagit avec la signification de la définition d'un autre Signe.

- Intégration de la contradiction entre connaissances des choses et connaissances des définitions
Nous étudions à cet endroit les contradictions potentielles entre perception componentielle et perception compositionnelle. Nous nous contentons d'indiquer, après avoir montré une contradiction particulière et naturelle, le mode de gestion que nous pouvons pratiquer.

- Intégration de l'inférence issue des connaissances sur les choses
Ce paragraphe qui aurait pu être placé avant le paragraphe précédent, correspond à la gestion de la prédication. Mais nous ne l'avons pas appelé ainsi car nous observerons dans le paragraphe que nous devons prendre en charge des situations bien plus générales que la prédication tout simple comme par exemple *manger(samouraï)*. En outre, ce paragraphe traite directement du Hiatus "dictionnaire à l'envers" en rapport avec l'absence d'organisation entre les concepts des quasi-définitions de la page 44 où nous affirmions qu'il faudrait bien qu'à un moment donné, dans le dictionnaire à l'envers, *personne+vendre+personne* puisse retourner quelque chose de pertinent.

---

[83] Peut-être parce que le trait de résoluble *Qu'on peut décomposer en ses éléments constituants* ne nous convient pas tout à fait.
[84] Ces cas sont extraits d'une série de cas que nous avons étudiés au-cours des années en raison du raisonnement très clair qu'ils autorisent du fait de leur nature plutôt transparente. Nous pourrons retrouver des traces de ce travail par exemple dans Dutoit [*2004*] pour l'unité du signe *i* ou dans Dutoit [2007] pour celle du signe *changer*.
[85] vx : *Décomposer (un agrégat, un organisme) par la séparation des parties.*
[86] Russel [2000] écrit : *Certains auteurs ont affirmé que les facultés perceptuelles et motrices constituent les parties les plus importantes de l'intelligence et que les capacités de "haut" niveau sont nécessairement parasites (il ne s'agirait que de simples extensions des facultés sous-jacentes). Il est vrai que l'essentiel de l'évolution et la plus grande partie du cerveau sont consacrés aux facultés perceptuelles et motrices, alors que l'IA s'est plus intéressée à des tâches plus faciles, telles que le jeu et l'inférence logique, qu'à la perception et à l'action dans le monde réel. Pensez-vous que l'intérêt de l'IA pour les capacités cognitives de haut niveau manque de pertinence?* Nous ne répondrons pas à cette question qui ne nous concerne pas. Cela posé, une fois dit que dans notre cas le monde réel n'est pas en cause, nous pouvons quand même ajouter que dans notre Monde abstrait, il y a beaucoup à faire sans s'occuper d'inférences logiques.

- Intégration de la syntaxe de la définition pour sauver une grammaire surfacique
Ce cas présente une résolution plus détaillée que le précédent. Il correspond au Hiatus "observations sémantiques" et observations dans le syntagme de la page 45.

Les cas auraient pu être présentés dans un ordre différent, mais s'agissant de cas complexes- c'est-à-dire de cas où plusieurs observations naissent en même temps, aucune organisation ne conviendrait tout-à-fait.

6.2.1.1 Intégration de la définition prenant une forme schématique

Nous commençons par cette famille de phénomènes du fait de leur caractère général qui implique une compréhension plutôt globale et avancée. Ce cas nous est venu par l'observation de l'article *riche* dans le Robert, qui nous semble tout-à-fait typique de son espèce. Nous n'étudierons pas ici tous les sens de *riche*. Considérons seulement le sens 3 suivant :

> *Qui contient de nombreux éléments, ou des éléments importants en abondance.*
> *Une riche collection de livres rares.*
> *Un sol, une terre riche.*
> *Aliment riche.*
> *Gaz* riche.*
> *Mélange riche (en carburant).*
> *Langue riche (en moyens d'expression)*
> *Rime* riche.* (Robert)

Nous avons alors pensé que l'énumération est incomplète. Elle devrait au moins se terminer par "..." ou etc. Par exemple, il y aurait dans le "etc" *thèse riche*. Alors deux questions se sont posées :
- dans ce cas, que vaut l'onomasiologie de "..." ou etc.
Cette question nous est suggérée par la Théorie Sens↔Texte qui ne travaille que par extension, réalisant ce que Jacques François appelle parfois une indexation féroce.
- qu'en est-il alors du statut théorique de la Théorie Sens↔Texte?
La question se pose en effet car si nous acceptons[87] comme le dit TST qu'*une des tâches primordiales de la linguistique théorique contemporaine est l'élaboration d'une théorie de la paraphrase langagière* [Mel'çuk, 1992, p10], il est clair que selon cette théorie les énoncés suivants devraient être paraphrases l'un de l'autre :

> *Jean a écouté un riche thèse*
> *Jean a écouté une thèse qui développe de nombreux arguments*

Nous observons que la Théorie Sens↔Texte ne peut rien traiter de ce cas pourtant au cœur d'une des tâches primordiales de la linguistique contemporaine puisque certainement responsable du plus nombre qui soit de paraphrases. Pourquoi voyons-nous ici un gisement considérable de paraphrases? Pour cette raison simple qu'il s'agit de toutes les paraphrases qui implique une certaine compréhension *intensionnelle*. Les paraphrases qui impliquent une certaine compréhension sont évidemment plus nombreuses que celles liés à des processus idiomatiques connus précisément et susceptibles d'être listés dans le dictionnaire.
Pour ce cas au moins, nous nous sommes sentis attirés par d'autres formes de linguistique.

---

[87] Voir dans 3.1.2 *L'enrichissement du modèle : le Dictionnaire Intégral (LDI)*, page 24, le paragraphe [A].

Le cas

Considérant l'adjectif *riche* et les *formes schématiques* de Culioli [1990], il m'apparaît clairement que de telles formes existent bien, au moins pour ce mot. Par exemple :

> *pour le locuteur, l'objet qualifié présente lui-même ou dans l'objet nécessaire d'une de ses relations prédicatives quelque chose*
> *qui existe en grand nombre ou en grande masse et ceci d'une manière favorable du point de vue du locuteur.*

Dans cette forme, le trait *quelque chose qui existe de manière favorable* est assez général à tous les exemples d'utilisation et ne pose pas de problème insurmonté. Par contre, le trait *quelque chose qui existe en grand nombre* est quant à lui très spécifique à la chose qualifiée:

une mine riche en or (en or, désirable pour le locuteur)
un homme riche (en moyens financiers, désirables pour le locuteur)
un sol riche (en éléments fertilisant, désirables pour le locuteur)
une thèse riche (en arguments, désirables pour le locuteur)
un style riche (en tournures, désirables pour le locuteur)
etc.

Il faut noter au plan linguistique, c'est-à-dire *in fine* au plan de la paraphrase, que les différentes compréhensions impliquées n'appellent précisément pas les mêmes paraphrases. En oubliant le cas particulier idiomatique *un homme riche* (ce cas pourra trouver une analyse assez voisine de celle de *changer* dans *je vais me changer*), tous les autres emplois se trouvent élucidés par la structure sémantique même des significations des noms qualifiés.
Pour revenir à ce qui nous intéresse, c'est-à-dire non pas sur une organisation du sens fondée sur des fonctions (comme c'est le cas avec la TST) mais à une organisation fondée sur des mondes méréologiques interconnectés, c'est-à-dire sur des organisations partie-tout, nous avons dans ce vocabulaire :

pour *mine* ce n'est pas la totalité *mine* qui vaut *beaucoup* mais les *minerais extraits*
pour *style* ce n'est pas la totalité *style* qui vaut *beaucoup* mais la *variété des moyens d'expression utilisés*
pour *sol* ce n'est pas la totalité *sol* qui vaut *beaucoup* mais ses *qualités nutritives pour les plantes*
pour *thèse* ce n'est pas la totalité *thèse* qui vaut *beaucoup* mais, par exemple, *l'ampleur, la profondeur et la variété des idées de la thèse*
etc.

Cette question pointe directement le problème de la référence de la quantification portée par *riche.*

Direction pour une solution structurale

En pratique, tous les exemples fournis ici présentent un certain degré de figement qu'un contrôle des emplois sur Frantext vérifie bien. Il est donc pertinent et souhaitable que le dictionnaire reflète ces emplois. Cela dit, ce qui nous occupe ici est l'impact de la *forme schématique* (marqueur[88]) supposée sur le signe cible (récepteur passif). Autrement dit, nous prétendons que le sens de *riche + nom* est précisément cet impact. Selon quel point de vue

[88] Voir note 71 sur puissance active/passive page 83.

thèse peut être valablement *riche ?*
Nous ne saurions trop dire ce que vaut *riche pélican* mais il nous semble bien que *riche thèse* signifie quelque chose qui se passe de tout contexte pour être supposé. Comment procéder?
La première chose à faire est d'ouvrir le dictionnaire au mot *cible* puisque nous savons déjà "tout" de *riche.* Ouvrons donc le dictionnaire pour le mot *thèse* :

> *1] Proposition ou théorie particulière qu'on tient pour vraie et qu'on s'engage à défendre par des arguments.*
> *2] Anciennt Proposition ou série de propositions que le candidat à un grade de bachelier, de licencié, de docteur, etc., s'engageait à soutenir.*
> *3] Ouvrage présenté pour l'obtention du doctorat.*
> *4] Philos. (Hegel) Premier moment de la démarche dialectique auquel s'oppose l'antithèse*, jusqu'à ce que ces contraires soient conciliés par la synthèse. [ROBERT]*

Les **points d'impacts**[89] immédiats sont dans :
1] .....défendre par ..... **beaucoup** ....d'arguments
2] .... série de ....***beaucoup de*** .... propositions
3] *?* peut-être, si la Structure en a besoin : Ouvrage ***qui coûte beaucoup*** ....présenté .....

Quelle pourrait être la solution informatique ? Une solution vraiment bonne ne peut pas être exposée avec le matériel dont nous disposons pour le moment. Disons toutefois qu'il existe un LCA intéressant entre *riche* et *.....défendre par ..... beaucoup ....d'arguments.*
Ce LCA est trouvé pour *riche, beaucoup.* Il s'agit du concept [quantité importante].
Cependant nous voyons bien que ce procédé n'est pas une solution suffisante puisqu'elle n'intègre pas directement la forme schématique.

Localisation de l'impact et formes résultantes
L'exemple impacte l'axe compositionnel à l'intérieur de la définition du mot qualifié par *riche.* En conséquence, il modifie l'émission componentielle dudit *mot* : par exemple, nous avons dorénavant une *thèse riche,* c'est-à-dire une thèse qui a non pas des arguments mais beaucoup d'arguments.
A compter de maintenant, nous traitons les nouveaux cas plus succinctement, en faisant l'impasse sur la situation théorique du cas dans l'état de l'art. D'une façon générale, les remarques précédentes resteraient valables.

6.2.1.2 Intégration de la contradiction entre connaissances des choses et connaissances des définitions

Le cas suivant est intéressant car il nous dit que selon ce qui émerge de la prédication et de la mémoire analogique, il ne résulte pas un même résultat. Ce cas considère aussi un risque lié à la résolution du Hiatus "dictionnaire à l'envers" en rapport avec l'absence d'organisation entre les concepts des quasi-définitions page 44, que nous considérons de fait comme comblé avec le matériel de 6.3.2 *La résolution du cheval blanc*, page 115

[89] Nous utilisons le mot *point d'impact* à regret car il est assez peu précis. Mais l'usage d'un néologisme formé d'un mot-valise savant n'améliorerait par la précision. Par ailleurs, l'emprunt d'un terme d'un auteur est difficile car cet emprunt ne viendrait pas, dans ces commentaires rapides, avec tout le matériel théorique dudit auteur. Cela n'apporterait que confusion. Le mot point d'impact signifie : lieu où une où un impact se produit ET forme de cet impact dans la Structure.

Le cas

Avec la seule proximité componentielle (voir chapitre 4.1.2.2.5, page 41) que nous avons définie, il est évident que le dictionnaire à l'envers peut faire des erreurs amusantes comme pour : *boisson de la vache*. En effet, notre dictionnaire à l'envers répondra : *le lait* et fera erreur. L'humain aussi est capable de cette erreur[90]. En plein amphithéâtre, présentant un article lors d'une conférence, nous avions posé la question :
- *Car après tout, qu'est-ce qu'elle boit la vache?*
Le modérateur de la conférence, avec son micro, répondit à vive voix pour l'assistance :
- *Mais du lait, évidemment, où est le problème*!
Puis, prenant quelques instants de recul, il s'esclaffa. Le problème était que notre Sémiographe ne pouvait pas corriger, même théoriquement, son erreur annoncée par la *Question 1 Comment enregistrer des concepts structurés dans le DAG ?* page 20.

Direction pour une solution structurale

Avec les outils présentés paragraphe 6.1, page 76, il est évident que nous disposons de tout le matériel nécessaire pour autoriser la perception de quelque chose comme *animal boire eau*. Ainsi, *vache* pourra *boire eau*. Mais autre chose doit-il déjà être fait? Nous disposons de tous les outils permettant de prioriser le compositionnel sur le componentiel. Doit-on aller dans cette direction? A court terme, je ne crois pas : cela sentirait le *c'est étudié pour*. Et à long terme? Peut-être, mais il faudra que dans tous les cas de figure, la Structure puisse localiser l'analyse faite par le modérateur ; autrement dit, il faut qu'elle soit capable de la reproduire.

Localisation de l'impact et formes résultantes

L'exemple impacte deux lieux sémantiques différents et produit deux formes.

6.2.1.3 Intégration de l'inférence issue des connaissances sur les choses

Le cas correspond au Hiatus "observations sémantiques" et observations dans le syntagme énoncé page 44. Il s'agissait d'un problème de *samouraï. Samouraï* mange-t-il ou ne mange-t-il pas?

Le cas

Un problème de *samouraï* pour nous est le suivant : comment pouvons-nous en décrivant *manger* d'une part, *samouraï* de l'autre, faire en sorte que des liens syntagmatiques acceptables apparaissent par exemple dans *bol et cuillère du samouraï*. Quelle méthode de travail pourrions-nous définir afin d'éviter des descriptions innombrables et anarchiques ?

Direction pour une solution structurale

Considérons *samouraï* et *manger* d'une part, et *samouraï* et *Sushi* de l'autre.
De *samouraï* à *manger*, la liste des LCA est vide dans la région componentielle. Nous savons cela depuis le paragraphe 4.2.2, page 45, du fait, entre autre que *samouraï* n'est pas *humain* en français mais est seulement *personne*. Nous observons qu'il n'est évidemment pas plus *mangeur* en français. En outre, même s'il se peut bien que nous trouvions un vague signal typique du *bruit de fond* inhérent au dictionnaire refermé sur lui-même, nous ne pouvons imaginer en faire quelque chose : il comporte bien trop de changements de *points de vue*.

[90] voir Korzybski [1933], pour une étude assez systématique de ce genre de choses.

Du côté de *samouraï* et *Sushi* les choses se passent bien mieux : *[Japon]* ressort immédiatement.
La solution du cas tient donc dans la mise sur le devant de la scène d'un certain point de vue qui manque. Cette information qui n'est pas définitoire de son contenu est compositionnelle au niveau de la perception : elle dépend strictement des prédications locales et est asymétrique. Cela veut dire que si quelque chose se passe d'un côté (terme en logique propositionnelle) du fait de l'action d'autre chose (prédicat en logique propositionnelle), alors nous obtenons un proposition (tout dans notre cas) qui entraîne l'émission d'une information vers, et c'est un minimum, les régions componentielles et les régions métalinguistiques pour le terme considéré. C'est précisément ce que nous avons déjà réalisé dans l'exemple le plus simple que nous pouvions trouver, paragraphe 6.1.2 *Intégration d'énoncés compositionnels métalinguistiques*, page 81 :
- nous avons d'abord vu une perception à l'œuvre sur la structure qui a laissé une première trace de ses bornes
- puis nous avons constaté une émission dans une région de l'interpréteur susceptible de réaliser des actions sur lui-même. Pour le cas qui nous concerne (*samouraï mange*), il nous faut voir que tous les énoncés suivants trouvent une solution disponible localement :

- d'un verbe : Le samouraï gloutonne[91]
- d'un nom : nourriture du samouraï
- d'un adjectif : samouraï gourmand
- d'une relation prépositionnel en "de" : repas du samouraï
- d'un adverbe utilisé en hypallage : samouraï regardant goulûment.
- phase en verbe : samouraï mâche
- conséquence en nom : digestion du samouraï
- argument objet en nom : bœuf braisé du samouraï
- instrument en nom : bol et cuillère du samouraï
- locatif en participe passé ou adjectif : Samouraï attablé
- etc.

Pour chacun de ces énoncés résolus[92], localement, *samouraï* est vu comme un *mangeur*[93]. C'est ce qu'il nous faut obtenir[94] en suivant les désintégrations et réintégrations compositionnelles et componentielles. Nous fournirons une piste plus détaillée paragraphe 6.3.2 *La résolution du cheval blanc*, page 115.

---

[91] Le caractère idiomatique de l'exemple est plus que discutable; c'est de fait un parti pris. Il rappelle que nous nous intéressons davantage à ce que veulent dire les choses qu'à la manière de les dire, cette dernière question relevant davantage de la grammaire et du bon usage que d'une sémantique interprétative.
[92] c'est-à-dire non au début du processus.
[93] Il devient membre de la classe des mangeurs comme une formule est devenu membre d'une classe de formule susceptible d'être invitée à s'exécuter par le contexte.
[94] Framenet de Fillmore fournit un assez bon exemple d'interconnexions des lieux que nous venons d'envisager. En cela, le suivi des travaux sur Framenet importe pour nous. Mais Framenet n'est pas une Structure. Il s'agit juste d'un réseau logico-sémantique conçu dans une vision particulière. Le besoin qu'il a de s'illustrer sur plusieurs milliers d'exemples pour chaque Frame créé le montre bien. Pour nous les exemples ne valent que s'ils reflètent quelque chose de particulier dans l'usage. Ils n'ont droit de citer dans le dictionnaire qu'en cela. Pour Fillmore, au contraire, les exemples font figure de modèle en tant que tels : ils sont la justification et ils déterminent le contour. Fillmore a besoin de ses exemples (au minimum dix exemples par morceau de description) pour espérer faire fonctionner ses frames du fait même qu'il manque de structure. Framenet de Fillmore ne peut exister comme WordNet de Fellbaum ou la TST de Mel'çuk que comme une extension de quelque chose. Ces modèles qui n'ont pas de contrainte structurale systémique ne peuvent en développer une comme ils le revendiquent parfois, depuis leur intérieur. Néanmoins, ces modèles proposent des matériaux, des points de vues que nous regardons avec convoitise et que nous aimerions avoir déjà intégrés.

Localisation de l'impact et formes résultantes

L'exemple impacte plusieurs chaînes compositionnelles et par retour émet régulièrement sur la bande componentielle qu'il modifie.

6.2.1.4 Intégration de la syntaxe de la définition pour sauver une grammaire surfacique

Le cas que nous allons étudié ici pourrait être localement traité par la TST à l'aide de fonctions lexicales. Mais il faudrait que la TST n'admette pas trop les termes, ou bien les accepte mais en faisant attention à bien respecter l'unité du signe du point de vue du terme (voir 6.1.4 *Intégration du terme*, page 87). L'observation des articles du DEC 1992 ne montrent cependant pas la prise en compte de cette contrainte (voir par exemple l'article *manière de parler*, page 233, du DEC 1992).

Le cas traite d'un problème posé par l'expression *pêcheur à la ligne.*

Le cas

Soient *pêcheur à la ligne*, *pêcheur au filet*, *pêche à la cuillère* etc. Ne retenons que *pêcheur à la ligne.*

Un système simple, œuvrant dans une seule cinématique, pourra réaliser l'une des erreurs suivantes :

- ou bien, il refusera la décomposition de l'expression semi-figée et sera inadapté à une application nécessitant une décomposition comme celle que nous avons montrée paragraphe 6.1.4 *Intégration du terme*, page 87. En refusant de casser *pêcheur à la ligne*, il considère alors *pêcheur à la ligne* comme *pomme de terre*. En plus d'échouer dans une application, il porte atteinte aux unités de structure, d'analyse, de résultat et de signe. En fait, ces travers vont ensemble.
- ou bien, par exemple, il est de niveau grammatical, et désambiguïse à l'aide d'une ontologie lexicale un peu faible : alors, il repère *personne* et *à la ligne*. Soit il sait qu'il ne sait pas rattacher un groupe prépositionnel (c'est normalement le cas s'il n'a pas la locution dans son lexique, et alors nous retrouvons l'erreur applicative et analytique précédente), soit il rattache sur :
- le nom, ce qui produit une erreur comique sur l'axe paradigmatique : *personne à la ligne*
- ou un verbe disponible et aimant la préposition *à*. Vraiment n'importe quoi pourra alors survenir.

Direction pour une solution structurale

Nous pourrions déjà commencer par réitérer ces observations pour, par exemple, *pêche à la ligne* ou *pêcheur à la cuillère*. Cet exercice nous aidera certainement à dégager que :

- *ligne* est *instrument du pêcheur*
- que *pêcheur* est *personne qui pêche*

et que finalement, *pêcheur à la ligne* peut se réécrire : *personne qui pêche à la ligne.*

En définitive, nous observons que *à la ligne* est complément instrumental d'un verbe d'action (*pêcher)* en même temps que *ligne* est défini par ce verbe (*instrument pour pêcher, de pêche, du pêcheur).*

Il nous faudrait maintenant représenter la Structure pour toutes ces observations. Nous ne le ferons pas ici car il nous semble que le graphe obtenu serait illisible.

Localisation de l'impact et formes résultantes

L'exemple utilise plusieurs chaînes compositionnelles (Frame dans l'emploi de Fillmore [1976]) ou componentielles (qualia chez Pustejovsky [1995]) œuvrant par combinaison à l'intérieur d'un même tout componentiel (Thème chez nous).

6.2.2 Conclusion

Avec ces exemples, nous comprenons qu'un même mot, même doté de tout ce que l'on voudra comme fonction inventée dans un micro-système, même si l'on considère la fonction MODIFIEUR elle-même, même encore si l'on considère la fonction FONCTION ne pourra jamais mettre en œuvre un autre mot pour les diverses tâches de perception impliquées dans les applications. Nous remarquons que cette conclusion est plutôt pessimiste.

Pourtant, il est possible d'être positif, en admettant une complexité supplémentaire. Il nous faut constater que les quatre exemples donnés dans ce paragraphe ont en commun qu'ils trouvent leur solution dans l'examen des définitions et potentialités du terme *passif*, celui qu'Aristote appellerait la *puissance passive* (voir note 71 page 83).

Pour que cet examen soit possible sans rompre l'unité de l'analyse, c'est-à-dire en se contentant de réflexes de perception, il faudrait que les définitions soient accessibles au-dessus de la structure des concepts. Techniquement, cela se fait dans le graphe à l'aide d'une structure miroir qui se crée ainsi :

pour toute relation r de A vers B **dans** la Structure
créer la relation r^{-1} de B vers A, **dans** la Structure miroir.

En développant cette structure miroir, nous réaliserions exactement l'étendue de notre étude : signe x au départ, signe x' à l'arrivée (voir paragraphe 2.1 *Positionnement de nos travaux,* page 7).

Par exemple, au plan fonctionnel, la structure miroir simplifierait grandement une opération que nous connaissons bien. Il s'agit du *dictionnaire à l'envers.* Dans le *dictionnaire à l'envers,* depuis 1996, nous générons d'abord tous les spécifiques d'un générique que nous évaluons plus ou moins individuellement par la suite. Cela nous a toujours semblé peu naturel et bien compliqué. Au regard de cette complication, une Structure miroir annule cette étape et fait gagner du temps : pour calculer *personne qui vend des hortensias,* il n'est plus nécessaire de générer exhaustivement toutes les personnes comme nous l'avions décrit.

La date (voir paragraphe 6.1.5, page 88) est une illustration du même mécanisme de structure miroir. En effet, chaque fois qu'une date est prouvée, toutes les composantes onomasiologiques (on est ici du côté du mot) sont rendues accessibles. C'est une fonction r^{-1}.

Mais il faut comprendre que si une structure miroir apparaît, une structure miroir de la structure miroir peut aussi apparaître. Dans un tel processus, l'intrication part d'un signe x, s'enrichit par des concepts et des événements divers, continue sur sa définition, qui devient alors x', et continue ce mouvement, qui aboutit le plus souvent non à un cercle mais à une spirale : ça ne boucle pas, ce qui est à la fois inquiétant pour le temps de calcul et stimulant[95] pour des questions d'apprentissage à long terme.

Considérant toutefois qu'il est beaucoup plus complexe de se représenter la Structure avec en son "haut" la structure miroir, qui a évidemment elle-même pour haut la Structure+1, qui a pour haut la structure miroir+1 etc... le tout formant la Structure dynamique, nous ne ferons plus allusion à une telle perspective dans ce mémoire.

---

[95] Voir note 2 page 8, sur la durée, dans le sens donné par Bergson.

Au contraire, nous allons considérer que nous avons dorénavant une Structure en tête, avec un bas et un haut bien clairs, sans miroir. Dans cette Structure, nous savons qu'un Signe peut agir sur un autre Signe d'une façon que nous allons essayer de décrire objectivement et formellement pour un ordinateur sans l'aide d'une multiplicité de fonctions qui finiraient par rendre compliqué et complexe un problème qui à l'origine est peut-être seulement complexe.

6.3 La définition d'une microsyntaxe pour élargir un peu le champ perceptif de la Structure

Le titre de ce paragraphe indique assez précisément les portée et motivation d'une définition d'un point de vue que nous avons choisi d'appeler *microsyntaxe*[96]. Il ne s'agit pas d'une théorie. Il ne s'agit pas plus d'une hypothèse. Il s'agit seulement d'un point de vue que nous pourrions avoir dans la perception d'une interaction entre deux Signes quelconques et que nous aimerions pouvoir transmettre à la Structure. En réalité, au point où nous en sommes, il pourra paraître que ce point de vue est presque tellement global qu'il en devient théorique. Cela ne nous importe pas. En effet, la seule chose que nous voyons maintenant, c'est que nous disposons d'une Structure extrêmement plastique et organisée capable de bénéficier sans effet délétère d'un :

- *élargissement de son champ perceptif.*

Nous donnons les directions prises par cet élargissement en suivant le plan suivant :
D'abord, nous introduisons un nouveau point de vue, c'est-à-dire un nouveau lieu de perception (chapitre 6.3.1 *Postuler la microsyntaxe*, page 109, ci-dessous).
Ensuite, nous étudions la résolution du *cheval blanc (*voir *6.3.2 La résolution du cheval blanc*, page 115).
Enfin, nous redonnons un peu de matière à un *cheval* qui a été dans le paragraphe précédent particulièrement dépouillé. Nous lui rendons un peu de cette matière du fait d'une exigence paraphrastique venant d'un énoncé et non du fait d'une croyance en l'animal *cheval* ou en le monde *notre-monde*. (voir 6.3.3 *Réintroduction du cheval : intégration de la chaîne microsyntaxique et potentialités*, page 126).
Nous concluons enfin sur la pertinence des postulats microsyntaxiques en terme de compatibilité avec les notions d'unité de Structure, d'Analyse, de Résultat et de Signe, qui sont pour nous au cœur de l'intégration structurale des points de vue compositionnels et componentiels.

6.3.1 Postuler la microsyntaxe

Dans ce paragraphe, nous commençons par des définitions que nous présentons d'abord

---

[96] Un des deux termes *microsyntaxe* et *microsémantique* est utilisé par la sémantique interprétative de Rastier et non l'autre. *Microsémantique* est ce terme commun. Nous le conservons par habitude bien qu'il marque chez nous un point de vue componentiel différent de celui existant chez Rastier. Par exemple, la seule manière que nous pourrions utiliser pour attacher un sème /animé/ au signe *père n.m.* serait de dire quelque chose à la Structure comme : *père* est */animé/* **dans** la théorie de Rastier et de nombreux autres théoriciens. C'est d'ailleurs ce que nous aurons comme résultat (de la même façon que nous avons dit qu'une instance de formule est un spécifique d'une classe de formule comportant un générique *formule*). Pour nous, */animé/* est un résultat (output) et non une donnée (input).

comme des postulats[97] (Définition de la microsyntaxe ci-dessous). Nous fournissons alors immédiatement des conditions qui nous permettraient de mieux accepter ces postulats (voir 6.3.1.2 *De la pertinence des postulats de la microsyntaxe*, page 113). Et nous situons le cas que nous allons étudier en détail dans une sorte d'état de l'art (voir 6.3.1.3 *Un corpus plus étendu de cheval blanc*, page 114).

6.3.1.1 Définition de la microsyntaxe

Nous fournissons ci-dessous les principales définitions (postulats). Nous ne faisons pas mention directe de travaux en relation en informatique. A notre sens, le principal inspirateur de ces définitions est vraisemblablement Culioli [1990] qui, par exemple, argumente :
- l'idée d'un *système de repérage* tandis que nous parlons de *système de localisation*.
Mais nous sommes encore très loin de pouvoir manier les principaux opérateurs de cette théorie du fait du niveau d'abstraction où ils se situent (en particulier l'articulation *quantification* Qnt et *qualification* Qlt).

Définitions complémentaires

Microsyntaxe : syntaxe qui opère essentiellement au niveau du syntagme et met en place les propriétés simples ou complexes prévues par des **scénarios définis dans le dictionnaire.**

NB : Le terme microsyntaxe est parfois [Avanzi 2007] réservé à l'étude des relations syntaxiques à l'intérieur de la phrase et s'oppose à la macrosyntaxe qui a une portée interphrastique. Dans notre emploi, le terme microsyntaxe a un sens tout-à-fait différent : il s'oppose à la syntaxe ordinaire qui établit une relation entre *pêcheur* et *à la ligne* pris comme des touts mais accepte des micro-relations entre des parties sémantiques de *pêcheur* et *à la ligne.*

Nous allons illustrer cette définition par un exemple que nous avons déjà discuté. Il s'agit de la question de la *vache* et de la *boisson* (voir paragraphe 6.2.1.2, page 104).
Nous nous rappelons que dans cet exemple du point de vue (pertinent) microsémantique, *boisson* + *vache* vaut *lait* tandis qu'au point de vue (pertinent) microsyntaxique *boisson* + *vache* vaut *eau* en tant que quelque que chose connaissant une propriété *boisson* c'est-à-dire connaissant une propriété (puissance active) telle que cette propriété appelle un propriétaire (puissance passive) capable de *boire*. **Selon** l'axe microsyntaxique et l'interprétation que nous venons de faire,
boisson-->*boire* est considéré comme un **marqueur syntaxique** pour *vache*.

Marqueur microsyntaxique : Signe capable d'appliquer à son environnement un programme particulier et identifiable[98]. Les marqueurs microsyntaxiques sont dotés de concepts particuliers en plus de leur concept componentiel. Les marqueurs microsyntaxiques ne sont pas des natures ou des classes ; ils sont des actes qui causent l'appartenance hypothétique d'instances à des natures ou des classes.

Par exemple, dans *boisson de la vache*, *boisson* cause une hypothèse d'appartenance

[97] Postulat : principe d'un système déductif qu'on ne peut prendre pour fondement d'une démonstration sans l'assentiment de l'auditeur (Robert). Ici le postulat concerne notre façon de voir la microsyntaxe.
[98] Nous donnons ici la note de Anne Nicolle : il s'appellerait *acteur* en langage de programmation. Mais comme nous ne sommes pas dans un langage de programmation nous préférons éviter cette dénomination.

ponctuelle, comme instance, de *vache* à la classe *buveur*. Nous sommes ici très loin de mécanismes de sélection booléens de traits statiques tels que proposés par Rastier (voir par exemple *base de règles* [Rastier 1994, p81]. Le processus casuel et dynamique proposé ici aboutit à l'émission d'une simple hypothèse dans la Structure. La figure suivante illustre d'une façon assez conceptuelle ce mécanisme.

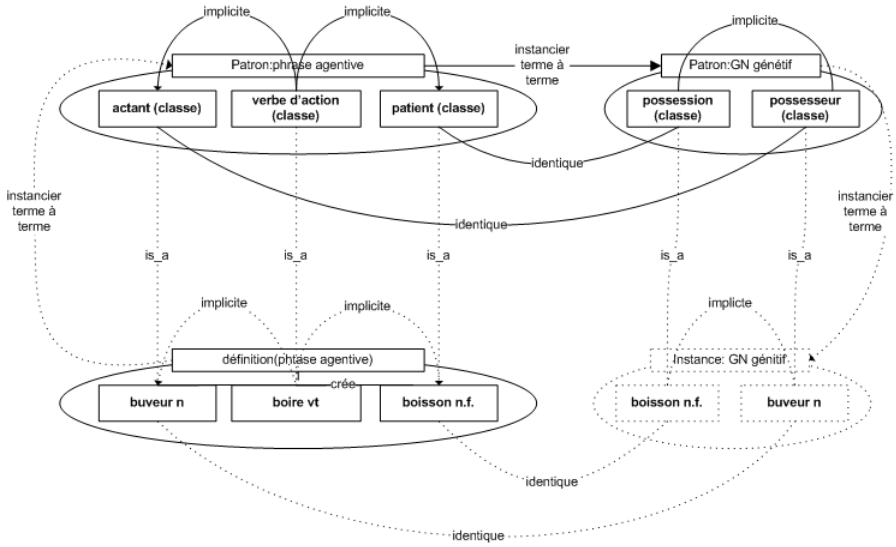


Figure 18 Les marqueurs *Buveur*, *boire* et *boisson* dotés d'une organisation supplémentaire dans le dictionnaire (représentation très peu détaillée).

Cette figure ne produit pas à proprement parler de concepts ou de symboles que nous n'ayons déjà présentés.
Tout en haut de la figure, "*instancier terme à terme*" stipule que le modèle de phrase agentive de droite produit un modèle de syntagme génitif. Il s'agit simplement d'une réification de Frames présents potentiellement dans les thèmes du DAG. Cette réification s'effectue depuis le modèle le plus complet (la phrase agentive) jusqu'aux modèles les plus elliptiques (dans notre cas un syntagme nominal). La réification est produite virtuellement dès la création de la phrase agentive au moyen des règles d'émission que nous avons vues. Le résultat est un syntagme qui finit par s'appeler (non représenté) syntagme dans la langue comme nous l'avons déjà vu pour une FORMULE (voir paragraphe 6.1.6, page 93).

Quel est le sens exact de la figure appliquée à un exemple particulier? Prenons :

La boisson du frigo

Avec ce graphe de la Structure, la Structure est amenée à examiner *frigo* du point de vue de *boisson.* En gros, la structure se demande[99] si *frigo est buveur.*
Une autre expression du même type, ailleurs dans le DAG, ferait "se demander" par la Structure si *boisson peut être contenue frigo.* Nous espérons évidemment que le deuxième

[99] En fait elle ne se demande rien : elle pose comme réflexe que le *frigo* est en place de *buveur*. Si rien ne vient valider cela, il ne se passera de toute façon pas grand chose de plus. Par contre, selon ce qui est connu, par exemple, si dans le texte il a été perçu quelque part *frigo* comme sujet de *boire* (c'est un texte décrivant un cartoon de Tex Avery), alors cela sera repris par une confirmation de l'événement.

résultat aboutira davantage que le premier. Mais pourquoi ce deuxième résultat ne serait-il pas trouvé par un LCA : après tout, il s'agit ici d'une partie de la définition de *frigo.*

Autres Définitions

Nous utiliserons par la suite les termes de *marqueur complexe* et de *marqueur simple* essentiellement pour organiser notre travail et notre propos.
Nous supposons seulement qu'un *marqueur complexe* comporte plusieurs *marqueurs simples*. C'était le cas pour l'exemple fourni avec *riche* (voir paragraphe 6.2.1.1 *Intégration de la définition prenant une forme schématique*, page 102) qui dans une certaine mesure marque d'un côté l'opinion et la faveur d'un locuteur particulier et de l'autre une articulation existant à l'intérieur de quelque chose.
Généralement, pour la suite, quand nous parlerons de *marqueur complexe* ce sera surtout pour dire que nous ne souhaitons pas en parler davantage : évidemment, pour nous comme pour tout constructiviste, nous ne pouvons montrer un objet complexe que si nous disposons en premier lieu de toutes ses parties.

Marqueur complexe : marqueur qui impacte plusieurs références à la fois.
Ces marqueurs sont toujours des hyper-arêtes d'ordre et de rang > 2 : ils n'ont pas d'ancrage simple dans le Signe.

Marqueur simple : marqueur microsyntaxique mettant en jeu une propriété simple du Signe.
Ces marqueurs se représentent directement dans le DAG : ils ne passent pas par la synthèse de plusieurs touts.

Cela posé, tous les Signes sont des marqueurs ou des marqués potentiels puisque la notion même de marqueur n'est qu'une invitation à prendre le point de vue d'une interprétation particulière. Par exemple, *couleurs* est un marqueur dans *couleurs du cheval* tandis que Rembrandt devient marqueur dans *couleurs de Rembrandt* (cas remarquable de l'actualisation d'une isotopie sémantique).

Nous disons que certains marqueurs peuvent agir en-dehors de leur champ sémantique. C'est par exemple le cas de *couleur* qui peut agit sur *vêtement* comme ici : *couleur du vêtement*. *Vêtement* n'est certainement pas dans le champ sémantique de *couleur*. Dans ce cas, nous appelons ces marqueurs des marqueurs forts : ils causent qu'un Signe se retrouve dans leur champ sémantique.

Nous appelons **marqueur fort**, un marqueur qui peut agir en-dehors de son champ sémantique (c'est le cas de *couleur* mais non celui de *Rembrandt).*
Le plus souvent, un signe se rencontre tantôt comme marqueur fort et tantôt comme marqueur faible : dans *le mot couleur* (voir chapitre 6.1.2, page 81), *couleur* est complètement repéré, localisé par *nom* et plus rien ne devrait s'échapper de lui. *Couleur* devient complètement marqué. D'une façon générale, nous utilisons aussi le terme *marqueur fort*[100] pour dire que

[100] Le premier sens de marqueur *fort* (agir en-dehors de son champ sémantique) correspond à la définition de *fort* suivante : qui a un grand pouvoir d'action. Le deuxième sens de marqueur *fort* (agir d'une façon claire) correspond à la définition de *fort* suivante : qui agit beaucoup ou efficacement. Evidemment, nous nous intéresserons d'abord aux marqueurs *forts* dans les deux sens du terme.

l'image de la marque, de l'impact est très claire quand le marqueur a agi.
Nous disons également que certains signes sont difficiles à imaginer comme marqueur fort (par exemple *personne*, mais sauf dans *cette personne* en parlant d'un *lapin* dans *Alice au Pays des Merveilles*), et que d'autres sont difficiles à concevoir comme marqueur faible (c'est notamment le cas de *vultuosité*).
Avant de conclure sur ces définitions, nous souhaitons donner encore une indication concernant la manière d'interpréter le mot *marqueur* :
- quand un marqueur en cause dans un co-texte est pris dans sa dimension faible, il pose toujours la question suivante : qu'est-ce qui permet de dire que ce Signe supporte telle ou telle prédication[101] (puissance passive)?
- quand un marqueur en cause dans un texte est pris dans sa dimension forte, il pose toujours l'affirmation suivante : il existe un marqueur faible qui supporte une certaine prédication (puissance active).

Le principal intérêt de la notion de marqueurs est lié au besoin de préciser le mot *sens* quand nous disons que tel *sens* est retenu. Le marqueur définit toujours un point de vue partant de quelque chose et portant sur une chose de classe ordinairement très différente et que le dictionnaire ne relie pas nécessairement. Dans les cinq premiers chapitres, nous n'avions pas besoin de la notion de marqueur puisque nous réfléchissions toujours dans le cadre de l'extension d'une définition. A contrario, le marqueur permet d'aborder plus ou moins la définition sur un plan intensionnel. Avec le marqueur, nous sortons du point de vue componentiel où le mot signifie dans tel ou tel co-texte et nous abordons le point de vue d'une pragmatique abstraite[102] dans lequel le mot agit directement sur la Structure. En définitive, le mot marqueur est utilisé pour montrer certains effets pratiques du *sens*[103].

6.3.1.2 De la pertinence des postulats de la microsyntaxe

Les postulats de la microsyntaxe nous sembleront corrects si et seulement si :
[A] Ils s'inscrivent dans l'une des questions générales laissée dans ce document.
[B] ils respectent les quatre principes d'unité que nous avons proposés : C1, C2, C3 et C4.

Concernant [A], ce critère est d'emblée satisfait puisqu'il figure en entrée : en effet; nous disposons de la question de la page 20. Voici ce qu'était cette question :
Question 4 Comment décrire dans notre DAG des données non componentielles qui appartiennent aux connaissances naïves de telle manière que cette description soit clairement distincte des données componentielles que nous avons déjà représentées.

Concernant [B], dans les paragraphes précédents, nous avons à plusieurs reprises fait des remarques permettant de circonscrire le champ d'une réponse. S'agit-il, à la manière de CYC [Lenat, 1999] d'essayer de déclarer toutes les connaissances d'une petite fille sous une forme logico-déductive? Ou bien s'agit-il d'insérer le trait */animé/* dans le Signe *samouraï* comme le fait Rastier?
Pour essayer d'obtenir une réponse précise à cette question, nous avons choisi de nous inscrire

[101] Par exemple, qu'est-ce qui permet de dire que le *frigo* peut *boire?*
[102] Voir note 69 page 78.
[103] Comme il nous a été posé plusieurs fois des questions pratiques sur le risque d'explosion combinatoire lié à la multiplication des perceptions, et particulièrement de cette perception d'un effet pratique du sens, nous avons ajouté une courte note en annexe sur les caractéristiques de notre moteur.

comme Directeur de Recherche Associé au laboratoire de linguistique de l'Université de Caen (le CRISCO). En effet, puisque de nombreux linguistes utilisent abondamment des traits comme [animé] ou [inanimé], et que ces traits nous semblent bien inaccessibles, il nous a fallu les interroger pour en comprendre l'origine ontogénétique.

Nous avons également réfléchi à l'exemple de référence le plus simple qui soit de telle manière que nous ressentions pleinement que cet exemple n'est pollué d'aucune interaction entre langage et monde, ni d'aucun présupposé. Nous avons alors posé la question que voici : pourquoi *blanc*, à un moment d'une analyse, prend-il *place*, en suivant des motifs compositionnels et componentiels strictement définitoires dans *quelle* sachant l'énoncé

Quelle est la couleur du cheval blanc d'Henri IV?

C'était mi-2001. Nous disposons depuis quelques pages de tous les éléments pour y répondre. Mais il semble qu'il faille exprimer complètement cette réponse. Nous rappelons que l'important est la forme prise par la réponse.
La forme de notre réponse sera-t-elle celle d'un lien avec un système logico-déductif externe à la Structure et dont nous aurions organisé précisément les inférences en fonction d'un but inconnu du Système et que nous poursuivons (trouver la bonne réponse)? Ou bien la forme de notre réponse sera-t-elle simplement une **Structure** qui s'est **formée** indépendamment de tout but exogène, laquelle Structure aurait une ontogenèse endogène à la perception d'une interrogation (*quelle)* et serait ontogénétiquement **obligée** de chercher à répondre du fait de cette même perception [104]?

En bref, la forme de notre réponse saura-t-elle respecter :
C1 : l'unité de la Structure : une seule Structure
C2 : l'unité de l'Analyse : une seule Analyse.
C3 : l'unité du Résultat : une seule Synthèse.
C4 : l'unité du Signe : un seul Signe.

C'est sur quoi il nous faut juger. Donnons-nous maintenant un petit corpus pour asseoir ce jugement sur une variation de faits.

6.3.1.3 Un corpus plus étendu de *cheval blanc*

Voici une expérimentation utilisant l'agent conversationnel de Virtuoz tel que nous pouvons le tester sur *http://www.virtuoz.com/fr/*.

[104] Comme c'est le cas, quand on sait le calculer, pour l'énoncé "*calculer 3+4*5 "*. Comme c'est aussi le cas pour l'énoncé : *combien de lettres a le mot samouraï?* Voir la note 86 page 97 et réfléchir à une réponse à donner à Russel.

N°	Nous	L'agent	Nous pensons
Q1	Quelle est la couleur du cheval blanc d'Henry IV?	Et bien c'est blanc, vous venez de me le dire ! Connaissez-vous VirtuOz ?	Bien.
Q2	Jean a un cheval blanc. Marie a un cheval noir. Quelle est la couleur du cheval de jean?	Je n'ai pas de pantalon ! Connaissez -vous VirtuOz ?	Bon, alors plus facile
Q3	Quelle est la couleur d'un cheval blanc?	Je n'ai pas de pantalon ! Connaissez -vous VirtuOz ?	Alors uniquement la question 1? je réessaie.
Q4	Quelle est la couleur d'un cheval blanc?	Et bien c'est blanc, vous venez de me le dire ! Connaissez-vous VirtuOz ?	Bien.

Tableau 5 Un corpus plus étendu de questions

Comme nous le voyons, il nous est assez difficile de nous faire une idée rapide de la performance d'un agent conversationnel sur le Web. Les réponses sont-elles partiellement aléatoires? Les erreurs masquent-elles des refus polis de perdre du temps à effectuer un calcul stupide ? Pour élargir notre enquête, nous avons fait d'autres essais (de *white horse)* avec plusieurs chatterboots anglais comme celui de http://www.artificial-solutions.com/ ou de http://www.verbots.com/. Les choses n'ont pas été beaucoup plus convaincantes. Nous n'avons pas testé Eliza mais nous avons testé Alice, développé initialement par Richard Wallace. Le résultat est souvent *noir*. Alors, évidemment, nous comprenons bien que tous les agents précédents soient des IA avec leur vie propre. Et nous sommes donc absolument certains que nous ne travaillons aucunement sur une IA[105].
En effet, tout ce qui nous intéresse de Q1 à Q4 est l'obtention d'une structure qui présente dans le graphe contextuel du signe *quelle* quelque chose qui ressemble à un fléchage vers *blanc.*
Abordons maintenant notre résolution du *cheval blanc.*

6.3.2 La résolution du cheval blanc

Maintenant que nous sommes certains que la réalisation d'un agent conversationnel ne nous intéresse pas en soi, et que nous sommes certains que ce qui intéresse relève uniquement de la méthode, examinons une solution qui semble bien fonctionner de Q1 à Q3. Nous examinerons ensuite les propriétés C1, C2, C3 et C4 de cette solution et nous concluons sur le point de vue de la déclaration des connaissances pragmatiques (ce que les mots font) dans la Structure que nous organisons. Nous conclurons finalement sur la méthode elle-même.

[105] Voir la note 86 sur les agents réflexes, page 96.

6.3.2.1 Notre façon de résoudre le problème

Commençons par nous faire une idée plus précise sur :
- un rapport qui existe entre définition et information élémentaire.
- le dictionnaire et les générateurs qu'il contient
- la structure méréologique

Nous étudierons ensuite le graphe obtenu à l'exécution et nous concluons la question technique pour reprendre la discussion conceptuelle et justifier que l'élargissement de la perception n'entraîne pas de difficultés particulières. Bien au contraire.

6.3.2.2 D'un rapport définition du dictionnaire et information élémentaire

Concernant *couleur* et *vert,* Le Robert écrit :

R1) vert *adj.* de couleur verte
R2) couleur *n.* propriété de la surface d'un objet
R3) surface *n. f.* partie extérieure (d'un corps) qui le limite en tout sens.

Arrêtons-nous ici et reformulons sans trop nous occuper de tenir un style ce que nous avons compris. Nous décidons de supprimer toutes les parenthèses des définitions du Robert mais nous aurions pu également décider de créer de nouvelles parenthèses. Nous avons :

1) vert *adj.* valeur d'une couleur
2) couleur *n. f.* propriété d'une surface
3) surface *n. f.* partie d'un corps
4) corps *n. f.* objet matériel

En lisant cette chaîne, nous avons d'abord l'impression d'une articulation ne comprenant que deux nœuds comme dans : *Caen --> Calvados --> France --> Europe --> ... --> Univers.*
Mais précisément --> n'est pas exactement rien ; nous pouvons le remplir avec par exemple *est situé dans*. *Est situé dans* est tout sauf vide.
Est situé dans est de plus une **relation transitive** : Caen *Est situé dans* France.
Si nous essayons d'appliquer la transitivité à notre chaîne de relation, nous remarquons que *de* n'est pas transitif. En effet, *os* de *chien, chien* de *garde* ne donne pas *os de garde* : je ne peux pas remplacer ce qu'il y a derrière *de* par la définition de ce contenu.
C'est absolument naturel puisque, pour faire court, dans A *de* B, *de* dit de considérer A dans la vue de B et ne dit pas grand chose d'autre.
Ainsi, avec ces définitions, nous avons des changements de *points de vue*. Toutefois, pouvons-nous espérer trouver, ne serait-ce que pour nous aider, une forme de continuum dans ces changements? Nous recherchons cette trace en supprimant l'hypéronyme suivant et en montrant ce qui reste dans la définition précédente :
1 et 2 --> vert adj. : valeur d'une couleur d'une surface
2 et 3 --> couleur n : propriété d'une surface d'un corps
3 et 4 --> surface n : partie d'un corps matériel
Et nous opérons encore une fois :
1, 2 et 2, 3 --> vert adj. : valeur d'une couleur <u>d'une surface d'une surface</u> d'un corps *
Ca ne va trop : on ne peut pas supprimer le mot *propriété*.
Essayons avec l'élément suivant :
2, 3 et 3,4 --> couleur n : **propriété d'une surface <u>d'un corps d'un corps</u> matériel ***
Ca ne va pas du tout.

Essayons encore :
1, 2 et 3, 4 --> vert adj. : valeur d'une couleur d'une surface d'un corps matériel
Cela semble tout à fait intelligible.
En définitive, nous avons le droit de :
- supprimer partie pour mettre un possesseur à la place
- supprimer propriété pour mettre la dénomination de la propriété à la place
En **1, 2 et 3, 4,** nous retrouvons deux représentations qui nous sont intuitives, familières et que nous allons utiliser :
[A] une connaissance du type VALEUR – PROPRIETE – ENTITE
valeur d'une couleur d'une surface
[B] une connaissance du type PARTIE – TOUT
surface d'un corps matériel
Mais ce "modèle" n'est pas encore tout-à-fait correct. Pour qu'il le soit, il faudrait que l'on représente [B] selon [A] et réciproquement [A] selon [B]. Nous ne le ferons pas ici car même si cela introduit une rupture C1-C2-C3-C4, cette rupture peut passer inaperçu pour notre propos.
Nous allons maintenant nous intéresser à [A].
[A] est très intéressant parce qu'il représente ce que nous appelons souvent une *information,* c'est-à-dire un renseignement sur quelque chose. Nous avons choisi de l'appeler *définition* du fait que du point de vue des instances, il définit complètement une information observable et ne va pas au-delà du domaine de cette information. En effet, si nous avons un énoncé : *cheval blanc*, *blanc* marque par (sa) définition *surface* tandis qu'on ne sait pas du tout si *cheval* doit plutôt marquer [animal], [cheval], [animé] ou n'importe quelle sorte d'autre chose. Tandis que *cheval* n'indique en rien de quelle manière il faut le prendre (en-dehors de lui-même), l'un des sens de *blanc* fournit par sa définition la façon unique dont il doit être pris et comment nous devons gérer ses instances :

blanc - *couleur – surface* (dénommée *cheval*).

C'est ce qui nous intéresse, et cela peut nous intéresser dans un autre point de vue, par exemple dans le cas du traitement de la polysémie[106] :

cheval :
1] *animal*
2] *équitation (faire du cheval).*

6.3.2.3 La mise en œuvre technique : tableau noir méréologique et génération de grammaire

Nous nous posons maintenant les trois questions suivantes :
1°) que faire avec notre définition [A] dans le cadre du dictionnaire ?
2°) la définition [A] et la Structure ?
3°) que faire avec notre définition [A] dans le cadre des données créées au fur et à mesure de l'enregistrement dans la Structure des faits d'un texte?

[106] Voir 6.3.3 *Réintroduction du cheval : intégration de la chaîne microsyntaxique et potentialités*, page 122.

6.3.2.3.1 La définition d'une information et le dictionnaire.

Nous remarquons qu'il est très facile d'ajouter dans notre dictionnaire ce petit concept structuré qu'est [A]. Il ne s'agit que de trois éléments pris dans un de nos concepts : le thème (voir chapitre 2.2, page 11).
Pour ces trois positions, dans le cas de la propriété COULEUR, nous disposons de :

- certaines valeurs sous la forme d'une liste d'adjectifs[107]
- une propriété : *couleur*
- une entité : *surface.*

Nous n'avons aucunement à nous intéresser aux chevaux puisqu'ils ne sont pas définis par *couleur*. Evidemment, si nous rencontrions dans le dictionnaire une race de cheval défini par *blanc,* nous serions ravis de noter convenablement, comme nous le faisons le plus souvent possible, cette information du dictionnaire. En relation directe avec ce que nous venons de dire, prévoyant l'apprentissage sur les choses simples du dictionnaire, nous plaçons la définition dans :

- le concept thème, à l'endroit où l'on met ordinairement les mots définis par une couleur (par exemple : *poisson rouge*)
- le générateur de grammaire adapté à la famille d'information concernée.

6.3.2.3.2 Le dictionnaire génère les formes paraphrastiques de l'information utiles à la perception de cette dernière

Malheureusement, dans le discours, l'information ne prend jamais la forme complète [A]. C'est pourquoi, nous devons générer dans le dictionnaire certaines apparences de [A] dans les textes.
Pour le schéma informationnel :

blanc - *couleur* – *surface* (dénommée *cheval*).

nous générons des concepts structurés sous la forme de LCA et correspondant grosso-modo à une grammaire. Nous générons par exemple :

1) couleur de N
2) N de couleur [adj. de couleurs]
3) N blanc
4) N [avoir, possède] couleur [adj. de couleurs]

Le générateur de grammaire produit dans Lexidiom les graphes connexes et reliés entre eux par des hyper-arêtes. Ces hyper-arêtes décrivent la façon dont l'information (les éléments valeur – propriété – entité) seront placés les uns par rapport aux autres (en faisant attention au calcul automatique des LCA) dans la Structure représentant l'unification du texte et du dictionnaire.

Dans ces conditions, nous n'imaginons pas comment il serait possible qu'un texte comme :

Jean a un cheval blanc
Quelle est la couleur du cheval de Jean?

---

[107] qui ressemble à un type énuméré en langage objet; mais qui peut prendre d'autres formes.

puisse passer dans la Structure sans réveiller les instances intéressantes. Pour ce texte, nous avons les éléments suivants :
- *blanc couleur surface_cheval ->instance placée [cheval]*$_{classe}$
- *couleur surface_cheval ->père du précédent et placé dans [cheval]*$_{classe}$
- *quelle couleur cheval ->fils du précédent et placé de plus dans [question]*$_{classe}$
- *quelle couleur cheval -> père de couleur surface_cheval* : il surveille.

puisque tous les phénomènes sont liés entre eux par des relations composés/composantes incluant les tokens et créant des LCA.

La figure suivante représente assez grossièrement les générateurs dans le dictionnaire.

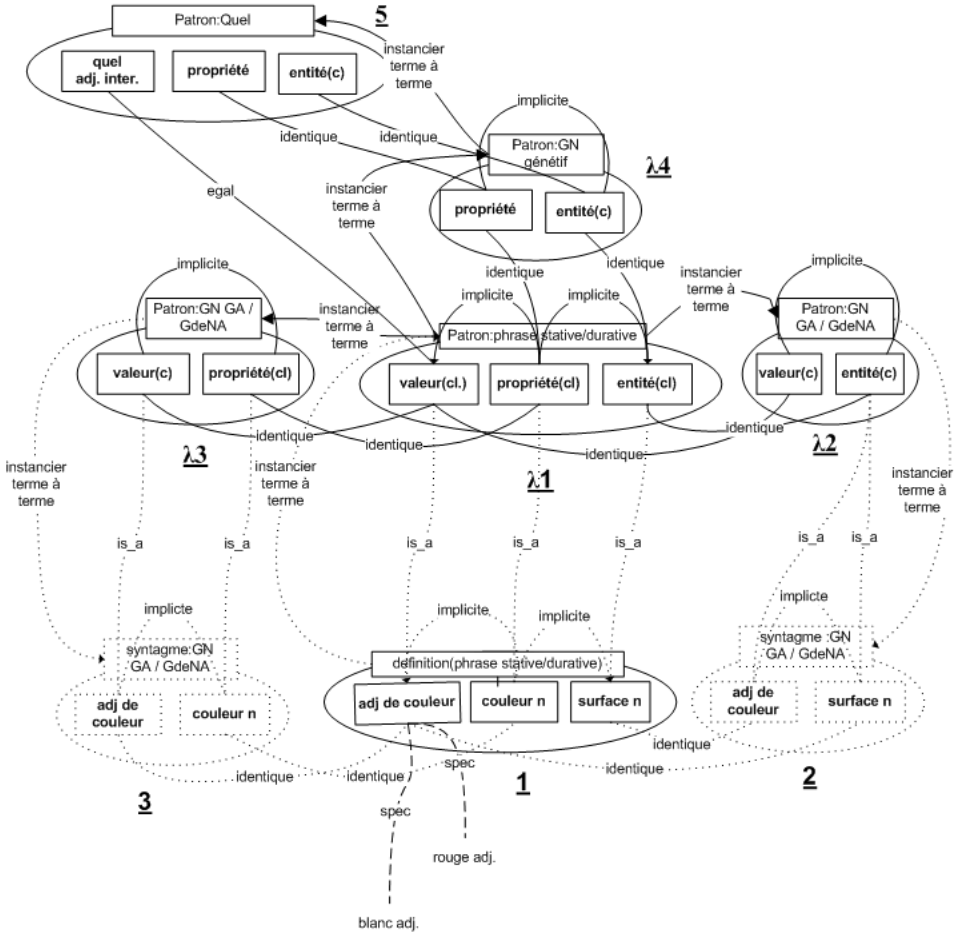


Figure 19 *Quel*, *couleur* et *blanc* dotés d'une organisation supplémentaire (représentation très peu détaillée).

Dans cette figure, le lexicographe s'est contenté de déclarer la définition d'une information, en bas, et de dire qu'il s'agit d'une information.
Le haut est la référence à un automate qui génère les LCA du bas. Il est possible d'utiliser Lexidiom + Sémiographe pour décrire ces automates. C'est ce que nous avons fait dans le cadre du projet : IVOMOB page 48. Mais l'usage est moyennement adapté et l'étude d'une articulation avec des outils comme Nooj [Silberztein 1999] serait certainement bienvenue.
Cette génération qui ne devrait pas comporter trop de combinaisons (nous n'avons que trois éléments en entrée, dont dans ce cas deux constantes) dépend du lexique de la langue et de la nature de la propriété. Par exemple, pour la propriété *poids* de l'entité *corps*, nous aurons des adjectifs comme *lourd* ou *léger* (ces classes existent déjà dans le dictionnaire), et des formes

plus compliquées comme *de XXX kg, d'un grand poids etc. ;* ces dernières formes incluent souvent des expressions régulières qui compliquent la perception.
Enfin, concernant *quel*, le mot est tout en haut. En effet, sa définition le
- fait se déclencher dans les cas suivants
  * + entité : *quel cheval* (!?) ou *quel beau cheval* (!?)
  * + propriété + entité : *quelle est la couleur du cheval ?*
- et le place en position de surveillance[108] des événements :
  * + propriété connue + entité : *cheval d'un beau blanc*

6.3.2.3.3 Le Dictionnaire, les instances et la Structure

L'ensemble des agents du Sémiographe sont intégrés au texte, totalement mêlés à ce dernier pour repérer les événements qui se produisent non pas seulement au niveau des concepts généraux mais également au niveau des instances et dans un maillage concept_instance.
Par exemple, si le tableau noir reçoit un premier mot d'un nouveau texte, comme *Token1-cheval*, il enregistrera :

- il existe *texte*.
- *texte* a pour père : *Token 1*
- *phrase* a pour père : *cheval*
- *Token 1* a pour père : *cheval-Token 1*
- *cheval* a pour père : *cheval-Token 1*
- *cheval* a pour père : *cheval* (lui-même, pour certains problèmes particuliers)
- *cheval* a pour père et est Générique : $[cheval]_{classe}$

etc.

En définitive, dès la réception de "cheval", le Sémiographe actualise toutes les informations du dictionnaire qu'il a à propos de *cheval* (le mot-sens) et à propos *cheval-Token 1* (l'occurrence). Cette actualisation des connaissances dans la Structure est réalisée en-dehors de l'espace des tokens sauf pour le mot_sens *cheval* lui-même. En effet, le Token_cheval ne peut pas encore instancier de phénomène du concept *cheval* du fait que, en tant que mot isolé, *cheval* n'est pas encore $[animal]_{classe}$, $[cheval]_{classe,}$ $[cheval]_{thème,}$ $[équitation]_{synonyme}$ ou autre.
A contrario, dès qu'un LCA impliquant *cheval_mot_sens* sera trouvé, tout le maillage token / concept se créera régulièrement, et produira une sorte de Sémiographe local du *token* : ce Sémiographe local permet de rendre endogène la surveillance dans la structure de toute occurrence de phénomènes futurs qui pourraient se produire le concernant, ou concernant un point de vue à son propos.

6.3.2.3.4 Première conclusion sur la résolution de *cheval blanc*

Nous ne pouvons détailler tous les éléments et tous les points de vue que la résolution de la question du *cheval blanc* implique chez nous. En particulier, il serait tout à fait incohérent de tenter une synthèse d'une présentation de la conjonction des points de vue qui s'est formée dans la Structure. Tout ce que nous pouvons faire se réduit à une énumération d'observations dans le genre de :
- il existe (forcément, puisque nous l'avons décrit) dans le graphe de telle instance X, du point de vue d'une autre instance, prise de telle manière particulière, un certain nombre de LCA.

---

[108] agent fondé sur un but selon la terminologie de Russel [2000]

Par contre, il est possible de limiter cette présentation de la résolution en n'abordant pas l'impact détaillé d'objets que nous n'avons pas encore introduit. Nous faisons l'impasse sur les reconnaissances et effets de certains mots comme, pour notre exemple, *est, la, de, le, Henry* et *IV*. Concernant *le* et *la,* il nous faudrait incorporer la note 73 page 86, qui est assez abstraite. Concernant *Henry IV,* il semble qu'il existe une littérature suffisamment abondante sur les entités nommées. Concernant *de,* nous avons observé une prise en compte fort locale dans la *Figure 18,* page 111 ci-dessus. Nous ne pouvons reprendre toutes ces questions.

Nous considérons maintenant que nous disposons d'un langage élémentaire et d'un dictionnaire intégrant des milliers de lieux qui ne demandent qu'à fournir leur localisation et des milliers de processus qui ne demandent qu'à se déclencher, aboutir et émettre des événements que la Structure écoutera pour se laisser modifier en conséquence, c'est-à-dire pour organiser méréologiquement toutes les conséquences du signal qu'elle reçoit, sous la forme de la définition précise de nouveaux lieux, tant dans l'axe compositionnel strict, c'est-à-dire l'axe des instances, qu'au plan componentiel strict, c'est-à-dire l'axe métalinguistique du dictionnaire, qu'au plan du maillage de ces deux directions qui se réalise en prenant la forme d'une décomposition componentielle de niveau instance de toute instance qui s'est composée à un niveau conceptuel. Cette décomposition nous fournit en quelque sorte des concepts-instances disponibles pour le calcul de LCA propres au co-texte en train d'être lu, pour tous les points de vue analytiques connus par la Structure.

Nous avons aussi décrit les actions et sens de *quelle,* de *couleur* et de *blanc.* Nous comprenons que la description que nous en avons faite aujourd'hui, ne sera jamais reprise demain, sauf pour être améliorée c'est-à-dire sauf pour devenir plus précise et plus respectueuse de ce que veulent dire ces mots. Nous avons observé toutefois que ce sens est considéré globalement, d'une manière holistique, recherchant *in fine* un ancrage dans le phénomène, -l'information-, et non pas dans la réalité ou dans l'abstraction des catégories.
Nous affirmons que la solution décrite pourra :
- réussir d'une façon en quelque sorte irrémédiable l'épreuve des questions Q1 et Q3 (Tableau 5 page 115).
Mais que peut cette solution sur la question Q2 du même tableau? La structure gère-t-elle la coréférence au niveau des LCA des instances?

6.3.2.3.5 Exemple de graphe des instances et des "ontologies" d'instance ; calcul de la question Q2 du Tableau 5 page 115

Nous rappelons ci-dessous Q2 :

> Q2 : Jean(Jean1) a un cheval (cheval1) blanc. Marie a un cheval (cheval2) noir.
> Quelle est la couleur du cheval (cheval3) de jean(Jean2)?

La Structure est créée avec les trois idées suivantes :
- nous savons ce que veulent dire *quelle, couleur* et *blanc :* ils sont définis par une information et s'unifient par défaut avec le dictionnaire componentiel.
- nous n'avons aucune idée de ce que signifient *Jean* et *cheval* puisque dans l'état de nos connaissances rien ne permet de les désambiguïser dans le co-texte. Ils ne s'unifient pas avec le dictionnaire componentiel[109].
- nous développons toutes les conséquences certaines de nos croyances (ou connaissances)

[109] dans l'état supposé de nos connaissances.

selon un ordre partie-tout[110].

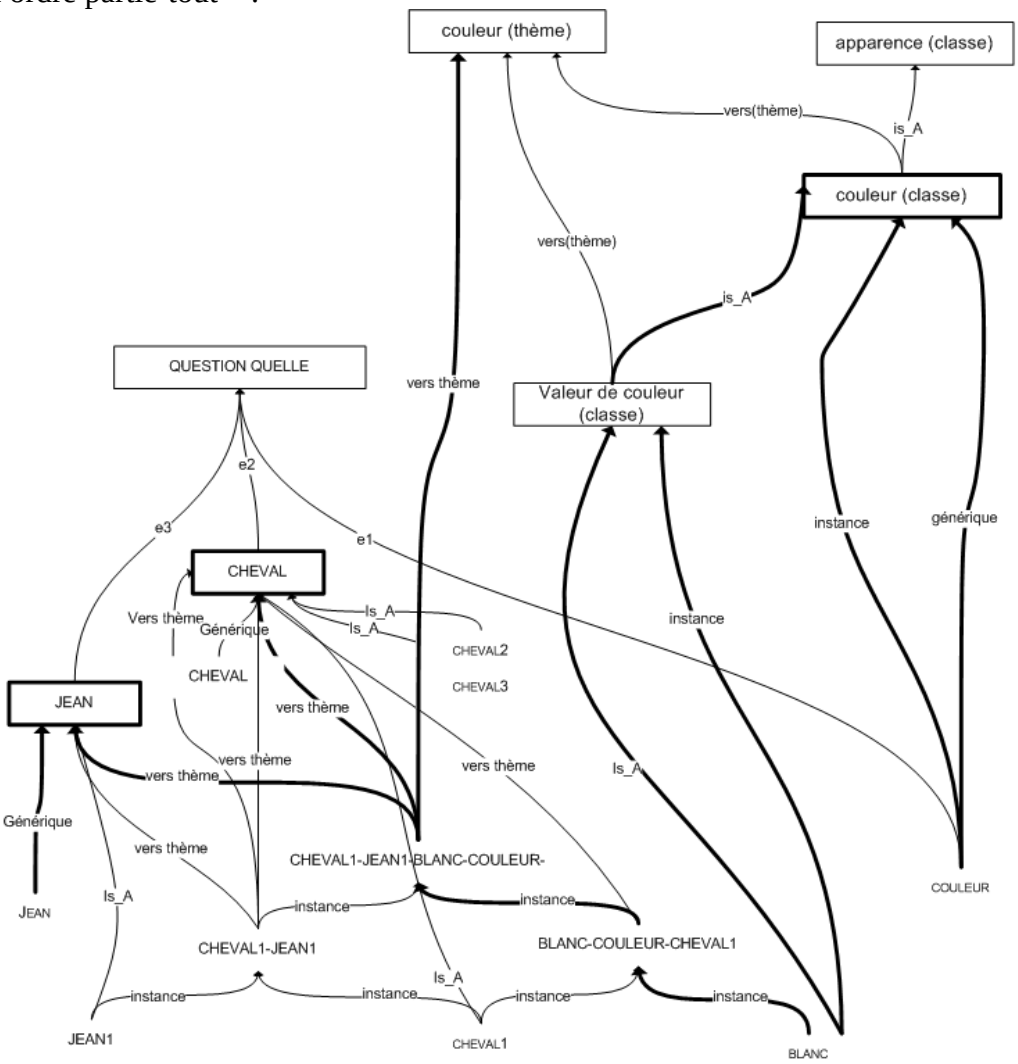


Figure 20 ***Graphe des instances pour la résolution de Q2.***

Cette figure présente un graphe des instances des mots de la phrase dans la Structure qui s'est créée. Dans la figure, nous trouvons des libellés en majuscules. Ces libellés indiquent que nous présentons des instances de tokens de la phrase et non des occurrences de mots dans le dictionnaire.
Ainsi nous lisons JEAN1 qui est une sorte de spécifique du concept d'instance [JEAN] ayant pour générique d'instance JEAN.
Nous avons fait de même pour CHEVAL1 si bien que CHEVAL1 est connu seulement comme spécifique du concept d'instance [CHEVAL] qui a pour générique d'instance CHEVAL.
Par contre *blanc* et *quelle* ont été traités différemment puisqu'ils ont été unifiés avec leur sens dans le dictionnaire. L'unification de *blanc* avec l'information

valeur de couleur – couleur – cheval

[110] C'est toujours dans cet axe que les croyances se manifestent. Comme dit Sartre (note 59, page 72), face à un cube, je ne vois qu'une partie et je crois en un tout.

a produit l'instance

BLANC-COULEUR-CHEVAL1.

Nous faisons l'hypothèse que N1 de N2 donne toujours un point de vue concernant N1 selon N2. Ainsi, après unification, nous avons écrit CHEVAL1-JEAN1
Observant deux instances liées l'une à l'autre (ici par CHEVAL1), nous pouvons créer une instance plus complexe : CHEVAL1-JEAN1-BLANC-COULEUR.
Enfin, nous rattachons toutes ces instances aux concepts d'instance [JEAN] et [CHEVAL] par la relation vers Thème.
Finalement, nous observons que Q2, à l'instar de Q1 et Q3, se résout sans la moindre difficulté : dans tous les mondes possibles de l'interprétation du co-texte et quels que soient les sens de *cheval* et de *Jean*, *blanc* est la meilleure réponse pour une question en dictionnaire à l'envers (voir paragraphe 5.1.2.2, page 55) portant sur le texte : *quelle est la couleur du cheval de Jean?* Nous pouvons aussi dire que cette réponse est l'unique *localisation* possible pour *quelle*.
Avec cette figure, examinons d'autres questions éventuellement intéressantes :

[A] *quelle est la couleur du cheval de Marie?*
[B] *quelle est l'apparence du cheval de Jean ?*
[C] *quelle est la couleur du gentil cheval de Jean?*
[D] *quelle est la couleur de l'équidé/la monture de Jean?*
[E] *quelle est la couleur du cheval de l'homme?*

[A] *Quelle est la couleur du cheval de Marie?*
Si le texte ne comporte aucune information sur ce cheval, *blanc* sortira également mais avec une réponse pénalisée par une non-saturation de la question du côté de Marie (voir la *différence componentielle* dans 4.1.2.2.4, page 41 et *le dictionnaire à l'envers* dans 5.1.2.2, page 55). A l'inverse, si le texte comporte une information sur le cheval (*noir*) de Marie, alors ce cheval arrivera en tête pour la même raison.

[B] *quelle est l'apparence du cheval de Jean ?*
Du fait que *couleur* est considéré comme unifié avec le dictionnaire, le générique *apparence* supposé ici ne modifie en rien le résultat.

[C] *quelle est la couleur du gentil cheval de Jean?*
Blanc également, mais avec une moins saturation de la question (voir [A].

[D] *quelle est la couleur de l'équidé/la monture de Jean?*
Dans cette situation, la structure dynamique construite ne peut rien répondre :
COULEUR + EQUIDE + JEAN ne donne rien du fait que l'unification de l'occurrence de *cheval* du texte avec les connaissances sur *cheval* du dictionnaire n'est pas réalisée.

[E] *quelle est la couleur du cheval de l'homme?*
Même situation que E mais cette fois-ci pour *Jean :* l'occurrence *Jean* du texte n'est pas encore unifiée avec l'occurrence *Jean* du dictionnaire.

[F] *quel est le cheval de Jean?*
La requête devient CHEVAL_générique d'instance ET Jean, et la solution est double : la solution retourne également CHEVAL1 et CHEVAL3.

Pouvions-nous prévoir ces trois échecs ? Il semble que oui puisque dans cette réponse partielle que constitue la *Figure 20*, nous avons porté atteinte à :
- C1 l'unité de la Structure : perte de la référence dans l'axe componentiel
- C2 l'unité de l'Analyse : l'analyse n'a pas été faite dans l'axe componentiel.
- C3 l'unité du Résultat : *cheval1* et *cheval3* sont ambigus
- C4: l'unité du Signe : la fonction référentielle des noms est abandonnée, les génériques sont perdus etc.

6.3.2.4 Une résolution incluant la gestion de la coréférence

Avant de tenter de résoudre la coréférence, il nous faut maintenant insister sur un point important :
- il n'était pas nécessaire pour résoudre Q2, [A], [B] et [C] de résoudre la coréférence. La résolution de la coréférence n'est importante que pour la résolution des questions [D], [E] et [F] qui ne sont pas dans notre corpus.

Avant de quitter cette question finalement assez générale de *couleur de cheval blanc*, nous faisons *comme si* le problème de la coréférence était résolu, par exemple au moyen de connaissances du genre :

cheval ET couleur --> cheval_animal
cheval ET blanc --> cheval_animal

La figure suivante donne alors la Structure où nous trouvons les manifestations sur l'axe componentiel :

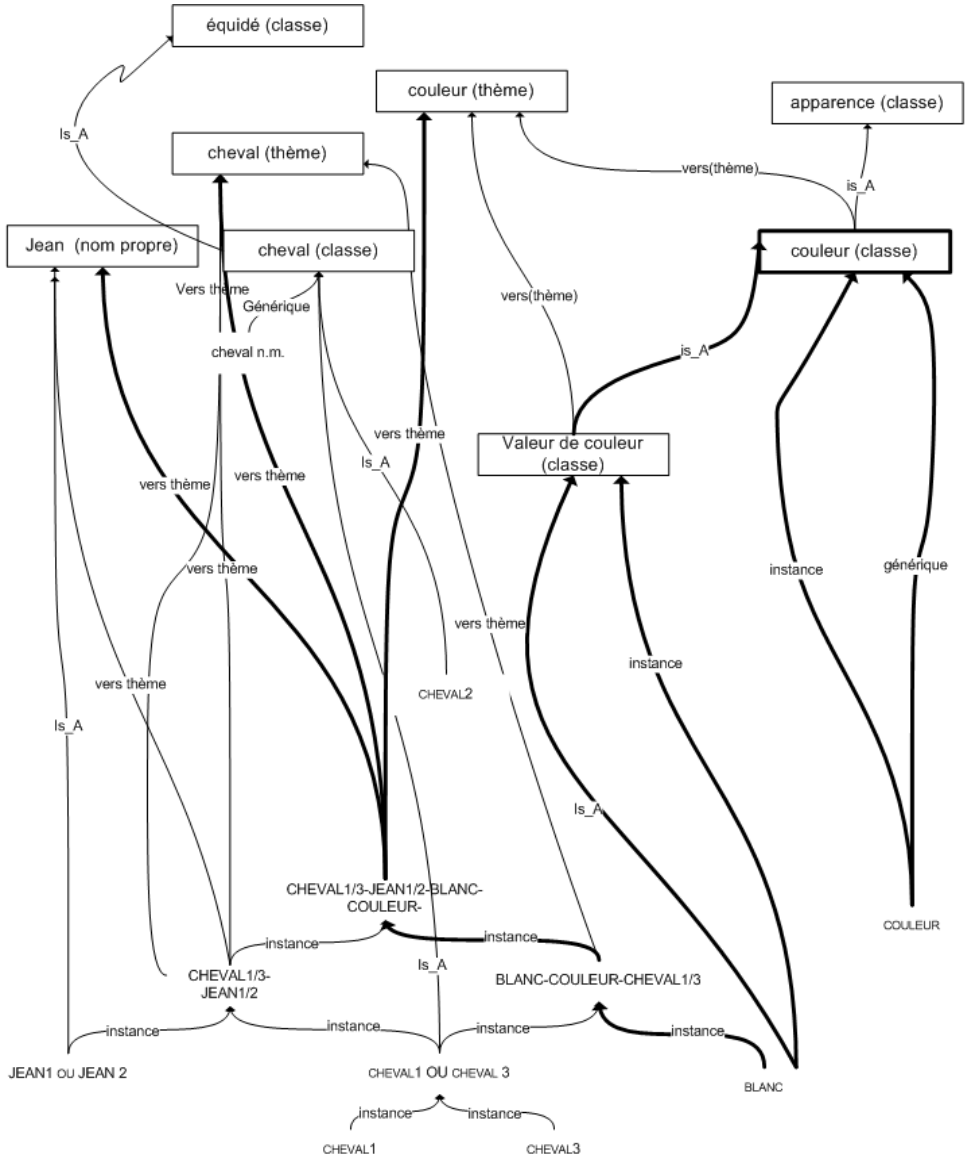


Figure 21 ***Graphe des instances réifiées dans l'axe componentiel pour la résolution de Q2.***

Dans cette figure, les formes sont voisines de celles de la *Figure 20*. En fait, la figure comporte certainement derrière la *Figure 20* : pourquoi supprimer les liens créés alors qu'ils ne sont aucunement remis en cause? Dans cette figure, les fonctionnalités sont étendues et la prise de risque est rendue minimale du fait de l'effet d'accumulation des relations et donc des LCA : nous sommes toujours certains de rester au plus près des questions qui seraient posées.

6.3.2.5 Conclusion sur la résolution

Nous avons effectué des variations sur la question du *cheval blanc*. Nous avons remarqué à quel point il est intéressant de rester près des phénomènes attestés du texte et de ne remonter aux catégories comme $[\text{cheval}]_{\text{classe}}$ que progressivement et en extension des liens attestés de plus bas niveaux : ainsi, nous préservons les relations d'ordre des LCA.
Nous avons aussi observé que les résolutions sont assez simples et sont de toutes manières accessibles à un ordinateur. En particulier, il est certain que les questions Q1 à Q3 (Tableau 5 page 115) ne posent pas vraiment de difficultés.

En rétablissant les unités, à travers un postulat de résolution de la référence, nous avons aussi trouvé des solutions dans des cas un peu plus complexes.
L'ensemble des calculs que nous avons faits ont finalement réussi dans la mesure où ils ont respecté :
- C1 l'unité de la Structure
- C2 l'unité de l'Analyse
- C3 l'unité du Résultat
- C4: l'unité du Signe.

Nous comprenons que cette condition est importante si nous souhaitons explorer de nouvelles épreuves pour notre Sémiographe. Ces épreuves sont :
* le suivi de la coréférence
* l'extraction d'information
* la question-réponse.

Avant de conclure sur ces perspectives, nous proposons de nous demander dans quelle mesure *cheval* pourrait finalement supporter *couleur* ou *blanc*.

6.3.3 Réintroduction du cheval : intégration de la chaîne microsyntaxique et potentialités

Nous savons le cheval que nous avons laissé dans le paragraphe précédent particulièrement désincarné. Nous nous sommes dits : *pourquoi en faire un équidé-mammifère-animal-monture* alors qu'il n'est rien de tout cela dans ce que l'énoncé nous commande de percevoir pour résoudre ses questions. Au fond, le cheval que la Structure a produit, du fait de l'énoncé, est un immuable qui pourra s'adapter à tout contexte où il servirait simplement de référence. Il est Signe conçu comme pur signifiant, récepteur unique de la référence, et prend dans le monde de la Structure le statut que prend la planète Vénus dans le monde de la Matière [Kripke, 1972]. De même que l'Etoile du Matin et l'Etoile du Soir sont les mêmes dans tous les mondes possibles, il est immuable dans tous les mondes où il construit la référence. Comme le nom propre est le désignateur rigide du phénomène, *cheval*, dans ce cas, est désignateur rigide pour toutes les mondes possibles le concernant.
Nous notons alors une idée essentielle. Cette idée est qu'il existe une juste mesure de l'interprétation qui se limite aux frontières des nécessités interprétatives. Il ne s'agit nullement ici de récompenser la paresse. Il s'agit juste d'inviter à éviter l'erreur fatale de celui qui surrinterprète, surtout s'il fait cela sans s'en rendre compte. En quelque sorte, Korzybski [1933] a pour thèse principale l'idée que la surrinterprétation et la scolastique aristotélicienne[111]-la catégorisation- sont les parents de tous les malentendus. Dans nos modèles d'analyse des langues, la surrinterprétation est le risque récurrent. Peut-être avons-nous si peur qu'un système manque d'information que nous le gavons de catégories au risque de perdre le sens du mot et de devoir discuter sans espérance de solution sur des conjonctures

[111] la scolastique aristotélicienne et non l'aristotélisme : nous aurions pu développer tout le chapitre 6 en utilisant les trois distinctions cardinales d'Aristote : essence-accident/acte-puissance/forme-matière. Il y aurait eu alors transposition du propos laquelle n'aurait pas été un gage de simplification,. Mais le point de vue d'Aristote, en particulier la lutte contre les sophistes et les paradoxes est certainement voisin du nôtre : si nous pouvons dire n'importe quoi, nous ne pouvons pas penser n'importe quoi. Pour scolastique, nous reprenons le mot de Claude Bernard : *La scolastique veut toujours un point de départ fixe et indubitable* […] *elle l'emprunte à une source irrationnelle quelconque, telle qu'une révélation, une tradition* (citation empruntée au Robert). Les scolastiques n'ont retenu d'Aristote que les catégories, qui permettent sans contrainte ni système, de poser toutes les convictions que l'on voudra. Ils n'ont retenu que la sophistique.

de paradoxes que nous avons créés de toute pièce. *Cheval* est *animé.* Ah? Et quid de *cheval* est *mort.* Comment se construit la classe ? Comment se transforme la classe? Et comment change-t-on de classe?
Nous espérons continuer à contribuer à cette réflexion pour la langue et non, dans notre cas pour la logique modale. D'une façon générale, en cloisonnant les lieux comme cela se doit, c'est-à-dire en laissant des frontières partout de telles manières qu'elles ne demandent qu'à être franchies, et en articulant les points de vue congruents entre eux au moyen d'une agrégation méréologique de faits, nous ne sommes aucunement en logique. Nous sommes même dans une sorte d'inverse de la logique où, au lieu de partir de prémisses exogènes pour tirer des conclusions parfois compliquées, nous élaborons au fil de l'eau les postulats du texte, et nous tentons de résoudre seulement à la condition que le texte le postule un problème qui a de toute façon été conçu. Nous savons qu'il y a des régions pour les dénominations, les génériques, les hyperonymes, les métonymes, les parties componentielles, les parties compositionnelles et toutes sortes d'autres régions qui sont également des repères conçus par le Signe en action. Et nous ne sommes intéressé que par ces régions-là.
Voilà maintenant que *cheval* devrait supporter *couleur*. Nous ne parlons pas de *bai* ou d'*alezan* qui supporte une couleur particulière et qui les définit. Nous parlons de *cheval.* Et comment alors *cheval* peut-il avoir *couleur* qui ne le définit pas ?
Nous savons que nous avons besoin de ce lien, du fait par exemple des questions D et E page 123 non résolues à cause d'une inconsistance componentielle de la Structure que nous avons voulu pour dégager, dans ce texte, l'existence d'un lieu spécifique à savoir celui d'une information bien formée. Mais tout de même, nous pourrions observer que dès l'occurrence de *monture* dans *quelle est la monture de Jean?* la Structure réévalue automatiquement le statut de *cheval* par la prise en compte d'une nouvelle détermination. Alors, dans quel cas le besoin d'un *sens* de *cheval* se ferait vraiment sentir?
Nous n'avons pas d'exemple clair. Peut-être, après tout, cela pourrait être un soulagement pour nous de savoir que la Structure aura désambiguïsé sur une étiquette de sens et éliminé du fait de *blanc* le sens *cheval_équitation.*
Nous proposons finalement de dire la chose suivante :
- puisque *blanc* asserte *surface*, quels sens de *cheval* supportent *extériorité d'un corps matériel* comme possession.
Cette question revient juste à créer un nouveau lieu, qui fera apparaître un nouvel LCA et de nouvelles dissymétries potentielles, à l'interface des axes componentiels et informationnels.
Ce lieu élimine *équitation* et ce n'est peut-être pas si mal. A ce jeu de l'effeuillage, il ne resterait plus que *viande_cheval* et *cheval_animal,* si cela importe vraiment. Pour *cheval* avoir *couleur,* c'est seulement permettre de réaliser *couleur* ; cette solution partant des faits est comme celle que nous avons donnée pour *bras,* qui, en se mariant à *grue* ne sert, en-dehors de l'idiome et de la chose, qu'à réaliser une partie de son programme *chose allongée*. *Grue* n'y peut rien, et s'en moque. Si nous disions de ce travail qu'il est une *grue*, nous ne devrions pas entendre qu'il a un *bras* mais nous devrions entendre autre chose.
La relation que nous posons est accessoire et accidentelle du point de vue de *cheval* et absolument obligatoire du point de vue de *couleur*. Pour *couleur,* dans le vocabulaire d'Aristote, nous avons une puissance active. Pour *cheval, ,* dans le vocabulaire d'Aristote, nous avons une puissance passive. Cette formulation ne porte pas atteinte à la sémasiologie de *cheval* : elle la reflète exactement. D'un *autre* point de vue, la relation fournit un moyen d'assurer, du fait de son émission dans l'axe componentiel (puisque *cheval_blanc,* alors *cheval_blanc* dans l'axe componentiel), une meilleure continuité de la Structure dans les axes congruents information et signification que nous avons définis. En tant que telle, elle aboutit à nouveau aux régions stables et platoniques de l'axe componentiel. Elle n'est plus alors limitée au seul phénomène capable de résoudre Q1, Q2 et Q3 mais retrouve l'intelligibilité nouménale

capable de résoudre tout ce qui concerne et est concerné par un *cheval* qui se construit et qui prend forme. Les choses ne sont des choses connues que par abduction sur leur totalité[112].
Nous présentons ci-dessous la forme que prend cette relation telle qu'elle ne porte pas trop atteinte à la sémasiologie de *cheval.*
Le dictionnaire définit:

- corps : partie matérielle des êtres animés (sens : doué de vie)
- volume : partie de l'espace qu'occupe un corps
- surface : partie extérieure d'un volume, qui le limite en tous sens.
- couleur : caractère de la surface d'un objet qui ...

La figure suivante présente le résultat et une intégration avec les classes componentielles existantes. Une relation *peut avoir,* en bas, manifeste en soi un changement de point de vue : il s'agit du point de vue imposé par *blanc,* par exemple, sur *cheval.* Cette relation non-inhérente à *cheval* ne vaut que tant qu'elle est sollicitée par un contexte, par exemple *blanc,* pour valider l'émission d'une information dans la Structure. Du côté de cheval, il s'agit juste de dire que ce changement de point de vue est toléré par *cheval*. C'est tout ce que nous pouvons. C'est juste ce que nous voulons.

[112] Revoir le mot de Sartre, note 59, page 72.

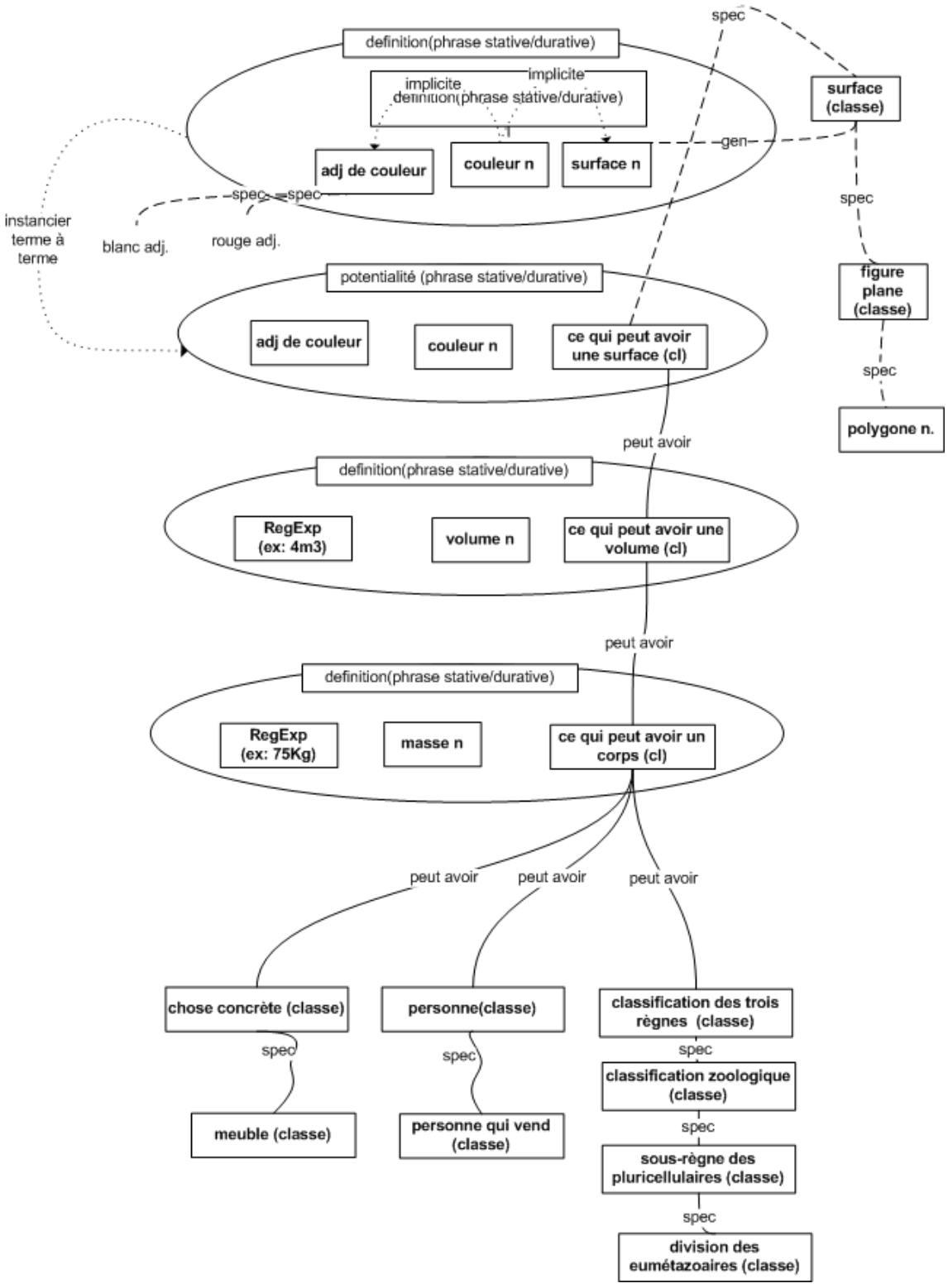


Figure 22 *Landgrave*, *samouraï*, *vache* et *cheval* comme *corps*, *volume* ou *surface*

Dans la *Figure 22*, pour *cheval*, POUVOIR AVOIR *corps*[113], c'est nécessairement POUVOIR AVOIR

[113] Dans la figure *corps* est considéré comme *tout objet matériel caractérisé par ses propriétés physiques*.(Le

volume ou POUVOIR AVOIR *surface*
Dans un autre point de vue de cette figure, nous pourrions montrer l'aspect définitoire : nous aurions *corps* A *volume* et *volume* A *surface*. De même, toujours dans une autre figure, nous aurions :

corps, **générique de tout** corps
volume, **générique de tout** *volume*
surface, **générique de toute** *surface.*

Encore, dans un autre point de vue, nous trouverions des données comme :

œil voir surface
appareil photo photographier surface
personne_qui_mesure mesurer surface,
surface avoir aire

et toutes sortes d'autres choses aux natures assez approximatives mais prévues dans le Dictionnaire de Langue.
Etant donnée la nature de ces choses qui nous intéressent et le contenu actuel de Lexidiom (26 langues pour ce vocabulaire courant), il s'agit au fond du développement d'une ressource pour des calculs sémantiques compositionnels et componentiels en environnement multilingue.
Cette ressource n'est pas idéologique et n'affirme l'Être ou le Phénomène qu'en cela qu'il est lexical. Pour le reste, elle ne construit que dans la mesure où elle sait conserver les unités que nous avons proposées :
- L'unité de la Structure qui assure que toute nécessité lexicale d'origine morphologique, grammaticale, sémasiologique, informationnelle ou du métadiscours puisse s'exprimer à travers un impact sur un autre ou sur le même point de vue
- L'unité de l'Analyse qui assure que les conséquences méréologiques de chaque origine et de chaque origine sur toute origine sont produites
- L'unité du Résultat qui assure que toute décision incorpore, selon une fonction d'utilité basée sur la durée[114], tous les résultats susceptibles de s'être produits lors de l'Analyse
- L'unité du Signe qui assure la consistance élémentaire de l'ensemble.

6.4 Conclusion

Dans ce chapitre, nous avons fait l'hypothèse qu'en plus que les mots entretiennent entre eux des relations lexico-sémantiques qui engendrent d'autres mots, ils peuvent être conçus comme Signes qui engendrent des actions. En opérant ainsi, nous sommes passés franchement de l'espace du dictionnaire à celui du discours.
Nous nous sommes alors posé une double question. La première question a concerné la nécessité de prendre en compte ces actions. La deuxième question a concerné la possibilité pratique de prise en compte de ces actions. Nous nous sommes alors donné six cas d'intégration tout-à-fait courants :
- Intégration de la morphologie compositionnelle
- Intégration d'énoncés compositionnels métalinguistiques
- Intégration d'une grammaire syntagmatique

Robert) ; il s'applique donc à *meuble*. Cela n'empêche évidemment *cheval* de disposer aussi de *corps : partie matérielle des êtres animés* .(Le Robert).
[114] Voir Bergson note 2 page 8.

- Intégration du terme
- Intégration d'une date
- Intégration d'une formule

L'examen de ces cas a toujours révélé que l'intégration ne devient possible que si nous conservons l'intégrité du Signe et que si nous leur permettons d'êtres actifs.

A travers ces six cas, nous avons conclu sur l'impérieuse nécessité et l'apparente faisabilité pratique de cette prise en compte des actions. Dans le même temps, nous avons mis au point une méthode d'analyse des problèmes d'intégration posés. Cette méthode indique qu'il faut regarder la Structure qui se construit, le mode analytique de cette construction, la nature du résultat au plan qualitatif et le statut du Signe dans le discours. Mais une méthode n'est pas seulement une liste de regards permettant d'aborder un sujet. Une méthode, normalement, inclut des points de contrôle que nous devons examiner pour pouvoir progresser. Nous avons défini ces points de contrôle en disant que toute atteinte à l'Unicité de la Structure, de l'Analyse, du Résultat ou du Signe comprend d'une façon endogène une source d'échecs ou de contradictions dans le développement de tout chemin que nous pourrions suivre depuis cette atteinte.

Munis de cette méthode, nous nous sommes alors consacré à l'analyse du plus petit des cas que nous pouvons imaginer où nous devrions bien voir que les Signes actent, et de quelle manière ils le font. Ce cas élémentaire est celui du *cheval blanc d'Henry IV*. Nous avons alors introduit la notion d'information. Cette notion d'information était sous-entendue dans les six cas précédents. Mais à ce moment, il n'y avait pas encore vraiment besoin de la désigner par le mot *information*. Cela pouvait passer inaperçu. Avec le *cheval blanc*, nous nous sommes d'emblée retrouvé au pied du mur. C'est la vertu de cet exemple. En prenant le *cheval blanc*, nous avons regardé une manière d'état de l'art de ceux qui se consacrent au titre principal de leur activité professionnelle à ce genre de résolution. Nous avons resitué aussitôt notre travail en disant que le but qu'ils cherchent est tout-à-fait annexe dans notre recherche qui se consacre uniquement à ce que veulent dire les Signes. En adoptant ce point de vue des *mots* qui sont des Signes qui signifient, nous avons alors immédiatement trouvé une solution à notre problème. Peu importe que cette solution dépasse ou non, actuellement ou potentiellement, l'état de l'art. En soi, c'est la forme prise par la solution du point de vue de la méthode qui a retenu notre attention. En particulier, cette forme en établissant des connexions méréologiques entre perception, dénomination, signification et information est devenue très générale : elle a fini par unifier signe et signification dans une dynamique dont nous ne pourrons connaître un jour la puissance fonctionnelle qu'en réalisant le travail de déclaration des informations[115] du dictionnaire, d'une manière structurale, qui nous reste à accomplir. Nous pensons alors aux fourmis de Langton qui, dessinant la complexité, aboutissent toujours à des formes régulières et nous sommes confiants sur les résultats que notre fourmi qui sait construire une Structure saura faire émerger si nous lui en laissons le temps. Au fond, nous disposons maintenant, certes à traits grossiers, des plans d'une machine qu'il serait intéressant de pouvoir expérimenter. Mais, le problème qui se pose reste la transformation de ce plan en une véritable machine.

En effet, la réalisation d'une telle machine nécessite des moyens qui, sans être énormes, suppose au moins un bon projet ANR. Il s'agit pour obtenir un bon projet ANR de résumer

[115] Nous rappelons que *information* désigne ici *définition d'une information* : il ne s'agit que d'un gabarit et non d'une connaissance actuelle sur le monde qui nous entoure et qui nous importe autant qu'elle importe à la langue, c'est-à-dire qui importe aucunement. Mon expérience du dictionnaire et de la modélisation me laisse l'impression qu'il y a très peu de gabarits d'information de base, l'abondance ne résultant que d'une composition méréologique des gabarits de base.

d'une façon intelligible, en deux fois trois pages (verrou et état de l'art), ce que nous avons dit ici en 130 pages sachant que le verrou est assez global[116] et que l'état de l'art considéré est plutôt pluriel[117]. Nous supposons qu'une Habilitation à Diriger des Recherches pourra nous aider à rendre plus acceptable ces deux fois trois pages qu'il nous faudrait écrire, en fournissant un lien sur le texte d'une HDR soutenue.

[116] la cognoscibilité de la signification?

[117] il va de différentes pratiques à des théories sans pratique actuelle en informatique ou en linguistique théorique.

7 CONCLUSION

L'ensemble de nos travaux porte sur la structure du langage à travers l'observation du dictionnaire qui, chez nous, emporte la compréhension automatique des textes. En observant le caractère pluriel des "méthodes" d'accès au contenu des documents textuels, nous concluons qu'il ne s'agit point précisément de méthode mais plutôt de techniques qui toutes utilisent un point de vue intéressant. Notre travail ne consiste finalement qu'à produire une méthode permettant de réunir ces points de vue.

Après quelques années consacrées à faire une sorte d'analyse de l'existant du contenu du dictionnaire, nous avons pris en entrée cette analyse de l'existant pour essayer de comprendre comment nous l'avons effectuée. Nous avons alors ouvert un nouveau dossier qui nous a amené à définir les principes de la conception elle-même. Pour montrer comment nous en sommes arrivés à ce point, nous avons été obligés dans les cinq premiers chapitres de ce dossier d'habilitation de retracer le parcours effectif que nous avons eu, en soulignant les résultats trouvés et les questions qui se sont posées depuis ces résultats.

Le point commun à tous ces résultats est qu'ils sont toujours issus d'un choix délibéré de travailler sur la langue générale, depuis toutes les "ambiguïtés" possibles, dans le cadre de tous les utilisations atteignables depuis chaque résultat atteint. Ce choix répond à une double motivation : il traduit une certaine croyance du fait que les applications servent d'abord à fournir un éclairage particulier sur l'organisation du dictionnaire et la conviction que la multiplication des tâches réalisables depuis un même processus est une garantie concernant la qualité de l'organisation elle-même sur, essentiellement pour nous, la compétence plus que la performance.

En terminant la présentation de chacune de nos étapes ou de chacune de nos applications, nous avons mis en évidence des limites, des questions à résoudre et des perspectives.

A un moment, vers 2000, le problème des questions prit une forme systémique, chaque question renvoyant sur une autre, et aucune d'elles ne pouvant se résoudre sans que l'autre le soit.

Nous avons donc cherché à répondre d'une façon systémique à un problème systémique, et nous croyons avoir progressé d'une façon remarquable dans la formulation d'une réponse, qui est évidemment elle-même systémique. Mais nous avons observé que cette réponse n'est pas uniquement systémique : elle est aussi naturelle et applicable.

Une indication de succès de cette réponse est l'observation que de nouvelles

applications deviennent désormais accessibles. Il s'agit pour l'essentiel de ce que nous percevons pour le moment du Question-Réponse, de l'Extraction d'Information et du suivi de la coréférence. En effet, toutes ces applications sont devenues maintenant inhérentes à une Structure dont nous ne pourrons plus les soustraire. Le résultat atteint semble certes intriqué et complexe mais nous le préférons à quelque chose qui serait entortillé et rempli de contradictions.

Notre travail sur le Dictionnaire nous a amené à embrasser la plupart des problématiques du traitement automatique des langues. C'est après tout le moins que nous puissions faire puisque nous prétendons que nous produisons une ressource linguistique susceptible de participer à la résolution de problèmes spécifiques et variés de traitement automatique des langues. Cette activité de production de ressource et de système nous a fait rencontrer des spécialistes de la linguistique et de l'informatique, particulièrement en traduction automatique, en intelligence artificielle et en recherche et indexation d'information. Ce mémoire ne reflète pas assez combien ces collaborations pluridisciplinaires sont fructueuses pour la définition d'une méthode et pour le suivi d'un objectif. La confrontation avec des techniques d'apprentissage automatique qui est à peine décrite dans ce document – nous ne pouvons pas tout décrire de ce que nous avons fait – nous permet de gagner du temps sur certaines tâches, comme la fabrication d'un dictionnaire intégral multilingue ou celle d'une grammaire surfacique, et devrait un jour être encore utilisée pour progresser dans la caractérisation des informations contenues dans le dictionnaire, pourvu que nous ayons déclaré préalablement qu'il s'agit effectivement d'une information. Le dialogue avec des spécialistes confrontés à des problèmes concrets nous a aidés à comprendre les verrous qu'ils rencontrent et qui sont endogènes à leur point de vue. Tout cela nous a permis de définir une méthode pour approfondir notre étude sur la structure éventuelle du langage. Nous ne voyons rien pour le moment qui s'oppose à l'approfondissement de cette étude puisque nous n'avons rencontré aucun paradoxe. Nous pensons que la définition deviendra intensionnelle chaque fois qu'elle sera mise à l'épreuve des énoncés. Mais pour le moment nous sommes conscients que nous ne restons qu'à des *prolégomènes au développement d'une pragmatique abstraite naissant de l'observation du dictionnaire de langue*, tout éventuel développement futur nécessitant le développement d'un instrument.

Dans la mesure où cette HDR sera soutenue, nous serions heureux de faire connaître à quelques étudiants intéressés les points de vue de ce texte. Il s'agit d'intéresser quelques jeunes qui seraient exigeants dans la prise en compte de la complexité aux problématiques soulevées dans ce travail le plus tôt possible de telle manière qu'ils aient le plus tôt possible une certaine vision globale des interactions de phénomènes qui se passent assurément et pour lesquelles ils devront le plus souvent choisir entre casser le Signe ou maintenir l'unité du Signe.

Il ne s'agit pas ici d'activité d'enseignement mais d'une activité ponctuelle d'enseignant-chercheur avec des étudiants avant thèse. C'est après tout ce que nous avons fait avec certains de nos stagiaires qui sont maintenant chercheurs ou enseignants-chercheurs. Mais nous n'inviterons jamais un étudiant à prendre de front les problèmes que nous soulevons comme nous le faisons. L'expérience est aussi une affaire de temps et comme seule l'expérience permet de mesurer les risques, si un jour quelqu'un se lance dans le développement d'un moteur méréologique complet, ce sera du seul fait de l'exercice de sa propre liberté.

Enfin, s'il s'agissait un jour à nouveau d'enseigner, il me semble que j'ai toujours pratiqué selon la répartition suivante : 90% de technique et 10% de recul sur la technique. Sans les 10% restants, nous ne ferions que formatage. En tant qu'enseignant, je pourrai ainsi apporter de mon expérience dans les projets, les pratiques, l'entreprise etc.

8 ANNEXE : MULTIPLICATION DES INFERENCES ET RISQUE COMBINATOIRE

L'exposé présente globalement un aspect linguistique. Mais du fait de quelques interrogations techniques que nous avons eues, nous ajoutons un commentaire technico-fonctionnel concernant la gestion d'une grande quantité d'inférences.

Face au *flux* d'information suivant:
3
3 +
3 + *4* (...?)

nous ne pouvons nous empêcher de calculer 7 quand bien même ce 7 pourrait être faux du fait d'un prolongement en * par exemple.
Nous venons de faire une erreur que notre outil de test BabySemio fait également dans le cas général[118]. Voyons maintenant comment cette erreur pourrait être réparée. Cette correction nous permettra de mieux accepter l'idée que

multiplier les points de vue et le meilleur moyen de préciser chaque point de vue.

Nous avons vu au chapitre 6.1, tout un ensemble de cas qui réduisent la combinatoire du seul fait de leur présence en tant que point de vue. Nous avons étudié :

- la Formule (voir paragraphe 6.1.6, page 93), pour laquelle nous avons su attendre la complétude dans la Structure avant d'effectuer le moindre calcul ou de fournir une seule dénomination linguistique.

- les locutions très figées (voir chapitre 6.1.1, page 77) : à un moment, il y a *pomme de terre*

[118] Notre structure exemple chapitre *Figure 17 Définition d'un TOUT précis et balisage.* page 92 ne ferait pas cette "erreur", mais cela n'importe aucunement.

qui s'accomplit et nous savons du fait de l'exemple sur la Formule que nous pourrons marquer ce terme quand nous le souhaiterons. Pourrons-nous cependant détruire quand nous le désirerons les autres sens de *pomme?*
La question du moment renvoie entre autre à la prise de risque et à la gestion de la mémoire, et la question de la manière est technique. Nous répondrons fonctionnellement à ces deux questions dès que nous aurons fini notre énumération.
Voici le restant de cette énumération pris dans ce chapitre :
- le métalangage (voir paragraphe 6.1.2, page 81).
- le syntagme (voir paragraphe 6.1.3, page 84).
- les dates (voir paragraphe 6.1.5, page 88).

A] La question technique

L'ensemble de la Structure comprend le Texte en train d'être analysé et les extraits du Dictionnaire Intégral participant à cette analyse. La Structure est une manière de tableau noir manipulée par deux agents :

1) l'agent de lecture du Texte et de consultation du dictionnaire ; cet agent perçoit un token, cherche tout ce qui a trait à ce Token dans le dictionnaire, le transforme en Token-Signe, c'est-à-dire en une forêt de DAG, et communique son résultat à la Structure qui fait ce qu'elle a à faire avec le signe arrivant et son propre état. La Structure incorpore le Signe et produit une nouvelle Structure.

2) l'agent de nettoyage qui supprime de la Structure tout ce qui ne correspond pas ou plus à certains critères. Les critères sont exogènes, purement applicatifs et peuvent être booléens (atteindre un certain seuil) ou relatifs (être meilleurs que quelque chose d'autre).

B] La question du moment

A quel moment l'agent de nettoyage se met-il à travailler?

L'agent de nettoyage répond à trois événements :
- un paramètre global qui est une fonction aléatoire : cet agent gère le temps et indique à quel rythme la Structure doit être nettoyée
- une exécution sur ordre de la Structure (place mémoire)
- une exécution sur réception d'une émission particulière d'un concept.

Nous avons maintenant une méthode permettant d'augmenter la perception sans conserver trop d'hypothèses non réalisées ou de résultats intermédiaires jugés peu intéressants du point de vue des critères : nous sommes finalement capable de faire disparaître de la Structure tout point qui ne contient pas un grand nombre d'événements créés en réaction du co-texte et selon des points de vue différents. Nous croyons qu'avec une telle approche, plus nos ordinateurs seront puissants, plus nous pourrons abaisser nos critères tout en élargissant encore nos points de vue.

9 PUBLICATIONS ET DISSEMINATION

Une dizaine de formations (40H) aux dictionnaires et Traitements de la langue.
Trois (co-)organisations de journée TAL.
7 participations comme expert à des jurys d'appel d'offres recherche.
12 participations à des comités de lecture.
6 conférences invité.
Environ 60.000 visiteurs/jour d'Alexandria.
Plusieurs travaux de recherche (DEA, thèse) utilisent les ressources ASP d'Alexandria (par ex. enseignement du français à Ottawa) ou le Sémiographe (par ex. dépouillement d'enquête à Grenoble).

22 publications (page suivante)

Conférence Internationale (13)

Conférence Internationale à comité de lecture sur article complet (7)

1. D. Dutoit [1992], *A set theoritic approach to lexical semantics*, Computational Linguistics (CoLing, Nantes)
2. Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, Maria Grigoriadou [2002], *Balkanet: A multilingual Semantic Network for Balkan Languages*, In Proceedings of the First International WordNet Conference, Mysore India.
3. D. Dutoit , T. Poibeau [2002] : *Inferring knowledge from a large semantic network*, full paper, acte de Conference on Computational linguistics, COLING TAIWAN
4. Dutoit D, P. Nugues [2002] *A lexical network and an algorithm to find words from definitions*, acte de European Conference on Artificial Intelligence, ECAI, LYON.
5. D. Dutoit, P. Nugues , P. de Torcy: [May 2003], *The Integral Dictionary : a lexical network based on computational semantics*, Springer Ed., ICCSA International Conference on Computational Science and its Applications, Calgary, Canada
6. D. Dutoit, Y. Picand , P. de Torcy, Roger G. [2003] *Natural Language Processing and Multimedia Browsing, Concrete and Potential Contributions*, European Symposium on Ambient Intelligence - Eindhoven, The Netherlands.
7. F Soufflet, S Le Huitouze, Korpipaa P, D Dutoit, P Ten Hagen, F Kuijik, O Guye, JR Vigouroux, L Chevallier [2003]: *Multimedia browsing*, European Symposium on Ambient Intelligence - Eindhoven, The Netherlands.

Conférence Internationale à comité de lecture sur article complet – Workshop (1)

1. D. Dutoit , T. Poibeau [2002] *Generating extraction patterns from a large semantic network and an untagged corpora*, acte de Conference on Computational linguistics, COLING, TAIWAN.

Conférence Internationale à comité de lecture sur proposition de résumé (5)

1. D. Dutoit [2000] *A text->meaning->text dictionary and process*, acte de Language resource and evaluation, LREC.
2. D. Dutoit , T. Poibeau [2002] *Evaluating resource acquisition tools for information extraction*, full paper, acte de Language resource and evaluation, LREC, Las Palmas
3. D. Dutoit, P. Nugues [2002] *The right word*, full paper, acte de Language resource and evaluation, LREC, Las Palmas
4. D. Dutoit, P. Nugues , P. de Torcy [2004] *The Integral Dictionary: An Ontological Resource for the Semantic Web*, full paper, acte de Language resource and evaluation, LREC, Barcelona

Conférence Internationale Invité (1)

1. D. Dutoit [April 1998], *Linguistique et apprentissage automatique*, 10th european conference on Machine Learning.

Conférence nationale (8+)

Conférence Nationale à comité de lecture (2+)

1. Dutoit D, T. Poibeau [2002]: *Évaluer l'acquisition semi-automatique de classes sémantiques*, acte de TALN.
2. D. Dutoit, P. de Torcy, Y. Picand [2004] *Quelques contenus généraux au service des documents*, 17 pages, Conférence Internationale sur le Document Electronique, La Rochelle (CIDE 7), France.

3. Plusieurs journées Atala

Conférence Nationale Invité (3)

1. D. Dutoit [1993], Le dictionnaire intégral [1999], journée d'études du Centre National d'Etudes pédagogiques
2. D. Dutoit, Le sémiographe [1999], présentation à la journée Outils pour le Tal organisée par le groupe de recherche Information-interaction-intelligence en association avec l'Atala
3. J. François., D. Dutoit, Compte-rendu de *Sémantique et traitement automatique du langage naturel* [2006], de Patrice ENJALBERT (dir.), publié chez Lavoisier / Hermès Science Publications [2005]. Publication au bulletin de la Société de Linguistique de Paris (ILF).

Revue Nationale à comité de lecture sur article complet (3)

1. D. Dutoit [1991], *Dicologique : un nouveau type de dictionnaire*, revue La banque des mots.
2. D. Dutoit, J. François [2007], *Changer et ses synonymes majeurs entre syntaxe et sémantique*, Le classement des verbes français en perspective, Revue Langue Française, édition Larousse, France.
3. T. Poibeau, D. Dutoit [2008], Automatic extraction of paraphrastic phrases from small size corpora, Linguisticae Investigationes.

10 BIBLIOGRAPHIE

A. Abeillé [1993]: *Les nouvelles syntaxes : Grammaires d'unification et analyse du français*, Armand Colin, Paris, Chapitre 3.

A. Abeillé, L. Clément, A. Kinyon [2000]: *Building a treebank for French*, in proceedings First Conference on Linguistic Resource.

A. Abeillé, L. Clément, R. Reyes ²1998]: *Talana annotated corpus: the first results*, in proceedings First Conference on Linguistic Resource.

J.-MI Adam [1999] *Linguistique textuelle*, des genres de discours aux textes, Nathan Université.

J.-M. Adam [1994] *Le texte narratif*, Nathan Université.

J.-M. Adam [1997] *Les textes ; types et prototypes*, Nathan Université.

E. Agirre, G. Rigau [1996] *An experiment on Word Sense Disambiguisation of the Brown Corpus using WordNet*, MCCS-96-291.

D. E. Appelt, J. R. Hobbs, J. Bear, D. Israel, M. Tyson [1993] *Fastus : A finite-state Processor for information extraction from Real-world Text*, In proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI), Chambéry.

Aristote [1969] *Organon, Les catégories*, Ed. J. Tricot.

S. Auroux [1991] *La philosophie linguistique d'Antoine Culioli*, in La théorie d'Antoine Culioli, Ouvertures et Incidences, Ophrys.

M. Avanzi, A. Lacheret-Dujour [2007], *Micro-syntaxe, macro-syntaxe : une prosodie toujours transparente ? L'exemple des périodes asyndétiques en français parlé*, http://www2.unine.ch/webdav/site/structuration_periodes/shared/articles_AM/AM_2007_AL-parataxe.pdf
Baker, F. Collin, C. J. Fillmore, J. B. Lowe [1998]: *The Berkeley FrameNet project*. In Proceedings of the COLING-ACL, Montreal, Canada (disponible sur http://www.icsi.berkeley.edu/~framenet/)

R. Barthes [1964] *Éléments de sémiologie*, Éd. Gonthier.

R. Barthes [1970] *L'Empire des signes*, Éd. Skira.

H. Béhar, M. B. [1995] *La nébuleuse des sentiments,* in L'analyse thématique des données textuelles, L'exemple des sentiments; sous la direction de F. Rastier, Collection "Etudes de sémantique lexicale", Didier Erudition, 1995, Paris.

G. Benoît [1991] *Formalisation dynamiques des relations prédicatives*, in La théorie d'Antoine Culioli, Ouvertures et Incidences, Ophrys.

H. Bergson [1907] *la pensée et le mouvant,* 15ème édition PUF, collection Quadrige Grands textes.

J. Bernhardt [1972] *Aristote,* in La philosophie, *sous la direction de François Chatelet,* Marabout Histoire, réédition 1979.

P. Beust [1998] *Contribution à un modèle interactionniste du sens*, Thèse de l'Université de Caen.

W.J. Black, J. McNaught, G.P. Zarri, A. Persidis, A. Brasher, L. Gilardoni, E. Bertino, G. Semeraro, P. Leo [2000], *A semi-automatic system for conceptual annotation, its application to resource construction and evaluation*, Second International Conference on Language Resources and Evaluation LREC.

L. Bossard [1994] *Conception et développement d'un analyseur morpho-syntaxique*. Mémoire de DEA, encadrement D. Dutoit, Université de Caen.

D. Bourigault [1994] *Lexter, Un logiciel d'extraction de terminologies, Application à l'acquisition des connaissances à partir des textes*, Thèse EHESS.

D. Bourigault, B. Habert [1998] *Evalutation of terminology extractors : principles and experiments, Proceedings of the* First International Conference on Language Resources & Evaluation (LREC).

T. Brants, S. Skut, H. Uskoreit [1999] *Syntactic annotation of a german newspaper corpus*. In Treebank Workshop, Paris, Atala.

Bresnan et Kaplan 1981, *Lexical functional grammar ; a formal system for grammatical representation, The mental representation of grammatical relations*, MIT Press, Cambridge, Massachussetts.

E. Brill [1992] *A simple rule based part of speech tagger,* Third Conference on Applied Natural Language Processing, pages 152-155, Trente, Italie.

E. Brill. [1995] *Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging*. Computational Linguistics, 21[4]:543--565.

É. Brunet, *Peut-on mesurer la distance entre deux textes ?*, Corpus, Numéro 2, *La distance intertextuelle* - décembre 2003, mis en ligne le 15 décembre 2004.
lien : URL : http://corpus.revues.org/document30.html. Consulté le 03 septembre 2008.

P. Cadiot et F. Nemo [1997] *Propriétés extrinsèques en sémantique lexicale*, French Language Studies 7.

L. J. Cahill, G. Gazdar [1999] *The polylex architecture : multlingual lexicons for related languages*, T.A.L., Traitement automatique du langage, volume 40, numéro 2, pp. 3-25.

N. Calzolari [1998] *An overview of written language Ressources in Europe : a few Reflexions,* Facts and a Vision, *Proceedings of the* First International Conference on Language Resources & Evaluation (LREC).

J. P. Caput [1969] *Dictionnaire des verbes français*, Librairie Larousse.

N. Catach [1984] *La phonétisation automatique du français, Les ambiguïtés de la langue écrite*, Édition du CNRS.

M. Chambreuil, A. Ben Gharbia, P. Gamallo Otero, *variations sur la compositionnalité montaguienne*, revue TAL, volume 39, numéro 1.

N. Chomsky [1957] *Syntactic structures*. The Hague, Mouton & co., traduction 1969, Structures syntaxiques (Trad. M. Braudeau, Éditions du Seuil, Paris).

N. Chomsky, G. A. Miller [1971] *L'analyse formelle des langues naturelles* (Trad. Ph. Richard & N. Ruwet, Mouton/Gauthier-Villars)

N. Chomsky [1976] *Le langage et la pensée* (Trad. Louis-J. Calvet, Petite Bibliothèque Payot.)

F. Chatelet [1972] *La philosophie, sous la direction de François Chatelet,* Marabout Histoire, 5 tomes, réédition 1979.

P. Constant [1990] *Analyse syntaxique par couches*, Thèse ENST informatique.

M. Cori, S. David, J. Léon [2002] *Pour un travail épistémologique sur le TAL* (Revue Traitement Automatique des Langues, Vol. 43, N°3.

B. Courtois [1990] *Un système de dictionnaires électroniques pour les mots simples du français*, Langue Française, N°87.

D. N. Christodoulakis [2000] *Design and Development of a Multilingual Balkan WordNet*, http://www.ceid.upatras.gr/Balkanet/.

A. Culioli, J.-P. Desclés, K. Kabore, D.E. Kouloughli [1981] *Systèmes des représentations linguistiques et métalinguistiques* : *Les catégories grammaticales et le problème de la description des langues peu étudiées*, Université de PARIS, Collection ERA 642.

A. Culioli [1990] *Pour une linguistique de l'énonciation.* Opérations et représentations. Tome 1. (OPHRYS)

Collins et Quillian 1969, *Retrieval time from semantic memory*, Journal of verbal learning and verbal memory. 8, 240-247.

C. Copeland, J. Durand, S. Krauwer, B. Maegaard [1991] *The Eurotra Linguistic Specifications, in Studies in Machine Translation and Natural Language Processing*, Commission of the European Communities, 2 vol.

L. Danlos, J. Véronis [1997] *Un demi-siècle de traitement automatique des langues : présentation*, T.A.L., Traitement automatique du langage Volume 38.

L. Danlos [1988] *Les expressions figées*, Langage.

J.-P. Desclés [1982] *Quelques réflexions sur les rapports entre linguistique et mathématiques,* Penser les mathématiques, Seuil, Paris.

L. Dini, V. Di Tomaso, F. Segond [1998] *Word Sense Disambiguation with Functional relations*, Proceedings of the First International Conference on Language Resources & Evaluation (LREC).

F. Droeskeke, M. Hallin, Cl. Lefevre [1987] *Les graphes par l'exemple*, Coll. Ellipses.

J. Dubois, M. Giacomo, L. Guespin, C. Marcellesi, J.-B. Marcellesi, J.-P. Mével [1973] *Dictionnaire de linguistique*, Librairie Larousse.

J. Dubois, Françoise Dubois-Charlier [1990] *Incomparabilité des dictionnaires*, Langue Librairie, N°87.

O. Ducrot, T. Todorov [1972] *Dictionnaire encyclopédique des sciences du langage*, Paris, Le Seuil.

O. Ducrot, J. M. Schaeffer [1995] *Nouveau dictionnaire encyclopédique des sciences du langage*, Paris, Le Seuil.

C.-A. Duhamel & C. Balaz [1993] *Le gros dico des tout petits, 3000 mots racontés pas les enfants, de A comme Avion, à Z comme zèbre,* Ed. Jean-Claude Lattés, Le livre de Poche.

D. Dutoit [1991] *Dicologique : un nouveau dictionnaire de la langue française*, Ed. du CILF, La banque des mots.

D. Dutoit [1992] *A set theoritic approach to lexical semantics,* COLING.

D. Dutoit, K. Laus, Amadeo Cappelli [1993] *Cristal : Conceptual retrieval of information using a semantic dictionary for Access in Three Languages* (Cristal Project), http://www.hltcentral.org/projects/detail.php?acronym=CRISTAL.

D. Dutoit [1995] *Rapport sur l'extraction des fiches signalétiques du MOURRE,* Rapport interne Memodata.

D. Dutoit , T. Poibeau [2002]: *Inferring knowledge from a large semantic network,*, full paper, acte de Conference on Computational linguistics, COLING TAIWAN

D. Dutoit , T. Poibeau [2002]: *Generating extraction patterns from a large semantic network and an untagged corpora*, , acte de Workshop, COLING, TAIWAN.

Dutoit D, P. Nugues [2002]: *A lexical network and an algorithm to find words from definitions*, acte de European Conference on Artificial Intelligence, ECAI, LYON.

Dutoit D, T. Poibeau [2002]: *Évaluer l'acquisition semi-automatique de classes sémantiques*, acte de TALN.

D. Dutoit , T. Poibeau [2002]: *Evaluating resource acquisition tools for information extraction*, full paper, acte de Language resource and evaluation, LREC, Las Palmas

Dutoit D, P. Nugues [2002]: *The right word*, full paper, acte de Language resource and evaluation, LREC, Las Palmas

Dutoit D, P. Nugues , P. de Torcy [2003]: *The Integral Dictionary : a lexical network based on computational semantics*, Springer Ed, ICCSA International Conference on Computational Science and its Applications, Calgary, Canada

Dutoit D, Y. Picand , P. de Torcy, Roger G. [2003]: *Natural Language Processing and Multimedia Browsing, Concrete and Potential Contributions*, European Symposium on Ambient Intelligence, - Eindhoven, The Netherlands.

F Soufflet, S Le Huitouze, Korpipaa P, D Dutoit, P Ten Hagen, F Kuijik, O Guye, JR Vigouroux, L Chevallier [2003]: *Multimedia browser*, European Symposium on Ambient Intelligence, - Eindhoven, The Netherlands.

G. Deleuze [1973] *A quoi reconnaît-on le structuralisme?* In : CHATELET F., Ed, *Histoire de la philosophie, idées, doctrines*, tome VIII, Paris : Hachette Littératures.

Dutoit D, P. Nugues , P. de Torcy [2004]: *The Integral Dictionary: An Ontological Resource for the Semantic Web, ,* full paper, acte de Language resource and evaluation, LREC, Barcelona

D. Dutoit, P. de Torcy, Y. Picand, [2004] *Quelques contenus généraux au service des documents*, 17 pages, CIDE 7 Conférence Internationale sur le Document Electronique, La Rochelle, France.

D. Dutoit, J. François [2007] *Changer et ses synonymes majeurs entre syntaxe et sémantique*, in Le classement des verbes français en perspective, Revue Langue Française, Larousse éd., France.

P. Enjalbert, B. Victorri [1994] *Du langage au modèle*, T.A.L. vol .35, no. 1, pp 37-65,.

P. Enjalbert [1989] *Notes préliminaires à une théorie opérationnelle du sens*, Intellectica., n°8.

J. Euzenat, P. Valtchev [2004] *Similarity-based ontology alignment in OWL* -lite. In Proc. 16th European Conference on Artificial Intelligence (ECAI), Valencia (ES), pp. 333–337.

E. van Loenen, [1998] *The ambience project*, http://www.extra.research.philips.com/euprojects/ambience/

C. Fellbaum [1998] *WordNet : An Electronic Lexical Database*, edited by Christiane Fellbaum, M.I.T. press.

J. Fodor [1981] *Introduction: Some Notes on What Linguistics is About*. In Block, Ned, ed. Readings in the Philosophy of Psychology, Volume 2. Cambridge, Mass.: MIT Press.

C. Fillmore [1968] *Case Grammar.*

C. Fillmore [1976] *Frame Semantics.*

France Telecom [1996] *Cahier des charges du Centre de Langage Naturel de 2° génération*, Annexe technique, 31/12/1996.

T. Fontenelle [2000], *A bilingual electronic dictionary for frame semantics*, Proceedings of second International Conference on Language Resources & Evaluation (LREC).

J.-J. Franckel, Daniel Lebaud [1991] *Lexique et opérations. Le lit de l'arbitraire*, in La théorie d'Antoine Culioli, Ouvertures et incidences, Ophrys.

J. François [2007] *Pour une cartographie de la polysémie verbale*, Société Linguistique de Paris, Peeters, Leuven, Paris.

F. L. G. Frege [1982] sens et dénotation.

C. Fuchs, Laurence Danlos, Anne Lacheret-Dujour, Daniel Luzatti, B. Victorri [1993] *Linguistique et traitements automatique des langues*, Hachette supérieur.

G. et al [1979] *Pragmatics: Implicature, presupposition, and logical form*. New York: Academic

G.et al [1979] *Generalized Phrase Structure Grammar*, Cambridge, MA: Harvard University Press.

Genelex [1993] *Rapport sur la couche sémantique*, rapport utilisateur 1994-1.

Genelex [1994] *Rapport sur le couche morphologique*, rapport utilisateur 1994-2.

Genelex [1994] *Rapport sur le multilinguisme*, rapport utilisateur 1994-3.

Y. Genthillomme [1994] *Panorama sur le Dictionnaire Explicatif et Combinatoire : retombées pédagogiques. In* Dictionnaire Explicatif et Combinatoire du français contemporain *(DEC)*, Recherche Lexico-sémantiques III, Presses de l'Université de Montréal, Québec.

R. Ghiglione, Agnès Landré, Marcel Bromberg, Pierre Molette [1998] *L'analyse automatique des documents*, DUNOD.

B. Godart-Wendling, F. Ildefonse, J.-C. Pariente, I. Rosier [1998] *Penser le principe de compositionnalité : éléments de réflexion historiques et épistémologiques*, T.A.L, volume 39.

G. Grefenstette [1995] *Comparing two Language Identification Schemes*, JADT 1995, 3rd International conference on Statistical Analysis of Textual Data, Rome.

A. J. Greimas [1966] *Sémantique structurale*, Paris, Larousse.

G. Gross [1990] *Définition des noms composés dans un lexique-grammaire*, Langue Française, Larousse.

M. Gross 1975, sur http://sites.univ-provence.fr/delic/lexiques_syntx.html

M. Gross [1990] *Le programme d'extension des lexiques électroniques*, Langue Française, Larousse.

G.Gross, M.Gross, M. Mathieu-Colas, D. Meunier, E. Roche [1991] *Rapport du LADL*, L2/91.

G. Gross, R. Vivès [2001] *La description en termes de classes d'objets et l'enseignement des langues*, Revue Langue Française N°131, Armand Colin.

A. Guiller, C. Leclère [1992] *La structure des phrases simples en français - Constructions transitives locatives*, Librairie Droz, Genève – Paris.

H. Hiz [1964] *The role of paraphrase in Grammar*, Washington, D.C., Georgetown University Press.

I. Prodanof, A. Cappelli, L. Moretti [2000], *Reusability as easy adaptability : a substantial advance in NL technology*, Proceedings of second International Conference on Language Resources & Evaluation (LREC).

N. Journet [2000] *Le langage est-il naturel ?*, Sciences Humaines, Hors-série Le langage, N°27.

O. Jouve [1997] *Manuel d'utilisation de Sampler*, manuel technique, CISI.

M. Kay 1979, *Functional Grammar. Proceedings of Fifth Annual Meeting of the Berkeley Linguistics Society*. Berkeley, CA: U. C. Berkeley.

A. Kilgariff [1998] *SENSEVAL : an exercice in evaluating WSD programs*, First International Conference on Language Resources and Evaluation ELREC.

G. Kleiber [1997] *Sens, référence et existence : que faire de l'extra-linguistique ?*, Langages, n° 127, page 9-37.

G. Kleiber [1994] *Contexte, interprétation et mémoire: approche standard vs approche cognitive*, Langue française 103, p. 9-22.

A. Korzybski [1933] *Science and sanity, an Introduction to Non-Aristotelician Systems and General Semantics. Introduction dans : Une carte n'est pas le territoire prolègomène aux systèmes*, Coll. Premiers secours, L'éclat.

S. Kripke [1970] *La logique des noms propres*, (Les Editions de Minuit, 1982 Trad. Pierre Jacob et F. Recanati).

Z. Harris [1968] *Mathematical Structures of Language*. (Wiley, New York)

C. Laclère [1990] *Organisation du lexique-grammaire des verbes français*, Dictionnaires électroniques du français, Langue Française, Larousse.

E. Laporte [1990] *Le dictionnaire phonémique DELAP*, Langue Française, Larousse.

Douglas B. Lenat [1999] *From 2001 to 2001: Common Sense and the Mind of HAL*, disponible à l'adresse Internet : http://www.cyc.com/halslegacy.html.

Douglas B. Lenat, R. Guha [1990] *Building large knowledge based systems*, Reading, Addison Wesley.

F. LEVY [1994] *Approches sémantiques*, TAL, 35(1-2).

A. Lentin, [1992] *Naissance et premiers pas de l'ATALA : quelques souvenirs et quelques réflexions*. TAL, 33(1-2):7-23.

C. de Loupy, Marc El-Beze, Pierre-F. Marteau [1998] *Word Sense Disambiguation Using HMM tagger*, Proceedings of First International Conference on Language Resources & Evaluation (LREC).

C. de Loupy, Marc El-Beze [2000] *Using few clues can compensate the small amount of resources available for Word Sense Disambiguation*, Proceedings of second International Conference on Language Resources & Evaluation (LREC).

J. Lyons, [1978] *Eléments de sémantique*, Coll. "Langue et langage", Larousse Université, Trad. J. Durand.

M. Marcus, M.-A., B. Marcinkiewicz, Santorini [1993] *Building a large annotated corpus of English : the penn treebank*, Computational Linguistics, 19[2] 313-330.

R. Martin [1983] *Pour une logique du sens*, Paris, Presses Universitaires de France.

Y. Mathet [2000] *Etude de l'expression en langue de l'espace et du déplacement : analyse linguistique, modélisation cognitive et leur représentation informatique*, Université de Caen, thèse de doctorat.

I. Mel'cuk [1986] *Dictionnaire explicatif et combinatoire du français contemporain*, Presses de l'université de Montréal, Québec.

I. Mel'cuk [1992] *Dictionnaire Explicatif et Combinatoire du français contemporain* (DEC), Recherche Lexico-sémantiques III, Presses de l'Université de Montréal, Québec.

I. Mel'cuk & A. Polguère[1995] *Introduction à la lexicologie explicative et combinatoire*, Coll. Champs linguistiques, Ed. Duculot.

I. Mel'cuk, Sylvain Kahane [1999] *Synthèse des phrases à extraction*, T.A.L., Traitement automatique du langage volume 40, numéro 2, pp. 25-85.

G. A. Miller [1998] Foreword, paru dans *An WordNet Electronic Lexical Database*, edited by Christiane Fellbaum.

R. Montague [1970] *The Proper Treatment of Quantification in Ordinary English*, The Journal of Philosophy.

R. Montague [1970] *English as a Formal Language*, The Journal of Philosophy,

J. Piaget [1972] *Épistémologie des sciences de l'homme*. Éd. Gallimard.

MUC-7 [1998] *Proceedings of the Seventh Message Understanding Conference*, http://www.muc.saic.com

Multilex [1993] *Linguistic description of the multilex standard*, Boulogne-Billabncourt, Cap Gemini Innovation.

I. Niles, A. Pease [2001] Towards a Standard Upper Ontology, in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001.

P. Pagès [1992] *Analyse morphologique automatique du français. Extraction des verbes et mise en valeur morpho-sémantique de la dérivation*. Thèse, INALCO, PARIS III.

D. Péchoin (sous la direction de) [1991] *Thésaurus Larousse*, des mots aux idées, des idées aux mots, Éd. Larousse.

J.-M. Pierrel [2000] *Ingénierie des langues*. Hermes Science Europe. ISBN 2-7462-0113-5.

S. Ploux, B. Victorri [1998] *Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes*, T.A.L., Traitement automatique du langage, vol. 39, p 161-182.

T. Poibeau, Adeline Nazarenko [1999] *L'extraction d'information, une nouvelle conception de la compréhension de texte* ? T.A.L., ., Traitement automatique du langage, volume 40, numéro 2, pp. 87-115.

T. Poibeau [2002] *Extraction d'information à base de connaissances hybrides*, thèse de doctorat soutenue le 8 mars 2002 à l'Université de Paris Nord.

H. Poincaré [*1905*] *Les mathématiques et la logique*, Premier article, Article publié dans la Revue de Métaphysique et de Morale, p. 815-835.

H. Poincaré (*1906-I*) : *Les mathématiques et la logique*, Deuxième article, Article publié dans la Revue de Métaphysique et de Morale, p. 17-38.

H. Poincaré (1906-II) : *Les mathématiques et la logique*, Troisième article, Article publié dans la Revue de Métaphysique et de Morale, p. 294-317. (Les articles sont disponibles sur www.ac-nancy-metz.fr/enseign/philo/textesph/ LES_MATHEMATIQUES_ ET_LA_ LOGIQUE.doc)

A. Popescu-Belis [1999] *Évaluation numérique de la résolution de la référence : critiques et proposition*, T.A.L., ., Traitement automatique du langage, volume 40, numéro 2, pp. 117-146.

C. Poirier, Y. Mathet, P. Enjalbert [1998] La compositionnalité à l'épreuve des faits, à travers un projet de compréhension automatique des constats d'accidents.

B. Pottier [1964] Vers une sémantique moderne. Travaux de sémantique et de littérature.

B. Pottier [1992] *Théorie et analyse en linguistique*, Coll. Hachette Supérieur.

James Pustejovsky [1995] *The generative lexicon*, Cambridge, Mass. : MIT press.

Quillian [1968] *Semantic Memory*, in M. Minsky (ed.), Semantic Information Processing, pp 227-270, MIT Press.

F. Rastier [1981] *Le développement du concept d'isotopie*, postface de M. Arrivé, contre-notes de J. Courtès, coll. Actes sémiotiques, Document du Groupe de Recherche Semio-Linguistiques, E.H.E.S.S.-C.N.R.S. Institut National de la Langue Française.

F. Rastier [1987] *Sémantique Interprétative*, coll. Formes sémiotiques, PUF.

F. Rastier [1990] *Signification, sens et référence du mot*, HERMES (Aarhus).

F. Rastier [1991] *Sémantique et recherches cognitives*, Paris, PUF.

F. Rastier, M. Cavazza, A. Abeillé [1994] *Sémantique pour l'analyse,* Paris, MASSON.

F. Rastier [1995] *La sémantique des thèmes ou le voyage sentimental*, parue dans L'analyse thématique des données textuelles, l'exemple des sentiments, Didier Erudition.

P. Resnik [1995] *Disambiguating Noun Groupings with Respect to WordNet Senses*, Proceeding of 3rd Workshop on Very Large Corpora.

P. Resnik, D. Yarowski [1997] *A perspective on word sens disambiguation techniques and their evaluation*, Proceeding of the SIGLEX Workshop « tagging text with lexical semantics : what, why and how ?, pp. 79-86, Washington, D.C.

A. Rey [1976] *Théorie du signe et du sens*, Paris, Klincksieck, 2 volumes.

Stephen D. Richardson, William B. Dolan, Lucy Vanderwende [1998] *Mindnet : acquiring and structuring semantic information from text*, Coling.

F. Rivenc [1989] *Introduction à la logique,* Petite bibliothèque Payot.

R. Rivière, D. Dutoit [1993] *Un phonétiseur automatique du français pour la correction,* Mémoire de Maîtrise, Université de Caen.

S. Russell, P. Norvig [2006] Intelligence artificielle, 1184 pages, 2° édition, Pearson Education, France.

Le Robert [1993] *Dictionnaire alphabétique et analogique de la langue française*, Ed. Le Robert.

G. Sabah [1998] *Le sens dans le traitement automatique des langues,* T.A.L., Traitement automatique du langage, vol. 38, n°2, pp.91-133

G. Sabah [1988] *L'intelligence artificielle et le langage*, vol. 1, Paris, Hermès.

G. Sabah [1989] *L'intelligence artificielle et le langage*, vol. 2, Paris, Hermès.

B. Sagot, D. Fišer [2008] *Construction d'un wordnet libre du français à partir de ressources multilingues*, TALN 2008, Avignon.

P. Saint-Dizier [1999] *Alternations and verb semantic classes for French analysis and class formation*, Predicative forms in Natural Language and in Lexical Knowledge bases, p. 1-52, Kluwer academic publisher, printed in the Netherlands.

P. Saint-Dizier [1999] *An introduction to the lexical semantics of predicative forms*, Predicative forms in Natural Language and in Lexical Knowledge bases, p. 139-170, Kluwer academic publisher, printed in the Netherlands.

R. Schank [1972] *Conceptual dependency : a theory of natural language understanding*, Cognitive psychology, vol. 3, p. 552-631.

R. Schank [1975] *Conceptual Information Processing.* (Elsevier, New York.)

R. Schank, Goldman, Rieger et Riesbeck, 1975, *Inference and Paraphrase by Computer,* Journal of the ACM (JACM).

J. Searle [1980] *Minds, Brains, and Programs. Behavioral and Brain* Sciences 3, notre copie : http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html.

M. Silberztein [1990] *Le dictionnaire électronique des mots composés*, Langue Française, N°87.

M Silberztein [1993] Dictionnaires électroniques et analyse automatique des textes. Paris : Masson.

M. Silberztein [1999]. Traitement des expressions figées avec INTEX. Linguisticae Investigationes (n° spécial « Analyse lexicale et syntaxique : le système Intex). pp. 425—449.

J. F. Sowa [1984] *Conceptual Structures.* Information Processing in Mind and Machine, 1984, Addison Wesley, Reading, MA.

S. Stamou, K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou, [2002] *A multilingual Semantic Network for Balkan Languages*", In Proceedings of the First International *WordNet* Conference, *Mysore* India.

C. Stratulat [1997] *Analyse syntaxique utilisant des motifs*, Rapport de stage de DEA, Université de Caen.

P. F. Strawson [1959] *Individuals: An Essay in Descriptive Metaphysics*. Trad. fr. *Les individus : Un essai de métaphysique descriptive*, Paris, Le Seuil, 1973.

H. Tardieu, A. Rochfeld, R. Coletti [1983] *La méthode Merise : tome 1*, Éditions d'organisation.

L. Tesnière [1959] *Eléments de syntaxe structurale, deuxième édition revue et corrigée*, Paris, Librairie C. Klincksieck, 1966.

L. Timbal-Duclaux [1989] *L'expression écrite,* Les éditions ESF.

J. Tuominen, [2000] *Opening Mobile Platforms for the Development of Component-Based Applications (Vivian),* http://www-nrc.nokia.com/Vivian/

A.E. Van Vogt [1969] *Postface au Monde du Non-A*, Coll. Omnibus, Presses de la Cité.

J. Vergne [1999] *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur*. Analyse syntaxique automatique non obligatoire, Greyc, Université de Caen, mémoire d'habilitation à diriger des recherches.

J. Vergne [1995] *Une syntaxe des langues concrètes*, Greyc, Université de Caen.

J. Véronis, N. Ide [1990] *Word Sense disambiguation with very large neural networks extracted from machine readable dictionaries*, Coling.

J. Véronis [1991] *Fusion de dictionnaires en vue de la création de grandes base de connaissances,* pages 117-130, Avignon 1991, Le traitement du langage naturel et ses applications.

J. Véronis [1995] *MULTEXT : Étiquetage grammatical multilingue.* Modèle lisible à l'adresse : *http://www.Ipl.univ-aix.fr/projects/multext/LEX/LEX2.html*.

B. Victorri, Catherine Fuchs [1996] *La polysémie : construction dynamique du sens*, HERMES.

B. Victorri [1998] *Dynamical construction of meaning : a challenge for Artificial Intelligence*, RFJA'98.

B. Victorri, Catherine Fuchs [1999] *Le sens grammatical*, Langues, LAROUSSE.

P. Vossen [1999] *Final report Deliverable D041, Work Package 0,* EuroWordNet, LE2-4003, LE4-8328.

P. Vossen, Laura Bloksma [1998] *Categories and Classifications in EuroWordNet,* Proceedings of the First International Conference on Language Resources & Evaluation.

J. Chauché, V.Prince, S. Jaillet, M. Teissire [2003] *Classification automatique de textes à partir de leur analyse syntaxico-sémantique*. Proceedings of TALN'2003, Batz-sur-mer. Vol I. Pp 45-55.

Weizenbaum [1966], voir http://i5.nyu.edu/~mm64/x52.9265/january1966.html

Y. Wilks, *Does anyone really still believe this kind of thing?* In K. Sparck Jones and Y. Wilks, editors, Automatic Natural Language Parsing, pages l82-l89, Ellis Horwood Limited, 1983.

Y. Wilks, Brian M. Slator, Louise M. Guthrie [1996] *Electric Words : dictionaries, Computers and Meanings,* ACL-MIT Press series in natural-language processing.

Y. Wilks [1997] *Sense Tagging : Semantic Tagging with a lexicon*, cmp-lg/9705016 .

Y. Wilks, Mark Stevenson [1998] *Word Sense Disambiguation using Optimised Combinations of Knowledges Sources* ?, cmp-lg//9806014 .

Y. Wilks [1999] *Is Word Sense Disambiguation just one more NLP task ?,* CL/990230 .

L; Wittgenstein [1961], Tractatus logico-philosophicus, suivi de Investigations philosophiques, Paris, Gallimard.

Xiaobin Li, Stan Szpakowicz, Stan Matwin [1995] *A WordNet-based Algorithm for Word Sense Disambiguisation*, Proc. of 14th International Joint Conference on Artificial Intelligence, p.1362, p.137

Mickael Zoch [2006], *Capitalisation d'une ressource en or : le dictionnaire,* TALN, Leuven.