Pour une instrumentation informatique du sens

Vincent PERLERIN, Pierre BEUST
GREYC CNRS UMR 6072 – ISLanD & Pôle Modescos de la MRSH
Université de Caen
14032 Caen Cedex, France
{pierre.beust, vincent.perlerin}@info.unicaen.fr

Mots-clefs : informatique, évaluation, système anthropocentré, ressources lexicales

Résumé

En informatique, la question du sens a longtemps été abordée sous l'unique biais du tout automatique. Le développement des travaux pluridisciplinaires au sein des sciences cognitives a permis de remettre en question le point de vue exclusivement calculatoire en affirmant la place prépondérante du sujet humain. Dans cette optique, nous proposons un modèle pour une instrumentation informatique du sens résolument tournée vers l'utilisateur. Ce modèle a donné lieu à différentes réalisations logicielles instrumentalisant des formes d'analyses sémantiques utiles pour des tâches relevant par exemple de la recherche documentaire et utilisant des ressources lexicales légères car non exhaustives.

I. Introduction

Si l'histoire du T.A.L. (Traitement Automatique des Langues) est très récente, ses applications actuelles trouvent leur place dans la vie quotidienne de millions d'utilisateurs à travers le monde. Des correcteurs orthographiques aux logiciels d'aide à la traduction, des moteurs de recherche sur l'Internet aux systèmes d'analyse de la parole, les domaines de recherche où l'informatique doit traiter du matériau linguistique sont très nombreux. Étonnamment, l'approche pluridisciplinaire informatique/linguistique ne semble pas indispensable dans nombre de ces travaux. Dès la fin de la seconde guerre mondiale, les pionniers du T.A.L. (originellement Traduction Automatique des Langues) envisageaient par exemple la traduction comme un simple codage/décodage. Le fonctionnement de la langue dite naturelle n'était pas distingué du fonctionnement d'autres systèmes de signes comme les codes des programmes informatiques ou les équations mathématiques. Si les échecs de ces premiers travaux ont permis des avancées considérables dans certains domaines, une tradition perdure : l'apport des travaux de Linguistique est parfois considéré comme accessoire pour certaines tâches touchant pourtant de très près à la langue. Ainsi, de nombreuses techniques intégrant essentiellement la logique et les mathématiques (comme la DRT[1] ou l'ASL[2] par exemple) parviennent à proposer des représentations des informations contenues dans une phrase ou un groupe de phrases sous forme d'assertions logiques. Ces représentations sont ensuite utilisées par les ordinateurs pour, par exemple, tenter de répondre à des questions dont les réponses se trouvent dans un ensemble de documents ainsi traités. Dans ce type d'approches, très courantes en T.A.L., les quantités de connaissances à manipuler lorsque l'on souhaite traiter des documents tout-venant sont très grandes mais ne peuvent jamais atteindre l'exhaustivité. La description du contenu sémantique d'une phrase ou d'un document par un formalisme donné n'est pas sans problème (Cavazza, 1996) : quelles sont les informations à prendre en compte, quelles sont celles irrémédiablement perdues lors de la transcription … ?

Depuis une trentaine d'années, un bon nombre de recherches sur le sens en Informatique ont profondément changé de nature (Victorri 1998). Les informaticiens ont appris à connaître les autres disciplines des sciences du langage, à travailler avec elles, à se méfier des annonces trop rapides et à envisager les problèmes langagiers dans des recherches à long terme. En retour, les autres disciplines ont trouvé dans l'expérimentation informatique des situations objectives de mise à l'épreuve de leurs modèles. Ces relations entre l'informatique et les autres disciplines au sein des sciences cognitives ont permis d'envisager d'autres voies pour les modèles touchant au matériau linguistique. Des chercheurs tentent en particulier de prendre en compte de la notion d'interprétation, en tant qu'assignation par un agent d'un sens à une unité linguistique (mot, phrase, texte, ensemble de textes…). Cette notion pose entre autre, la question de l'implication des utilisateurs dans les processus automatisés et insiste en cela sur la variabilité du sens. Les pratiques langagières propres à un individu,

[1] La D.R.T. (Théorie de la Représentation du Discours) initiée par H.Kamp (Kamp, 1981), permet de construire une représentation d'un texte à partir de structures (D.R.S.) qui traduisent approximativement les propositions.

[2] L'A.S.L. (Analyse Sémantique Latente) est une méthode d'analyse de textes permettant de trouver des corrélations entre termes au sein d'un texte (Landauer et Dumais, 1997).

la façon dont il parle d'un sujet ou même ses affinités avec un domaine précis, ne sont pas prises en compte par les techniques exposées ci-dessus. Partant de ce constat, nous avons choisi la voie de l'anthropocentrisme dont le principe fondateur est le suivant : « [dans un système anthropocentré] *l'homme n'est plus en charge d'entrer dans un monde informatique quasi-autosuffisant mais c'est à la machine de se construire autour des besoins de l'utilisateur pour mieux l'assister* » (Thlivitis, 1998). Dans ce cadre, nous tentons le plus possible de rendre personnalisables les données utilisées par la machine. L'utilisateur a pour charge de compléter ou réviser des ressources (essentiellement lexicales) nécessaires au fonctionnement des programmes, l'ordinateur a pour but de les utiliser le plus efficacement possible (en fonction du travail d'implémentation des informaticiens) et de renseigner en retour l'utilisateur. Dans nos travaux, nous ne cherchons pas à représenter le sens ou de le synthétiser dans un formalisme. Nous ne cherchons pas à fournir de résultat unique mais de maintenir des conditions d'interaction dans lesquelles le sens est négocié entre l'utilisateur et la machine.

Dans cet article, après avoir exposé les principes généraux de notre modèle, nous nous attarderons sur les instrumentations logicielles qui en sont issues. L'atelier formation CNRS *«Variation, construction et instrumentation du sens »* qui s'est tenu à Tatihou en juillet 2002, nous a donné l'occasion de mener une première expérience sur la question cruciale de l'évaluation de telles instrumentations. Il s'agissait en l'occurrence d'évaluer la capacité d'utilisateurs potentiels à s'approprier des principes de construction différentielle de ressources lexicales. Nous ferons un premier bilan sur les résultats cette expérience.

II. Un modèle pour l'instrumentation du sens

Nous ne parlons ou n'écrivons pas tous de la même façon. Nous ne faisons pas tous une interprétation identique d'un même texte. Nos cultures, nos histoires, nos langues, nos façons de parler des choses qui nous entourent, sont différentes. Pourtant, si le langage et ses emplois recèlent des parts propres à chaque individu, il permet d'informer, de parler de soi, de parler de la langue, d'échanger, de transmettre, d'entendre[3]... Le langage a donc une dimension sociale (partagée) et une dimension propre à chaque être. François Rastier (Rastier et al., 1994) a donné à ce propos une partition des degrés de systématicité de la sémantique unifiée (le système fonctionnel de la langue, les normes sociales et les normes idiolectales). Le système fonctionnel de la langue, qui correspond au dialecte, pose les normes de la langue. Les normes sociales correspondant au sociolecte, découlent de considérations essentiellement culturelles. Enfin, les normes idiolectales liées au locuteur-individu, représentent l'ensemble de ses régularités personnelles ou « normes individuelles » dans ses actes langagiers. Notre problématique de recherche en Informatique concerne la mise au point d'outils d'assistance à l'utilisateur dans des tâches langagières médiatisées par l'informatique qui tiennent compte le plus possible des spécificités propres à cet utilisateur ou au groupe d'utilisateurs[4] à qui il appartient. On entend par tâches langagières, les situations où l'interaction homme-machine demande une prise en compte de la dimension sémiotique d'un matériau linguistique (énoncés, phrases, textes, corpus …). C'est par exemple le cas dans le dialogue homme-machine, le résumé automatique ou encore l'indexation de documents textuels. En ce qui nous concerne, nous cherchons plus spécifiquement à appliquer nos propositions la recherche documentaire sur l'Internet qui constitue une activité langagière à part entière (Perlerin, 2001). Les applications des analyses et des outils que nous proposons peuvent cependant être utiles tout aussi bien à la création d'interfaces de lecture rapide, de logiciels d'aide à l'interprétation ou au classement de documents.

Nos propositions pour l'instrumentation de la dimension sémiotique des productions langagières consistent en un modèle informatique d'analyse de matériau linguistique ayant pour but d'objectiver une interprétation. Ce modèle met l'accent sur la cohésion thématique du matériau linguistique (la présence redondante de termes en rapport avec un domaine donné dans une entité textuelle définie, la façon dont ces termes apparaissent à

[3] Nous renvoyons le lecteur surpris par ce dernier verbe à la lecture (Coursil, 2000).

[4] Il ne s'agit pas ici de considérer tant une communauté linguistique géographiquement située qu'un groupe socialement marqué à l'intérieur duquel des mots ont un emploi commun attesté permettant une intercompréhension optimale. A propos de l'immutabilité et la mutabilité du signe, Saussure écrivait «*… s'il l'on veut démontrer que la loi admise dans une collectivité est une chose que l'on subit, et non une règle librement consentie, c'est bien la langue qui en offre la preuve la plus éclatante.* » (Saussure, 1915 p.104). Mais si le signe échappe à notre volonté, le sens se construit et trouve sa source dans le consensus. C'est ce consensus au sein d'une communauté donnée qui devra servir de ressource pour les analyses effectuées à partir de notre modèle. Ainsi la distinction utilisateur et groupe d'utilisateurs pourra dans certaines configurations être considérablement réduite si ce n'est devenir totalement obsolète. L'expérience décrite à la fin de cet article participe pleinement à notre réflexion sur ce point.

l'intérieur de cette entité…) et sur les représentations lexicales qui permettent de la mettre en évidence (les termes en question). Il s'inscrit dans le courant des approches du sens et de la signification issus des travaux en linguistique de Harris, Rastier ou encore Mel'cuk qui avancent que la construction de lexiques (à des fins de traitements automatiques ou d'analyses « manuelles ») doit être fondée sur une étude des usages des mots attestés dans des corpus (textes, dialogues…). A la différence d'autres travaux en TAL, la distinction entre sens et signification est pour nous primordiale. Le sens se détermine relativement au contexte et à la situation, au sein d'une pratique sociale alors que la signification ou le signifié d'une unité linguistique, est défini en faisant abstraction des contextes ou des situations possibles. Tout signification est ainsi un artéfact. Les ressources lexicales utilisées dans notre modèle, ne sauraient alors provenir que de matériau linguistique « réel », et être attestées en tant que tel par un utilisateur en fonction d'un corpus, d'un document ou encore d'une pratique sociale donnée. Une signification, provenant par exemple d'un dictionnaire ou d'un thesaurus construit *ex nihilo*, ne peut rendre compte que difficilement d'une utilisation précise dans une situation donnée et encore moins d'une pratique langagière idiolectale. Même dans l'approche différentielle et componentielle que nous adoptons[5] (nous reviendrons sur ces principes plus tard), nous ne comptons que sur des lexiques restreints pour aboutir à des résultats utiles à certaines tâches touchant au matériau linguistique et souffrant à l'heure actuelle du manque de prise en compte de l'utilisateur dans les dimensions exposées ci-dessus.

Le modèle que nous proposons cherche à produire des instrumentations qui rendent compte d'un processus *d'aller-retour* entre les usages des signes linguistiques et leurs représentations lexicales, c'est-à-dire une articulation de deux dimensions des productions langagières : l'axe syntagmatique (axe de la chaîne linguistique) et l'axe paradigmatique (axe des représentations lexicales évoquées par la chaîne) (Figure 1).

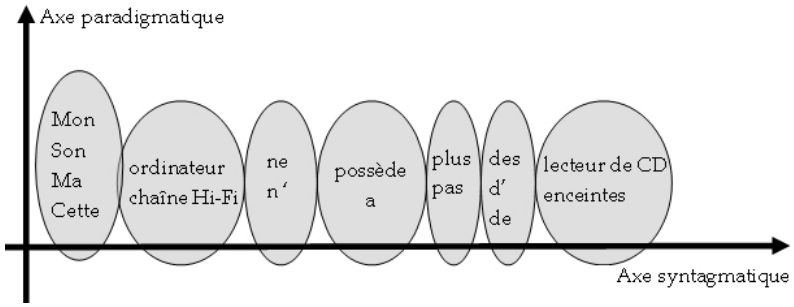


Figure 1. Axe paradigmatique, axe syntagmatique.

Pour pouvoir tenir compte des spécificités des utilisateurs et donner une dimension interprétative à nos analyses, notre modèle s'inscrit dans la voie épistémologique de l'anthropocentrisme. On ne cherche à représenter dans les structures paradigmatiques utilisées que ce que l'utilisateur y juge pertinent. En d'autres termes, nous n'utilisons que les ressources lexicales restreintes aux besoins de l'utilisateur. Ces ressources s'avèrent suffisantes pour dégager la cohérence thématique (ou l'absence de celle-ci) au sein d'un document ou d'un corpus de documents et pour procéder à des analyses utiles pour des tâches relevant par exemple du classement ou du réordonnancement de documents. On évite ainsi le problème insurmontable classiquement posé en TAL, de savoir jusqu'où détailler les représentations lexicales. Il en résulte le principe d'une sémantique légère dans le sens où l'on peut produire des instrumentations qui indiquent avec relativement peu de ressources certains aspects de la dimension sémiotique d'une forme linguistique sans chercher à fournir une représentation du sens qui demanderait des ressources lourdes et nécessairement incomplètes. Notre but n'est pas de considérer la sémiotique des matériaux linguistiques dans leur globalité. Nous mettons en œuvre des analyses dont les résultats suffisent pour les tâches automatiques ciblées. Ces analyses proposées à partir notre modèle tendent à rendre compte de l'articulation syntagmatique/paradigmatique à l'œuvre dans l'interprétation d'une chaîne. Nous appropriant les propositions de la sémantique interprétative de Rastier (Rastier, 1987), nous analysons cette articulation en terme de dynamique sémique en analysant la récurrence d'un même trait sémantique (appelé sème) au sein d'une entité textuelle.

*1. Le **skipper** et son **trimaran** restaient **encalaminés** dans le **pot-au-noir**.*

Dans l'exemple 1 extrait de (Rastier, 1991, p. 220), les mots en gras portent tous le trait sémantique */navigation à la voile/*. On peut noter au passage qu'un mot peut bien sûr porter plusieurs sèmes. Par exemple, on peut

[5] et contrairement à certaines positions sur le sujet (Cavazza, 1996 p.56).

associer à *skipper* les traits /humains/, /navigateur/… Depuis les travaux de Greimas (Greimas, 1983), les récurrences de traits sémantiques dans les textes sont appelées **isotopies**. Elles caractérisent les langues naturelles car on ne les retrouve pas dans d'autres formes d'écriture (formules mathématiques, programmes informatiques …). C'est pour cette raison que l'isotopie est un concept central de notre modèle.

Nous proposons une technique de coloriage thématique qui s'apparente à un surlignage électronique pour mettre en évidence les fortes récurrences de traits sémantiques et donc la présence d'isotopies. Cette technique de visualisation par coloriage thématique exploite l'isotopie en tant que certaines isotopies, appelées isotopies génériques (i.e. une récurrence d'un sème générique), indexent des mots dans différents thèmes. Dans l'exemple 1, l'isotopie générique portée par les mots *skipper*, *trimaran*, *encalaminés* et *pot-au-noir* correspond au thème de la navigation. Le coloriage thématique consiste alors à affecter une couleur à chaque isotopie générique et à surligner électroniquement les mots du texte sur lesquels elles s'appuient. On peut alors examiner avec le coloriage les différentes répartitions des isotopies au long du texte, leurs alternances et leurs enchaînements. De ce point de vue, le coloriage est aussi une méthode pour rendre objectif (et donc partageable) certains aspects fondamentaux des interprétations que l'on peut produire.

Compte tenu des phénomènes d'homonymie et surtout de la polysémie des langues naturelles qui touche massivement les domaines lexicaux courants, certains mots peuvent appartenir à plusieurs thèmes révélant des domaines bien distincts. Ce serait par exemple, le cas du mot *avocat* que l'on pourrait aussi bien affecter à une classe sémantique du champ lexical des aliments qu'à une classe sémantique du vocabulaire juridique. Pour le coloriage, il faut alors déterminer quelle est la couleur à attribuer à un mot rencontré dans un texte si ce mot appartient à plusieurs thèmes. Dans un tel cas, l'heuristique considérée est celle qui tend à prolonger le plus possible les isotopies du texte (c'est-à-dire à favoriser la redondance naturelle des langues). Ainsi, parmi ses couleurs possibles, on attribuera au mot la couleur la plus représentée dans l'unité textuelle considérée (en cas d'égalité du nombre de sème, c'est la dernière couleur utilisée dans l'unité qui sera affectée, l'évaluation de cette technique est cours). Comme on le montre dans les deux exemples suivants, ceci constitue une méthode de désambiguïsation des mots polysémiques.

2. *Au marché, j'ai acheté des poireaux, des concombres et des avocats.*
Où les mots *poireaux*, *concombres* et *avocats* portent par exemple le sème /comestible/.

3. *En sortant du tribunal, j'ai vu mon avocat.*
Où les mots *tribunal* et *avocat* portent par exemple le sème /juridique/.

Une analyse interprétative en terme d'isotopies demande naturellement une représentation componentielle des significations des mots (par la suite on parlera de lexie qui est une notion plus générale que le mot[6]), c'est-à-dire une représentation en terme de sèmes. Le point de vue classique en linguistique est de considérer que le sème est une paraphrase métalinguistique descriptive d'un aspect d'un signifié (Pottier, 1974) (exemple : /comestible/ ou /sert à s'asseoir/…).

Dans notre modèle, nous proposons de considérer le sème comme une entité qui ne soit pas uniquement descriptive mais plus généralement comme une entité structurante. Ainsi, le sème n'est pas vu comme une propriété autosuffisante des significations, mais comme une relation oppositionnelle dans laquelle une lexie s'inscrit au même titre que celles qui ont des significations proches. Par exemple, plutôt que de considérer que la lexie *avocat* porte le sème /defenseur/ qui est a priori sans relation avec le sème /qui requiert une peine/ porté dans la lexie *procureur*, nous considérons que les deux lexies portent un même sème oppositionnel actualisé de deux façons différentes. Ce sème est une opposition entre *défendre* et *requérir* qui struture des rôles possibles (ce qu'on appelle un domaine d'interprétation) dans le domaine juridique.

Ce point de vue oppositionnel sur le sème est mis en application dans notre modèle appelé Anadia[7]. Anadia est une méthodologie de représentation componentielle et différentielle (dans la lignée de travaux de Greimas et de Rastier) fondée sur la notion de valeur saussurienne (Saussure, 1915). Elle établi qu'un signifié est déterminé par un jeu d'oppositions avec d'autres signifiés sémantiquement proches. Il s'agit avec Anadia d'isoler les sèmes pertinents pour différentier les termes du domaine entre eux afin de les organiser dans ce que l'on appelle des dispositifs.

[6] « pomme de terre » ou « femme du monde » sont par exemple des lexies.

[7] Pour de plus amples détails sur Anadia, nous renvoyons à (Nicolle et al, 2000) ainsi qu'à l'URL http://www.info.unicaen.fr/~perlerin/recherche/anadia.

Un dispositif Anadia est principalement constitué de tables. Une table est une structure qui sert à identifier, à différentier et/ou à regrouper des lexies selon la combinatoire de certains attributs ; un attribut consistant en un sème opposant des valeurs dans un domaine d'interprétation. Par exemple, la table suivante identifie, regroupe et différentie des lexies du domaine de la météorologie :

Phénomènes	**Évaluation**	**Axe**
	bien	agitation
calme plat – pétole	mal	agitation
zone de turbulence	pas connoté	agitation
soleil	bien	couverture nuageuse
ciel voilé – nuageux	mal	couverture nuageuse
nuage	pas connoté	couverture nuageuse
douceur – doux	bien	température
glacial – polaire – étouffant – gel	mal	température
chaud – froid – torride	pas connoté	température
anticyclone	bien	pression
dépression	mal	pression
	pas connoté	pression

Figure 2. Table des phénomènes dans le dispositif de la météorologie

Cette table (Figure 2) est construite par la combinatoire des valeurs des attributs [Évaluation] (opposant les valeurs 'bien', 'mal' et 'pas connoté') et [Axe] (opposant 'agitation', 'couverture nuageuse', 'température' et 'pression'). La colonne de gauche de la table contient les lexies du domaine décrit à l'aide des attributs [Évaluation] et [Axe]. La structure de cette table est le reflet des choix de son auteur. Par exemple, on constate qu'il n'a pas distingué les lexies *ciel voilé* et *nuageux* car elles figurent sur une même ligne de la table alors que par exemple, les lexies *anticyclone* et *dépression* sont clairement différenciées mais considérées comme sémantiquement proches dans le sens où elles ne présentent qu'une différence de valeur sur l'attribut [Évaluation][8].

En général la construction d'un dispositif Anadia ne donne pas lieu à une table unique. C'est un cas très particulier car le plus souvent, parmi les différences pertinentes qu'il convient de faire, on peut en isoler certaines qui sont subordonnées à d'autres. Celles-ci ne sont alors pertinentes que pour certaines valeurs d'oppositions d'autres différences. Par exemple, dans une structuration du champ lexical de la météorologie, l'attribut [Rapport à l'activité] opposant les valeurs 'rôle' et 'profession' est subordonné à l'attribut [Rapport au domaine]. En effet, il n'est pertinent que pour la valeur 'agent, activité' de cet attribut car envisager notamment la profession d'un objet, d'un influent ou d'un phénomène qui sont les autres valeurs de l'attribut [Rapport au domaine], serait un non-sens. Il s'ensuit que si des attributs sont subordonnés à un autre, ces derniers ne peuvent donner lieu à une même table au même niveau de la structuration microsémantique. Il y a plusieurs tables à prévoir dans le dispositif dont certaines pourront être issues d'une sous-catégorisation d'une ligne d'autres tables. Ainsi, dans notre description du champ lexical de la météorologie, la table des agents est subordonnée à la table des entités du domaine (Figure 3 et Figure 4).

Entités du domaine	**Rapport au domaine**
	objet
	agent, activité
	influent
	phénomène

Figure 3. Sous-catégorisation d'une ligne d'une table. (On peut noter que toutes les lignes des tables Anadia ne sont pas nécessairement renseignées. Pour la table « Entités du domaine », l'utilisateur n'a trouvé aucun représentant lexical dans le corpus étudié pour la combinaison de valeurs d'attribut présente.)

Agents	**Action**	**Rapport à l'activité**
	intervient	rôle
prévisions météo	étudie, analyse	rôle
	intervient	profession
météorologue, bulletin météo	étudie, analyse	profession

[8] Le logiciel *Anadia* permet d'éditer graphiquement ce genre de proximité sémantique grâce aux calculs de graphe appelés Topiques.

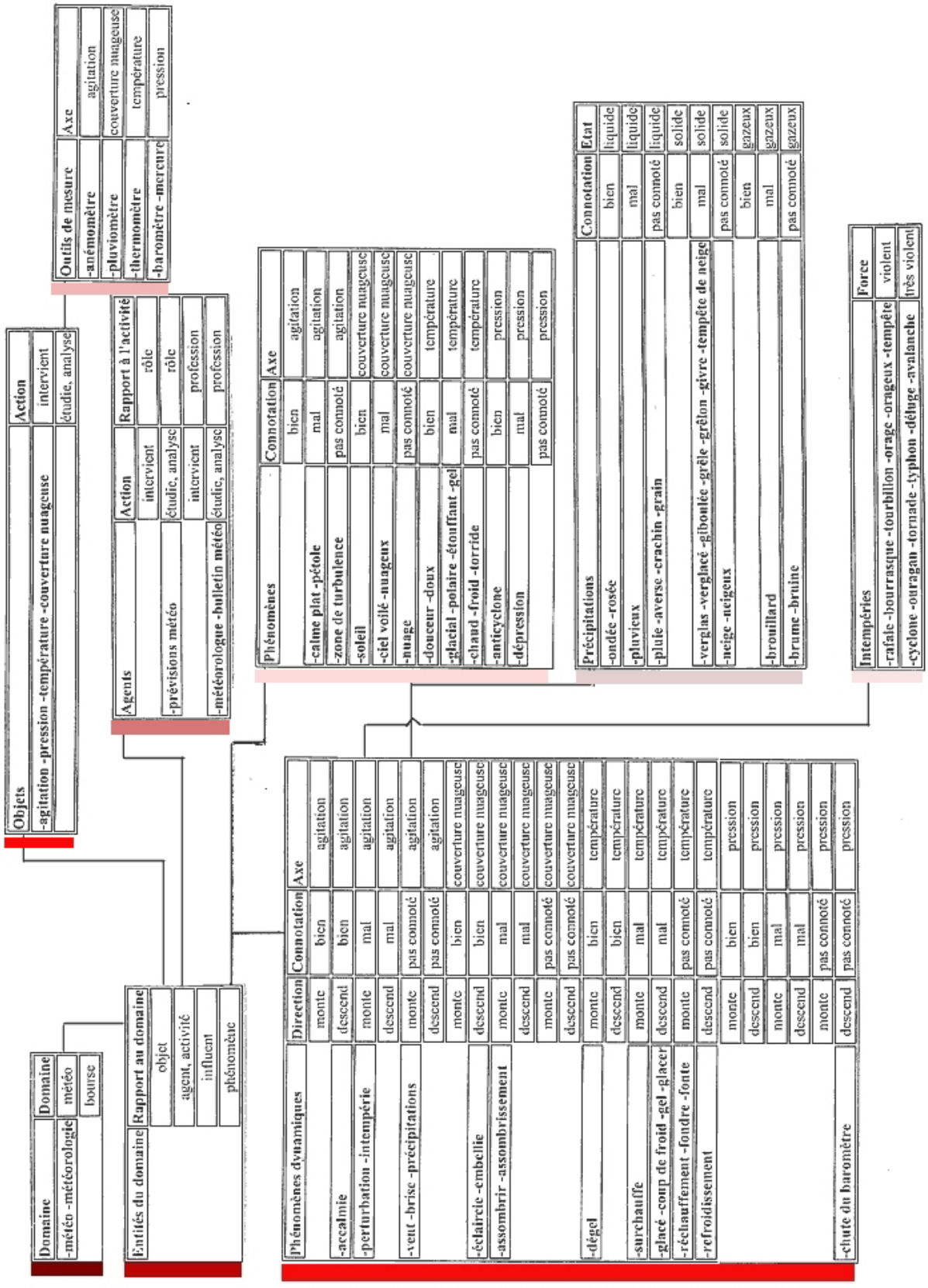


Figure 4. Dispositif en rapport avec la météorologie.

Cette relation de subordination entre tables[9] permet de concevoir des dispositifs où les lexies du domaine sont organisées des plus générales aux plus déterminées. En se rapportant au cadre théorique de la Sémantique Interprétative (Rastier, 1987), on retrouve avec cette relation de subordination les notions de sèmes génériques et de sèmes spécifiques. Pour deux lexies apparaissant dans une même table, les attributs de la table qui permettent de les différencier sont des sèmes spécifiques tandis que les attributs des tables de plus niveaux qui ne les différencient pas puisqu'ils ont en commun les mêmes valeurs, sont des sèmes génériques.

Afin de mettre en œuvre une analyse des distributions thématiques par le moyen d'un coloriage, on affecte des couleurs aux tables des dispositifs (Figure 4). Une heuristique d'attribution de couleurs aux tables a été choisie : on affecte une couleur par dispositif et on décline cette couleur en divers dégradés parmi les différentes tables du dispositif. Il s'ensuit qu'un coloriage thématique produit à l'aide de dispositifs ainsi colorés peut montrer deux choses : les différentes formes de répartition des thèmes (alternances des différentes couleurs) et pour chaque thème les différentes occurrences dans les lexies du thème des sèmes génériques et spécifiques (répartition des diverses nuances d'une même couleur).

L'analyse par coloriage thématique de différents documents d'un corpus et la modélisation des significations qu'elle requiert permet de projeter sur chaque document les représentations construites à partir de l'ensemble du corpus. L'analyse d'un document montre donc en quoi il fait partie d'un corpus homogène ou non. Le modèle insiste en cela sur la nature intertextuelle du sens en tant que l'analyse sémantique d'un document tient compte des propriétés sémantique d'un plus haut niveau que le document. Nous argumentons ici pour un point de vue non compositionnel du sens : l'homogénéité du corpus n'est pas déterminée par le sens des documents qui le composent, c'est la dimension sémantique de chaque document qui dépend du corpus. Les corpus influence les parcours interprétatifs dans lesquels peuvent s'inscrire les documents. C'est le principe d'Intertextualité.

Dans le travail de prise en compte d'un corpus nouveau, l'application de l'analyse interprétative est amorcée par les résultats d'une analyse distributionnelle (i.e. essentiellement statistique) indiquant les fréquences d'occurrences des lexies et des thèmes pertinents. Il s'agit de comptabiliser toutes les lexies du corpus et de les classer par ordre du nombre d'apparition pour faciliter leur extraction. Il s'ensuit une démarche cyclique où la modélisation lexicale permet de pointer sur des extraits de corpus des phénomènes de dynamique sémique remarquables (comme par exemple des emprunts métaphoriques d'un thème à un autre – voir IV.a). Ces phénomènes donnent en retour la possibilité d'enrichir ou de réviser les représentations des significations. C'est ainsi que notre modèle d'instrumentation du sens est le modèle d'une activité sémiotique plus que d'un résultat. Cette activité est une boucle a priori non limitée dans le temps entre le corpus et l'utilisateur qui permet d'intégrer différents types d'analyses dans la constitution du parcours interprétatif de l'utilisateur.

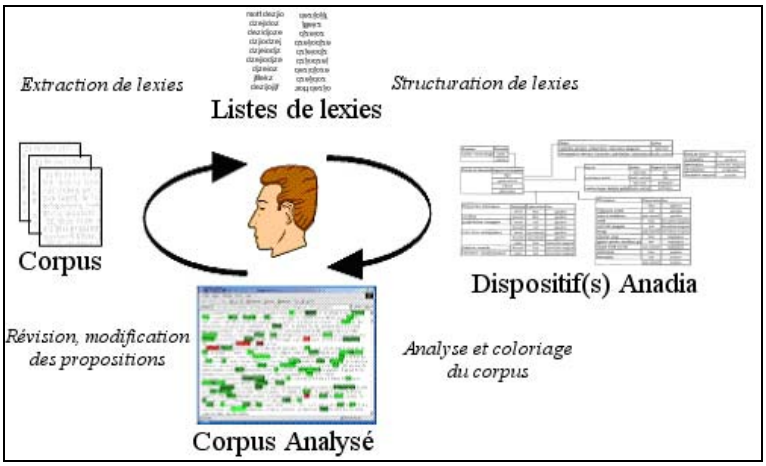


Figure 5. Démarche cyclique de création et révision des dispositifs Anadia.

[9] La relation de subordination entre tables est plus générale que la relation *est-un* des réseaux sémantiques. Elle indique un point de vue sur une signification d'une ou plusieurs lexies et, comme le montrent notamment les cas de polysémie, ce point de vue n'est pas exclusif.

Le modèle Anadia et les principes de coloriage et de repérage d'isotopies décrits précédemment sont utilisés par notre équipe dans des tâches nécessitant une instrumentation anthropocentrée. Une de ces tâches est la veille technologique et plus spécifiquement la recherche documentaire (i.e. les activités de recherche d'informations ou de documents sur un sujet plus ou moins précis). L'un des problèmes les plus important dans ce domaine (outre le fait que les particularités langagières des utilisateurs ne sont pas pris en compte) est le nombre de documents retournés par les systèmes informatiques (logiciels dédiés, moteurs de recherche sur l'Internet…) en fonction d'une requête exprimée par l'utilisateur. La plupart du temps, la liste de documents proposée est si longue que son exploration représente un travail impossible à effectuer à la main : l'apport de l'informatique pour la tâche est donc partiellement remis en question. L'utilisation de dispositifs Anadia créés par l'utilisateur sur les domaines qui l'intéresse, permet de mettre en place un filtrage et un réordonnancement de ces listes en fonctions des isotopies repérées (ou non) au sein des documents et des structures des représentations sémantiques envisagées par l'utilisateur. Nos propositions dans ce domaine ont été détaillées dans (Nicolle et al., 2002). Les principes de coloriage seront bientôt intégrés à cette étude pour permettre de proposer des interfaces personnalisées de parcours de la liste filtrée (Figure 6).

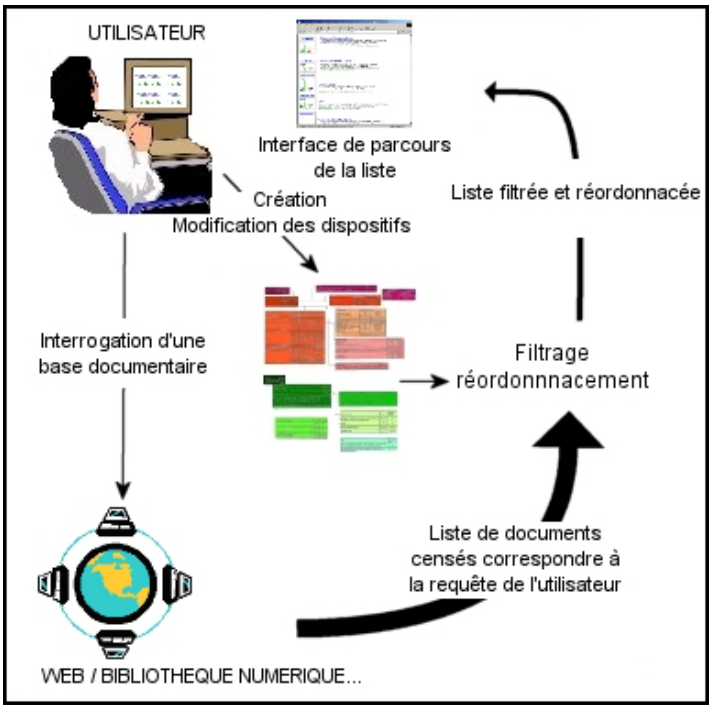


Figure 6. Principes pour la recherche documentaire ou la veille technologique

Dans cette première partie, nous avons pu voir dans quelle problématique se placent nos travaux et quelles sont les voies épistémologiques choisies. Pour rendre concrètes nos propositions et pour les évaluer, nous avons dû mettre au point un certain nombre de logiciels répondant aux exigences de notre modèle.

III. Des outils pour la sémantique des langues

Comme nous l'avons vu à travers la présentation de notre modèle, différentes étapes doivent être réalisées pour amorcer ou poursuivre le processus d'analyse proposé, qu'il soit utilisé dans le cadre de la recherche documentaire ou pour une toute autre tâche. La première étape à réaliser est communément appelée dans la communauté du TAL « extraction de termes ». Le logiciel créé par nos soins et dédié à cette tâche s'appelle *MemLabor* (Perlerin, 2002). Il utilise principalement des méthodes statistiques pour assister son utilisateur dans l'analyse des différentes distributions de mots significatives dans un corpus. Pour pouvoir structurer ou tester la validité des lexies retenues après cette étape, nous proposons *ThemeEditor* (Beust, 2002), qui permet de mettre en place un coloriage thématique de corpus sans structuration profonde des lexies, ces dernières n'étant stockées en machine que sous forme de paquets de mots rassemblés en thèmes. *ThemeEditor* permet à l'utilisateur de constituer des classes de lexies et d'évaluer visuellement et statistiquement les répartitions de ces classes sur un ensemble de documents électroniques. Enfin, pour pouvoir structurer les lexies en dispositifs Anadia et procéder à des analyses interprétatives fines en terme d'isotopies, nous avons créé le logiciel *Anadia* et un certain nombre de modules d'analyse externes. Le logiciel *Anadia* est dédié à l'assistance de son utilisateur dans une tâche de description fine de la sémantique (une description microsémantique) d'un ou

plusieurs champs lexicaux. Les modules qui l'accompagnent permettent un coloriage du corpus en fonction des lexies des dispositifs crées.

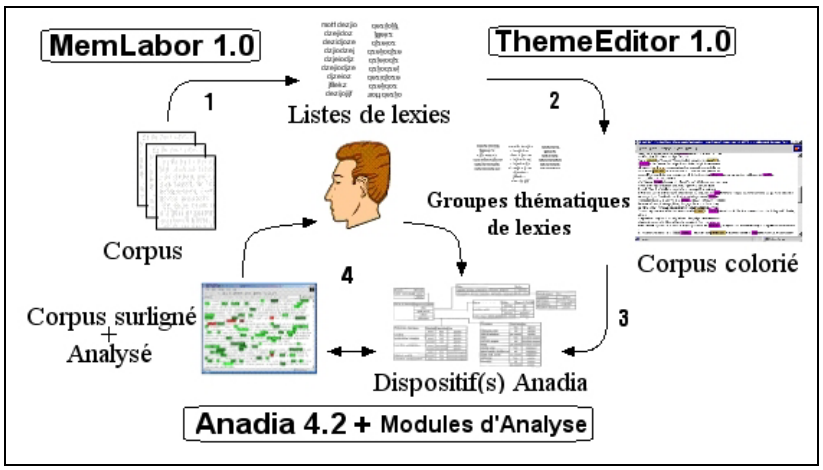


Figure 7. Les logiciels *MemLabor*, *ThemeEditor* et *Anadia* dans le processus d'analyse d'un corpus et de création de dispositifs Anadia.

1. Le corpus est analysé statistiquement. Le logiciel propose une liste de lexies issue du corpus.

2. Les lexies choisies par l'utilisateur sont regroupées par thèmes pour une première analyse de leur distribution dans le corpus.

3. Les thèmes construits sont structurés en dispositifs Anadia par l'utilisateur à l'aide du logiciel éponyme. Les modules d'analyse permettent de surligner le corpus et de procéder à des analyses en terme de répartition des thèmes et des isotopies repérées au sein des documents.

4. Le processus cyclique est enclenché : les dispositifs Anadia peuvent être modifiés en fonction des résultats des analyses du corpus ou des résultats de la tâche pour laquelle ils ont été créés (recherche documentaire, analyse d'un fait de langue particulier…)

Les logiciels que nous développons[10] et expérimentons ont tous pour objectif de permettre à un utilisateur d'instrumentaliser une analyse de l'usage des productions langagières. Chacun à leur manière, les trois outils présentés sont des logiciels anthropocentrés dédiés à la dimension sémantique du matériau linguistique. Ces logiciels sont des exemples de ce que l'on nomme des logiciels d'étude (Nicolle, 2001). Un logiciel d'étude est un programme informatique qui permet de tester des hypothèses. En ce qui nous concerne, ces hypothèses portent sur le sens et la signification. Ces réalisations informatiques sont confrontées, soit à du matériau linguistique dans le cadre d'analyses automatisées, soit à une condition d'interaction homme-machine dans le cadre de constructions de ressources.

I.a. Memlabor

Le logiciel *Memlabor* (Perlerin, 2002) permet de gérer des corpus de documents électroniques sous forme d'une description XML[11] (des fichiers de textes, des fichiers de lexies - mots grammaticaux d'une langue, lexies d'un domaine…) et des programmes qui prennent ces corpus en entrée et produisent des résultats (comptages, graphiques…) conservés avec le corpus. On peut y ajouter des programmes (classes, méthodes ou exécutables) tout en conservant les techniques de manipulation utilisant XML. Les fichiers peuvent être partagés

[10] et que nous mettons à disposition gratuitement sur Internet (aux adresses http://www.info.unicaen.fr/~perlerin et http://www.info.unicaen.fr/~beust).

[11] Le langage XML (http://www.w3c.org/xml) permet de structurer des documents électroniques à l'aide de balises personnalisables telles que <corpus><journal>Libération</journal><date>10/02/98</date></corpus>.

entre plusieurs corpus puisqu'ils ne sont pas modifiés par les traitements et qu'ils peuvent être référencés par une URL[12] ou un emplacement sur un support de stockage (disque dur, Cd-rom…).

Dans sa version actuelle, *MemLabor* propose cinq types de travaux sur corpus (ou sur des sous-ensembles du corpus) : la transformation automatique du format des documents du corpus en TXT[13] (depuis XML ou HTML[14]), un découpage en mots paramétrable, un calcul de type Zipf (voir ci-dessous), une segmentation en paragraphes et une recherche de cooccurrences d'ensemble de lexies. Pour le présent article, seul le calcul de type Zipf et la recherche de cooccurrences sont des fonctions primordiales. Il s'agit ici d'utiliser *MemLabor* pour aider un utilisateur à extraire d'un corpus des lexies qu'il juge pertinentes, c'est-à-dire en rapport avec les thèmes qu'il veut explorer. En TAL, l'extraction de termes pertinents est une voie de recherche explorée depuis de nombreuses années. Cette tâche est en particulier très utile pour l'indexation de documents à l'aide de mots-clefs[15]. On distingue principalement trois types de méthodes pour l'extraction de termes en TAL : les méthodes linguistiques basées sur des règles syntaxiques (par exemple : (Smadja et McKeown, 1990)), les méthodes statistiques basées sur les redondances de séquences de mots (par exemple : (Lebart et al., 1998)) et les méthodes intégrant les propositions des deux premiers types (par exemple : (Bourrigault, 1994)). *MemLabor* est un logiciel dédié à l'aide à l'extraction de termes, il ne s'agit pas ici de proposer une extraction automatique : on parle alors d'extraction supervisée. *MemLabor* met en place un calcul statistique très simple : un calcul de type Zipf. Il n'utilise que des ressources extrêmement limitées : la liste des mots grammaticaux de la langue du corpus à explorer. Ces listes sont finies et peu importantes (Giguet, 1998), donc facile à manipuler. Il n'a pas pour objectif de rivaliser avec les techniques de pointe dans le domaine de l'extraction de termes mais il est suffisant, dans une approche anthropocentrée, pour une première appréhension d'un corpus.

Zipf a observé (Zipf, 1949) que la fréquence d'utilisation des mots décroît de manière quasi linéaire et que le produit $f.R$, soit la fréquence f d'un mot multipliée par le rang R de ce mot est à peu près constant (cette constante dépend du texte ou de l'ensemble de textes considéré). *MemLabor* permet d'effectuer un calcul de type Zipf sur l'ensemble des documents d'un corpus ou sur un sous-ensemble de documents d'un corpus. Ce calcul donne lieu à la création d'un fichier rassemblant les formes graphiques découvertes classées par ordre décroissant de leur nombre d'occurrences au sein des textes. *Memlabor* présente la liste des lexies repérées avec leur nombre d'occurrences au sein de l'ensemble du corpus ainsi que des représentations graphiques en histogrammes et en log/log des résultats (respectivement les fenêtres du centre et de droite dans la Figure 8 – l'interface proposée est semblable à celle de l'*applet* d'Emmanuel Giguet (Giguet, 1998) proposant un calcul similaire *en ligne*[16]). Ces graphiques permettent outre de vérifier la validité de la loi de Zipf sur le corpus considéré, de repérer d'éventuelles irrégularités inhérentes à des corpus hétérogènes (c'est-à-dire rassemblant des textes utilisant des vocabulaires très différents). Dans ce cas, la représentation en histogramme n'est pas d'allure logarithmique et la droite de régression – en rouge dans l'interface du programme - de la représentation en log/log peut être absente de la fenêtre, le nuage de points étant trop disséminé.

[12] Une URL (*Uniform Ressource Locator*) est l'adresse électronique d'une ressource ou d'un fichier disponible sur l'Internet (ex : http://www.w3c.org.xml)

[13] Le cigle TXT désigne le format de fichiers textes ne contenant pas d'informations sur la mise en forme des caractères (gras, italique…) et des paragraphes.

[14] HTML (Hyper Text Mark-up Langage) est un format de présentation et de structuration des pages de l'Internet avec des balises et des liens hypertextes.

[15] Les mots-clefs sont des mots censés refléter le (ou les) thèmes d'un document. Ils sont utilisés en particulier pour l'indexation les pages Internet par les moteurs de recherche.

[16] http://www.info.unicaen.fr/~giguet/java/zipf.html

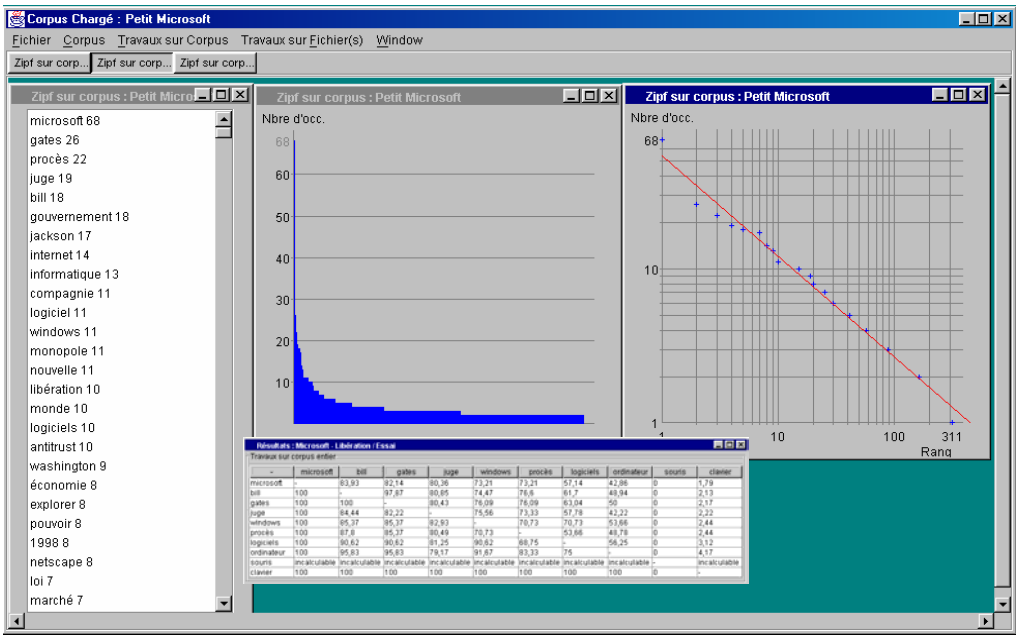


Figure 8. Copie d'écran de *MemLabor*
Calcul de type Zipf et cooccurrence de lexies sur un corpus.

Dans l'exemple proposé, *MemLabor* a été utilisé pour faire un calcul de type Zipf sur un corpus de 25 articles du journal Libération traitant du procès contre la compagnie Microsoft pour position de monopole en 1998. Dans la fenêtre de droite, on peut voir la liste des mots proposés avec pour chacun d'entre eux son nombre d'occurrences à l'intérieur du corpus. Cette liste a été filtrée à l'aide d'une liste des mots grammaticaux du français (et de certaines formes adjectivales et verbales[17]). On retrouve tout en haut de la liste proposée, des mots qui ont bien directement rapport avec le corpus (Par exemple, *Jackson* qui apparaît 17 fois dans le corpus, est le nom du juge chargé de l'affaire).

Selon un principe bien connu, les listes d'occurrences produites par un calcul de type Zipf permettent grossièrement de reconnaître tout d'abord les mots grammaticaux (plus fortes occurrences) puis par la suite les mots représentatifs du domaine de spécialité (mots d'occurrences moyennes) et enfin les mots peu redondants du corpus (occurrences faible et hapax). En filtrant les listes d'occurrences avec une liste de mots non significatifs connus et en permettant des calculs sur les cooccurrences, *MemLabor* facilite l'extraction des principales classes terminologiques évoquées dans le corpus. Le calcul de type Zipf est une première étape pour l'aide à l'extraction des lexies. La cooccurrence de ces dernières au sein des documents (ou d'entités linguistiques plus petites telles que le paragraphe, la phrase ou un contexte syntaxique défini) est une source d'information importante pour les filtrer ou les classer de façon supervisée. *MemLabor* peut effectuer ce type de calcul avec comme grain d'analyse le document. Les résultats permettent d'extraire des sous-listes de lexies pertinentes et pouvant être rassemblées au sein d'un même groupe thématique (Perlerin, 2002).

Le logiciel *Memlabor* a été utilisé à plusieurs reprises dans les travaux de recherche menés au sein de notre équipe. Il a notamment fourni une aide dans la constitution de listes de mots-clés à des fins de recherche documentaire (Perlerin 2002) grâce à l'analyse des distributions et des cooccurrences sur un corpus choisi à cet effet.

1.b. ThemeEditor

ThemeEditor (Beust 2002) est un logiciel d'étude[18] qui permet à son utilisateur de tester, de créer et de mettre à jour des listes de mots relevant d'un même champ lexical en examinant leur répartition dans un corpus à

[17] Cette liste de 614 formes, appelée *stop-list* ou « anti-dictionnaire », a été créée à partir de celle proposée par Jean Véronis de l'Université Aix Marseille I, sur http://www.up.univ-mrs.fr/~veronis/data/antidico.txt.

[18] Disponible à l'URL http://users.info.unicaen.fr/~beust/ThemeEditor.html

l'aide d'une technique de coloriage. De même que (Pichon et Sébillot, 1999), nous appelons ici « thèmes » ces listes de mots qui dénotent les sujets abordés dans un texte ou dans un corpus.

L'outil que nous proposons s'inscrit dans le même courant d'étude que le logiciel *PASTEL* développé par (Tanguy, 1997). A la différence de *PASTEL* conçu pour la visualisation des isotopies d'un texte, l'outil que nous proposons est conçu pour l'étude de corpus de documents électroniques. De plus *ThemeEditor* est avant tout dédié à la construction de classes sémantiques. L'analyse de leur répartition dans le matériau textuel n'est pas une fin en soi mais une façon d'aider cette construction.

La méthode de construction des classes sémantiques par le coloriage thématique est essentiellement manuelle et anthropocentrée. C'est la différences avec des systèmes qui proposent automatiquement des thèmes par analyse des voisinages de mots (Pichon et Sébillot, 1999) ou bien qui proposent des ontologies issues de calculs de distances basés sur une analyse morpho-syntaxique, par exemple le système *ASIUM* de (Faure, 2000). Comme pour *PASTEL*, notre méthode est basée sur une analyse interprétative de textes électroniques. Partant d'un corpus de textes, le système permet d'enrichir des thèmes par sélection de lexies dans des listes d'occurrences de type Zipf. Ces thèmes permettent en retour de produire automatiquement le coloriage des isotopies du corpus (Figure 9). L'utilisateur peut réitérer ce processus autant fois que nécessaire à la lumière des statistiques issues du coloriage (par exemple le nombre de mots surlignés, le classement des principaux thèmes du corpus, le pourcentage de couverture d'un thème dans le corpus… Figure 10). Plus le corpus d'étude est thématiquement homogène, plus il y a de redondance concernant les thèmes visés et donc plus il est facile de les constituer avec la méthode proposée.

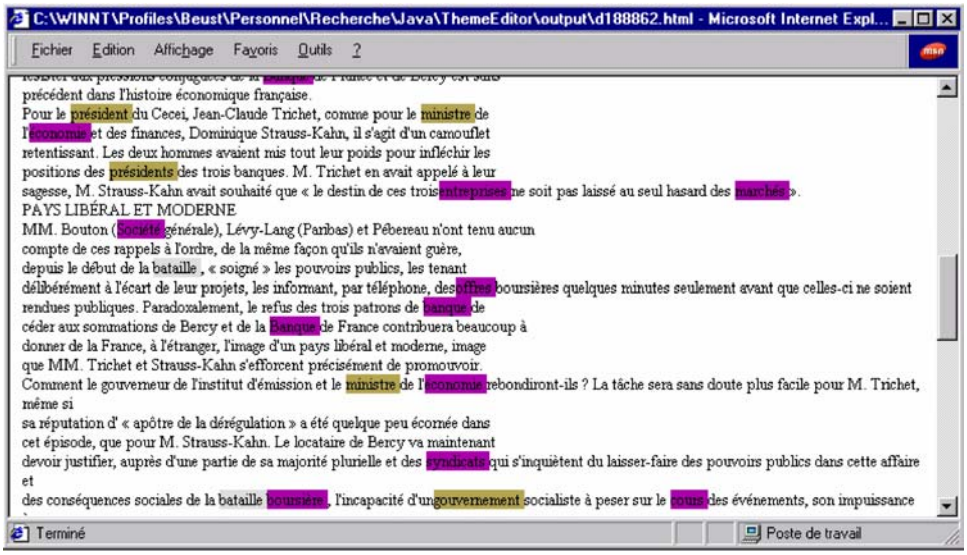
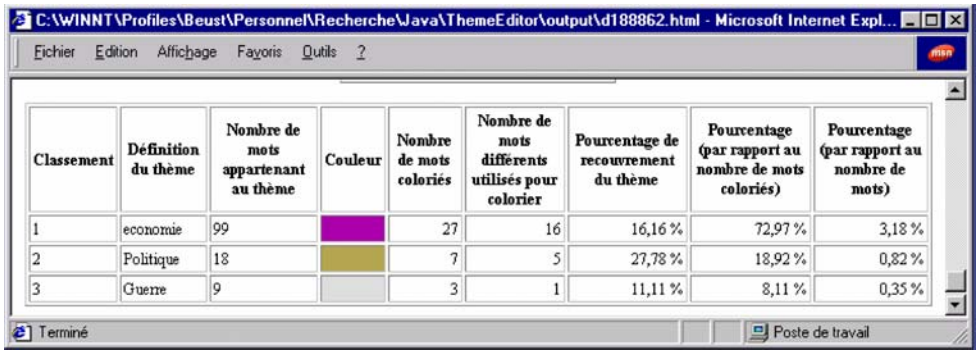


Figure 9. Exemple de coloriage thématique produit par *ThemeEditor*.



Classement	Définition du thème	Nombre de mots appartenant au thème	Couleur	Nombre de mots coloriés	Nombre de mots différents utilisés pour colorier	Pourcentage de recouvrement du thème	Pourcentage (par rapport au nombre de mots coloriés)	Pourcentage (par rapport au nombre de mots)
1	economie	99		27	16	16,16 %	72,97 %	3,18 %
2	Politique	18		7	5	27,78 %	18,92 %	0,82 %
3	Guerre	9		3	1	11,11 %	8,11 %	0,35 %

Figure 10. Statistiques de coloriage produites par *ThemeEditor*.

Nous avons effectué une première expérimentation de l'outil que nous avons exercé sur un petit corpus de 336 textes électroniques totalisant 260 418 mots. Les principaux thèmes que nous avons construits dans cette expérimentation montrent en quoi il s'agit effectivement d'un corpus homogène. Ces thèmes sont l'aviation (133 lexies dans le thème) et l'économie (153 lexies dans le thème).

I.c. Anadia et modules d'analyse

Il est un bon nombre de tâches pour lesquelles une représentation lexicale en liste de mots, comme les thèmes construits avec *ThemeEditor* ou *MemLabor*, n'est pas suffisante mais où les relations lexicales entre les termes (hyperonymie, hyponymie, synonymie ou encore antonymie par exemple) sont déterminantes. C'est le cas par exemple des tâches de recherche documentaire sur Internet. Le logiciel d'étude *Anadia* est un outil qui a pour but d'aider son utilisateur à mettre évidence la structure lexicale de terminologies, dont une première description en extension peut être produite avec *Memlabor* ou *ThemeEditor*. Ici, les objectifs ne sont plus tout à fait les mêmes. On passe d'une étude de macro-structuration d'un vocabulaire limité à l'identification de listes de termes à une étude de micro-structuration lexicale des domaines de spécialité étudiés.

Dans sa dernière version en Java, le logiciel Anadia offre des outils de production, d'analyse, de modification et de visualisation des dispositifs. Il assure toutes les étapes algorithmiques et offre un environnement d'interaction pour les étapes de choix qui sont de la responsabilité de l'usager. Il permet également de compléter les mots présents dans les tables par un calcul automatique ou semi-automatique de leurs flexions grâce à une interface avec la base de données MAHTLex du laboratoire IRIT de l'Université Paul Sabatier de Toulouse[19]. La construction des tables à l'aide du logiciel facilite leurs modifications éventuelles tout en vérifiant la cohérence des relations de sous-catégorisation au sein des dispositifs. Les dispositifs créés par l'utilisateur sont enregistrés dans des fichiers structurés grâce au langage XML de sorte à pouvoir être utilisés dans des chaînes de traitements ultérieurs (cf. partie suivante). Ils pourront, en particulier, être utilisés par les modules d'analyse de corpus que nous avons implémentés. Une fois les dispositifs construits, coloriés et mis en référence avec MAHTLEX (cf. partie précédente), les mots du corpus présents dans les dispositifs sont traités par un ensemble de programmes en Java et de feuilles de transformation XSLT[20], qui assurent le coloriage de chaque document du corpus et la réalisation d'histogrammes représentant graphiquement la proportion de chacune des tables de chaque dispositif dans ces documents (Figure 11). Tous les histogrammes sont regroupés au sein d'une même page HTML. Cette page permet une vision macroscopique de la distribution des domaines et des tables des dispositifs correspondants à l'intérieur du corpus dans son entier. Un système de liens hypertextes permet d'accéder directement aux documents coloriés depuis les histogrammes pour étudier de façon microscopique la dynamique des sèmes des articles sélectionnés. Ces pages HTML permettent également de naviguer aisément au sein du corpus et représentent à elles seules une interface de lecture rapide efficace.

Le corpus surligné est stocké en machine au format XML. Cette technique permet de récupérer aisément des données statistiques sur la répartition des thèmes et des isotopies au sein d'un ou plusieurs documents, ou au sein d'un ou plusieurs paragraphes. Ces données statistiques sont utilisées entre autre en recherche documentaire, pour décider de l'adéquation d'un document avec les représentations sémantiques proposées par l'utilisateur.

[19] http://www.irit.fr/ACTIVITES/EQ_IHMPT/ress_ling.v1/accueil01.php - Institut de Recherche en Informatique de Toulouse.

[20] XSLT (Extensible Stylesheet Language) est un langage informatique permettant entre autre, la transformation d'un document au format XML en HTML en séparant les données de leur mise en forme.

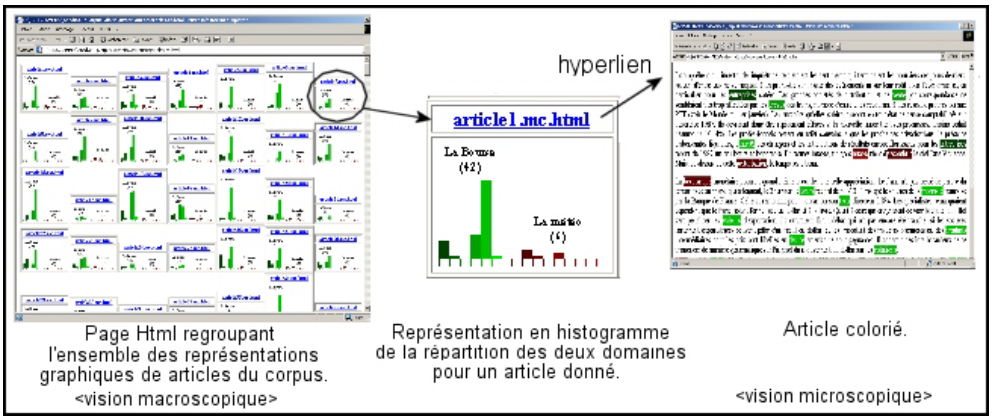


Figure 11. Coloriage d'un corpus et création d'une interface de lecture rapide à partir de dispositifs Anadia

IV. Modélisation lexicale personnelle : une expérimentation

a. Utilisation du modèle pour la caractérisation et la détection des métaphores

i. Principes

En recherche documentaire en particulier, et dans pratiquement tous les travaux qui touchent au TAL en général, les tropes sont à la base de nombreuses difficultés. Spécifiquement, la métaphore est un fait de langue qui induit une multiplicité des interprétations et qui de fait, est difficilement prédictible. C'est parce que la métaphore opère un transfert de sens que ces lexicalisations doivent être repérées et en tous cas, pas interprétées d'une façon littérale. Les systèmes à base de mots-clefs, les plus couramment utilisés en recherche documentaire, se heurtent souvent à ce phénomène car il est pratiquement impossible de l'aborder de façon non littérale, uniquement au niveau du mot. Partant de ce constat, nous avons choisi d'entamer des recherches sur ce fait de langue en collaboration avec Stéphane Ferrari du GREYC (Perlerin et al., 2002). Cette étude a pour objectif la caractérisation des métaphores et l'élaboration d'une aide à leur interprétation. Nous nous intéressons en particulier aux métaphores conceptuelles telles que décrites dans (Lakoff et Johnson, 1980).

Pour la métaphore, la communauté linguistique s'accorde à différencier la source de la cible, par exemple dans la phrase :

« Si c'est non, il y aura une bourrasque monétaire.»[21]

« *bourrasque* » est la source de la métaphore et « *un mouvement boursier brutal* » (ou une paraphrase analogue) en est la cible. Pour les métaphores conceptuelles, on dépasse le stade lexical et l'on parle de domaine source et domaine cible. Par exemple, des termes et expressions originellement dédiés à la guerre sont souvent utilisés lorsqu'on parle d'un débat (exemples : « *rester sur ces positions* », *« tirer à boulet rouge* »…). Cet exemple fameux nommé par Lakoff *Argument is War* est loin d'être unique. Les métaphores conceptuelles acquièrent l'épithète « conventionnelles » lorsqu'elles sont très redondantes dans un domaine donné comme la bourse décrite en des termes de météorologie par exemple. C'est justement cette métaphore conceptuelle conventionnelle que nous avons choisi d'étudier en terme de dynamique sémique sur un corpus dans lequel elle avait déjà été repérée (Ferrari, 1997). Notre démarche expérimentale se fonde sur une observation du fait de langue sur un corpus électronique à l'aide d'outils informatiques. Le corpus étudié est constitué de l'ensemble des articles relatifs à la bourse entre 1987 et 1989 dans le journal *Le Monde sur CD-ROM*. Il contient notamment de nombreux emplois de la métaphore conceptuelle « météorologie boursière ». Ce corpus est d'autant plus intéressant qu'il contient aussi des emplois non métaphoriques de termes en rapport avec la météorologie : un des intérêt de cette étude est de mettre en évidence les différences entre les types de propagations sémiques (dissimilation, assimilation) et les formes sémiques (enchevêtrement, alternance, …) pour les emplois métaphoriques et non métaphoriques de termes en rapport avec la météorologie.

[21] Extraits des *Nouveaux chiens de garde*, Serge Halimi, Editions Liber-Raisons d'Agir 1997.

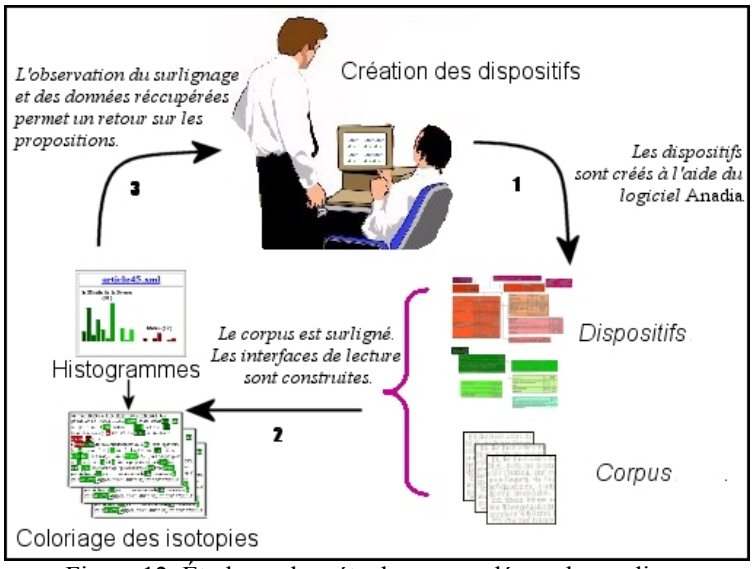


Figure 12. Étude sur la métaphore: une démarche cyclique

Le point de départ de cette étude est la constitution de dispositifs Anadia à partir d'une liste issue d'un calcul de type Zipf opéré sur un corpus au format XML (étape 1 sur la Figure 12). C'est le processus classique de constitution de dispositifs tel qu'il a été décrit précédemment. La composition de ces dispositifs nous permet d'amorcer la démarche cyclique représentée dans la Figure 12 : l'interface de lecture rapide du corpus décrite dans la partie précédente est construite automatiquement et le corpus est surligné (étape 2 sur la Figure 12). En fonction de l'observation des articles surlignés, les dispositifs initiaux peuvent être modifiés. Nous revenons ainsi à la première étape de la démarche en affinant au fur et à mesure nos propositions de représentation des domaines (étape 3 sur la Figure 12). Nous rappelons à ce sujet que les dispositifs construits pour quelque tâche que ce soit, en tant qu'issus d'un processus anthropocentré, n'ont jamais valeur d'universalité, n'ont pas pour vocation d'être exhaustifs et ne sont, à chacune des étapes de leur élaboration, que le reflet d'un point de vue sur un domaine exprimé dans un but précis, pour une tâche donnée.

ii. Résultats

Le premier résultat remarquable obtenu lors de cette étude, par ailleurs toujours en cours, concerne directement le modèle Anadia. Au cours des retours sur les dispositifs sur la bourse et la météorologie construits pour le coloriage du corpus, nous nous sommes aperçus qu'un certain nombre d'attributs choisis pour des tables appartenant à l'un des deux dispositifs pouvaient tout aussi bien être utilisés dans l'autre alors que cela était impossible pour d'autres attributs. Par exemple, l'attribut [Direction] ayant pour valeurs 'monte' et 'descend' ou l'attribut [Évaluation] ayant pour valeurs 'bien', 'mal' ou 'pas connoté' étaient communs aux deux dispositifs bourse et météorologie (Figure 13 et Figure 14), alors que l'attribut [Axe] ayant pour valeurs 'agitation', 'couverture nuageuse', 'température' et 'pression' ne pouvait vraisemblablement pas intervenir dans le dispositif de la bourse.

PHENOMENES DYNAMIQUES DE LA BOURSE	Direction	Évaluation
	monte	bien
	descend	bien
	monte	mal
dévalorisation, dévaluer, krak, minikrach	descend	mal
hausse des cours, inflation	monte	pas connoté
baisse des cours, déflation	descend	pas connoté

Figure 13. Table des phénomènes dynamiques - dispositif "Bourse"

PHENOMENES DYNAMIQUES DE LA METEOROLOGIE	Direction	Évaluation	Axe
	monte	bien	agitation
accalmie	descend	bien	agitation
perturbation, intempérie	monte	mal	agitation

	descend	mal	agitation
	monte	pas connoté	agitation
vent, brise, précipitations	descend	pas connoté	agitation

Figure 14. Extrait de la table des phénomènes dynamiques - dispositif "Météorologie"

Une nouvelle distinction est alors apparue : les sèmes communs et les sèmes propres. Les sèmes communs sont des sèmes qui peuvent être potentiellement utilisés dans plusieurs dispositifs et dont la valeur n'est pas directement en corrélation avec domaine envisagé dans le dispositif (comme par exemple [Direction] et [Évaluation]). Les sèmes propres sont des sèmes dont l'emploi semble circonscrit au domaine du dispositif (comme par exemple [Axe] dans le dispositif de la météorologie). En d'autres termes, un sème propre n'a pour vocation que d'être utilisé pour un domaine précis alors qu'un sème commun pourra être partagé par plusieurs. L'usage d'un sème dit commun dans plusieurs dispositifs appelés à être utilisés pour une même tâche permet le repérage d'isotopies faisant intervenir ce même attribut et donc de pouvoir repérer des isotopies trans-dispositifs indiquant des formes sémantiques plus profondes (comme celles produites par les sèmes /armée/ et /Église/ dans son étude sur *Le rouge et le noir* de Stendhal dans (Rastier, 1987)). Pour notre étude sur la métaphore, cette propriété devient primordiale car elle nous permet de retrouver des résultats connus mais jamais formalisés comme tels. Il s'agit en l'occurrence d'absences de lexicalisation dans un domaine source d'une métaphore conceptuelle pour certaines valeurs sémantiques possiblement comblées par une lexicalisation de valeur analogue dans le domaine cible. Pour la météorologie boursière, nous n'avons pas trouvé dans notre corpus de lexie du domaine boursier pouvant apparaître dans la table « *Phénomènes dynamiques de la bourse* » (Figure 13) pour la combinaison de valeurs 'monte' et 'mal' correspondant respectivement aux attributs [Direction] et [Évaluation]. Pour cette même combinaison de valeurs dans le dispositif en rapport avec la météorologie (table «Phénomènes dynamiques de la météorologie », Figure 14) nous avons trouvé dans le corpus, le mot « perturbation » qui pouvait être inséré dans un dispositif en rapport avec la météorologie. Ce mot était justement utilisé plusieurs fois de façon métaphorique dans notre corpus :

> *4. Les professionnels restent en effet convaincus que les prochaines privatisations, la présence d'abondantes liquidités, l'intérêt des étrangers et les anticipations de résultats encore florissants pour les entreprises feront de 1987 un cru boursier honorable. En termes imagés, un gros nuage vient d'assombrir le ciel Rue Vivienne. Mais, au-dessus de cette* ***perturbation****, le temps reste beau.*
> (article 1 – Corpus *Le Monde sur Cd-rom* – domaine boursier)

Confirmant ainsi le statut conceptuel de la métaphore étudiée, des phénomènes analogues ont été observés avec d'autres lexies comme par exemple, « bourrasque » présente dans le dispositif de la météorologie avec les attributs/valeurs [Direction] 'monte', [Évaluation] 'mal', [Axe] 'agitation' et [Force] 'violent' :

> *5. La* ***bourrasque*** *financière a inspiré tous les discours que M. François Mitterrand a prononcés, le jeudi 29 octobre, lors des premières étapes de sa visite dans la Loire. A la mairie de Saint-Etienne, le chef de l'Etat a appelé les Français à " se grouper et se regrouper quand la tornade passe ".*
> (article 145 - Corpus *Le Monde sur Cd-rom* – domaine boursier)

> 6. *"COURAGE, fuyons... " Tel était le slogan en vogue ces jours derniers sous les lambris du palais Brongniart, où la* ***bourrasque*** *monétaire a fait s'envoler nombre d'investisseurs et autant d'espoirs de nouveaux records.*
> (article 10 - Corpus *Le Monde sur Cd-rom* – domaine boursier)

L'exemple 4 montre l'intérêt de l'analyse en terme de dynamique sémique des entités textuelles mettant en jeu ce type de lexies pour détecter ou non un emploi métaphorique et éventuellement proposer une aide à leur interprétation. Dans cet extrait du corpus, nous sommes en effet en présence de lexies mettant en place une isotopie en rapport avec la météorologie (supportée par les mots *nuage*, *ciel* et *temps*) alors que celles-ci sont clairement employées de façon métaphorique, l'emploi métaphorique étant textuellement marqué par « *En termes imagés »* en début de phrase.

b. Évaluation

L'évaluation de modèles anthropocentrés est chose délicate. Comme nous avons pu déjà le préciser, nos travaux s'inscrivent dans un courant du TAL influencé par les Sciences Cognitives où l'on préfère la coopération système/utilisateur à l'automatisation. Dans nos propositions, la machine n'est plus l'organe central du modèle, elle tient un rôle en rapport avec ses capacités premières : le calcul et la manipulation rapide sur des masses de

données. Elle n'est pas une entité omnisciente, un factotum mystérieux sans capacités d'interprétation, elle est source d'assistance et de suggestion, un compagnon personnel pour l'aide à l'interprétation et à la manipulation de documents textuels. Dès lors, l'évaluation de ces techniques supporte difficilement la mesure car cela impliquerait une possible quantification qualitative de l'interprétation qui est par essence ni absolue, ni universelle. La confrontation de notre modèle à des tâches déjà pourvues de techniques d'évaluation (comme la recherche documentaire avec les mesures : *rappel* et *précision*[22]) ne doit pas laisser entendre que ces techniques, par ailleurs souvent critiquées, lui sont applicables. Dans une approche anthropocentrée, le problème de l'évaluation se trouve déplacé. Il ne s'agit pas tant d'évaluer une application que l'efficacité et la faisabilité d'une interaction entre un utilisateur humain et un agent logiciel. Certaines propositions, en particulier dans le domaine du dialogue Homme-Machine préconisent déjà des modalités d'évaluation adaptées à des conditions d'interaction entre l'utilisateur et le système (*taux de compétence* et *taux d'efficacité* - (Luzzati, 1996)). Cependant, notre modèle basé sur des principes anthropocentrés, ne peut souffrir de telles évaluations sans réflexion préalable : une adaptation est nécessaire pour pouvoir évaluer les caractéristiques de l'anthropocentrisme. Nos réalisations informatiques et les modèles qui en sont à l'origine ne sauraient donc être évalués que par l'intermédiaire d'expériences et de discussions contradictoires sur des résultats obtenus en laboratoire ou en condition réelles. L'atelier formation du CNRS « *Variation, construction et instrumentation du sens »* a été pour nous l'occasion de mettre en place une toute première évaluation. Nous souhaitions tester la capacité des participants à s'approprier les principes généraux du modèle Anadia en leur proposant de construire dans un temps imparti, un dispositif sur un sujet précis (la bourse) afin de pouvoir comparer les résultats. Cette expérience s'est déroulée au cours de deux séances de deux heures trente chacune et avec un total de 8 participants d'horizons différentes (linguistique, psychologie, ergonomie, informatique, microbiologie, sciences cognitives…) repérés par les codes suivants : SM, BA, MS, AA, JR, AP, IK et MC. Après un exposé d'environ une heure sur nos propositions dans le domaine (précisément celles présentes dans cet article), nous avons fourni aux participants une liste de 216 lexies issue de notre corpus *Le Monde sur CD-ROM*. Cette liste avait été obtenue à partir d'un calcul de type Zipf sur l'ensemble des articles traitant de la bourse et de l'économie de laquelle nous avions enlevé tous les éléments non verbaux et non substantivaux (extrait en Figure 15) et dont nous avions sélectionné arbitrairement 216 représentants. Les consignes données aux participants se bornaient à leur demander de construire un dispositif selon leur façon propre de parler du domaine. Pour cette tâche, ces derniers avaient à titre d'exemple le dispositif de la météorologie construit par nos soins pour nos recherches sur la métaphore (Figure 4).

achat	dépression	OPA
acheteur	dévalorisation	opérateur
action	dévaluation	or
actionnaire	dévaluer	palais Brongniart
affaire	devise	parité
agent de change	dividende	participations
argent	échange	pertes
analyste	économiste	petits porteurs
baisse des cours	fonds	place boursière
back office	entreprises	place financière
banque	front office	portefeuille

Figure 15. Extrait de la liste de lexies fournie lors de l'expérience.

Les participants ont presque tous travaillé sur papier car ils n'étaient pas placés devant un poste de travail informatique pour créer leur dispositif. Des machines sur lesquelles se trouvait installé le logiciel Anadia, étaient à leur disposition dans la salle, mais il ne leur était pas explicitement demandé de s'en servir (aucune interdiction n'a pas non plus été formulée). Seule l'un d'entre eux a utilisé son propre matériel, en l'occurrence un logiciel (*Inspiration* de Inspiration Software Inc.) originellement destiné à la réalisation de diagrammes et de schémas d'organisation.

[22] Le rappel et la précision sont deux mesures entre autres utilisées dans le cadre de l'évaluation de systèmes de recherche documentaire sur des corpus préalablement traités par des experts. Le rappel est égal au rapport du nombre de documents retrouvés sur le nombre de documents pertinents à partir d'une requête donnée et la précision est égale au rapport du nombre de documents pertinents retrouvés sur le nombre de documents trouvés dans le corpus pour une requête donnée. Ces mesures sont par exemple critiquées du fait du rôle prépondérant des experts décidant de la validité d'un document en fonction d'une requête.

A l'issue des deux séances d'expérience, aucun des participants des deux groupes n'a pu créer un dispositif entier (nous reviendrons sur les raisons plus tard). Ils ont cependant tous au moins proposé des groupes de lexies, parfois précisé les différences qu'ils considéraient effectives au sein de ces groupes et créé des tables Anadia avec un ou plusieurs attributs. Pour analyser ces résultats et éventuellement dégager des classes d'équivalence, des régularités ou des singularités, nous avons effectué un certain nombre de calculs (Figure 16). Pour les groupes de lexies formés, qu'elles soient réunies dans des tables ou non, nous avons calculé le nombre de mots communs 2 à 2 (NMC). Nous avons également calculé le pourcentage de mots communs entre les groupes (pourcentage d'appartenance des mots du groupe 1 dans le groupe 2 : G1/G2= NMC(G1,G2)/Card(G2)) et le taux de recouvrement des groupes entre eux (T(G1,G2) = (G1/G2 + G2/G1)/200).

G1	G2	NMC	G1/G2	G2/G1	c(G1)	c(G2)	T(G1,G2)
AA0	JR1	8	88,89%	100,00%	8	9	0,94
AA3	MC0	8	88,89%	80,00%	10	9	0,84
AA3	JR0	10	62,50%	100,00%	10	16	0,81
JR0	MC0	9	100,00%	56,25%	16	9	0,78
...	...	...	...	...	...	...	...

Figure 16. Extrait du tableau relatif aux groupes de mots d'une table entière. Les résultats sont classés par ordre décroissant de taux de recouvrement des groupes 2 à 2. (ex : AA0 représente les mots de la table n°0 du participant AA, ce groupe rassemble 89% des mots du groupes JR1)

[AA0]	**Direction**
[AA0a] dévalorisation – dévaluation – krach – mini-krack – baisse des cours – dévaluer- déflation	Descend
[AA0b] inflation	Monte

Figure 17. Table en rapport avec les phénomènes dynamiques de la bourse du participant AA.

[JR1]	**Direction**	**Connotation**
[JR1a] dévaluation – dévalorisation – dévaluer – baisse des cours - déflation	descend	-
[JR1b] krach – mini-krach	descend	mal
[JR1c] inflation	monte	mal
[JR1d] hausse des cours	monte	-

Figure 18. Table en rapport avec les phénomènes dynamiques de la bourse du participant JR.

[MC0]	**Connotation**	**Rapport à l'action**
[MC0a] actionnaire – boursicoteur – investisseur – porteur - petit porteur,	sans	rôle
[MC0b] bénéficiaire	bien	rôle
[MC0c] analyste - agent de change - économiste	sans	profession
	bien	profession

Figure 19. Table en rapport avec les acteurs de la bourse du participant MC.

[AA3]	**Action**	**Rapport à l'activité / Résultat**
[AA3a] actionnaire - souscripteur	achat	investissement
[AA3b] bénéficiaire	reçoit	retour sur investissement
[AA3c] boursicoteur	achat/vente	investissement

[AA3d] porteur - petit porteur	-	-
[AA3e] analyste - économiste	analyse	Etude/ observation
[AA3f] opérateur	-	-
[AA3g] agent de change	traitement opération	travail

Figure 20. Table en rapport avec les acteurs de la bourse du participant AA.

Parmi les résultats obtenus des participants, 20 tables entières ont pu être soumises aux calculs exposés ci-dessus. Les tables sont comparées 2 à 2 ce qui représente 380 groupes de 2 tables à considérer dans les calculs. Au final, 13 groupes de 2 tables (6,84%) présentent un taux de recouvrement supérieur à 0,5 parmi lesquels 4 groupes de 2 tables (2,11%) présentent plus de 10 mots communs. 54 groupes de 2 tables (28,84%) présentent un taux de recouvrement non nul et 12 groupes de 2 tables (6,32%) présentent plus de 5 mots communs.

En comparant ces résultats de calculs au matériel fourni par les participants, on peut apprécier le fait que les tables ayant 2 à 2 un taux de recouvrement le plus important, concernent majoritairement les acteurs du monde boursier comme les tables MC0 et AA3 (Figure 19 et Figure 20) par exemple. Ce sont ces tables que l'on rencontre généralement en haut du tableau classant l'ensemble des tables formées par ordre décroissant de taux de recouvrement. Par exemple : AA3, MC0, JR0, AL1 et SG5 ont 2 à 2 des taux de recouvrement supérieur à 0,7 et sont toutes relatives aux lexies pouvant faire référence à des personnes physiques (*boursicoteur*, *agent de change*…). Le deuxième groupe de tables qui présentent des taux de recouvrement important (au moins supérieur à 0,3) recèle majoritairement des tables en rapport avec les « *phénomènes* » boursiers. Par exemple, AL3, JR1, AA0 et SG2 ont 2 à 2 majoritairement des taux de recouvrement supérieur à 0,3 et sont toutes des tables relatives aux lexies pouvant faire référence à des phénomènes dynamiques (*dévaluation, déflation, baisse et hausse des cours*…) (voir Figure 17 et Figure 18 pour AA0 et JR1). Les tables proposées se conforment principalement aux divisions à teneur ontologique proposées dans le dispositif sur la météorologie : lieux, acteurs, phénomènes tout en présentant des particularités remarquables propres à chacun des participants. On peut noter également que les deux sujets ayant fait l'objet du plus grand nombre de tables quasi-identiques (phénomènes et *acteur*) regroupent des lexies facilement identifiables comme appartenant à ces sujets. Au contraire, les lexies de lieux et/ou institutions par exemple (*entreprise*, *banque*…) semblent plus difficiles à différencier et apparaissent donc plus rarement dans des tables ressemblantes d'un participant à l'autre. A ce propos, on peut noter que les consignes précisant qu'il était demandé de construire un dispositif en rapport avec la bourse, les lexies comme *entreprise* et *banque* n'étaient pas forcément identifiées comme faisant classiquement partie de ce domaine. Il semble qu'un consensus se soit dégagé autour des mots faisant directement référence à la bourse et que les mots relevant plutôt de l'économie ou des influents de la bourse aient été plus difficiles à classer et à différencier au sein de tables vu le sujet du dispositif.

Des calculs similaires à ceux déjà exposés ont été effectués sur les lignes des tables construites et sur les groupes de mots non structurés en tables. Pour les lignes, les calculs ont porté sur 44 lignes comparées 2 à 2 (donc 1892 lignes à considérer dans les calculs) : 60 groupes de lignes (3,10%) présentent un taux de recouvrement supérieur à 0,5 et 54 groupes (9,51%) un taux de recouvrement non nul. On peut observer également que seuls 5 groupes de lignes (0,53%) présentent plus de 5 mots communs. Parmi les lignes à un seul mot, 9 seulement sont parfaitement identiques. Des similarités apparaissent pour des lignes à plusieurs mots (exemple : AA0 et JR1, Figure 17 et Figure 18 avec la même valeur d'attribut 'descend' pour [Direction]) mais ces cas restent rares au vue de l'ensemble des tables construites. Pour les groupes de mots non structurés, une dizaine (sur 29) a des taux de recouvrement égaux à 100% mais aucun n'est parfaitement identique d'un participant à un autre.

Au vu de tous ces résultats et après entretien avec les participants, nous avons pu estimer tout d'abord que l'expérience présentait un certain nombre de défauts. Le premier est certainement le temps imparti trop court pour la réalisation du travail demandé. L'absence du corpus d'origine et donc l'impossibilité de revenir sur un texte faisant intervenir les mots proposés a également été ressentie comme un handicap par les participants. Ces variables seront donc à redéfinir pour une autre expérience. Il est important de noter que ce retour sur corpus est envisagé dans la démarche cyclique de construction des dispositifs. Une expérience sans corpus permettait simplement de tester la faisabilité de la construction de tels matériaux, d'apprécier la capacité des participants à amorcer ce processus cyclique. En l'occurrence, nous avons constaté que la méthode de construction des dispositifs s'acquiert rapidement et que les principes qui la régissent sont facilement assimilables. Les différences et les points communs découverts au sein des travaux rendus par les participants nous encouragent à poursuivre dans notre voie : les utilisateurs intègrent leur propre sensibilité par rapport à un domaine (cette

sensibilité pouvant relever d'une méconnaissance totale de ce domaine[23]) tout en se conformant aux usages qu'ils ont pu rencontrer des mots proposés. On peut trouver trace des degrés de systématicité de la sémantique unifiée tels qu'exposés dans (Rastier et al., 1994) (normes idiolectales, sociolectales et dialectales) à l'intérieur du dispositif et la dimension sociale (donc partagée) de la langue qui ne peut pas être absente de ce type de construction (Nicolle et al., 2002, p.61). Cette expérience nous a également permis de revenir sur nos propositions logicielles. Les participants ont en effet souvent regretté l'absence de support informatique pour la construction des dispositifs et ont émis le souhait de pouvoir, au sein d'un même logiciel, interroger le corpus de base et créer les ressources. Nous tiendrons donc compte de cette remarque pour les prochaines versions du logiciel *Anadia*. Nous avons pu constater différentes méthodologies de construction des dispositifs chez les participants : certains constituant d'abord des classes de mots pour chercher ensuite à en différencier les représentants, d'autres cherchant d'abord les différences pertinentes pour ensuite créer les tables et y faire figurer les mots du domaine. Ce constat devra également être considéré pour nos réalisations informatiques futures.

L'expérimentation que nous avons menée lors de l'atelier «*Variation, construction et instrumentation du sens* » (Tatihou, Juillet 2002) nous a permis d'apprécier la capacité d'utilisateurs potentiels à s'approprier les principes du modèle que nous proposons. Bien que la tâche de description des significations ne soit pas triviale, comme l'a souligné par exemple C. Kerbrat-Orecchioni[24], nous avons pu constater que les utilisateurs arrivent à formuler dans un temps raisonnable des représentations lexicales reflétant leur point de vue sur le domaine proposé. Ceci constituait l'une des hypothèses que nous cherchions à estimer. Durant les deux séances de cette expérimentation et, à travers les différences et les points communs entre les dispositifs et les groupes de mots fournis par les participants, reflet de leurs capacités interprétatives, nous avons pu considérer à sa juste valeur la dimension sociale et partagée du langage latent en chaque locuteur. Cela a renforcé notre point de vue résolument anthropocentré.

Conclusion

Le travail que nous avons présenté dans cet article indique une étape dans nos recherches sur l'instrumentation du sens. Beaucoup de points soulevés donnent à entrevoir de larges champs de recherche qu'il nous faudra explorer. Par exemple, nos études sur la métaphore nous poussent en ce moment à réfléchir sur la nature des sèmes en jeu dans les tropes. Cependant, au fur et à mesure, qu'avancent nos travaux certaines hypothèses semblent être attestée. Par exemple, l'appropriation d'une méthodologie de micro-structuration lexicale différentielle est tout a fait envisageable pour des utilisateurs non familiers avec un travail qui relève de la terminologie. De même, il nous apparaît essentiel d'envisager dans les interactions homme-machine les instrumentations du sens avec un point de vue résolument anthropocentré. Ce courant de l'anthropocentrisme est selon nous tout à fait prometteur et plus encore, dans le cadre du développement de logiciels d'études, lorsqu'il est mis en oeuvre conjointement aux méthodes distributionnelles. Nous avons pu en tirer également comme conséquence, qu'il n'est pas nécessaire de constituer des ressources exhaustives pour instrumentaliser efficacement des tâches en rapport avec la sémantique des langues naturelles, comme par exemple la réalisation d'interfaces de lecture rapide, la mise en place d'aide logicielle à l'interprétation ou encore la réalisation de systèmes de filtrage en recherche documentaire.

Dans le cadre des systèmes anthropocentrés, une question cruciale reste en suspens : l'évaluation. Évaluer la compétence interprétative d'une machine dans une interaction particulière avec un sujet humain, c'est-à-dire évaluer l'aspect « naturel » d'une interaction homme-machine, est de toute évidence un problème de taille qui ne serait être un enjeu strictement informatique du fait de la place prépondérante du sujet, de l'interaction et plus largement de la culture et de la société. Ce genre d'évaluation interroge toutes les dimensions des sciences cognitives. C'est une question passionnante dont l'intérêt dépasse de loin le classement des modèles en fonction de leur taux de succès. Il faudra certainement consacrer beaucoup d'investissements en temps et en ressources humaines pour mettre au point, dans le cadre de recherches pluridisciplinaires, des protocoles d'évaluation

[23] A ce propos, on constate qu'une mauvaise connaissance d'un domaine n'empêche en rien les sujets d'avoir une compétence langagière sur celui-ci. On peut en déduire une différence fondamentale entre les connaissances ontologiques sur un domaine et son lexique.

[24] « un des problèmes majeurs que pose (…) la description des structurations lexicales réside dans le fait qu'elles tiennent à la fois des systèmes diacritiques (non hiérarchiques) et des systèmes taxinomiques (hiérarchiques) » (Kerbrat-Orecchioni, 1988)

nouveaux de ces nouvelles formes d'instrumentation du sens anthropocentrées. C'est ce à quoi nous cherchons à contribuer à travers nos travaux.

Bibliographie :

Beust P. (1998). *Contribution à un modèle interactionniste du sens*. Thèse de Doctorat d'informatique de l'Université de Caen. (http://users.info.unicaen.fr/~beust/These/these.html)

Beust P. (2002). *Un outil de coloriage de corpus pour la représentation de thèmes*. JADT 2002 : 6emes Journées internationales d'Analyse statistique des Données Textuelles. Saint Malo.

Bourrigault D. (1994). *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*. Thèse de Doctorat en mathématiques, informatique appliquée aux sciences de l'homme, École des Hautes Études en Sciences Sociales de Paris.

Cavazza M. (1996). *Sémiotique textuelle et contenu linguistique*. Intellectica, n°3, 1996/2.

Coursil J. (2000). *La fonction muette du langage*. Petit-Bourg : Ibis rouge éd., Schoelcher, Presses universitaires créoles-GEREC.

Faure D. (2000). *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système* ASIUM. Thèse de Doctorat en Informatique de l'Université de Paris Sud.

Ferrari S. (1997). *Méthode et outils informatiques pour le traitement des métaphores dans les documents écrites*. Thèse de doctorat en Informatique de l'Université de Paris XI.

Giguet E. (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse de Doctorat en Informatique de l'Université de Caen Basse-Normandie.

Greimas A. J. (1983). *Du sens II*, *Essais sémiotiques*. Editions du Seuil : Paris.

Hejlmslev L. (1943). *Prolégomènes à une théorie du langage*. Editions de Minuit (1968) : Paris.

Kamp H. (1981). *A theory of truth and semantic representation*. In Gronendijk/Janssen/Stokhof (eds), Formal Methods in the Study of Language, Part 1. Mathematisch Centrum.

Kerbrat-Orecchioni C. (1988). *Sémantique*. In Encyclopedia Universalis, 693-699. Understanding and creating sentences. *American psychologist*. Vol. 18, 735-751.

Lakoff, G. et Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press, 5. Chicago.

Landauer T. K., Dumais S. T. (1997). *A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge*. Psychological Review, 104, 211-240.

Lebart L., Salem A., and Berry L. (1998). *Exploring textual data*. Kluwer Academic.

Luzzati D. (1996). *Le dialogue verbal homme-machine, études de cas*, Editions Masson : Paris.

Nicolle A. (2001). *La question du symbolique en informatique*. Colloque Interdisciplinaire en Sciences Cognitives ARCO'2001. Lyon. (en cours de publication)

Nicolle A., Beust P., Perlerin V. (2002). *Un analogue de la mémoire pour un agent logiciel interactif*, Revue In Cognito, n°21, p. 37-66.

Perlerin V. (2002). *Memlabor, un environnement de création, de gestion et de manipulation de corpus de textes*. RECITAL 2002, pp 507-516, tome1. Nancy.

Perlerin V. (2001). *La recherche documentaire, une activité langagière*, RECITAL 2001. pp 469-479, tome1. Tour.

Perlerin V., Ferrari S. et Beust P. (2002). *Métaphore et Dynamique. 2ème Journées de Linguistique de Corpus*. Lorient –à paraître.

Pottier B. (1974). *Linguistique générale, théorie et description*. Klincksieck, Paris.

Pichon R. et Sébillot P. (1999). *Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience*. TALN 1999, pages 279-288. Genève.

Rastier F. (1991). *Sémantique et recherches cognitives*. Presses Universitaires de France : Paris.

Rastier F., Cavazza M., Abeillé A. (1994). *Sémantique pour l'Analyse*. Masson : Paris.

Rastier F. (1987). *Sémantique interprétative*. Presses Universitaires de France : Paris.

Saussure F. de (1915). *Cours de linguistique générale*. Editions. Mauro-Payot : Paris (1986).

Smadja F. et McKeown M. (1990). Automatically *extracting and representing collocations for language generation*. 28th Annual Meeting of the Association for Computational Linguistics (ACL'90), Pittsburgh.

Tanguy L. (1997). *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle informatique de la sémantique interprétative.* Thèse de Doctorat en Informatique de l'Université de Rennes I.

Thlivitis T. (1998). *Sémantique interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension des textes*. Thèse de Doctorat en Informatique de l'Université de Rennes 1.

Victorri B. (1998). *La construction dynamique du sens : un défi pour l'intelligence artificielle*. RFIA'98.

Zipf G.K. (1949). *Human Behavior and the Principle of least effort : an introduction to Human Ecology*, Mass: Addison-Wesley, Reading.